NATAL, 2018

Realização: SBC

Organização: UFRN

anais 2018

XXXVIII CONGRESSO DA SOCIEDADE BRASILEIRA DE COMPUTAÇÃO

7º BraSNAM –BRAZILIAN WORKSHOP ON SOCIAL NETWORK ANALYSIS AND MINING

CENTRO DE CONVENÇÕES | NATAL•RN | 22 A 26 DE JULHO DE 2018

#COMPUTAÇÃOESUSTENTABILIDADE

CSBC 2018

anais 2018

XXXVIII CONGRESSO DA SOCIEDADE BRASILEIRA DE COMPUTAÇÃO
CENTRO DE CONVENÇÕES | NATAL•RN | 22 A 26 DE JULHO DE 2018
#COMPUTAÇÃOESUSTENTABILIDADE

Coordenador Geral

Francisco Dantas de Medeiros Neto (UERN)

Comissão Organizadora

Bartira Paraguaçu Falcão Dantas Rocha (UERN)
Camila Araújo Sena (UERN)
Everton Ranielly de Sousa Cavalcante (UFRN)
Felipe Torres Leite (UFERSA)
Ilana Albuquerque (UERN)
Isaac de Lima Oliveira Filho (UERN)
Priscila Nogueira Krüger (UERN)

Realização

Sociedade Brasileira de Computação

Organização

Universidade do Estado do Rio Grande do Norte

CSBC 2018
XXXVIII Congresso da
Sociedade Brasileira de Computação

Apresentação

Estes anais registram os trabalhos apresentados durante o XXXVIII Congresso da Sociedade Brasileira de Computação (CSBC 2018), realizado em Natal-RN, de 22 a 26 de julho 2018. O evento teve como tema central a Computação e Sustentabilidade, pois se compreende que o avanço da computação e as questões ambientais devem caminhar lado-a-lado, tendo em vista que as técnicas computacionais necessitam ser usadas para possibilitar o desenvolvimento sustentável, e, desse modo, equilibrar as necessidades ambientais, econômicas e sociais.

Organizar o maior evento acadêmico de Computação da América Latina foi um privilégio e um desafio. Foi enriquecedor promover e incentivar a troca de experiências entre estudantes, professores, profissionais, pesquisadores e entusiastas da área de Computação e Informática de todo o Brasil. Ao mesmo foi desafiador termos que lidar, principalmente, com às dificuldades impostas pelo momento de crise que o nosso Brasil vem enfrentando. Uma crise que afeta diretamente nossas pesquisas e, consequentemente, o desenvolvimento e inovação do nosso amado Brasil.

Por meio de seus 25 eventos, o CSBC 2018 apresentou mais de 300 trabalhos, várias palestras e mesas-redondas. O Congresso ainda abrigou diversas reuniões, que incluem a reunião do Fórum de Pós-Graduação, a reunião do CNPq/CAPES, a reunião dos Secretários Regionais SBC, a reunião das Comissões Especiais e a reunião do Fórum IFIP/SBC.

O sucesso do CSBC 2018 só foi possível devido à dedicação e entusiasmo de muitas pessoas. Gostaríamos de agradecer aos coordenadores dos 25 eventos e aos autores pelo envio de seus trabalhos. Além disso, gostaríamos de expressar nossa gratidão ao Comitê Organizador, por sua grande ajuda em dar forma ao evento; e, em especial, à equipe da Sociedade Brasileira de Computação (SBC), por todo apoio.

Por fim, reconhecemos a importância do apoio financeiro da CAPES, do CNPq, do CGI.br, do Governo do Estado do Rio Grande do Norte, da Prefeitura Municipal do Natal, da Prefeitura Municipal de Parnamirim, da CABO Telecom, da ESIG Software e Consultoria, da DynaVideo e do SENAI.

Natal (RN), 26 de julho de 2018.

Chico Dantas (UERN)
Coordenador Geral do CSBC 2018

Anais do CSBC 2018

BraSNAM 2018
VII Brazilian Workshop on
Social Network Analysis and Mining

BraSNAM 2018
VII Brazilian Workshop on
Social Network Analysis and Mining

Apresentação

É um prazer anunciar o VII Brazilian Workshop on Social Network Analysis and Mining (BraSNAM). O workshop acontecerá juntamente com o 38o Congresso da Sociedade Brasileira de Computação (CSBC), que é o evento oficial da Sociedade Brasileira de Computação. O objetivo é discutir avanços recentes em mineração e análise de redes sociais. A evento será realizado em Natal (RN), nos dias 24 e 25 de julho de 2018. O BraSNAM 2018 reunirá pesquisadores e profissionais interessados na análise de rede social e áreas afins, e promoverá a colaboração e intercâmbio de ideias e experiências.

O estudo sobre redes sociais tem suas origens em comunidades sociais, educacionais e empresariais. O interesse acadêmico nessa área tem crescido desde a metade do século XX. O crescente aumento no número de usuários Web estimula a interação entre pessoas, a disseminação de dados, a troca de informação e também aumenta significativamente os dados disponíveis. A partir da mineração e análise de redes sociais, podemos criar ou enriquecer soluções aplicadas a: identificação de grupos (nocivos ou não), identificação de influência, detecção de necessidades, propagação de informações, fluxo da informação, identificação de rumores, fomentação de alianças, entre outros. Tais soluções podem ser aplicadas a vários cenários como ecossistemas de software, marketing, segurança, emergência, gestão de equipes, cidades inteligentes e outros relacionados "Computação e Sustentabilidade", tema do CSBC 2018. Além disso, este evento tem como objetivo servir como um facilitador na troca de conhecimento e colaboração entre academia e empresas, ligando pesquisadores e profissionais que trabalham na área.

Em sua sétima edição, o BraSNAM recebeu um número significativo de trabalhos, totalizando 61 artigos submetidos, dos quais foram aceitos 17 artigos completos e 10 artigos resumidos. O processo de revisão foi double blind review. Destaca-se ainda que todos os autores dos artigos completos serão convidados a submeterem versões estendidas para a iSys - Revista Brasileira de Sistemas de Informação - CESI/SBC.

O BraSNAM oferecerá uma palestra intitulada "Nem Sempre Se Vê Mágica no Absurdo", proferida pelo Prof. Altigran Soares da Silva (IComp/UFAM). Além disso, teremos o painel "Marcando Presença: Como a Computação Social Transforma o Mercado de Software?", formado por: Prof. Adriano Bessa Albuquerque (BANOR/UNIFOR, Moderador), Prof. Jair Leite (DIMAp/UFRN), Profa. Tanara Lauschner (IComp/UFAM) e Prof. Davi Viana (UFMA). Promovendo uma maior discussão e troca de ideias, teremos os seguintes "Bate-Papos": "Smart Cities and Social Networks" a ser proferido pelo Prof. Fábio Kon (USP), "Computação Transparente" a ser proferido pela Profa. Cláudia Cappelli (UNIRIO), "Big Social Data e Eleições: Fake News, Discurso de Ódio, Rastreabilidade e Os Outros" a ser proferido pela Profa. Jonice Oliveira (UFRJ) e "Segurança da Informação em Mineração e Redes Sociais" a ser proferido pelo Prof. Avelino Zorzo (PUCRS).

Nós gostaríamos de agradecer aos organizadores do CSBC 2018, em especial ao coordenador geral, Chico Dantas (UERN); ao diretor de publicações da SBC, Prof. José Viterbo; ao periódico iSys e aos seus editores-chefes, Prof. Rodrigo Santos (UNIRIO) e Prof. André Freire (UFLA); aos membros do Comitê Diretivo do BraSNAM; e aos membros do Comitê de Programa, revisores, palestrantes, painelistas, autores e participantes dessa sétima edição do BraSNAM. Tenhamos todos um excelente evento!

Rodrigo Pereira dos Santos (UNIRIO)
Raimundo Santos Moura (UFPI)

Coordenadores do BraSNAM 2018

Comitê de Organização

Coordenação Geral / Coordenação do Comitê de Programa
Rodrigo Pereira dos Santos (UNIRIO), Raimundo Santos Moura (UFPI)

Comitê Diretivo

Daniel Figueiredo (UFRJ), Fabrício Benevenuto (UFMG), Giseli Lopes (UFRJ), Jano Moreira de Souza (UFRJ), Jonice Oliveira (UFRJ), Juliana Valério (UFRJ), Li Weigang (UnB), Mirella Moro (UFMG), Renata Galante (UFRGS), Roberto Imbuzeiro Oliveira (IMPA), Rodrigo Santos (UNIRIO)

Comitê de Programa

Adriano César Pereira (UFMG), Aline Paes (UFF), Ana Paula Couto da Silva (UFMG), Anderson Ferreira (UFOP), Artur Ziviani (LNCC), Bernardo Estácio (PUCRS), Bernardo Pereira Nunes (PUC-Rio), Cássio Prazeres (UFBA), Catarina Costa (UFAC), Claudia Cappelli (UNIRIO), Claudia Werner (UFRJ), Claudio de Farias (UFRJ), Claudio Pinhanez (IBM Research), Daniel Figueiredo (UFRJ), Daniel Lichtnow (UFSM), Daniel Menasche (UFRJ), Davi Viana (UFMA), Eduardo Bezerra (CEFET-RJ), Eduardo Nakamura (UFAM), Elisa Huzita (UEM), Elizeu Santos-Neto (Google), Emanuel Coutinho (UFC), Fabiola Souza (UFU), Fabrício Benevenuto (UFMG), Fábio Basso (Unipampa), Flavio Horita (UFABC), Genaina Rodrigues (UnB), George Valença (UFRPE), Giseli Lopes (UFRJ), Gislaine Camila Leal (UEM), Gustavo Guedes (CEFET/RJ), Heitor Costa (UFLA), Hernane Pereira (SENAI CIMATEC), Humberto Marques (PUC-Minas), Igor Steinmacher (UTFPR), Igor Wiese (UTFPR), Isabela Gasparini (UDESC), Ivaldir Junior (UFPE/Softex Recife), Jairo Souza (UFJF), Jefferson Ebert Simões (UNIRIO), Jonice Oliveira (UFRJ), José Falazo Oliveira (UFRGS), José Maria David (UFJF), Josiane Prol (PUCRS), Juliana Valério (UFRJ), Jussara Almeida (UFMG), Leandro Silva (Mackenzie), Leila Wetzel (UFF), Li Weigang (UnB), Luciana Roman (Embrapa), Luciano Digiampietri (USP), Luis Rivero (UFAM), Luiz André Paes Leme (UFF), Luiz Henrique Merschmann (UFLA), Marco Casanova (PUC-Rio), Maria Gilda Esteves (UFRJ), Michele Brandão (UFMG), Mirella Moro (UFMG), Nazareno Andrade (UFCG), Paulo Sérgio Santos (UFRJ), Pedro Olmo Vaz de Melo (UFMG), Raimundo Moura (UFPI), Renata Araujo (UNIRIO), Renata Galante (UFRGS), Ricardo Prudêncio (UFPE), Rodrigo Santos (UNIRIO), Sean Siqueira (UNIRIO), Sergio Manuel Serra da Cruz (UFRRJ), Thiago Pardo (USP), Valdemar Vicente Graciano Neto (UFG), Victor Stroele (UFJF)

Revisores Externos

Everton Gomede (UEL), Jaqueline de Oliveira (PUC Minas), Juliana Fernandes (IFPI/UNIRIO), Luiz Fernando Assis (USP), Rebeca Schroeder (UDESC), Sidgley Andrade (UTFPR), William Christhie (UFMG)

Painéis / Palestras

Nem Sempre Se Vê Mágica no Absurdo
Altigran Soares da Silva (UFAM)

Apesar do grande interesse da academia e do mercado em torno de temas ligados à chamada Ciência de Dados, é notório o fato de que não menos do que 80% do tempo e esforço em projetos nesta área são despendidos com tarefas ligadas à preparação dos dados a serem analisados. De fato, tarefas como coleta, extração, deduplicação, integração de dados, embora cruciais no processo, estão pouco relacionadas às atividades típicas de Ciência de Dados, como, análise, mineração de padrões, geração de modelos, etc. Neste momento, em que novos aspectos como ética, conformidade legal, reprodutibilidade científica, qualidade de dados e viés algorítmico estão emergindo de forma decisiva, este esforço tende aumentar ainda mais. Nesta palestra, pretendo discutir como métodos, técnicas e ferramentas típicas de Engenharia de Dados podem ajudar a reduzir este esforço para que a mágica prometida pela Ciência de Dados não deixe de acontecer por causa do absurdo dos dados em estado bruto. Como exemplo concreto, apresentarei alguns resultados recentes de minha pesquisa relacionada a métodos para viabilizar a análise de sentimentos em textos opinativos escritos por usuários.

Bio: Altigran Soares da Silva é professor associado do Instituto de Computação da Universidade Federal do Amazonas (IComp/UFAM) onde atua como pesquisador, professor e orientador na graduação, mestrado e doutorado. Concluiu seu doutorado em Ciência da Computação pela UFMG em 2002. Seus interesses de pesquisa envolvem Gerência de Dados, Recuperação de Informação e Mineração de Dados com ênfase no ambiente da World-Wide Web e Mídias Sociais. Sobre estes temas, tem coordenado e participado de dezenas de projetos de pesquisa que resultaram em mais de 100 publicações científicas em periódicos e anais de conferência de boa qualidade nestas áreas. Foi coordenador de comitês de programa de conferências no Brasil e no exterior, tendo participado também como membro de comitês técnico de programa em cerca de 50 conferências e workshops internacionais. Exerceu entre 2007 e 2009 a Pró-reitoria de Pesquisa e Pós-Graduação da UFAM. No triênio 2011-2013 foi o Coordenador Adjunto da área de Computação na CAPES e é atualmente membro do CA-CC do CNPq. Entre 2005 e 2015 foi membro da diretoria da Sociedade Brasileira de Computação (SBC), sendo atualmente membro do conselho da Sociedade. É co-fundador de empreendimentos de tecnologia, entre eles a Akwan Information Technologies, adquirida pela Google Inc. em 2005, e a Neemu.com, empresa de tecnologia para varejo on-line que é líder no e-commerce brasileiro e que foi adquirida pela Linx Sistemas em 2015. Em 2013 uma tese de doutorado sob sua orientação recebeu o Primeiro Lugar no Concurso de Teses e Dissertação da Sociedade Brasileira de Computação (SBC) e Menção Honrosa no Prêmio CAPES de Teses. Recebeu também em 2013 o prêmio de Sócio Destaque da SBC por sua atuação junto às Comissões Especiais da sociedade, contribuindo para o aperfeiçoamento do Qualis CAPES de Conferências na área de Ciência da Computação. Em 2015 foi ganhador de um dos "Google Research Awards in Latin America" como orientador.

Smart Cities and Social Networks
Fábio Kon (USP)

We will discuss the basic concepts and challenges in the field of Smart Cities, focusing on the synergies and conflicts with Social Networks. The discussion will start by presenting a few examples of interesting applications of Smart Cities, with a particular focus on improving city

government and supporting Evidence-Based Public Policy making. We will then reflect on how research on Social Networks Analysis and Mining could further support the work of city officials in everyday operations, short-term planning, and long-term policymaking.

Bio: Fabio Kon is a Full Professor of Computer Science at the University of São Paulo. His research interests include Smart Cities (http://interscity.org), Big Data Processing, Data Science, Distributed Systems, and Startup Ecosystems. Fabio has been in the Scientific Program Committees of a dozen high-impact international conferences in the past 15 years and is currently the Editor-in-Chief of the SpringerOpen Journal of Internet Services and Applications. Fabio is an ACM Distinguished Scientist and a Special Advisor to the Scientific Director for Innovation programs at the São Paulo Research Agency.

Computação Transparente
Cláudia Cappelli (UNIRIO)

As organizações têm sido avaliadas pela sua capacidade de fornecer conhecimento confiável e transparência sobre suas operações, desempenho e resultados. O objetivo é melhorar a visão e o entendimento das pessoas sobre processos, serviços e informações para conscientizar, reduzir a omissão, possibilitar o controle, facilitar a pesquisa e aumentar a confiança. Nesse sentido, a transparência organizacional torna-se uma preocupação importante, principalmente ao projetar sistemas de informação que interagem com pessoas. Com o uso cada vez mais crescente da inteligência artificial, e principalmente do aprendizado de máquina, cada vez mais disseminado nas organizações, há cada vez mais sistemas, aplicativos e outros artefatos computacionais, que indicam decisões que podem ser tomadas pelo usuário ou, ainda, sistemas que tomam suas próprias decisões. No entanto, o entendimento de como a ação desses sistemas é definida pode ser fundamental em muitas situações. Este tema tem sido investigado para agregar de forma sistemática valores como auditabilidade, adaptabilidade, acessibilidade, usabilidade, compreensão, correção e consistência. Dessa forma, para implementar transparência, é necessário abordar como o software deve lidar com esse conceito. Tudo isso mostra que a transparência é uma nova e importante preocupação ao projetar um software que automatiza processos, processa informações e interage com pessoa. Assim, configura-se um duplo desafio: (i) quais são os aspectos de transparência que devem ser analisados ao projetar um sistema de informação?; e (ii) como representar e projetar a forma como um sistema de informação deve lidar com eles? Vamos falar sobre isso?

Bio: Claudia Cappelli é Professora Adjunta IV e membro do Programa de Pós-graduação em Informática da Universidade Federal do Estado do Rio de Janeiro (UNIRIO). Bolsista de Produtividade Desen. Tec. e Extensão Inovadora do CNPq – Nível 2. Doutora em Ciências - Informática pela PUC-Rio (2009). Jovem Cientista FAPERJ. Pesquisadora do Instituto Nacional de Ciência e Tecnologia em Democracia Digital (INCT-DD). Mestre em Sistemas de Informação pela Universidade Federal do Rio de Janeiro (2000). Graduada em Informática pela Universidade do Estado do Rio de Janeiro (1985). Realizou estágio Pós-Doutoral junto ao Programa de Pós-Graduação em Informática da Unirio (2010). Gerente da Área de Arquitetura Corporativa e Planejamento de Tecnologia do Citibank e da Telemar por 8 anos. Coordenou durante 12 anos o NP2Tec (Núcleo de Pesquisa e Prática de Tecnologia) e por 2 anos o CiberDem (Núcleo de Pesquisa em CiberDemocracia) ambos na Universidade Federal do Estado do Rio de Janeiro (UNIRIO). Representante do Comitê de Transparência Organizacional da UNIRIO. Revisora de diversos periódicos nacionais e internacionais. Representante Institucional da Sociedade Brasileira de Computação (SBC) na UNIRIO.

Diretora de Articulação com Empresas na SBC. Coordenadora do projeto Meninas Digitais da SBC no Rio de Janeiro (Digital Girls in Rio). Atua na área de Sistemas de Informação, principalmente nos seguintes temas: Gestão de Processos de Negócio, Arquitetura Corporativa, Gestão de TI, Transparência Organizacional e Governo Digital.

Big Social Data e Eleições: Fake News, Discurso de Ódio, Rastreabilidade e Os Outros
Jonice Oliveira (UFRJ)

Devido ao rápido desenvolvimento da computação social e à multiplicação das mídias sociais, grande parte das interações é mediada pela tecnologia da informação e ocorre no mundo digital. Com isto, temos uma maior variedade de informação e opiniões sendo disseminadas de uma maneira muito mais rápida. Além das questões de variedade, velocidade, volume e veracidade do que é criado e disseminado, grande parte deste conteúdo, perfil de seus titulares e conhecimento inferido está sob o controle de instituições privadas. Consequentemente, o usuário perde o controle da sua informação: a gerada e a consumida. Como tal cenário pode impactar na formação de opinião? Em uma época de renovação de governos (como está acontecendo em toda a América Latina), como a ausência de transparência sobre dados sociais pode afetar a democracia? Como nós - cidadãos - podemos lidar com este problema? Como nós - profissionais de Computação - podemos ajudar a combatê-lo?

Bio: Jonice Oliveira obteve o seu doutorado em 2007 na área de Engenharia de Sistemas e Computação, ênfase em Banco de Dados, pela COPPE/UFRJ. Durante o seu doutorado recebeu o prêmio IBM Ph.D. Fellowship Award. Na mesma instituição realizou o seu Pós-Doutorado, concluindo-o em 2008. Desde 2009 é professora do Departamento de Ciência da Computação da UFRJ e atualmente é coordenadora do Programa de Pós-Graduação em Informática (PPGI-UFRJ). Coordena o Laboratório CORES (Laboratório de Computação Social e Análise de Redes Sociais), que conduz pesquisas multidisciplinares para o entendimento, simulação e fomento às interações sociais. Sua principal área de pesquisa é Computação Social, mais especificamente nos temas de Análise de Redes Sociais, Big Social Data, Suporte à Decisão e Recomendação. Possui uma larga experiência em tais áreas, com mais de 220 artigos, dezenas de orientações e envolvimento (como membro ou como líder) em projetos de pesquisas nacionais e internacionais. Mais detalhes em: http://www.joniceoliveira.net/

Segurança da Informação em Mineração e Redes Sociais
Avelino Zorzo (PUCRS)

O uso de dados pessoais em redes sociais tem trazido muita discussão neste ano, principalmente, em relação a problemas de vazamento de informações e do seu uso inadequado. Este aspecto tem despertado a atenção das pessoas em relação a privacidade das informações e também de sua integridade. Enquanto privacidade está relacionado com a liberação do acesso às informações para quem queremos, integridade está relacionado com que as informações não sejam alteradas. Neste bate-papo conversaremos sobre tendências em termos de segurança das informações compartilhadas em redes sociais.

Bio: Avelino Zorzo é associado da Sociedade Brasileira de Computação (SBC). Possui graduação em Ciência da Computação pela Universidade Federal do Rio Grande do Sul (1986-1989), mestrado em Ciência da Computação pela Universidade Federal do Rio Grande do Sul (1990-1994), doutorado em Ciência da Computação pela University of Newcastle Upon Tyne

(1995-1999) e pós-doutorado na área de segurança no Cybercrime and Computer Security Centre da Newcastle University (2012-2013). Atualmente é professor titular da Faculdade de Informática (FACIN) da Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS), avaliador de condições de ensino do Ministério da Educação, consultor ad hoc do CNPq, CAPES e da FAPERGS. Atuou como diretor da FACIN/PUCRS entre 2005 e 2011; como Coordenador Adjunto para Programas Profissionais da CAPES/MEC (2014-2018), como diretor adjunto de treinamento e ensino da SUCESU-RS entre 2008 e 2011; membro da diretoria da ASSESPRO-RS entre 2008 e 2011; membro do conselho técnico-consultivo da SOFTSUL; membro do Comitê de Ética em Pesquisa da PUCRS; como Diretor de Educação da Sociedade Brasileira de Computação (SBC), (2015-2017); e, como Diretor de Articulação com Empresas da SBC (2013-2015). Tem experiência na área de Ciência da Computação, com ênfase em Software Básico, atuando principalmente nos seguintes temas: segurança de sistemas, tolerância a falhas, teste de software, sistemas operacionais e modelagem analítica de sistemas confiáveis.

Marcando presença: como a Computação Social transforma o mercado de software?
Adriano Albuquerque (BANOR/UNIFOR) - moderador
Jair Leite (DIMAp/UFRN), Tanara Lauschner (IComp/UFAM), Davi Viana (UFMA)

Nos últimos anos, as áreas de Mineração e Análise de Redes Sociais vêm ganhando um maior destaque devido à quantidade e diversidade de dados que podem ser analisados, à capacidade de processar e resolver análises complexas de uma maneira mais eficiente, ao desenvolvimento de novas soluções para visualizar redes cada vez mais complexas e à aplicação de seus conceitos em outras soluções. Com o subsídio dessas áreas, podemos aperfeiçoar métodos ou técnicas de apoio à contextualização, filtragem, recomendação, identificação de pares, análise de cenários, personalização da informação – questões que os sistemas computacionais atuais, independente do cenário onde são aplicados, precisam lidar. Neste painel discutiremos a geração de novas ideias, perspectivas para a área de análise e mineração de redes sociais e aplicações na Computação e contará com a presença de pesquisadores de instituições e empresas que atuam na área de redes sociais e computação social.

Bio:

Adriano Albuquerque possui graduação em Bacharelado em Ciências da Computação pela Universidade Federal do Ceará (1990), mestrado em Informática Aplicada pela Universidade de Fortaleza (2001) e doutorado em Engenharia de Sistemas e Computação pela Universidade Federal do Rio de Janeiro (2008). Atualmente é professor da Universidade de Fortaleza. Tem experiência na área de Ciência da Computação, com ênfase em Engenharia de Software, atuando nos seguintes temas: qualidade de software, avaliação e melhoria de processos de software, métricas, normas ISO e modelos de maturidade de software. Atua como implementador e avaliador oficial dos modelos de maturidade de software: CMMI e MPS.BR.

Jair Cavalcanti Leite é professor titular do DIMAp atuando em disciplinas e projetos na área de Design de Software, Arquitetura de Software e Interação Humano-Computador. Seus interesses de pesquisa incluem, dentre outros, linguagens para apoio ao design e desenvolvimento de software tais como IMML, XSED e SysADL. Concluiu 16 orientações de alunos de pós-graduação e publicou diversos artigos em conferências nacionais e internacionais. É formado em Ciência da Computação pela UFPB, com pós-graduação em Informática na PUC-Rio, obtendo os títulos de Mestre (M.Sc.) em 1991 e de Doutor (D.Sc.)

em 1998. Em 2004/05 foi pesquisador visitante na Universidade de Lancaster, Inglaterra, e em 2013, no IRISA, Université de Bretagne-Sud, onde trabalhou na equipe que desenvolveu a linguagem SysADL. Foi coordenador dos cursos de Ciência da Computação e do Programa de Pós-graduação em Sistemas e Computação da UFRN e do curso de graduação em Engenharia de Software, da UFRN. Desde 2010, atua na equipe que elaborou o projeto do Instituto Metrópole Digital da UFRN, liderou o grupo que criou o curso de Bacharelado em Tecnologia da Informação e atualmente ocupa o cargo de Diretor de Projetos. É ainda um ativo colaborador da Sociedade Brasileira de Computação (SBC), onde foi Secretário Regional para os estados do RN, PB, PE e AL, coordenador da Comissão. Especial de Interação Humano-Computador, coordenador geral dos eventos IHC 2006, SBSC 2006 e Web Media 2006, e do WEI 2018; e atual membro da Comissão de Educação.

Tanara Lauschner é Engenheira Eletricista (UFAM-1998), Mestre em Ciência da Computação (UFMG-2002), Doutora em Informática (Puc-Rio 2010). Professora da Universidade Federal do Amazonas desde 2002, tendo trabalhado previamente em empresas do distrito industrial de Manaus e em instituições de pesquisa privadas. Atua no movimento de Mulheres e é Coordenadora do Programa Cunhantã Digital que visa atrair meninas do ensino médio e fundamental para a computação, é Diretora do Instituto de Computação da UFAM e conselheira titular do Conselho Gestor da Internet (CGI.br).

Davi Viana é Doutor e Mestre em Informática pelo Programa de Pós-Graduação em Informática da Universidade Federal do Amazonas. Graduado em Ciência da Computação pela Universidade Federal do Amazonas (UFAM). Atualmente é Professor Adjunto A da Universidade Federal do Maranhão e atua no curso de Engenharia da Computação, no Programa de Pós-Graduação em Ciência da Computação (PPGCC) e na Diretoria de Difusão do Empreendedorismo. Tem experiência nas áreas de qualidade de software, melhoria processo de software (MPS), Ensino em Engenharia de Software, Engenharia de Software Experimental e startups de software. Já participou de implementações de programas de MPS com ênfase na adoção de modelos de maturidade (MPS.BR).

Sumário

Artigos Completos

Artigos Resumidos

Análise da estrutura da rede de frames da FrameNet Brasil

Míria Bóbó[1], Victor Ströele[1], Ely Edison da Silva Matos[2], Regina Braga[1], Fernanda Campos[1], José Maria N. David[1], Tiago Timponi Torrent[2]

[1]Departamento da Ciência da Computação – Universidade Federal de Juiz de Fora (UFJF)
Caixa Postal 20.006 – 36.016-970 – Juiz de Fora – MG – Brazil

[2]FrameNet Brasil – Universidade Federal de Juiz de Fora (UFJF)
Caixa Postal 20.006 – 36.016-970 – Juiz de Fora – MG – Brazil

{miria.luisa,victor.stroele,jose.nazar}@ice.ufjf.br,

{ely.matos,regina.braga,fernanda.campos,tiago.torrent}@ufjf.edu.br

Abstract. *FrameNet Brazil is a computational lexicography project that aims to use frames for the description of the meaning of the word. It generates a network of frames that bind to specific relationships. The process of creating or updating a frame can pass through more than one Linguistic expert at different times, which can generate a change in the network structure that does not reflect the conceptual model of the project. The objective of the work is to make an analysis of the frame network structure of FrameNet Brazil and to identify critical points that may reflect in violations of the constraints of the its conceptual model. The results found were analyzed by a specialist in the field of Linguistics and his considerations indicate that the proposed solution is viable in the analysis of the network of FrameNet Brazil.*

Resumo. *A FrameNet Brasil é um projeto de lexicografia computacional que tem como objetivo usar frames para a descrição de significados de palavras. Ela gera uma rede de frames que se ligam com relacionamentos específicos. O processo de criação ou atualização de um frame pode passar por mais de um especialista da Linguística, em momentos diferentes, o que pode gerar uma mudança na estrutura da rede que não reflete o modelo conceitual do projeto. O objetivo do trabalho é fazer uma análise da estrutura da rede de frames da FrameNet Brasil e identificar pontos críticos que possam refletir em violações de restrições do modelo conceitual da mesma. Os resultados encontrados foram analisados por um especialista da área da Linguística e as suas considerações indicam que a solução proposta é viável no que se refere à análise da rede da FrameNet Brasil.*

1. Introdução

Redes complexas podem ser definidas como grafos que apresentam topologias complexas [Barabási et al. 2000]. Elas podem ser usadas como abstrações para representar situações do mundo real onde existam relacionamentos entre pares de entidades envolvidas, como por exemplo palavras de uma frase e as ligações entre elas.

A FrameNet Brasil[1] é um laboratório de Linguística Computacional, cuja missão é desenvolver soluções computacionais para problemas de Processamento de Linguagem

[1]http://www.ufjf.br/framenetbr/

Natural, usando Semântica de *frames*. O projeto central do laboratório é a manutenção da FrameNet[2], que é uma rede complexa formada por *frames* - estruturas cognitivas que definem situações, objetos ou eventos através de cenários - e seus relacionamentos.

A criação de *frames* é feita por linguistas, através de softwares específicos. Contudo, devido a características inerentes ao projeto, um *frame* pode ser atualizado por diferentes especialistas. As mudanças efetuadas podem violar as restrições do modelo conceitual da FrameNet, visto que parte do processo de criação e/ou atualização dos *frames* apela para a intuição do linguista [Fillmore et al. 2003b]. Assim sendo, este artigo se propõe a fazer uma análise da estrutura da rede de *frames*, com o objetivo de identificar situações que possam refletir em violações de restrições do modelo conceitual da FrameNet Brasil. As características identificadas foram avaliadas por um especialista como uma forma de avaliação da solução proposta.

Como contribuições deste trabalho podemos destacar: (i) a concepção de uma abordagem para análise da FrameNet Brasil, possibilitando a identificação de pontos críticos que devem ser analisados pelos especialistas; (ii) a caracterização dos elementos dessa rede; e (iii) o uso de métricas de redes complexas para extrair informações sobre a estrutura da rede.

O restante do artigo está organizado da seguinte forma: na seção 2 são contextualizados o projeto FrameNet bem como os conceitos associados a ele. O Modelo Conceitual é descrito na seção 3 e a proposta do artigo na seção 4. As considerações finais e trabalhos futuros são apresentados na seção 5.

2. Referencial Teórico

Nesta seção é apresentada a FrameNet, projeto precursor da FrameNet Brasil, e são definidos os conceitos necessários para o entendimento deste artigo.

2.1. FrameNet

A FrameNet, também chamada de FameNet de Berkeley, é um

> "projeto de lexicografia computacional que extrai informações sobre links semânticos e sintáticos vinculados às palavras em inglês de grandes corpus de textos eletrônicos, usando procedimentos manuais e automáticos, e apresenta esta informação em uma variedade de relatórios baseados na web" [Fillmore et al. 2003a, p. 1].

Ela foi criada em 1997, liderada por Charles J. Fillmore, e surgiu do cruzamento da Semântica de *Frame* com Lexicografia (geração de dicionário). Atualmente ela foi expandida para outras línguas como Espanhol, Alemão, Sueco, Letão, Japonês, Chinês, Coreano e Português brasileiro.

O projeto tem como objetivo organizar as descrições lexicográficas por *frames* e usar os dados extraídos de corpus para descobrir todas as funções semânticas e propriedades gramaticais das palavras - *unidades lexicais* - que evocam o mesmo *frame* [Salomao 2009].

A Semântica de *Frames*, ou semântica da compreensão, foi gerada como uma abordagem para solucionar problemas da semântica lexical (dar significados as

[2]http://webtool.framenetbr.ufjf.br/index.php/fnbr/report/frame/main

palavras) [Salomao 2009]. A ideia central é que os significados das palavras devem ser descritos em relação aos *frames* semânticos, que são "representações esquemáticas das estruturas conceituais e padrões de crenças, práticas, instituições, imagens, etc., que fornecem uma base para uma interação significativa em uma determinada comunidade de fala" [Fillmore et al. 2003a]. Neste sentido, os *frames* são pacotes de conhecimento que moldam e permitem que os humanos deem sentido às suas experiências [Fillmore and Baker 2010]. Os *frames* que são evocados se baseiam no conhecimento que temos sobre os fenômenos e sua associação com os valores culturais. Um exemplo é descrito a seguir:

Mary foi convidada para a festa de Jack. Ela se perguntou se ele gostaria de um brinquedo.

Os *frames* evocados na frase acima estão ancorados ao verbo `convidar` - que indica um relacionamento envolvendo um anfitrião, um convidado e uma ocasião - e ao substantivo `festa` - que evoca um evento social geralmente com um anfitrião, convidados e uma ocasião. O trecho *festa de Jack* remete a uma festa na qual o Jack é o anfitrião ou na qual Jack é celebrado. Não se verificam implicações linguísticas que evoquem diretamente o *frame* `festa de aniversário` porém o substantivo *brinquedo*, a preocupação *se o Jack iria gostar de um* e os outros detalhes fornecidos pela linguagem, permitem que o leitor infira que o Jack é o aniversariante, que brinquedo é o presente de aniversário, que a Mary é a convidada e assim por diante.

2.2. Estrutura semântica da FrameNet

A FrameNet é constituída por *unidades lexicais*, *frames*, *elementos de frame* e *relacionamentos*. Nesta seção serão apresentados, de forma detalhada, cada um destes elementos, que são usados na descrição da rede formada pela FrameNet Brasil.

i) *Unidade Lexical (UL)*
Uma *Unidade Lexical - UL* é a palavra quando lhe é atribuída um dos seus significados, ou seja, é um emparelhamento de uma palavra com um significado que pertence a um *frame* [Ruppenhofer et al. 2016]. Uma palavra com quatro significados é tratada como quatro unidades lexicais e, na maioria dos casos, ela pode pertencer a mais de um *frame* [Fillmore et al. 2004]. Dizemos que a palavra *evoca* um *frame* quando o significado dela é baseado no *frame*.
A meta é que toda *UL* evoque um *frame* porém, ela deve destacar algum elemento desse *frame* de forma particular [Salomao 2009]. Por exemplo, o *frame* `Aplicar-Calor` é o que descreve uma situação envolvendo cozinha, comida e um instrumento de aquecimento, e ele é evocado pelas palavras *assar, cozer, ferver, secar, borbulhar, corar, dourar, grelhar, vapor, etc*. Essas palavras são chamadas de *Unidade Lexicais*.

ii) *Frames*
Um *Frame* é uma "estrutura conceitual que descreve um tipo particular de situação, objeto ou evento, juntamente com seus participantes e adereços" [Ruppenhofer et al. 2016]. É um sistema de conceitos relacionados de modo que para compreender qualquer um deles, é necessário compreender o sistema como um todo. Ele é composto por elementos que ajudam a completar o seu significado

(*Elementos de Frame*), evocado por uma *Unidade Lexical* e se liga a outros *frames* com relacionamentos específicos.

iii) *Elementos de Frames (EF)*
Os *Elementos de Frame - EF* são atributos usados como etiquetas para as palavras ou frases que estão na construção gramatical com as *UL* que evocam o *frame* [Fillmore et al. 2004]. São os papéis semânticos das entidades envolvidas em cada *frame* [Fillmore et al. 2003a].
Os *EF* existem na estrutura do *frame* porém, podem ou não estar representados na frase em que o *frame* é evocado. Por exemplo o *frame* `Danificar` - que é definido como um *Agente* que afeta um *Paciente* de modo que este mude para um estado não-canônico - possui como principais *EFs* os atributos:

- *Agente:* A entidade consciente, geralmente uma pessoa, que realiza a ação intencional que resulta no dano ao *Paciente*;
- *Paciente:* A entidade que é afetada pelo *Agente*, para que esteja danificada;
- *Causa:* Um evento que leva ao dano do *Paciente*.

iv) *Tipos Semânticos*
O principal objetivo de usar *Tipos Semânticos* na FrameNet é de demonstrar informações que não são bem representadas na hierarquia de *frames* [Fillmore et al. 2003a]. Eles também podem ser usados para expressar importantes diferenças semânticas entre *ULs* que se repetem em vários *frames* [Ruppenhofer et al. 2016]. Por exemplo, as *ULs* de um *frame* podem possuir avaliações positivas - *louvar* do *frame* `Julgamento`, *gostar* do *frame* `Sujeito_experimentador` - ou negativas - *criticar* do *frame* `Julgamento`, *detestar* do *frame* `Sujeito_experimentador`.

Atualmente a FrameNet cobre mais de 13.000 *Unidades Lexicais*, distribuídas em mais de 1.200 *Frames* e atestadas por mais de 200.000 frases anotadas.

2.3. Relacionamentos

A FrameNet é uma rede de *frames* em que as ligações ocorrem através de relações específicas (*Herança, Perspectiva, Uso, Subframe e Precedência*). As relações são usadas para melhorar a compreensão dos *frames* e proporcionar robustez (visto que *frames* semanticamente similares podem estar associados, apesar de estarem separados) [Ruppenhofer et al. 2016]. A Figura 1 ilustra relacionamentos entre *frames* e seguir são descritos cada um deles.

i) *Herança*
É o relacionamento mais forte na FrameNet [Ruppenhofer et al. 2016]. Ocorre entre um *frame* pai e um *frame* filho, onde *frame* filho herda todos ou parte dos *EFs*, *subframes* e tipos semânticos do pai. Os *EFs* do *frame* filho não têm, necessariamente, os mesmos nomes dos *EFs* do pai e podem ser adicionados outros, dadas as especificidades do *frame* filho. Por exemplo, o *frame* `Fornecer` herda do *frame* `Dar` e, além de possuir os *EFs Tema* e *Destinatário*, ele também especifica que o doador é um *Fornecedor* e que tem que haver o *Objetivo* do tema proposto.

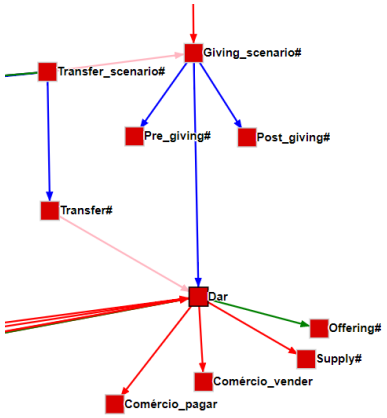


Figura 1. Relacionamentos do *frame* `Dar` com os outros *frames*

ii) *Uso*
É a relação usada quando parte do cenário evocado pelo *frame* filho se refere a um *frame* pai. É um tipo de relacionamento de herança, onde o *frame* filho pode usar (herdar de) múltiplos *frames* pai. Por exemplo, o *frame* `Oferta` usa o *frame* `Dar`, já que o *Ofertante* oferece um *Tema* para um *Potencial_Destinatário* e somente quando o *Potencial_Destinatário* aceita a oferta é que a `Transferência` ocorre.

iii) *Perspectiva*
Esta relação é semelhante a de *Uso* e consiste em indicar, pelo menos, dois pontos de vista de um *frame* neutro. Por exemplo, o *frame* `Dar` é uma perspectiva do *frame* `Transferência`, uma vez que os verbos *dar* e *receber* expressam o ponto de vantagem da cena (do *Doador* ou do *Destinatário* respectivamente).

iv) *Subframe*
Essa relação é usada para representar sub-eventos ou subpartes de um *frame* complexo. Os *subframes* comumente se referem à sequência de passos ou transações e podem ser descritos, separadamente, por *frames*. Por exemplo, o *frame* `Dar` juntamente com os *frames* `Pré_Doação` e `Pós_Doação` são *subframes* do *frame* `Cenário_Doação`.

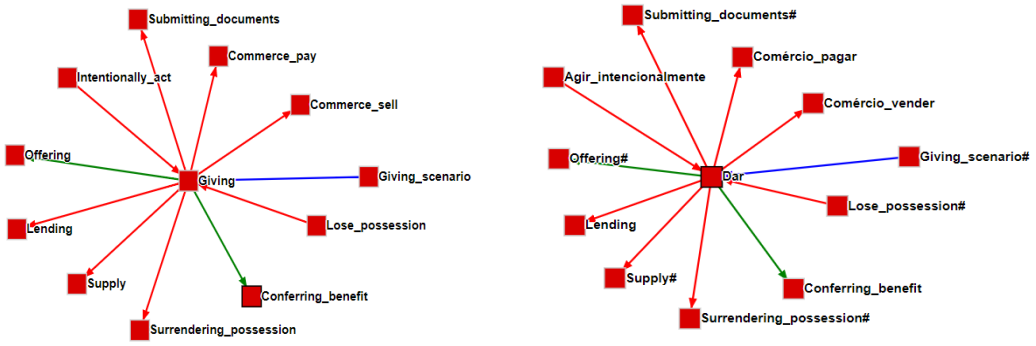
v) *Precedência*
Essa relação ocorre apenas entre dois *subframes* de um *frame* complexo para especificar a sequência dos eventos de um certo cenário. Este é o único relacionamento que pode formar ciclos. No *frame* `Cenário_Doação`, o relacionamento *Precedência* é usado para conectar o *frame* `Pré_Doação` ao *frame* `Dar` e o *frame* `Dar` ao *frame* `Pós_Doação`, pois eles devem ocorrer em ordem.

2.4. FrameNet Brasil

A extensão do projeto inicial para outras línguas inclui o uso do estrutura definida no Inglês - *Unidades Lexicais, Frames, Elementos de Frames e suas conexões* - além da implementação de novas funcionalidades. A FrameNet Brasil focou no desenvolvimento de um léxico, de um *Construction*, de um banco de dados trilíngue para o domínio do

esporte e turismo, e de uma FrameNet Brasil WebTool.

a) Léxico
O Léxico é uma expansão da FrameNet e possui uma rede que relaciona *frames* a *unidades lexicais*. Alguns *frames* não foram alterados no processo de expansão da FrameNet para a FrameNet Brasil, como é o caso do *frame* `Giving` que em português pode ser evocado pela *unidade lexical* `Dar`, onde seus *elementos de frame* (entre outras características) se mantiveram intactos. A Figura 2 mostra o mesmo *frame* evocado pelas *unidades lexicais* `Giving` e `Dar`.



(a) Unidade Lexical `Giving` (b) Unidade Lexical `DAR`

Figura 2. Frame `Giving`

b) *Construction*
O *Construction* é um repositório de construções gramaticais do Português brasileiro. As construções se assemelham a estrutura dos *frames*, contendo uma definição em prosa e um conjunto de *Elementos de Construção* que são os componentes da construção. A principal contribuição do *Construction* da FrameNet Brasil foi a criação de dois tipos de relacionamentos e 5 tipos de restrições na base de dados. Foi gerado o relacionamento *Herança* entre os construtores e o relacionamento *Evoca*, que liga o construtor ao *frame* que ele evoca.

c) Base de dados trilíngue
Uma FrameNet de domínio específico multilingual foi desenvolvida, como prova de que *frames* podem ser usados como representações interlinguais em um dicionário eletrônico trilíngue (português-inglês-espanhol) para o Turismo e para Copa do Mundo de Futebol [Torrent et al. 2014]. O dicionário (`http://framenetbrazildictionary.com`) possui 128 *frames* trilíngues, 1.125 *unidades lexicais* e mais de 13.000 frases anotadas gerando, automaticamente, uma lista de tradução equivalente para todos os verbos e substantivos que indicavam eventos . A Figura 3 mostra a interface do aplicativo do dicionário (`http://www.dicionariodacopa.com.br/`).

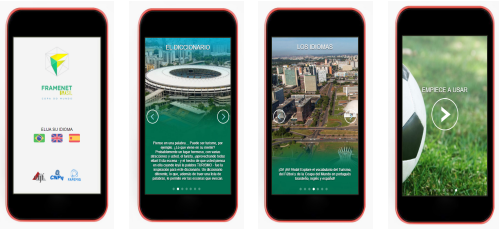


Figura 3. Telas do aplicativo dicionário trilíngue

d) FrameNet Brasil WebTool
O WebTool é o sistema de gerenciamento e anotação do banco de dados. Usa um banco de dados relacional, atualmente implementado no MySQL, que mantém os mesmos conceitos e estrutura usados na FrameNet, para facilitar a migração e o alinhamento de dados.

3. Modelo conceitual da FrameNet Brasil

A FrameNet procura capturar os *insights* humanos em estruturas semânticas de forma eficiente [Baker et al. 1998]. O processo começa com a caracterização do *frame* que a *UL* evoca, definindo o tipo de entidade, situação ou objeto que ele representa, escolhendo os rótulos dos *EFs* e a lista de *ULs* vinculadas a ele [Fillmore et al. 2003b].

Quatro etapas de processamento foram necessárias produzir o banco de dados da FrameNet Brasil:

- *Preparação:* descrições iniciais dos *frames, ULs* e *EFs* e verificação do padrão sintático de cada um, para o uso em consultas do subcorpus e anotação [Baker et al. 1998];
- *Extração de subcorpus:* geração de boas frases de exemplos através de ferramentas da linguística computacional;
- *Anotação:* marcação (a mão) dos *EFs* detectados no subcorpus e identificação de padrões de exemplos e frases com problemas;
- *Escrita da Entrada:* adição dos dados às tabelas do banco.

Os anotadores humanos são os que escolhem os termos que farão parte da lista de *ULs*, examinam o uso dos *EFs* e determinam os contextos sintáticos e colocacionais do significado do *frame* [Fillmore et al. 2003a] consultando dicionários eletrônicos e tesauros. Eles podem, em qualquer etapa de processamento, modificar uma decisão anterior (tomadas ou não por eles) com base em evidências do corpus e continuar o processo a partir desse ponto [Fillmore et al. 2003b]. Essas mudanças também podem implicar na alteração de *frames* já existentes e seus relacionamentos, podendo gerar um *loop* de alterações.

As relações entre os *frames* são implementadas no banco de dados, contendo as tabelas *frame, ULs, EFs* e *relacionamento* [Fillmore et al. 2004]. As tabelas e as relações entre elas procuram refletir a base teórica do projeto [Fillmore et al. 2003a].

4. Análise da Rede de *frames* da FrameNet Brasil

Nesta seção é apresentada a estrutura da rede de *frames* da FrameNet Brasil e as métricas de análise em redes complexas, com o intuito de mostrar como essa análise se reflete no

modelo conceitual da rede. Usou-se o software *Gephi*[3] para análise e visualização da rede e o software *yEd*[4] para a geração das árvores de *frames* (que serão descritas na seção 4.4).

4.1. Caracterização da rede de *frames*

A rede de *frames* é um grafo onde os vértices são os *frames* e as arestas são os relacionamentos entre eles. Possui 1359 nós e 1960 arestas. O grafo é dirigido pois os relacionamentos indicam uma certa hierarquia na rede. Existem *frames* mais genéricos que possuem muitos outros *frames* conectados a eles, gerando assim núcleos ou *clusters* bem definidos no grafo. Esses mesmos *frames* podem ser observados como pais das árvores geradas por suas conexões, onde quanto maior o nível do nó na árvore mais específico é o *frame*. Assim sendo, o grafo da FrameNet Brasil pode ser considerado como uma floresta de *frames* (vide seção 4.4) em que: os pais das árvores descrevem cenários mais gerais, os nós folha cenários mais específicos, os vértices podem se relacionar como irmãos e a altura da árvore pode ser um indicativo de quão bem estudado (aprofundado) foi o assunto em questão. A Figura 4 ilustra o grafo da FrameNet Brasil.

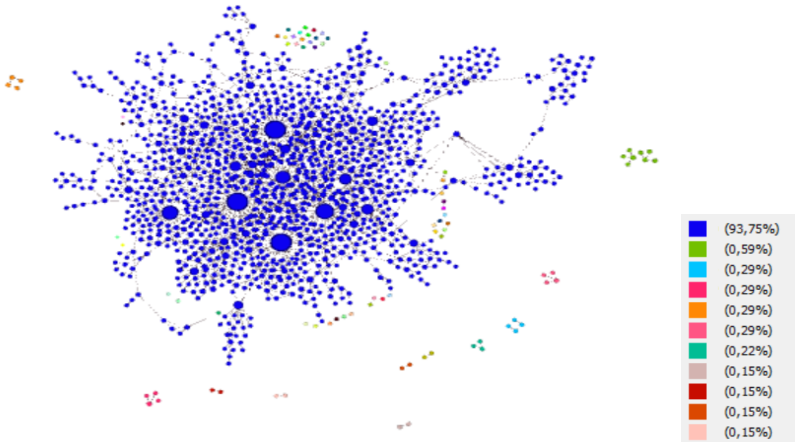


Figura 4. Grafo da rede de *frames* da FrameNet Brasil. O tamanho do vértice foi determinado pelo seu grau. As cores assinalam as componentes conexas.

4.2. Medidas de Centralidade

Foram aplicadas medidas de centralidade com o objetivo de detectar vértices de maior importância no grafo. A Tabela 1 lista essas medidas e o intervalo em que seus valores ocorrem.

Tabela 1. Medidas de centralidade e o intervalo de valores. O *Closeness, Betweenness* e o *Eigenvector* possuem os valores contidos entre [0,1]

Intervalo	Grau	G. Entrada	G. Saída	Closeness	Betweenness	Eigenvector
Menor	0	0	0	0.0	0.0	0.0
Maior	54	6	53	1.0	0.001601	1.0

[3]https://gephi.org/
[4]https://www.yworks.com/products/yed

O grau do nó indica a quantidade de relacionamentos que o vértice possui. Por ser um grafo dirigido, o valor do grau é a soma do grau de entrada e o de saída. O grau de saída foi usado para revelar vértices que poderiam apresentar um certa relevância no grafo, por possuírem valores maiores que o grau de entrada (como pode ser verificado na Tabela 1). Os vértices com maior grau de saída são os que possuem muitas relações de *herança* (por representarem cenários mais genéricos) e que, por essa razão, podem ser considerados como vértices críticos, visto que a sua remoção implicaria na fragmentação da árvore de *frames*, bem como do grafo. Foram identificados tanto vértices com grau (de entrada e de saída) nulo quanto vértices com grau de entrada maior que 1, e esses casos serão mais explorados adiante.

O *Closeness* foi usado para identificar os vértices mais centrais do grafo porém, os vértices com *Closeness* igual à 1 são os que possuem grau menor que 9; ou seja, são os vértices periféricos ou do penúltimo nível das árvores (pais das folhas). Isso ocorre por que o *Closeness* é uma medida que calcula a proximidade do vértice em relação aos outros que ele alcança, e nessa camada da rede os vértices alcançados são apenas os adjacentes.

O *Betweenness* foi usado para detectar os vértices *ponte* do grafo. O intervalo superior dessa medida é menor que 0,1. Isso acontece por que o grafo da FrameNet Brasil tende a uma estrutura de floresta de *frames* existindo poucos caminhos curtos que passam pelo mesmo vértice.

O *Eigenvector* foi usado para reconhecer vértices importantes com base na importância de seus vértices adjacentes, isto é, identificar vértices adjacentes a mais de um nó central. Contudo, como foi constatado anteriormente, os nós mais centrais deste grafo são os nós periféricos e, por esta razão, apenas um vértice apresentou o *Eigenvector* igual à 1.

4.3. Conectividade da rede

Visto que as medidas de centralidade não permitiram obter uma visão clara da estrutura da rede, optou-se por aplicar dois algoritmos de agrupamento (detecção de Componentes Conexas e de Comunidades) fornecidos pelo *Gephi*.

Para detecção de componentes conexas aplicou-se o algoritmo *Componentes Conexas* [Wasserman and Faust 1994]. Foram identificadas 60 componentes conexas, sendo que a maior engloba mais de 90% dos vértices (1274 *frames*). Nessa componente se encontram os vértices com maior grau e as árvores de *frames* mais altas. Das 59 componentes restantes 48 possuem apenas um vértice, isto é, existem vértices sem arestas no grafo o que indica que alguns *frames* estão sem relacionamentos na FrameNet Brasil. Estes *frames*, que representam 3% do total, deveriam ser analisados por especialistas para verificar se essa peculiaridade é aceita ou não semanticamente. A Figura 4 ilustra as componentes do grafo, onde as cores distinguem os vértices pertencentes a mesma componente.

Para detecção de comunidades aplicou-se o algoritmo de *Modularidade*, que consiste em agrupar os vértices de acordo com o valor da modularidade da partição [Blondel et al. 2008] (usou-se o valor de resolução=10 como parâmetro de entrada do algoritmo no *Gephi*). Foram detectadas 82 comunidades no grafo. Estas comunidades são grupos com ligações muito densas entres os vértices participantes. Os dois maiores grupos contêm dois dos vértices de maior grau e possuem mais de 90 membros, respectivamente.

Como pode ser observado na Figura 5, os vértices desconexos constituem também comunidades compostas de um único elemento.

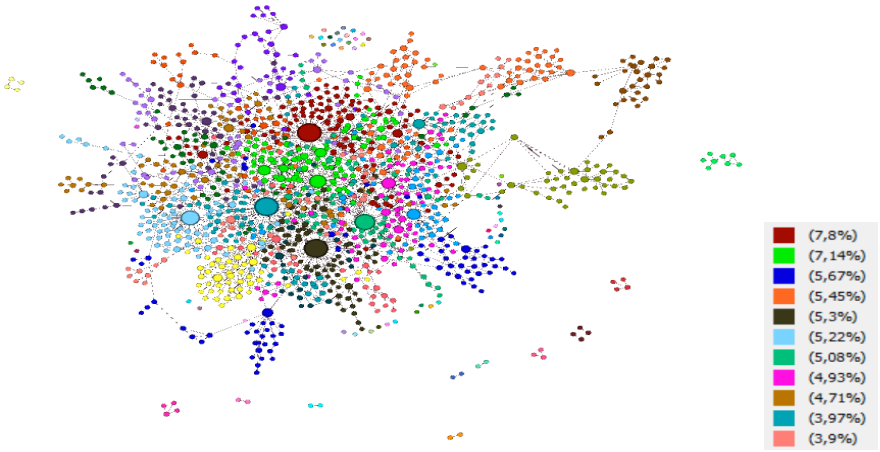
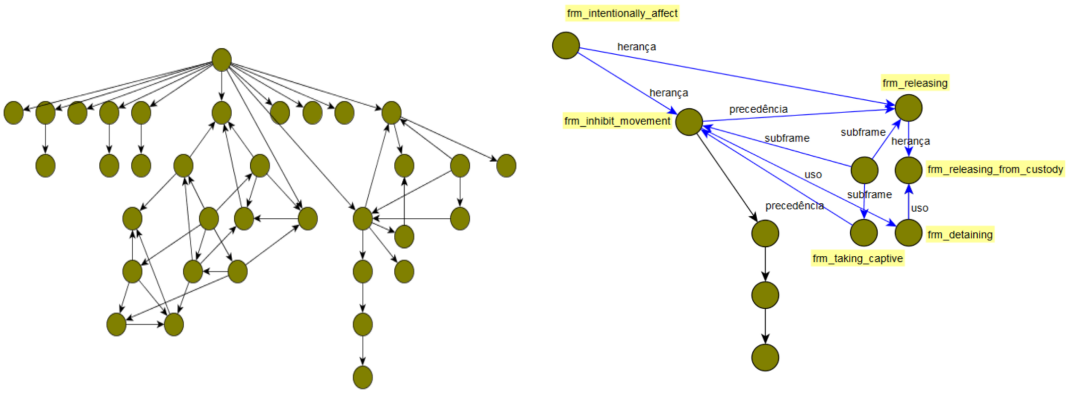


Figura 5. Comunidades do grafo

4.4. Árvores de *frames*

As ligações entre os *frames* que compõem um cenário geram uma estrutura hierárquica similar a de uma árvore, onde o nó pai é o *frame* mais geral e o nó folha o mais específico. A Figura 6(a) mostra a árvore gerada por uma das componentes conexas do grafo. Computacionalmente, os relacionamentos entre os *frames* violam as restrições da árvore (como a estrutura de dados) porém, semanticamente essas infrações podem fazer sentido para a descrição do assunto em questão (Figura 6(b)). Cabe a um especialista analisar e validar tais situações ou corrigir caso necessário.



(a) Árvore

(b) Relacionamentos entre os nós de uma das subárvore

Figura 6. Visualização em árvore de *frames* de uma das componentes conexas do grafo

4.5. Avaliação da proposta

As medidas e algoritmos apresentados nas seções anteriores foram empregados com o objetivo de identificar possíveis pontos críticos da rede. Foram detectados alguns casos,

a nível estrutural, e os mesmos foram encaminhados a um especialista em linguística computacional, para as respectivas avaliações. Doutor em Linguística Computacional, o especialista atua no projeto da FrameNet Brasil. Alguns casos de criticidade foram descartados, instantaneamente, e outros mereceram uma análise semântica mais profunda. Abaixo são descritos os casos e suas respectivas avaliações e/ou validações.

i) Medidas de centralidade
O grau do nó se revelou uma métrica válida para detecção de nós relevantes tanto a nível topológico quanto a nível semântico da rede. Nós com graus altos, quando eliminados da rede, tendem a gerar mais componentes conexas e o mesmo ocorre semanticamente, fazendo com que os *frames* adjacentes percam uma parte da informação associada a eles. Por exemplo, se o *frame* `Dar` for retirado da rede todos os *frames* adjacentes a ele perderão parte do significado (Figura 2). Esta medida foi validada de forma instantânea, contudo, o emprego do *Closeness, Betweennes* e *Eigenvector*, que são medidas que revelam o privilégio do nó em relação à sua localização, não se mostrou tão proveitoso. Os nós com *Closeness, Betweennes* e *Eigenvector* altos são, na maioria dos casos, nós periféricos da rede. Por serem casos atípicos à realidade computacional, essas medidas demandaram uma análise semântica mais acentuada. Todavia, foram descartadas por não oferecerem risco semântico à FrameNet.

ii) Conectividade
O algoritmo de detecção de componentes conexas evidenciou grupos pequenos de *frames* desconexos do restante da rede e até mesmo um número considerável de *frames* soltos (41), sem nenhum tipo de relacionamento. Quando apresentados, o especialista afirmou que estes casos podem ocorrer se o *frame* descrever um cenário muito específico. Todavia, essa subparte da rede não deixa de ser um motivo de alerta e receberá especial atenção dos linguistas do projeto.

iii) Árvore de *frames*
A estrutura de dados que mais se aproxima da rede de *frames* é a árvore, como já foi discutido na seção 4.4, porém nem todas as propriedades desta estrutura são respeitadas na FrameNet Brasil: filhos com mais de um pai, irmãos com ligações diretas (sem passar pelo pai), ciclos entre nós de uma subárvore dentre outros, são os casos que podem ser verificados nessa rede. O especialista afirmou que prender a FrameNet Brasil a uma estrutura tão rígida limita o projeto no que tange a representação dos mais variados cenários da vida cotidiana, bem como na associação de informações que agreguem valor semântico ao *frame*. No entanto, quando foram apresentadas as subárvores com restrições violadas, foram necessárias algumas análises semânticas para validar os casos. A Figura 6(b) ilustra um exemplo de *frames* irmãos se relacionando sem passar, necessariamente, pelo *frame* pai: os *frames* `inibir_o_movimento` e `liberar` *herdam* do *frame* `afetar_intencionalmente`, porém eles se ligam pela relação de *precedência* que define a ordem em que os fatos ocorrem.

5. Considerações Finais

Este trabalho apresentou uma análise da rede de *frames* da FrameNet Brasil, com o objetivo de identificar situações que pudessem estar em desacordo com o modelo

conceitual da mesma.

Foram aplicadas medidas de centralidade e algoritmos de agrupamento na análise da rede. Não foram encontrados casos explícitos de violação do modelo conceitual da FrameNet Brasil porém, foram identificados cenários atípicos que foram repassados aos profissionais do projeto para uma análise detalhada.

Como trabalhos futuros pretende-se avançar na análise da rede de *frames* usando métricas mais sofisticadas de redes complexas e usar a FrameNet Brasil como uma ferramenta de Análise de Sentimentos.

Referências

[Baker et al. 1998] Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998). The berkeley framenet project. In *Proceedings of the 17th international conference on Computational linguistics-Volume 1*, pages 86–90. Association for Computational Linguistics.

[Barabási et al. 2000] Barabási, A.-L., Albert, R., and Jeong, H. (2000). Scale-free characteristics of random networks: the topology of the world-wide web. *Physica A: statistical mechanics and its applications*, 281(1-4):69–77.

[Blondel et al. 2008] Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008.

[Fillmore and Baker 2010] Fillmore, C. J. and Baker, C. (2010). A frames approach to semantic analysis. In *The Oxford handbook of linguistic analysis*.

[Fillmore et al. 2004] Fillmore, C. J., Baker, C. F., and Sato, H. (2004). Framenet as a"net". In *LREC*.

[Fillmore et al. 2003a] Fillmore, C. J., Johnson, C. R., and Petruck, M. R. (2003a). Background to framenet. *International journal of lexicography*, 16(3):235–250.

[Fillmore et al. 2003b] Fillmore, C. J., Petruck, M. R., Ruppenhofer, J., and Wright, A. (2003b). Framenet in action: The case of attaching. *International journal of lexicography*, 16(3):297–332.

[Ruppenhofer et al. 2016] Ruppenhofer, J., Ellsworth, M., Petruck, M. R., Johnson, C. R., and Scheffczyk, J. (2016). *FrameNet II: Extended theory and practice*. Institut für Deutsche Sprache, Bibliothek.

[Salomao 2009] Salomao, M. M. M. (2009). Framenet brasil: um trabalho em progresso. *Calidoscópio*, 7(3):171–182.

[Torrent et al. 2014] Torrent, T., Salomão, M. M., Campos, F., Braga, R., Matos, E., Gamonal, M., Gonçalves, J., Souza, B., Gomes, D., and Peron, S. (2014). Copa 2014 framenet brasil: a frame-based trilingual electronic dictionary for the football world cup. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations*, pages 10–14.

[Wasserman and Faust 1994] Wasserman, S. and Faust, K. (1994). *Social network analysis: Methods and applications*, volume 8. Cambridge university press.

Análise das Interações Sociais em Comunidades Online de Aprendizado de Idiomas: um estudo de caso no Reddit*

Rafael Sales Medina, Ana Paula Couto da Silva, Fabricio Murai

[1]Departamento de Ciência da Computação
Universidade Federal de Minas Gerais (UFMG) – Belo Horizonte, MG – Brazil

{rafael.medina, ana.coutosilva, murai}@dcc.ufmg.br

Abstract. *Reddit is a online social network in which users can share information about mutual interests in specific communities (subreddits). Recently, communities focused on language learning have gained much popularity among users. These subreddits enable users to interact, regardless of their proficiency level on a specific language. Typical interactions include answering questions and sharing tips for facilitating the learning process. In this paper, we analyze four of these communities: EnglishLearning, French, German and Spanish. This analysis focuses on interactions between users, how discussion revolves around threads and linguistic traits of users belonging to different proficiency levels. Moreover, we highlight similarities and differences among these communities.*

Resumo. *Reddit é uma rede social online na qual os usuários podem trocar informação sobre interesses comuns em comunidades específicas (subreddits). Recentemente, comunidades voltadas para o aprendizado de idiomas vêm ganhando popularidade. Esses subreddits permitem que usuários interajam, independentemente do seu nível de proficiência. Interações típicas incluem a resolução de dúvidas e a troca de dicas para facilitar o aprendizado. Neste trabalho, analisamos quatro comunidades: EnglishLearning, French, German e Spanish. Esta análise foca em interações entre usuários, como as discussões se desenrolam em torno das threads e os traços linguísticos dos usuários que pertencem a diferentes níveis de proficiência. Além disso, ressaltamos as semelhanças e diferenças entre essas comunidades.*

1. Introdução

Nos últimos anos, as redes sociais online têm sido utilizadas para diversas finalidades, como manter contato com amigos antigos, fazer novas amizades, compartilhar atualizações [Joinson 2008] e até mesmo para formação de grupos de apoio virtuais, como aqueles voltados para emagrecimento [Pappa et al. 2017, Cunha et al. 2016] e para vítimas de abuso sexual [Andalibi et al. 2016].

Uma rede que permite o contato entre pessoas em torno de um tema de interesse comum é o Reddit[1], um site de fóruns com características de redes sociais onde os membros compartilham suas experiências e dúvidas sobre os mais diversos assuntos. O Reddit é organizado em comunidades (subreddits) e, em 2017, era formado por 1.204.126 comunidades com 900 milhões de comentários. Existem subreddits específicos para discutir

*The authors' work has been partially funded by CAPES, CNPq and FAPEMIG.

[1]http://www.reddit.com

programas de televisão[2], jogos de computador[3] e até mesmo para compartilhamento de tópicos relacionadas à saúde, como dicas de emagrecimento[4] e suporte mútuo em relação a problemas de saúde mental, como aqueles estudados em [De Choudhury and De 2014].

Em particular, comunidades específicas voltadas para o aprendizado de idiomas estrangeiros vem ganhando muita popularidade, como EnglishLearning[5] (será abreviada como English), French[6], German[7] e Spanish[8]. Esses subreddits permitem que pessoas com interesse em certo idioma interajam, compartilhando dúvidas, sugestões e dicas, tornando o processo de aprendizado mais dinâmico e interessante.

Como o aprendizado online e as redes sociais online atraem cada vez mais a atenção das pessoas no mundo inteiro, neste trabalho caracterizamos as atividades e interações dos usuários de subreddits voltados para o aprendizado de idiomas. Mais precisamente, nossas análises têm como objetivo responder as seguintes perguntas de pesquisa:

- **QP1:** As interações entre usuários em um subreddit são semelhantes àquelas em uma rede social tradicional?
- **QP2:** Como as publicações em um subreddit estão distribuídas em relação às threads?
- **QP3:** Existem diferenças linguísticas no texto de usuários com diferentes níveis de proficiência?

Para responder estas questões, modelamos a rede de usuários e suas interações dentro de um subreddit como um grafo, seguindo trabalhos recentes da literatura [Pappa et al. 2017]. Métricas como o *closeness*, o grau médio de entrada e saída e o coeficiente de clusterização são usadas para analisar os padrões de comportamento dos usuários que participam dos subreddits analisados. As publicações feitas pelos usuários são analisadas através da definição de árvores de discussão, que permitem investigar características importantes como a profundidade e a largura das *threads*. Adicionalmente, utilizamos a ferramenta LIWC (*Linguistic Inquiry and Word Count*)[9] para identificar diferenças linguísticas nos *posts* de membros do subreddit *German* associados a diferentes níveis de proficiência. Nossos resultados são importantes para a definição de novos modelos de aprendizado de idiomas apoiado por tecnologia considerando, por exemplo, a evolução da proficiência e o perfil de interação de usuários participantes destes subreddits.

Este artigo está organizado da seguinte forma: a Seção 2 descreve os principais trabalhos relacionados; a Seção 3 detalha os dados e métodos utilizados; os resultados da análise dos subreddits são apresentados na Seção 4; as implicações deste trabalho e trabalhos futuros são discutidos na Seção 5.

2. Trabalhos Relacionados

Há pelo menos duas décadas é possível encontrar trabalhos voltados para o aprendizado de idiomas apoiado por tecnologias [Zhao 1996, Warschauer and Healey 1998]. Esse campo

[2]http://www.reddit.com/r/rupaulsdragrace
[3]http://www.reddit.com/r/thesims
[4]http://www.reddit.com/r/loseit
[5]http://www.reddit.com/r/EnglishLearning/
[6]http://www.reddit.com/r/French/
[7]http://www.reddit.com/r/German/
[8]http://www.reddit.com/r/Spanish/
[9]http://liwc.wpengine.com/

de estudos é chamado de *Computer Assisted Language Learning* (CALL) [Levy 1997]. CALL engloba quaisquer tipos de aplicações que possam auxiliar no aprendizado de um idioma estrangeiro, como as redes sociais online, que são o foco deste trabalho. O interesse dos pesquisadores na área de redes sociais é relativamente recente, como descrito por [Zourou 2012]. Neste mesmo trabalho, a autora discute o estado da arte em relação ao uso de mídias sociais para o ensino de idiomas, mas sem especificar uma rede ou linguagem. Ela demonstra que as redes têm influência positiva no ensino e são bastante utilizadas por fomentarem participação dos usuários.

Os autores de [Arnold and Paulus 2010] analisam, sob a visão de estudantes, de um instrutor e de um observador externo, um caso prático em que uma turma real de aprendizado de idioma utilizou o sistema Ning, voltado para a criação de comunidades sociais. Este trabalho conclui, sob a visão do instrutor de idioma, que a utilização da comunidade social para discussão teve resultado positivo no ensino.

Sob a ótica de pessoas que estão aprendendo um novo idioma, a pesquisa realizada em [Lin et al. 2016] analisa o comportamento e desenvolvimento de usuários de redes específicas para aprendizado de idiomas. O estudo, feito por meio de questionários e acompanhamento de estudos de casos, conclui que as redes sociais voltadas para o aprendizado de idiomas apresentam resultados positivos, mas também limitações. Além disso, conclui-se que para que os usuários alcancem o sucesso esperado, é necessário que as redes ofereçam apoio, orientação e atividades bem estruturadas, de maneira a promover o engajamento e interação dos usuários.

Apesar de existirem abordagens à utilização de redes sociais para o auxílio do aprendizado de idiomas, os trabalhos citados anteriormente focam em redes criadas especificamente para o escopo de ensino e aprendizado. O nosso trabalho analisa como os usuários do Reddit, que é uma rede social composta por comunidades criadas em torno dos mais diversos tópicos, pode auxiliar no aprendizado de um determinado idioma. A partir do estudo as interações dos usuários, dos seus interesses e como os mesmos se organizam em torno de tópicos em comum, fóruns e comunidades poderão ser criados, visando um resultado mais positivo no aprendizado de novos de idiomas. Segundo o nosso conhecimento, este é o primeiro trabalho a investigar comunidades no Reddit que focam no aprendizado de idiomas.

3. Metodologia

Apresentamos abaixo os métodos utilizados neste trabalho para o estudo de subreddits focados no aprendizado de idiomas. Em particular, descrevemos as técnicas usadas para coleta e extração dos dados, modelagem da interação entre usuários e análise textual.

3.1. Coleta e extração dos *datasets*

Os dados do Reddit analisados neste trabalho foram obtidos a partir de *dumps* disponíveis na Web[10]. Coletamos todas as atividades realizadas por usuários nos subreddits English, French, German e Spanish entre janeiro de 2010 e dezembro de 2016. As atividades consistem em *posts* e comentários feitos pelos usuários das comunidades.

A Tabela 1 apresenta o número de usuários, o total de *posts* e comentários observados durante este intervalo, bem como os valores médios por usuário, para cada um

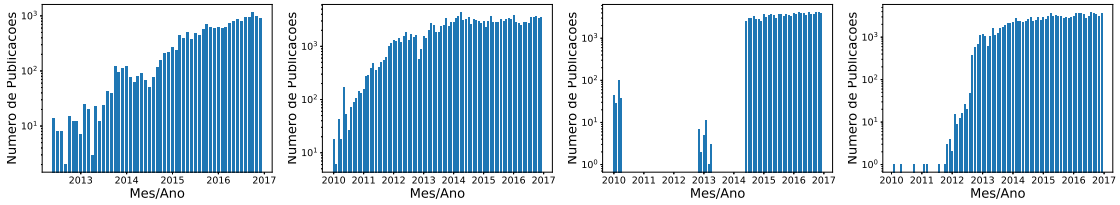
[10]http://files.pushshift.io/reddit/

dos subreddits. A comunidade voltada para ensino de inglês é aquela que tem o menor número de usuários. Um dos motivos que provavelmente contribui para que isto aconteça é que a maior parte dos visitantes do Reddit é de países de língua inglesa. Em dezembro de 2017, os três países seguintes correspondiam a mais de metade dos usuários ativos[11]: Estados Unidos (39,79%), Inglaterra (7,16%) e Austrália (3,46%).

	EnglishLearning	French	German	Spanish
Usuários	3.737	18.113	12.235	13.648
Posts	5.493	17.329	12.441	14.295
Comentários	11.967	145.700	96.265	113.843
Média de *posts* por usuário	1,47	0,96	1,02	1,05
Média de comentários por usuário	3,20	8,04	7,87	8,34

Tabela 1. Estatísticas dos subreddits no período entre 2010 e 2016.

A Figura 1 mostra o volume total de publicações nos subreddits a cada mês. Observa-se que os subreddits English e French já ganhavam popularidade desde 2010. No German houve algumas rajadas de atividade em 2010 e 2013, tendo-se observado um salto abrupto no volume de publicações durante 2014. No Spanish também observou-se um salto, embora menos pronunciado, em 2012. Em todos os casos, o volume de atividades em 2016 se manteve relativamente constante.



(a) English (b) French (c) German (d) Spanish

Figura 1. Volume mensal de publicações (*posts* e comentarios).

3.2. Modelo de Interação entre Usuários

Para cada subreddit, modelamos as interações entre os usuários como um grafo direcionado ponderado $G_d = (V, E_d, W_d)$, onde V é um conjunto de vértices, E_d é um conjunto de arestas e W_d é uma função que mapeia cada aresta $e \in E_d$ a um peso $W_d(e) \in \mathbb{R}$. Cada vértice representa um usuário que tenha publicado um *post* ou comentário no subreddit, e cada aresta indica uma interação entre dois usuários. As arestas são direcionadas: v aponta para u se o vértice v respondeu a um *post* ou comentário de um vértice u. Para cada aresta $e = (i, j) \in E_d$, o peso $W_d(e)$ é igual ao número de interações de i com j.

Definimos também o grafo não-direcionado ponderado $G = (V, E, W)$ induzido por G_d ao tornarmos as arestas em E_d não-direcionadas e fazermos $W(e) = W_d(e) + W_d(e')$, para $e = (i, j)$ e $e' = (j, i)$.

Utilizamos o grafo direcionado G_d para caracterizar um subreddit quanto à distribuição do volume de atividades dos seus usuários medido em termos (i) dos graus

[11] https://www.statista.com/statistics/325144/reddit-global-active-user-distribution

de entrada e (ii) de saída e (iii) do *closeness*. O *closeness* é uma métrica clássica de centralidade em redes que tenta capturar a importância relativa dos nós a partir da estrutura do grafo [Newman 2011]. Além disso, usamos também o grafo não-direcionado G para calcular a distribuição do coeficiente de clusterização dos nós.

A Tabela 3.2 apresenta os dados básicos dos grafos de interação G_d, incluindo a quantidade de vértices (usuários) e arestas, de componentes conexos e o tamanho do maior componente conexo de cada rede.

	EnglishLearning	French	German	Spanish
Vértices	3.737	18.113	12.235	13.648
Arestas	8.881	95.152	66.309	74.108
Número de componentes conexos	888	1.435	776	12.14
Vértices no maior componente	2.804	16.607	11.425	12.398
Arestas no maior componente	6.516	72.231	49.814	56.463

Tabela 2. Características dos grafos de interação.

3.3. Análise dos *Posts* e Comentários

Cada *thread* em um subreddit pode ser vista como uma árvore iniciada por um post (nó raiz) que pode ser respondido diretamente por comentários. Cada comentário pode, por sua vez, ser respondido por outros comentários. Denominamos **árvores de discussão** as árvores que reconstruímos ao mapearmos cada comentário em um subreddit ao seu "nó pai". Iremos calcular a profundidade e a largura destas árvores a fim de identificar os tópicos que despertam o maior interesse dos usuários. Além disso, considerando os *timestamps* associados aos nós, iremos mensurar a taxa de crescimento destas árvores.

Adicionalmente é possível analisar o texto das publicações com o auxílio da ferramenta LIWC (*Linguistic Inquiry and Word Count*)[12], que realiza a análise automatizada de textos em diversas línguas e os classifica em diferentes categorias. Essa análise permite relacionar características linguísticas do texto, baseadas nos resultados do LIWC, com a proficiência indicada pelo usuário em seu perfil.

4. Análise dos resultados

Nesta seção descrevemos os resultados obtidos através dos métodos descritos na Seção 3 e explicar como eles respondem as questões de pesquisa levantadas neste trabalho.

4.1. QP1: As interações entre usuários nesses subreddits são semelhantes àquelas em uma rede social tradicional?

A partir de G_d, calculamos a distribuição conjunta de graus de entrada e de saída dos vértices. O grau de entrada de um vértice é a quantidade de comentários que o usuário correspondente recebeu em suas publicações. Por outro lado, o grau de saída de um vértice é a quantidade de publicações feitas por um usuário. A distribuição conjunta é mostrada na Figura 2 através de um mapa de calor em escala log-log, onde a cor do ponto (i, j) indica a fração de vértices em G_d com grau de entrada $i - 1$ e grau de saída $j - 1$.

[12]http://liwc.wpengine.com/

Para todos os subreddits, observamos uma distribuição de cauda pesada em relação a ambas as distribuições marginais, além de uma forte correlação entre grau de entrada e grau de saída. Em sua grande maioria, usuários fazem e recebem poucos comentários, enquanto alguns poucos que fazem muitos comentários também recebem muitas respostas em suas publicações. Em particular, o subreddit English se destaca por apresentar pouquíssimos nós com graus de entrada ou saída maiores que 10^2, o que pode ser explicado por ter menos usuários que outras comunidades.

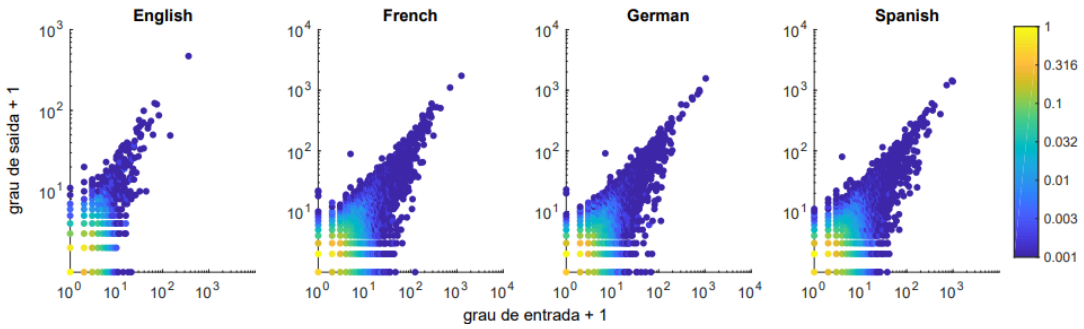
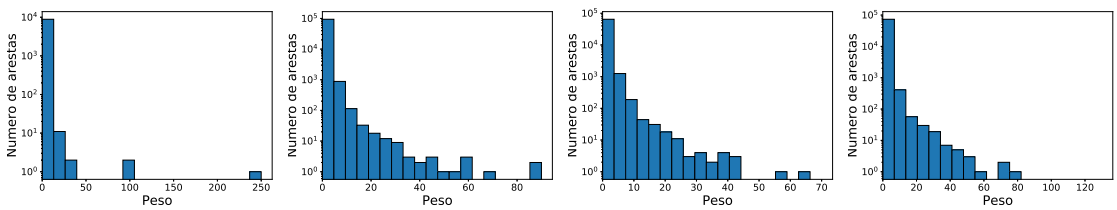


Figura 2. Distribuição conjunta do grau de entrada e saída. Cor indica a fração de participantes com dado grau de entrada e saída.

Embora a distribuição de graus de entrada e saída em um subreddit tenham cauda pesada assim como grande parte das redes sociais online, é natural ponderar se os grafos G_d e G exibem características locais semelhantes a essas redes. Para responder esta questão, mensuramos algumas destas características através das seguintes métricas:

- Pesos das arestas: medem a intensidade da interação entre pares de usuários. Definido como número de interações entre um par durante o intervalo considerado.
- Centralidade de *closeness*: mede o quão próximo um vértice está dos outros. Definido como o inverso da soma das distâncias entre um vértice e cada um dos outros vértices alcançáveis.
- Coeficiente de clusterização: mede o quão conectados estão os vizinhos de um nó. Definido como a fração de arestas existentes entre vizinhos de um nó dentre o máximo possível.

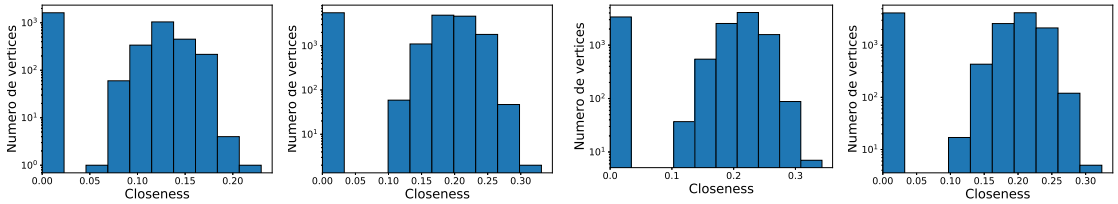
A Figura 3 mostra os histogramas de pesos nas arestas obtidos para cada subreddit. Observamos que a maioria arestas tem peso 1, indicando apenas uma interação entre dois usuários. Poucas arestas possuem peso elevado (p. ex., acima de 20), o que indica altos níveis de interação entre poucos pares de vértices.



(a) English (b) French (c) German (d) Spanish

Figura 3. Distribuição do peso das arestas.

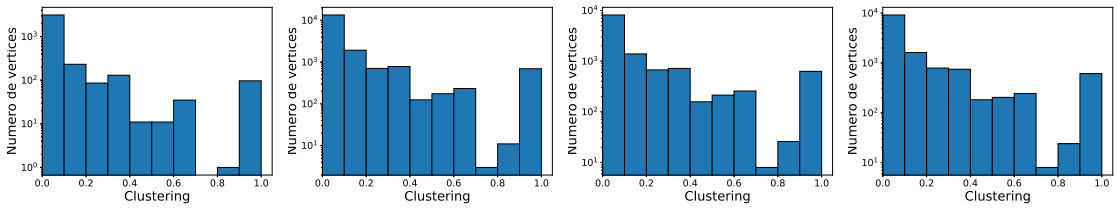
A Figura 4 mostra os histogramas da distribuição do *closeness* para cada subreddit. Observam-se valores baixos, indicando distâncias longas entre vértices. Isto corrobora a intuição de que os usuários não têm interesse específico em conhecer pessoas, e sim em compartilhar conteúdo de interesse à rede. As conexões são formadas conforme os tópicos de interesse comum são compartilhados.



(a) English (b) French (c) German (d) Spanish

Figura 4. Distribuição do *closeness*.

A Figura 5 apresenta os histogramas do coeficiente de clusterização nas redes. É possível observar que um número elevado de usuários apresenta este valor igual a zero e muitos apresentam um valor baixo para esta métrica, o que indica que as redes são esparsas e não apresentam muita formação de triângulos, como no caso de redes sociais tradicionais. Isso reflete novamente na principal característica do Reddit, que é centrado em conteúdo em vez de amizade entre usuários.



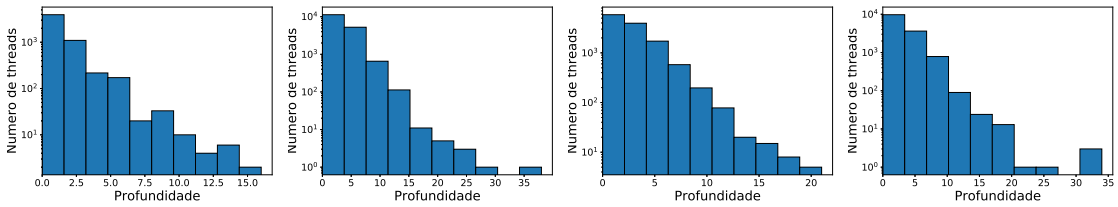
(a) English (b) French (c) German (d) Spanish

Figura 5. Distribuição do coeficiente de clusterização.

4.2. QP2: Como as publicações em um subreddit estão distribuídas em relação às *threads*?

Além da análise dos grafos, também foram analisados os *posts* e comentários dos subreddits através das árvores de discussão. Primeiramente, investigamos a distribuição da profundidade destas árvores. A profundidade é um indicador da progressão de discussões, já que uma árvore com muitos níveis indica que na *thread* correspondente houve pelo menos uma longa cadeia de comentários. A Figura 6 mostra a distribuição das profundidades para cada um dos subreddits.

Observa-se que a maioria das postagens tem a profundidade baixa e que são poucas as postagens que terminam em discussões longas. As árvores de profundidade 1 são aquelas cujos *posts* não receberam nenhum comentário. Por conveniência, mostramos a fração de *posts* respondidos em cada subreddit na Tabela 3. Os subreddits French, German e Spanish apresentam altos índices de respostas às dúvidas compartilhadas pelos membros.



(a) English (b) French (c) German (d) Spanish

Figura 6. Distribuição da profundidade das árvores de discussão.

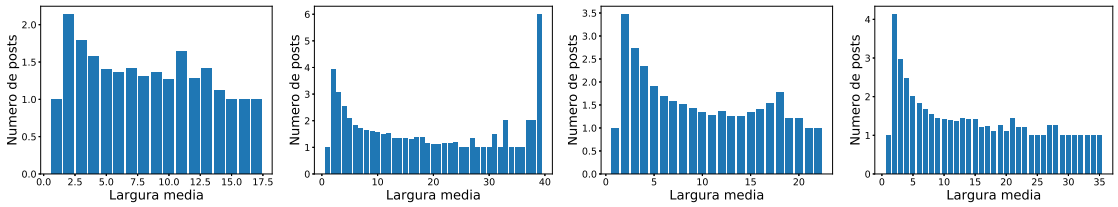
	English	French	German	Spanish
Com resposta	53,55%	83,88%	88,15%	81,56%
Sem resposta	46,45%	16,12%	11,85%	18,44%

Tabela 3. Porcentagem de *posts* com e sem respostas.

Para tentarmos compreender melhor o que leva uma *thread* a se prolongar por muitos níveis, foi realizada uma análise qualitativa daquelas com a maior profundidade considerando cada subreddit. Observamos que no subreddit *German*, a maior árvore é voltada ao compartilhamento de dicas de pronúncia, enquanto no *French* o tema está relacionado à dicas para melhora do vocabulário. Para o *English*, a *thread* de maior profundidade discute as diferenças entre o inglês coloquial e forma culta.

Na comunidade *Spanish* a árvore de discussão mais longa tem um tema menos voltado para proficiência e mais para a vivência, dado que a discussão gira em torno de como a imersão na cultura estrangeira é uma das melhores maneiras de se aprender um novo idioma. Esta análise sugere indícios de que *threads* de maior engajamento dos usuários tendem a ter como assunto principal conselhos para melhora da proficiência de um aluno.

Em seguida, investigamos a largura média das árvores da discussão, ou seja, a quantidade média de comentários que cada *thread* teve em cada nível de profundidade. A Figura 7 mostra a distribuição média da largura por nível de profundidade nos subreddits.

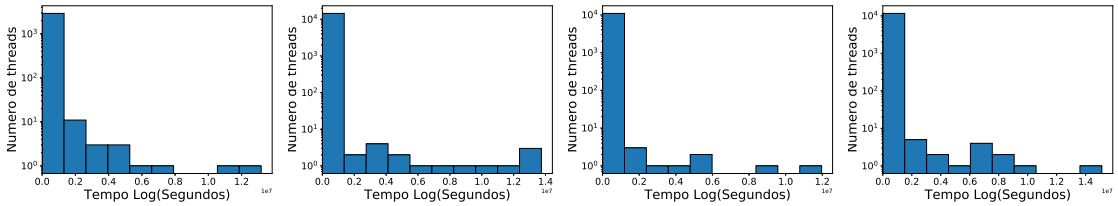


(a) English (b) French (c) German (d) Spanish

Figura 7. Largura média por nível das árvores de discussão (condicionado em profundidade $>$ nível).

Para permitir uma melhor comparação, na Figura 7 o primeiro nível indica a quantidade de *posts* nos subreddits. É possível observar que o engajamento dos usuários é muito elevado nos primeiros níveis de profundidade, ou seja, os usuários respondem diretamente ao post ou aos primeiros comentários. Isso pode indicar que as discussões não se prolongam por muitos comentários e são poucas as *threads* em que a interação dos usuários ocorre por muito tempo.

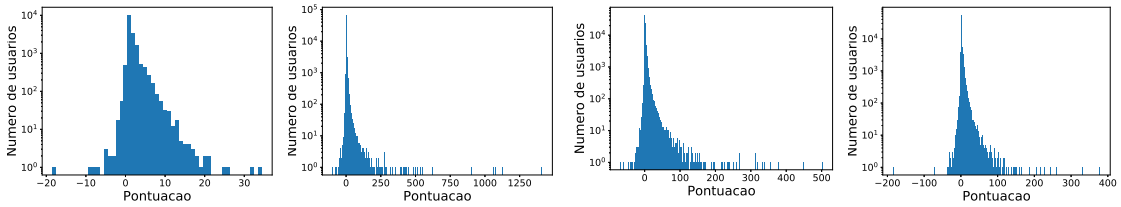
Outra métrica importante de engajamento dos usuários é o tempo decorrido até que um post seja respondido pela primeira vez. A Figura 8 mostra que, dentre os *posts* que receberam uma resposta, a grande maioria foi respondida rapidamente e que poucos ficaram dias até serem respondidos.



(a) English (b) French (c) German (d) Spanish

Figura 8. Distribuição do tempo decorrido até a primeira resposta em uma *thread*.

Uma outra métrica que pode ser avaliada está relacionada à pontuação dos *posts*, calculada pela diferença entre o número de *upvotes* e *downvotes*. O próprio Reddit ordena as publicações pela pontuação final: quanto maior, mais destaque tem o *post* e, consequentemente, mais usuários poderão interagir nesta *thread*. A Figura 9 mostra a distribuição da pontuação dos *posts* por subreddit.



(a) English (b) French (c) German (d) Spanish

Figura 9. Distribuição da pontuação das *threads*.

É possível observar que a pontuação possui valores baixos, em torno de zero. Isso quer dizer que os usuários não têm muito costume de votar nos *posts*, o que reforça a ideia de que as interações são voltadas para o conteúdo e não aprofundam nas discussões.

4.3. QP3: Existem diferenças linguísticas no texto de usuários com diferentes níveis de proficiência?

Para esta questão, iremos utilizar apenas o subreddit German, pois este solicita explicitamente aos usuários que adicionem *tags* (*flairs*) de proficiência a seus perfis. Consequentemente, o German possui um número muito grande de usuários que indicam seu nível de fluência quando comparado às outras comunidades.

O número de usuários do subreddit German com as *tags* de proficiência Iniciante, Intermediário, Avançado e Nativo é, respectivamente, 774, 583, 197 e 896. Um total de 9785 perfis de usuários não possui nenhum desses *tags*. A partir dos *posts* associados a usuários com certo nível de proficiência, fizemos uma análise textual usando a ferramenta LIWC. Avaliamos quatro propriedades: (i) o número de palavras no texto; (ii), o número de palavras por frase; (iii) a quantidade de palavras com mais de 6 letras; e (iv) a quantidade de palavras reconhecidas pelo dicionário do LIWC. Essas duas últimas propriedades assumem valores de 0 a 100.

A Tabela 4 mostra as médias e medianas de cada propriedade para cada grupo de *posts*. Os valores mais elevados para cada propriedade aparecem em negrito. Observa-se que usuários com alemão avançado ou de língua nativa tendem a usar mais palavras nos *posts*. Além disso, estes usuários tendem a usar mais palavras por frase. Considerando "Nativo" como mais proficiente que "Avançado", pode-se observar que o tamanho das palavras utilizadas cresce com a proficiência e que o uso de palavras do dicionário diminui com ela (possivelmente dando lugar a gírias e expressões idiomáticas).

		Iniciante	**Intermediário**	**Avançado**	**Nativo**
Contagem de palavras	Mediana	17,00	18,00	**29,00**	25,00
	Média	44,36	40.44	54,12	**58,77**
Palavras por frase	Mediana	7,00	8,30	**11,00**	10,25
	Média	7,67	8,98	**11,63**	11,44
Palavras grandes	Mediana	16,67	19,35	21,01	**21,05**
	Média	17,29	19,61	21,25	**22,39**
Palavras do dicionário	Mediana	**64,71**	64,58	62,96	59,04
	Média	**66,32**	65,46	63,27	60,67

Tabela 4. Resultados da análise textual utilizando a ferramenta LIWC.

Os resultados da análise textual das publicações indicam que existem diferenças na forma como as pessoas em diferentes níveis de proficiência escrevem no subreddit. Esse fato pode ser utilizado para estimar a proficiência do usuário, quando esta é desconhecida. Contudo, esta análise possui duas limitações: a proficiência é auto-declarada e, portanto, as médias e medianas podem estar super- ou subestimadas; apesar da comunidade solicitar o uso das *tags* explicitamente, 80% dos usuários não as tem em seus perfis, podendo o viés daqueles que as tem ser significativo sobre as métricas estudadas.

5. Conclusão

Neste trabalho analisamos as interações sociais em quatro comunidades do Reddit voltadas para o aprendizado de idiomas: English, French, German e Spanish. Para isto, coletamos *posts*, comentários, *upvotes*, *downvotes* realizados por usuários nestes subreddits entre janeiro de 2010 e dezembro de 2016. A partir destes dados, geramos um grafo de interação entre usuários e árvores de discussão, utilizados para responder três perguntas de pesquisa.

Em relação à **QP1**, concluímos que as interações dentro dessas comunidades diferem daquelas em redes sociais tradicionais. Embora a distribuição de graus de entrada e saída tenham cauda pesada, o volume de interação entre pares de usuários tende a ser pequeno, assim como a centralidade de *closeness* e o coeficiente de clusterização dos vértices. Embora a comunidade EnglishLearning tenha um número menor de usuários, ela possui características semelhantes aos outros subreddits em relação às interações entre usuários, respeitadas as respectivas escalas.

Em relação à **QP2**, observamos que a distribuição da profundidade das árvores de discussão não possui cauda pesada. Enquanto os três primeiros níveis abaixo da raiz têm largura média maior que 2 para todos os subreddits, exceto English, a maioria tem largura média muito próxima de 1. Essas duas observações indicam que as interações dentre de uma thread costumam ser entre pares de usuários, e não de muitos usuários

interagindo com uma mesma pessoa. Uma característica marcante do EnglishLearning é o alta proporção de *posts* sem resposta (46,45% dos *posts* analisados) em relação às outras comunidades (de 11,85% a 18,44%). É provável que isto esteja correlacionado ao baixo número de usuários na comunidade do inglês, embora não tenhamos investigado se há causalidade e, em que direção.

Em relação à **QP3**, baseado nas *tags* de nível de proficiência utilizadas no subreddit German, observamos que existem diferenças entre os textos de usuários com diferentes níveis de alemão quanto ao tamanho das palavras, uso de expressões idiomáticas, etc. Limitações desta análise incluem o fato da proficiência ser auto-declarada e de que muitos usuários não possuem *tags* de proficiência associadas a seus perfis. Como trabalhos futuros, pode-se investigar a evolução da proficiência dos usuários ao longo do tempo, à medida que eles interagem com a comunidade. Assim seria possível avaliar a eficácia da participação de comunidades online no aprendizado de idiomas.

Referências

Andalibi, N., Haimson, O. L., De Choudhury, M., and Forte, A. (2016). Understanding social media disclosures of sexual abuse through the lenses of support seeking and anonymity. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 3906–3918. ACM.

Arnold, N. and Paulus, T. (2010). Using a social networking site for experiential learning: Appropriating, lurking, modeling and community building. *The Internet and higher education*, 13(4):188–196.

Cunha, T. O., Weber, I., Haddadi, H., and Pappa, G. L. (2016). The effect of social feedback in a reddit weight loss community. In *Proceedings of the 6th International Conference on Digital Health Conference*, pages 99–103. ACM.

De Choudhury, M. and De, S. (2014). Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In *ICWSM*.

Joinson, A. N. (2008). Looking at, looking up or keeping up with people?: motives and use of facebook. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1027–1036. ACM.

Levy, M. (1997). *Computer-assisted language learning: Context and conceptualization*. Oxford University Press.

Lin, C.-H., Warschauer, M., and Blake, R. (2016). Language learning through social networks: Perceptions and reality.

Newman, M. (2011). Resource letter cs–1: Complex systems. *Am. J. Phys*, 79:800.

Pappa, G. L., Cunha, T. O., Bicalho, P. V., Ribeiro, A., Silva, A. P. C., Meira Jr, W., and Beleigoli, A. M. R. (2017). Factors associated with weight change in online weight management communities: A case study in the loseit reddit community. *Journal of Medical Internet Research*, 19(1).

Warschauer, M. and Healey, D. (1998). Computers and language learning: An overview. *Language teaching*, 31(2):57–71.

Zhao, Y. (1996). Language learning on the world wide web: Toward a framework of network based call. *Calico Journal*, pages 37–51.

Zourou, K. (2012). De l'attrait des médias sociaux pour l'apprentissage des langues–regard sur l'état de l'art. *Alsic. Apprentissage des Langues et Systèmes d'Information et de Communication*, 15(1).

Análise de Comunidades de Suporte a Transtornos de Saúde Mental do Reddit*

Bárbara Silveira, Ana Paula Couto da Silva, Fabricio Murai

[1]Departamento de Ciência da Computação
Universidade Federal de Minas Gerais (UFMG) – Belo Horizonte, MG – Brazil

{barbarasilveira, ana.coutosilva, murai}@dcc.ufmg.br

Abstract. *In the last years, online social networks have gained new functionalities and goals. Initially designed for fostering friendships and for exchanging images or videos, they started to connect people willing to share experiences related to health problems, such as obesity and depression. In this work, we characterize users from four Reddit communities centered on sharing experiences related to mental health problems, named: Depression, SuicideWatch, Anxiety and Bipolar. The focus of this paper lies on analyzing (i) these users activity, (ii) the social support provided by these communities and (iii) the experiences shared through posts and comments.*

Resumo. *Nos últimos anos, redes sociais online ampliaram suas funcionalidades e objetivos. Inicialmente focadas em fomentar amizades, troca de imagens ou vídeos, passaram a conectar pessoas dispostas a trocar experiências relacionadas à problemas de saúde, como por exemplo, obesidade e depressão. Neste trabalho, caracterizamos os usuários de quatro comunidades do Reddit centradas na troca de experiências relacionadas à problemas de saúde mental, intituladas: Depression, SuicideWatch, Anxiety e Bipolar. O enfoque principal deste artigo é na análise (i) da atividade destes usuários, (ii) do suporte social oferecido por estas comunidades, e (iii) das experiências compartilhadas através de posts e comentários.*

1. Introdução

Nos últimos anos, dados da Organização Mundial da Saúde (OMS) alertam para o aumento do total de pessoas no mundo que sofrem de algum tipo de transtorno de saúde mental. Como exemplo, um relatório global lançado recentemente pela mesma organização[1], aponta que o número de casos de depressão aumentou 18% entre 2005 e 2015: são 322 milhões de pessoas em todo o mundo, a maioria mulheres. No Brasil, a depressão atinge 11,5 milhões de pessoas (5,8% da população), enquanto distúrbios relacionados à ansiedade afetam mais de 18,6 milhões de brasileiros (9,3% da população).

Devido ao aumento de casos destes tipos de transtornos, políticas de saúde pública eficientes devem ser implementadas. No caso específico da depressão, a OMS é responsável pelo programa *Mental Health Gap Action Programme*, que visa ajudar os países

*The authors' work has been partially funded by the EUBra-BIGSEA project by the European Commission under the Cooperation Programme (MCTI/RNP 3rd Coordinated Call), Horizon 2020 grant agreement 690116, CAPES, CNPq and FAPEMIG.

[1]https://news.un.org/

a aumentar os serviços prestados às pessoas com transtornos mentais, neurológicos e de uso de substâncias, por meio de cuidados providos por profissionais de saúde que não são especialistas em saúde mental. A iniciativa defende que, com cuidados adequados, assistência psicossocial e medicação, dezenas de milhões de pessoas com transtornos mentais, incluindo depressão, poderiam começar a levar uma vida normal – mesmo quando os recursos são escassos.

Mais ainda, a combinação entre recursos escassos (principalmente entre países de economia mais instável), o estigma social associado aos transtornos mentais e muitas vezes a resistência em pedir ajuda, faz com que muitas pessoas que sofrem destes transtornos não sejam ajudadas da melhor maneira possível, levando a situações drásticas, como o suicídio. Menos da metade dos afetados no mundo (em muitos países, menos de 10%)[2] são diagnosticados corretamente em um quadro de depressão, por exemplo.

Este quadro geral faz com que novos recursos em busca de compreender e auxiliar os indivíduos afetados por estes transtornos sejam explorados. Por exemplo, podemos ressaltar o papel das redes sociais online. Inicialmente focadas em fomentar amizades, troca de imagens ou vídeos, passaram a conectar pessoas dispostas a trocar experiências relacionadas à problemas de saúde, como por exemplo, obesidade [Pappa et al. 2017] e depressão [Choudhury and De 2014, Kavuluru et al. 2016]. Outro trabalho [Lopes et al. 2014] utilizou uma rede social para identificar a percepção da promoção da saúde por grupo de profissionais da saúde. Neste contexto, nosso trabalho apresenta uma análise de comunidades direcionadas à discussão de transtornos de saúde mental. As observações produzidas por esta análise podem ajudar a guiar a realização mais eficiente de intervenções que auxiliem indivíduos que sofrem destas doenças.

O nosso estudo foca nas comunidades do Reddit, onde os membros compartilham suas experiências e dúvidas sobre os mais diversos assuntos. Uma característica importante é que os usuários podem permanecer anônimos, definindo identidades temporárias, encorajando os mesmos a discutir sobre assuntos mais delicados e compartilhar pensamentos e sentimentos que muitas vezes não são aceitos facilmente pela sociedade. O Reddit é organizado em comunidades (subreddits) e em 2017 era formado por 1.204.126 comunidades com 900 milhões de comentários. Pelo menos 25 subreddits estão relacionados à transtornos da saúde mental. Nosso trabalho foca na análise dos quatro subreddits com o maior número de usuários ativos [Gkotsis et al. 2016, Gkotsis et al. 2017]: Depression (/r/depression), SuicideWatch (/r/suicide) – será abreviada como Suicide–, Anxiety (/r/anxiety) e Bipolar (/r/bipolar).

A partir do estudo das atividades dos usuários e na exploração do conteúdo compartilhado, as nossas principais contribuições são:

- Caracterizar o volume de atividades dos usuários dentro das comunidades, através do conceito de árvore de discussão.
- Analisar as experiências dos usuários aplicando o RMN (Relationship Modeling Network), proposto por [Iyyer et al. 2016], buscando encontrar semelhanças na forma com que as pessoas se expressam quando discutem sobre transtornos de saúde mental.

[2]https://nacoesunidas.org/oms-registra-aumento-de-casos-de-depressao-em-todo-o-mundo-no-brasil-sao-115-milhoes-de-pessoas/

- Verificar indícios de apoio social oferecido por essas comunidades, a partir dos tópicos e descritores derivados pelo RMN.

Este artigo está organizado da seguinte forma: a Seção 2 descreve os principais trabalhos relacionados; a Seção 3 detalha os dados e métodos utilizados; os resultados da análise dos subreddits são apresentados na Seção 4; as implicações deste trabalho e trabalhos futuros são discutidos na Seção 5.

2. Trabalhos Relacionados

A aplicação de ferramentas para auxiliar pessoas que sofrem de diferentes transtornos de saúde mental, além de consultas presenciais, não é um paradigma novo. Por exemplo, a utilização de atendimento telefônico, onde voluntários prestam suporte é uma maneira de auxílio muito difundida em vários países.

No entanto, o avanço dos meios de comunicação, sejam por telefones celulares ou através da Internet, está revolucionando o suporte oferecido às pessoas que sofrem de transtornos como depressão e ansiedade. Assim, alguns trabalhos na literatura buscam entender como novas tecnologias são eficazes em minimizar os sintomas destas doenças. A seguir, apresentamos os trabalhos que são mais relevantes ao escopo da análise que fazemos neste artigo.

Os autores em [Althoff et al. 2016] analisam conversas, via SMS, de indivíduos auxilados por conselheiros ligados à organização *Crisis Trends*[3]. O estudo busca avaliar o comportamento dos conselheiros, uma vez que estes desempenham um papel crucial no apoio dos indivíduos. Uma das principais conclusões é que os conselheiros bem sucedidos em ajudar os indivíduos são mais sensíveis à trajetória da conversa, respondem às mensagens de forma mais criativa, sem usar frases genéricas e rapidamente identificam o foco do problema no qual o indivíduo se encontra, colaborando para a solução do mesmo. O estudo de [Gkotsis et al. 2016] investigou a linguagem dos *posts* do Reddit relacionados à saúde mental e identificou várias das características linguísticas dessas comunidades. Esses mesmos autores fizeram outro trabalho [Gkotsis et al. 2017] analisando as postagens do Reddit para desenvolver classificadores que reconheçam e classifiquem postagens relacionadas à doença mental, através da técnica de *deep learning*. Além disso, existem trabalhos [Souza et al. 2017, Wang et al. 2016] que abordam a caracterização de informações, através de extração de tópicos a fim de descrever o contexto analisado.

Considerando o uso do Reddit e comunidades direcionadas à saúde mental, o trabalho em [Choudhury and De 2014] apresenta uma análise do discurso feito por usuários de comunidades ligadas à saúde mental. Por exemplo, os autores investigam como o grau de desinibição nos comentários e *posts* feitos por usuários anônimos se difere daqueles feitos pelos usuários que se identificam. O anonimato, por meio de contas descartáveis, permite um maior *self-disclosure* em torno do tópico estigmático da saúde mental, uma vez que não existe a preocupação em ser identificado. Adicionalmente, postagens com menos inibição são as que possuem mais atenção e discutem problemas de relacionamentos e de saúde. Essas postagens reúnem maior apoio comunitário, através de votos e comentários. Os autores em [Kavuluru et al. 2016] focam a análise do subreddit SuicideWatch, com o objetivo de classificar, de forma automática, comentários que possam

[3]https://crisistrends.org/

impactar positivamente o comportamento de indivíduos com pensamentos suicidas (comentários *helpful specific*).

As análises apresentadas neste artigo diferem dos artigos descritos anteriormente em dois pontos principais. Primeiro, analisamos as atividades dos usuários em quatro comunidades, com o principal foco nos textos das publicações, isto é, dos *posts* e comentários. Avaliamos o engajamento dos usuários através das árvores de discussão construídas a partir das publicações. A segunda diferença é a aplicação do RMN (*Relationship Modeling Network*), proposto por [Iyyer et al. 2016], ao texto das publicações para modelar o papel dos usuários do Reddit nas comunidades das quais eles participam. Diferentemente do LDA (*Latent Dirichlet Allocation*) que associa documentos a um conjunto de tópicos latentes extraídos do corpus, o RMN tem como objetivo modelar a relação entre um par de entidades que aparece múltiplas vezes em diversos documentos ou contextos. Mais do que isso, o RMN modela a trajetória de um relacionamento dentro de cada contexto. Neste artigo, utilizamos o RMN para estudar a relação entre usuários do Reddit e suas respectivas comunidades.

3. Metodologia

Nesta seção descrevemos o conjunto de dados analisados (Seção 3.1), a maneira como os *posts* e comentários feitos pelos usuários serão analisados, através da criação de árvores de discussão (Seção 3.2), e a definição de descritores utilizando o *Relationship Modeling Network* (Seção 3.3).

3.1. Conjunto de Dados

Os dados de usuários do Reddit analisados neste trabalho foram recuperados online[4]. Existem mais de 25 subreddits que focam na discussão de transtornos mentais, entretanto selecionamos os quatro subreddits com o maior número de *posts* e comentários (desconsiderando o Opiates (/r/opiates), ligado à dependência de rémedios) [Gkotsis et al. 2016, Gkotsis et al. 2017]: Depression (/r/depression), SuicideWatch (/r/suicide), Anxiety (/r/anxiety) e Bipolar (/r/bipolar).

Todas as publicações dos usuários destas comunidades (*posts* e comentários) realizadas no ano de 2017 foram extraídas dos dados disponibilizados, totalizando 261.511 *posts* e 1.256.669 comentários de 207.683 usuários únicos. A Tabela 1 apresenta as principais características de cada um dos subreddits analisados.

	Depression	SuicideWatch	Anxiety	Bipolar
Usuários Únicos	43.322	13.940	105.879	44.542
Posts	44.288	24.349	145.072	47.802
Comentários	188.045	185.957	624.578	258.089
Posts por dia	121,34	66,71	397,46	130,96
Coméntarios por dia	515,19	509,47	1.711,17	707,09

Tabela 1. Estatísticas básicas dos subreddits analisados.

[4]http://files.pushshift.io/reddit/

3.2. Árvores de Discussão

Uma maneira de analisar a atividade gerada em torno dos tópicos iniciados pelas *threads* (em termos de *posts* e comentários) é a construção de árvores de discussão. Consideramos como árvore de discussão, toda a troca de informação que se inicia por um post (raiz da árvore), seguido de comentários feitos pelo próprio usuário que gerou o post ou por outros usuários na comunidade.

A análise da largura e profundidade destas árvores pode revelar quais são os tópicos que geram maior repercussão entre os usuários, bem como se existem usuários-chave que atraem a atenção dos demais membros da comunidade. Encontrar tais usuários em um subreddit é importante para projetar intervenções direcionadas que podem auxiliar na melhoria da saúde mental de um grande número de participantes.

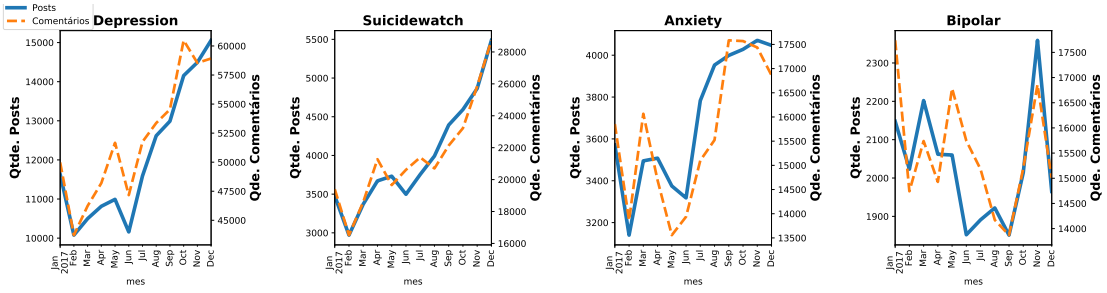
3.3. Relationship Modeling Network (RMN)

O RMN é uma rede neural recursiva projetada para modelar relacionamentos entre pares de entidades a partir de texto. Um relacionamento em um dado instante é representado como um vetor de pesos sobre K descritores. As entidades não precisam ser da mesma classe: por exemplo, neste trabalho, as entidades são os usuários e as comunidades nas quais eles interagem. Cada post ou comentário corresponde a um momento diferente.

As palavras são representadas por *embeddings* de dimensão P, ou seja, cada palavra w de um vocabulário $\mathcal{V}$ é um vetor em $\mathbb{R}^P$. Usuários e comunidades são representados por *embeddings* de dimensão U e C respectivamente. Assim como nos trabalhos anteriores, geramos os *embeddings* usando o GloVe [Pennington et al. 2014]. Os descritores obtidos pelo RMN são vetores em $\mathbb{R}^P$, permitindo que encontremos as palavras mais próximas a cada um deles. A representação de post (ou comentário) é denotada por $v_{post} \in \mathbb{R}^P$. Este vetor é igual a média dos *embeddings* das palavras que o post contém. As representações dos usuários e das comunidades são denotadas por v_{user} e v_{comm}.

Para cada post ou comentário, o RMN é alimentado com um vetor $v \in \mathbb{R}^{P+U+C}$ obtido a partir da concatenação de v_{post}, v_{user} e v_{comm}. Estes vetores são combinados através dos pesos da rede neural para obter uma representação $d_t \in \mathbb{R}^K$ para a relação entre o usuário e a comunidade naquele instante específico. O RMN utiliza um parâmetro de suavidade $\alpha \in (0, 1)$ para evitar mudanças bruscas na representação de uma mesma relação em instantes consecutivos, d_t e d_{t-1}. A matriz de descritores $R \in \mathbb{R}^{K \times P}$ é usada para tentar reconstruir o post v_{post}, fazendo-se $r_t = R^\top d_t$. Os parâmetros do RMN (pesos e matriz dos descritores) são treinados de forma a maximizar uma função objetivo que visa aproximar r_t e v_{post} e distanciar r_t de outros *posts* amostrados aleatoriamente. Veja [Iyyer et al. 2016] para mais detalhes sobre o funcionamento do RMN.

Os dados de entrada para o RMN foram pré-processados da seguinte forma. Primeiramente, foram removidos todos os *posts* e comentários marcados como *[deleted]* ou *[removed]*, as *stopwords* (aplicando a biblioteca NLTK) e as pontuações. Segundo, foram considerados somente os *posts* e comentários dos usuários que realizaram no mínimo 50 atividades (*posts*/comentários) em cada subreddit, seguindo a metodologia apresentada em [Wang et al. 2016]. Por fim, foram selecionadas todas as palavras que aparecem nos quatro subreddits analisados, buscando encontrar semelhanças na forma com que as pessoas se expressam quando discutem sobre transtornos de saúde mental. Assim, o sub-



(a) Depression (b) Suicide (c) Anxiety (d) Bipolar

Figura 1. Evolução do total de *posts* e comentários em cada subreddit.

conjunto de dados analisados pelo RMN engloba 18.020 palavras únicas, 25.101 *posts* e 401.428 comentários.

4. Análise dos Subreddits

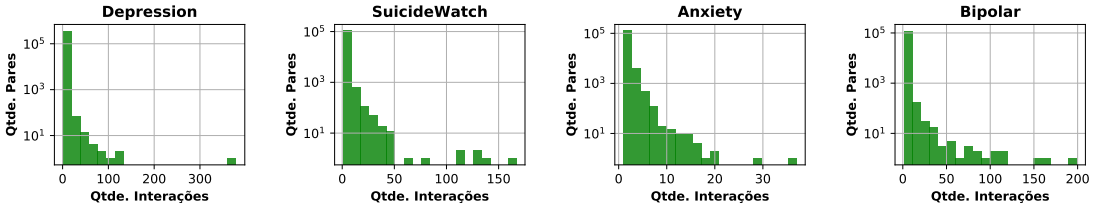
Nesta seção apresentamos a análise dos usuários dos quatro subreddits relacionados à transtornos de saúde mental. O primeiro conjunto de resultados descreve o volume de atividades dos usuários (Seção 4.1). A seguir, apresentamos uma análise detalhada das discussões dos usuários, dos tópicos que mais aparecem nos *posts* e comentários, bem como de alguns perfis de usuários extraídos da análise destes tópicos (Seção 4.2).

4.1. Atividades dos Usuários

A Figura 1 apresenta o volume mensal de *posts* e comentários em cada um dos subreddits analisados. É interessante notar que, exceto para o subreddit Bipolar, o número mensal de publicações manteve uma tendência de crescimento durante quase todo o ano de 2017. No caso da comunidade Anxiety, o mês de Maio revela uma diminuição no volume de *posts* e comentários, seguido de uma crescimento substancial. Outro ponto a ser ressaltado é que os picos de atividades nestas comunidades podem estar correlacionados com eventos reais que geram grande comoção entre pessoas. Por exemplo, se considerarmos o subreddit Suicide, entre os meses de Fevereiro e início de Abril o total de *posts* aumentou $\approx 16\%$ enquanto o total de comentários aumentou em $\approx 30\%$. Este crescimento coincide com a divulgação e lançamento da série *13 Reasons Why*, que gerou forte discussão em torno do tema de suicídios cometidos por adolescentes [5].

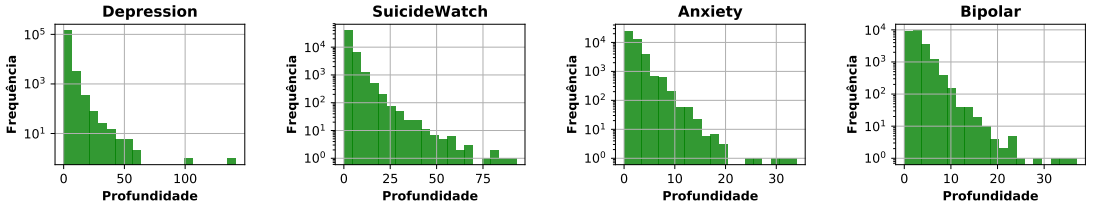
Uma outra maneira de quantificar o volume de atividades em um subreddit é mensurar a quantidade de interações entre pares de usuários. Dizemos que houve uma interação entre um par de usuários (i, j) quando i comenta em uma publicação de j ou vice-versa. A Figura 2 apresenta o histograma contendo o número de pares encontrados quantizados por números de interações. Podemos observar que a grande maioria dos usuários estabelecem poucos diálogos entre si (primeira coluna dos histogramas apresentados). Este resultado corrobora a filosofia da rede social do Reddit, onde o objetivo principal é o engajamento em torno de conteúdos em que os usuários possuem interesse, independentemente dos usuários que participam da discussão (dado que a identidade dos usuários sequer é revelada, em muitos casos).

[5]https://www.theatlantic.com/entertainment/archive/2017/08/13-reasons-why-demonstrates-cultures-power/535518/



(a) Depression (b) Suicide (c) Anxiety (d) Bipolar

Figura 2. Histograma de números de pares de usuários quantizados por volume de interação.



(a) Depression (b) Suicide (c) Anxiety (d) Bipolar

Figura 3. Histograma da profundidade das árvores de discussão.

4.2. Posts e Comentários

A seguir apresentamos as análises dos *posts* e comentários dos usuários sob dois pontos de vista: o primeiro, análise da árvore de discussão para identificar o engajamento dos usuários em pedir ajuda e serem atendidos por outros usuários; o segundo, através dos descritores que caracterizam os *posts* e comentários dos usuários, revelando, por exemplo, quais são os temas mais discutidos nestas comunidades e possíveis perfis dos usuários.

Árvores de Discussão

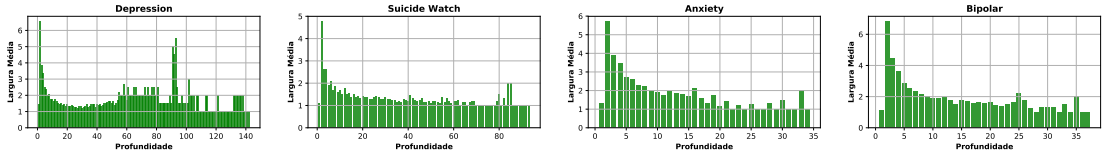
A Figura 3 mostra a distribuição da profundidade das árvores de discussão em cada um dos subreddits. Um ponto a ser ressaltado é que nos subreddits Depression e Suicide, quase a totalidade das árvores de discussão possuem profundidade de até 60. Nos subreddits Anxiety e Bipolar, a quase totalidade das árvores de discussão possuem profundidade de até 20 (1/3 da profundidade dos primeiros dois subreddits). Uma possível explicação para este resultado é que as discussões nos dois primeiros subreddits tendem a ter maior complexidade ou gravidade, fomentando, assim, maior suporte social entre os usuários.

A Tabela 2 traz uma investigação detalhada das árvores de maior profundidade de cada subreddit. É interessante notar que os usuários realmente buscam ajuda nestes tipos de comunidades, relatando abertamente os problemas pelos quais estão passando (veja abaixo, por exemplo, o post que gera a maior árvore de discussão no subreddit Suicide). Podemos ressaltar também a importância do papel de um membro que atua como moderador. Este perfil de membro pode motivar os demais a compartilharem o estado emocional, fazendo com que os outros usuários iniciem um processo de aconselhamento e ajuda.

Com o objetivo de melhor caracterizar as árvores que fomentam maior engajamento dos usuários, realizamos uma análise manual daquelas com as maiores profundida-

Subreddit	Publicações	Profundidade	**Análise**
Depression	769.650	142	Post feito pelo moderador da comunidade (“skyqween”) incentivando os usuários a relatarem como estão. Alguns usuários oferecem ajuda, gerando suporte social. Pelo total de *posts* e comentários feitos, podemos inferir que os usuários compartilharem seus estados emocionais quando são motivados.
Suicide	305.891	93	Usuário relata estar sofrendo de pensamentos suicidas. Um outro usuário estabelece um diálogo, questionando o estado emocional do usuário que fez o post, oferecendo ajuda.
Anxiety	232.333	34	Usuário relata ansiedade por ter feito uma entrevista de emprego e demais usuários trocam experiências similares, oferecendo ajuda.
Bipolar	210.306	37	Usuário pede ajuda e um diálogo é iniciado com um outro usuário.

Tabela 2. Análise das árvores com maior profundidade.



(a) Depression (b) Suicide (c) Anxiety (d) Bipolar

Figura 4. Largura média por nível das árvores de discussão (condicionado a profundidade $>$ nível). Nós de mesma profundidade estão no mesmo nível da árvore.

des. Através da inspeção de 16 árvores de discussão, obtivemos as seguintes conclusões:

1. Usuários que iniciam a árvore de discussão tendem a escrever *posts* procurando por ajuda. A partir disso, um diálogo se inicia, e se estabelece, na maior parte das vezes, com somente um usuário em específico.
2. Árvores de discussão de maior profundidade são aquelas em que as pessoas que estão dispostas a ajudar demonstram mais interesse no problema da pessoa que pediu ajuda. O interesse pode ser mostrado, por exemplo, através de perguntas que buscam entender melhor a situação real do usuário que pediu ajuda.
3. Usuários que iniciam uma árvore de discussão com um *post* formado por muitas palavras, tendem a ter árvores de discussão de maior profundidade onde o usuário que está ajudando faz comentários longos. Este resultado, de uma certa forma, é semelhante ao do artigo [Althoff et al. 2016], pois uma das características de um bom conselheiro é esclarecer problemas de ambiguidade. Assim, membros com perfil de conselheiros tendem a escreverem mais a fim de deixar bem claro o problema pelo qual o usuário que iniciou a árvore de discussão tem vivido.

Seguindo a análise das árvores de discussão, a Figura 4 mostra a largura média a cada nível de profundidade das árvores de discussão. A largura média é a quantidade de *posts* e/ou comentários em determinado nível dividido pela quantidade de árvores com profundidade de pelo menos o nível que está sendo considerado.

A partir da Figura 4 podemos observar que, exceto para o subreddit Depression, as maiores larguras médias estão concentradas nos primeiros níveis de profundidade. No caso do Depression, larguras médias maiores que 5 (cinco) são encontradas em níveis de maior profundidade (entre 85 e 95). Uma possível explicação para este resultados é a existência de moderadores que postam semanalmente incentivando os usuários a falarem dos seus problemas. Assim, um comentário ao *post* pode gerar uma árvore de grande profundidade e eventualmente existirão mais comentários em cada nível, resultando em larguras diferentes.

Tópico	Descritores
0	realistic practical perspective perception innate objective value ideal aligned reframe
1	**ideally hesitate available anytime soonest preferably appointments arrange offered pm**
2	indecision duress disco uncontrolled negating catastrophising accompanying aggravates curved eliminates
3	**selfish die kill neither fault blame killing anybody hurt abandon**
4	clearance excite fulltime relocating salary earn savings electrician lucrative funds
5	ah stardew btw soundcloud youtu portugal awsome goo huh ahaha
6	podcasts relaxing videos skate parks asmr movies distracting cafe music
7	**resolve heal acknowledge overcome reassure gently encourage seek confront subtly**
8	**world life beautiful future inside imagine reality person somehow feels**
9	**met dating university hs group dated girlfriends college friendships acquaintances**
10	**unconfident overreacting rly teenager immature teenagers labelled ashamed unloved embarassed**
11	headaches fog fatigue heavy intense aches periods levels exhaustion moderate
12	**worthless waste achievements loser useless meaningless pointless bullshit effort homework**
13	disorders discriminate individuals patients prevalence uneducated definitions diagnoses diagnose clinical
14	heater cocoa headless shrimp champagne sundae brittle tray 22yrs headlight
15	accuses memy recorder reorganized uninvited drifts inserting nearer unsuspecting clone
16	**hugs wishes rooting buddy xx thank bless hug hope luck**
17	moved sat went dropped stayed walked drove ran home discharged
18	bupropion buspar welbutrin zoom anxiolytic tolerated zoloft geodon lamotrigine topamax
19	blacklist cdn eood 270 relationshipadvice freecompliments nytimes 736x 4chan scientificamerican
20	etc often stress anxious makes sometimes physical constantly social deal
21	14th frasier coincide ummmm roughest 2001 alternating 13th noteworthy 30th
22	doc appt pdoc dr thursday gp monday wednesday tuesday yesterday
23	flunk arsed whatcha definitively islamic physics duh honors retake ged
24	**comment responses replying sounded thoughtful appreciate sincere commented reply responding**
25	habit push pushing drinking avoid start habits slow stop edge
26	yrs 68 euros 20m twelve centuries trillion 204 sixteen machines
27	hii sympathies onslaught awwwww arg sterile agh youuu glee sweety
28	**books book useful art published research recommend coloring journal bought**
29	journey vacation luck easter smoothly rough merry holidays zealand ride

Tabela 3. Descritores e seus termos principais.

Tópico	Contexto
1	Posts/Comentários que oferecem suporte social
3	Posts/Comentários que relatam pensamentos suicidas
7	Posts/Comentários que refletem encorajamento
8	Posts/Comentários que refletem melhora do usuário frente ao problema de saúde mental
9	Posts/Comentários que refletem problemas de relacionamento
10	Posts/Comentários que refletem problemas de usuários durante a adolescência
12	Posts/Comentários de usuários desesperançosos
16	Posts/Comentários de usuários de agradecimentos
24	Posts/Comentários que agradecem a uma ajuda
28	Posts/Comentários que sugerem atividades para amenizar uma condição emocional negativa

Tabela 4. Contexto de alguns descritores.

Descritores

A Tabela 3 apresenta os termos principais associados as descritores encontrados pelo RMN quando treinado com o texto de *posts* e comentários de todos os subreddits. Os subreddits são considerados em conjunto, pois desejamos obter descritores comuns a todas as comunidades. É interessante notar que o modelo captura um conjunto considerável de descritores *coerentes* (i.e., os termos principais dos descritores são relacionados), mostrando que os quatro subreddits analisados possuem um linguagem em comum e que está intimamente relacionada aos transtornos de saúde mental (veja, por exemplo, o descritor 3). Os resultados apresentados a seguir focam na análise dos descritores que estão em negrito na Tabela 3. A Tabela 4 lista possíveis interpretações dos *posts* e comentários que possuem maior probabilidade de estarem relacionados aos tópicos destacados.

Cada usuário em um subreddit tem um conjunto de comentários e *posts*, deno-

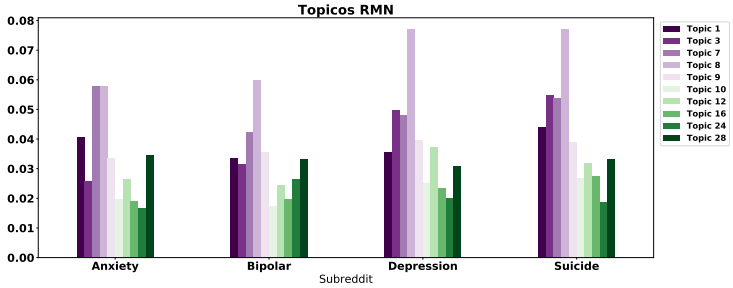


Figura 5. Probabilidade de cada post ou comentário ser associado a um tópico.

minados ***spans*** pelo RMN. Ao final da execução, o RMN associa a cada *span* um vetor de probabilidades sobre os tópicos. Para analisar os tópicos que compõem um subreddit, calculamos a média dos vetores associados aos respectivos spans. Esta média pode ser interpretada como a esperança da distribuição de probabilidade sobre os tópicos ao selecionarmos um *span* de maneira uniforme aleatória. A Figura 5 mostra as probabilidades médias associadas aos tópicos listados na Tabela 4 para cada subreddit. Observamos que:

1. O **Tópico 8**, dentre os 10 analisados, é o que ocorre com maior probabilidade nas quatro comunidades. Este tópico contém descritores positivos, dando indícios de que os usuários dos subreddits oferecem suporte e ajuda uns aos outros, principalmente através de palavras de otimismo e de perspectiva de um futuro melhor. O **Tópico 7**, relacionado a termos de encorajamento, ocupa posição de destaque em todos os subreddits.
2. O **Tópico 3** é o segundo mais frequente nas comunidades Depression e Suicide. Este tópico é negativo e remete a descritores de morte, ferimento, culpa, abandono, entre outros, sendo um resultado esperado em comunidades centradas na discussão de problemas de saúde mental relacionados à depressão e ao suicídio.
3. Os subreddits Depression e Suicide possuem perfis bem semelhantes, quando comparamos estes 10 tópicos. No entanto, dois tópicos os diferenciam: (i) o **Tópico 12** aparece com maior probabilidade no subreddit Depression. Este tópico remete a descritores com sentimento de inutilidade e desperdício, o que é um indício que as pessoas da comunidade Depression expressam mais seu sentimento de inutilidade e; (ii) o **Tópico 16**, que remete a descritores de agradecimento, aparece em maior probabilidade no subreddit Suicide, o que pode indicar que usuários desta comunidade respondem de maneira mais afetuosa ao apoio social oferecido.
4. O **Tópico 9**, que remete à lugares e grupos de relacionamentos (amigos, namoradas, universidade, encontros), aparece com maior probabilidade nas comunidades Depression e Suicide, o que sugere que os problemas de saúde mental dos usuários dessas comunidades podem ter sido desencadeados, principalmente, devido à relacionamentos problemáticos com os pares.
5. O **Tópico 10** é sobre insegurança e adolescência, sendo este assunto mais frequente na comunidade Suicide e possivelmente refletindo problemas de *bullying* comuns nesta fase da vida.
6. O **Tópico 28** tem a probabilidade parecida de ocorrência nas quatro comunidades, revelando que usuários tendem a indicar atividades que possam ocupar a mente (por exemplo, ler um livro).

Analisamos também os *posts* e comentários feitos pelos usuários, auxiliando em uma possível definição de perfis nas comunidades. Para isto, calculamos a média dos vetores associados aos *spans* de cada de usuário. Para cada tópico da Tabela 4, descrevemos a seguir o perfil dos cinco usuários cujas probabilidades associadas ao tópico é máxima. A seguir, iremos correlacionar o comportamento destes usuários com os tópicos encontrados pelo RMN:

1. Conforme descrito anteriormente, o **Tópico 1** se caracteriza por descritores que mostram uma tendência de suporte social. Dos cinco usuários com maior probabilidade média de relevância neste tópico, quatro fazem parte da comunidade Suicide e um da comunidade Depression. Ao mesmo tempo a probabilidade de relevância dos *spans* destes usuários em tópicos com descritores mais negativos (3, 10 e 12) é pequena. Provavelmente estes usuários tem um perfil de conselheiro, oferecendo ajuda aos demais membros destas comunidades.
2. O **Tópico 3** possui descritores com alto teor de negatividade. Dos cinco usuários com maior probabilidade média de relevância neste tópico, três fazem parte da comunidade Suicide e dois da comunidade Depression. Um ponto interessante é que, apesar destes usuários possuírem um discurso que tende a ser muito negativo, os mesmos possuem um probabilidade relativamente alta associada ao **Tópico 8**, que reflete um discurso de otimismo e possível melhora do transtorno sofrido. Este resultado de certa forma corrobora a análise de [Althoff et al. 2016], onde os autores concluem que pessoas que atravessam períodos de maior dificuldade são mais propensas a pensar no futuro e serem positivas, quando auxiliadas por outras pessoas que oferecem ajuda.
3. Os *spans* dos usuários mais relevantes para o **Tópico 10**, que está relacionado com problemas de insegurança e adolescência e que aparecem de forma mais pronunciada em Suicide e Depression, também exibem altas probabilidades associadas a tópicos de teor negativo (10 e 12). Isto sugere a presença de um número expressivo de adolescentes que buscam as comunidades online para relatar inseguranças e pensamentos suicidas.
4. Os *spans* de usuários mais relevantes para o **Tópico 28**, que é ligado a sugestões de atividades para amenizar uma condição emocional negativa, também possuem altas probabilidades em relação aos **Tópicos 7 e 8**, revelando um possível perfil de usuários que buscam compartilhar experiências de possíveis melhoras da saúde mental através de atividades e *hobbies*.

5. Conclusão e Trabalhos Futuros

O presente trabalho apresentou uma análise da árvore de discussão dos subreddits: Anxiety, Bipolar, Depression e Suicide. Analisando as 16 árvores de discussão com maiores profundidades percebe-se que a maior parte dos *posts* são pedidos de ajuda e, que na maior parte das vezes, mais de uma pessoa oferece ajuda. As árvores mais profundas, em sua maioria, são compostas por dois usuários (o que fez o *posts* e o que ofereceu ajuda).

Através do modelo RMN foram gerados tópicos que descrevem aquilo que é frequentemente discutido nas quatro comunidades. Tópicos relacionados a pensamentos suicidas são mais frequentes na comunidade de Suicide e Depression. Além disso, constatou-se que os subreddits Depression e Suicide são parecidos em relação a maior parte dos tópicos. Também foi possível analisar os usuários mais relevantes para cada tópico. Por

exemplo, o tópico 7, que está relacionado à superação de problemas, tem os usuários mais relevantes nas comunidades Bipolar, Suicide e Anxiety.

Como trabalho futuro pretendemos explorar as trajetórias dos usuários em termos dos descritores relacionados às suas publicações.Além disso, pretendemos analisar as características de sucesso em uma discussão.

Referências

Althoff, T., Clark, K., and Leskovec, J. (2016). Large-scale analysis of counseling conversations: An application of natural language processing to mental health. *Transactions of the Association for Computational Linguistics*, 4:463.

Choudhury, M. D. and De, S. (2014). Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In *International Conference on Web and Social Media*.

Gkotsis, G., Oellrich, A., Hubbard, T., Dobson, R., Liakata, M., Velupillai, S., and Dutta, R. (2016). The language of mental health problems in social media. In *Proceedings of the Third Workshop on Computational Lingusitics and Clinical Psychology*.

Gkotsis, G., Oellrich, A., Velupillai, S., Liakata, M., Hubbard, T. J., Dobson, R. J., and Dutta, R. (2017). Characterisation of mental health conditions in social media using informed deep learning. *Scientific reports*, 7:45141.

Iyyer, M., Guha, A., Chaturvedi, S., Boyd-Graber, J. L., and III, H. D. (2016). Feuding families and former friends: Unsupervised learning for dynamic fictional relationships. In *HLT-NAACL*, pages 1534–1544.

Kavuluru, R., Ramos-Morales, M., Holaday, T., Williams, A. G., Haye, L., and Cerel, J. (2016). Classification of helpful comments on online suicide watch forums. In *Proceedings of the 7th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, New York, USA.

Lopes, C. R. S., Cunha, M., Rodrigues, A., Vilela, A. B. A., Casotti, C. A., and Pereira, H. (2014). Identificando as representações sociais sobre promoção da saúde em uma rede social de trabalhadores de saúde. In *Proceedings of the III Brazilian Workshop on Social Network Analysis and Mining, Brasília-DF Brazil*.

Pappa, G. L., Cunha, T. O., Bicalho, P. V., Ribeiro, A., Silva, A. P. C., Meira Jr, W., and Beleigoli, A. M. R. (2017). Factors associated with weight change in online weight management communities: A case study in the loseit reddit community. *Journal of medical Internet research*, 19(1).

Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Souza, B. Á. S., Almeida, T. G. A., Menezes, A. A. M., Figueiredo, C. M. F., Nakamura, F. G. N., and Nakamura, E. F. N. (2017). Uma abordagem para detecção de tópicos relevantes em redes sociais online. In *Proceedings of the VI Brazilian Workshop on Social Network Analysis and Mining, São Paulo, SP, Brazil*.

Wang, A., Hamilton, W. L., and Leskovec, J. (2016). Learning linguistic descriptors of user roles in online communities. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 76–85.

Análise de Sentimentos em *Tweets* em Português Brasileiro

Daniel P. Kansaon[1], Michele A. Brandão[2], Saulo A. de Paula Pinto [1]

[1] Pontifícia Universidade Católica de Minas Gerais (PUC-MG) - Belo Horizonte, MG – Brasil

[2]Universidade Federal de Minas Gerais (UFMG) - Belo Horizonte, MG – Brasil

[1]{daniel.kansaon,saulo}@sga.pucminas.br,[2]micheleabrandao@dcc.ufmg.br

Abstract. *There are several studies on sentiment analysis for the English language. In the case of Brazilian Portuguese, the number of papers is smaller because there are not so many datasets available and methods to perform the analysis. This work presents a methodology to compare techniques that classify feelings expressed directly or indirectly in tweets in the Brazilian Portuguese language. In addition, seven classes of feelings are considered and identified in the tweets. The results are promising when classifying distinct feelings, as the best classifier achieves 85% of accuracy. On the other hand, relations between close feelings present results less than 70% of accuracy.*

Resumo. *Existem diversos trabalhos sobre análise de sentimentos para a língua inglesa. No caso do português brasileiro, a quantidade de trabalhos é menor por não existirem tantas bases de dados disponíveis e métodos para realizar a análise. Este trabalho apresenta uma metodologia para comparar técnicas que classificam sentimentos expressos diretamente ou indiretamente em tweets no idioma português brasileiro. Ademais, são consideradas e identificadas sete classes de sentimentos nos tweets. Os resultados são promissores ao classificar sentimentos distintos, pois o melhor classificador alcança 85% de acerto. Por outro lado, relações entre sentimentos próximos apresentam resultados inferiores a 70% de acerto.*

1. Introdução

Dados são coleções de valores que podem ser explorados e processados a fim de se obter padrões, associações, mudanças e anomalias em prol de algum objetivo. Nas últimas décadas, devido à globalização, a produção de dados na forma digital tem aumentado exponencialmente. Algumas estimativas apontam que a quantidade de informação produzida no mundo se duplica a cada vinte meses [Dwivedi et al., 2016]. Um bom processamento e manipulação dessas informações pode trazer benefícios a organizações, grupos de investidores e até mesmo pessoas que desejam extrair algum tipo de informação sobre um conjunto de dados [Dwivedi et al., 2016].

Com o avanço tecnológico e a popularização das redes sociais online, as pessoas utilizam cada vez mais essas redes como um grande aliado para a comunicação, fazendo com que as redes sejam uma grande fonte de informação. Devido a grande quantidade de dados disponíveis, as análises dessas redes podem trazer informações valiosas para as várias empresas que buscam informações sobre seus produtos, clientes e que procuram entender melhor o mercado e os concorrentes, por exemplo.

Uma das técnicas que auxiliam essas avaliações é a análise de sentimentos, que tem como ideia principal extrair e descobrir qual o sentimento expresso em um texto. Assim, é possível classificar a sua polaridade, que é a tarefa de determinar se um sentimento expresso em um texto é positiva, negativa ou até mesmo neutra [Cavalcanti et al., 2012]. Ademais, a análise de sentimentos é uma das técnicas aplicadas no contexto das redes sociais e está associada com a mineração de dados. Essa técnica pode ser empregada em diversas situações, principalmente, para entender o senso comum sobre um determinado acontecimento, como: eventos, economia e política.

Nesse contexto, este trabalho apresenta uma análise de sentimentos expressos pelos usuários em *tweets,* postagens realizadas no Twitter. A identificação dos sentimentos tem como apoio processos de mineração, que visam extrair informações relevantes, utilizando-se de métodos de classificação para determinar o sentimento expresso nos *tweets*. Assim, o foco deste trabalho é responder à pergunta: quais algoritmos de classificação podem melhor descobrir sentimentos em *tweets* em português brasileiro? Em outras palavras, uma pesquisa exploratória é realizada com objetivo principal de verificar o desempenho de diferentes algoritmos de classificação ao serem aplicados na análise de sentimentos semelhantes e distintos em *tweets* em português brasileiro.

O restante deste artigo está organizado da seguinte forma. A Seção 2 apresenta os trabalhos relacionados. A Seção 3 descreve a metodologia utilizada e a Seção 4 descreve os resultados obtidos nos experimentos, detalhando a aplicação de cada método. Já a Seção 5 discute os resultados experimentais obtidos. Finalmente, a Seção 6 apresenta conclusões e comenta sobre trabalhos futuros.

2. Trabalhos Relacionados

A mineração de dados juntamente com a análise de sentimentos tem ampla aplicabilidade em várias áreas do conhecimento, por exemplo, jogos, educação e política, e sempre busca informações significativas em grandes volumes de dados. Assim, existem diversos trabalhos recentes que enfatizam o uso da mineração de dados, especificamente, a aplicação de técnicas de classificação.

Nesse cenário, Bouazizi e Ohtsuki [2016] propõem uma abordagem para quantificação de sentimentos no Twitter. Tal trabalho destaca a importância da detecção de sentimentos em *tweets* e apresenta uma maneira de extrair diferentes sentimentos utilizando a técnica de classificação Random Forest. Foram coletados cerca de 40.000 *tweets* e atribuídos um sentimento para cada um, sendo eles: amor, felicidade, diversão, entusiasmo, alívio, ódio, raiva, tristeza, tédio preocupação, surpresa e neutralidade.

Ademais, Pandeye e Rajpoot [2016] analisam sentimentos através da comparação de diferentes técnicas de classificação como J48, Random Forest, Decision Tree, Random Tree, NaiveBayes, SimpleNaiveBayes, NaiveBayes e DecisionStump. Essa análise foi feita com um conjunto de dados sobre o consumo de álcool por estudantes de uma escola. O algoritmo que realizou a melhor classificação foi Decision Stump, com um resultado de 95,44% de acurácia. Além disso, Caetano et al. [2017] utilizaram a ferramenta SentiStrength para analisar sentimentos no Twitter sobre candidatos das eleições americanas. Tal estudo foi realizado com o objetivo de analisar

Tabela 1. Modelos diferentes de emoções básicas propostas por teóricos. Fonte: adaptada de Yadollahi et al. [2017].

Teórico	Ano	Emoções Básicas
Ekman	1972	raiva, desgosto, medo, alegria, tristeza, surpresa
Plutchik	1986	raiva, antecipação, desgosto, medo, alegria, tristeza, surpresa, confiança
Shaver	1987	raiva, medo, alegria, amor, tristeza, surpresa
Lovheim	2011	raiva, desgosto, angústia, medo, alegria, interesse, vergonha, surpresa

a homofilia (a tendência de pessoas se relacionarem com semelhantes) entre usuários. Já Garimella et al. [2018] realizaram um estudo a fim de quantificar controvérsias em redes sociais baseada em gráficos, tal estudo indica que os gráficos que apresentam temas controversos, possuem uma estrutura de cluster, onde indivíduos com opinião similar tendem a ampliar os argumentos um dos outros.

Em relação à comparação de diferentes métodos de análise sentimentos, Araújo et al. [2015] apresentam um estudo que compara oito diferentes métodos para detecção de sentimentos em *tweets* no idioma inglês. Já Araújo et al. [2016] propôs uma análise de sentimento multilíngue, onde é feita tradução automática de textos para o inglês, e posteriormente utiliza-se métodos existentes na língua inglesa para a análise de sentimento. Ademais, Yadollahi et al. [2017] realizam uma categorização cuidadosa de tarefas na análise de sentimentos e retratam a importância de uma taxonomia clara e lógica dentro dessa análise. Adicionalmente, métodos e melhorias do estado da arte para a análise de sentimentos em textos foram discutidos, além de aspectos teóricos ligados ao surgimento das emoções e sentimentos. Note que os sentimentos estão relacionados às emoções visto que são respostas a elas.

Nesse contexto, a Tabela 1 apresenta algumas emoções básicas que podem ser encontradas em todos os seres humanos, bem como a respectiva classificação de acordo com as teorias da emoção [Yadollahi et al., 2017]. Além disso, Bouazizi e Ohtsuki [2016] também classificam sentimentos considerando um conjunto de classes como: amor, feliz, tristeza e raiva. A análise de tais estudos auxiliou na escolha de sete sentimentos para identificar em *tweets* coletados neste trabalho: Triste, Chateado, Amor, Feliz, Raiva, Inveja, Ironia.

Em relação à comparação de várias técnicas de classificação, Garg e Khurana [2014] propõem um estudo em um contexto diferente da análise de sentimentos. O objetivo é utilizar algoritmos de classificação para projetar um modelo eficaz de detecção de intrusão, impedindo que as redes de computadores sejam invadidas. Os métodos foram avaliados usando 41 atributos e cerca de 94.000 instâncias para o conjunto de treinamento, além de 48.000 instâncias para o conjunto de teste. Por fim, foi apresentada uma lista de 45 métodos de classificação ordenados pelos seus respectivos resultados. Os cinco algoritmos com melhores resultados são Rotation Forest, Random Tree, Random Committee, Random Forest e IBK.

Figura 1. Etapas da metodologia.

Todos esses estudos mostram desempenhos diferentes para algoritmos de classificação a depender da base de dados utilizada. Assim, este trabalho contribui com o estado da arte ao considerar o idioma português brasileiro para realizar análise de sentimentos em *tweets*. Considerar tal idioma não é muito comum por não existirem muitas bases de dados disponíveis [Neri et al., 2012], além de ter poucos métodos e ontologias disponíveis no português brasileiro para realizar análises do texto. Outra contribuição é a metodologia proposta, que inclui coleta e processamento dos dados, identificação de diferentes classes de sentimentos, seleção dos algoritmos de classificação e descrição do treinamento dos classificadores.

3. Metodologia

Esta seção apresenta os processos de coleta, processamento e preparação dos dados para a realização da análise de sentimentos, bem como a descrição dos algoritmos de classificação selecionados e o processo de treinamento dos classificadores. Especificamente, este trabalho identifica sentimentos e emoções expressos em *tweets* através de métodos de classificação. Para atingir o objetivo, foi necessário realizar uma série de etapas para adaptar os dados ao formato compatível com o ambiente WEKA, que possibilita o uso de diferentes algoritmos.

A Figura 1 exemplifica as etapas principais da metodologia. Cada etapa pode ser executada mais de uma vez. As seções a seguir detalham cada uma dessas etapas.

3.1. Seleção dos Sentimentos

A definição das classes de sentimentos analisados durante a pesquisa tem como base os diversos sentimentos que podem ser expressos através dos *tweets*. As expressões podem ser identificadas de várias formas, através de *emojis*, *emoticons*, *hashtag*. A *hashtag* é composta por uma ou algumas palavras-chave relacionadas ao assunto abordado na postagem. Elas são representadas pelo símbolo de cerquilha (#), seguida da palavra-chave. Posteriormente, as *hashtags* são indexadas para utilização dos mecanismos de busca. Os *emojis* e *emoticons* são amplamente usados no Twitter e estão diretamente associados a emoções. Os *emojis* são exibidos como imagens claras e podem conter animações. Por outro lado, os *emoticons*, são formas mais simples de expressar emoção e, geralmente, são representados por uma sequência de caracteres, por exemplo: {:(, :D, :/} [Matsuda, 2017].

Os sentimentos analisados neste trabalho foram escolhidos através do estudo de outros trabalhos apresentados na Seção 2. Além disso, o estudo feito por Yadollahi et al. [2017], oferece uma contribuição importante para a escolha dos sentimentos aqui analisados, pois apresenta as emoções básicas que podem ser encontradas nos seres humanos (Tabela 1). Dessa forma, busca-se tweets que expressem os sentimentos: Triste, Chateado, Amor, Feliz, Raiva, Inveja, Ironia. Para coletar *tweets* que expressem

Figura 2. Detalhamento do fluxo de coleta dos dados.

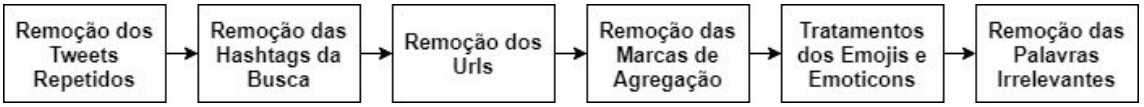


Figura 3. Etapas do processamento de cada *tweet*.

esses sentimentos, foi definido que o *tweet* precisa conter a *hashtag* com o nome do sentimento selecionado. Assim, são considerados os tweets que possuem as hashtags: #Triste, #Chateado, #Feliz, #Amor, #Raiva, #Inveja, #Ironia.

3.2. Coleta dos Dados

Após a seleção dos sentimentos, foi necessário realizar a extração e o armazenamento dos dados. A API oficial do Twitter fornece vários recursos para a coleta de *tweets*. Por meio dela, usuários que se cadastram como desenvolvedores conseguem coletar *tweets* a partir de parâmetros. A API possui algumas restrições para a coleta dos dados, por exemplo, a limitação do número de cem *tweets* como resposta para cada consulta realizada. Apesar disso, ela se apresenta suficiente para realização do estudo. Para considerar o *tweet* na análise, é necessário estar no idioma português brasileiro e possuir as *hashtags* definidas (#Triste, #Chateado, #Feliz, #Amor, #Raiva, #Inveja, #Ironia).

A Figura 2 apresenta o funcionamento do fluxo de coleta dos *tweets.* Toda a comunicação com a API é feita através do protocolo HTTP usando métodos de GET. Todas as respostas retornadas foram armazenadas na base de dados. A coleta foi feita entre os meses de agosto e outubro de 2017, contando com 12.814 *tweets* armazenados na base de dados[1], criada na linguagem SQL (SQL Server) e hospedada no servidor na nuvem Azure. Por se tratar de uma plataforma gerenciada, a base de dados na nuvem oferece alta disponibilidade sem precisar da instalação de qualquer software para a utilização. Todas as informações relevantes dos *tweets* foram armazenadas na base de dados, sendo elas: texto, *hashtag* de busca, autor, data de criação, linguagem, quantidade de URLs, quantidade de menções feitas, texto do URLs e tipo.

3.3. Processamento dos Dados

O processamento realizado tem como objetivo tratar os dados coletados removendo as informações que não contribuem para classificação dos sentimentos. Cada *tweet* contém uma grande quantidade de informações que são irrelevantes para a análise. Por isso, são removidas dos *tweets*. Nesta etapa da metodologia, foram necessários vários passos de

[1] Base de dados construída a partir da coleta de *tweets*: http://bit.ly/dataset-analise-sentimento-BraSNAM

processamento e tratamento dos dados, realizados de forma sequencial, conforme mostra a Figura 3.

Amor x Triste	**Feliz x Triste**	**Feliz x Chateado**	**Amor x Feliz**	**Triste x Chateado**

Figura 4. Relações de sentimentos.

Na coleta dos *tweets*, foi observado que a API do Twitter retorna informações repetidas. Com isso, é necessário remover todos os *tweets* repetidos na base de dados, para que não influencie o resultado final ao aplicar os algoritmos. Em seguida, as *hashtags* usadas para a coleta dos *tweets* são removidas dos textos, os quais possuem *links* compartilhados e até mesmo notações utilizadas para mencionar outros usuários. Assim, as marcas de agregação identificadas através da sigla RT ou @ nome do usuário são removidas juntamente com os *links* encontrados. O ambiente WEKA não possui suporte a caracteres Unicode. Assim, os *emojis*, por se tratarem de caracteres com formato específico, são transformados em palavras-chave, formadas pela junção da letra "E" maiúscula, completada posteriormente com uma palavra que representa o respectivo símbolo do *emoji*, exemplo: *ECoracao*, *EBravo*, *ESorridente*. Por fim, as palavras conhecidas como *stop words* (a, as, de, para, etc) também são removidas.

Além das etapas de processamento, um critério para nivelamento dos dados é considerado com base na quantidade de *tweets* coletados. É necessário que os sentimentos tenham uma quantidade suficiente de exemplos ou instâncias para serem analisados, pois um número pequeno de dados pode levar a um resultado não fidedigno. Para evitar essa situação, foi definida uma premissa: um sentimento apenas é analisado, se possuir um mínimo de mil *tweets*. Com isso, os sentimentos: Ironia, Inveja e Raiva foram excluídos. Após a exclusão de tais sentimentos, a quantidade total de *tweets* disponíveis passou para 9.631, assim, representando uma média de dois mil *tweets* para cada sentimento restante.

3.4. Preparação dos Dados

Após os tratamentos realizados, os dados coletados são separados em classes de sentimentos. Algumas relações entre sentimentos são criadas para avaliar se os métodos de classificação conseguem distinguir as diferenças entre os sentimentos. As relações são formadas visando comparar os sentimentos separando em: positivos x negativos, positivos x positivos e negativos x negativos. Assim, os sentimentos amor e feliz são considerados sentimentos positivos, já os sentimentos triste e chateado, representam os sentimentos negativos. Dessa forma, a Figura 4 apresenta as relações formadas para a análise de sentimentos.

Em seguida, é necessário exportar os dados para o formato "ARFF" (Attribute Relation Format File), do ambiente WEKA. Para isso, é criado um arquivo para cada relação de sentimentos selecionados anteriormente, com os requisitos exigidos pelo WEKA. Em tal arquivo, os dados foram divididos em dois atributos, o primeiro chamado "Descricao" que contém o texto do *tweet* a ser analisado e o segundo atributo, chamado de "Sentimento" que define qual classe de sentimento o *tweet* se enquadra.

3.5. Seleção dos Algoritmos de Classificação e Treinamento

Neste trabalho, os algoritmos de classificação baseados no modelo Naive Bayes são algumas das abordagens escolhidas. Esses algoritmos utilizam a probabilidade condicional para criar o modelo de dados a ser trabalhado. São considerados algoritmos como o Naive Bayes Multinominal, no qual a classe de um documento (*tweets*) é determinada não apenas pelas palavras existentes, mas também pela frequência que ocorrem [Witten et al., 2016]. Ademais, França e Oliveira [2014] utilizam o algoritmo Naive Bayes no idioma português brasileiro e apresentam resultados de até 90% de acurácia ao classificar polaridades expressas nos *tweets* relacionados aos protestos ocorridos no Brasil em 2013.

Além dos algoritmos baseados no Naive Bayes, alguns algoritmos que apresentaram bons resultados em outros trabalhos, também foram selecionados. Isso permite verificar se os resultados obtidos nos trabalhos se repetem na classificação de *tweets* no idioma português brasileiro. Em Garg e Khurana [2014], são avaliados o desempenho de vários métodos de classificação, conforme descrito na Seção 2. Dessa forma, foram escolhidos os algoritmos que apresentaram bons resultados e que são compatíveis com o formato dos dados trabalhados. Sendo eles: IBK, Forest e Random Committee [Garg e Khurana, 2014]. Em resumo, os algoritmos selecionados para a classificação dos *tweets* são: Naive Bayes, Naive Bayes Multinominal, Naive Bayes Multinominal Updateable, Sparge Generative Model, DMNB Text, Complement Naive Bayes, Bayesian Logistc Regression, IBK, Forest e Random Committee.

Para treinar o classificador, é comum usar um terço dos dados para testes e dois terços para treino, mas o problema dessa abordagem é que a parte usada para treino pode não ser representativa do problema e nem do conjunto de testes. A base de dados deve ser representada na proporção certa, pois se o problema não estiver bem representado, dificilmente qualquer algoritmo apresentará um bom resultado. Uma maneira de mitigar qualquer parcialidade causada pela amostragem, é utilizar o tipo de treinamento conhecido como validação cruzada (de dez partições ou vias). Nesse caso, os dados são divididos em dez partições de tamanhos aproximadamente iguais e cada uma delas é usada para testes e o restante é usado para treinamento. Esse processo é repetido dez vezes para que no final, cada instância seja usada uma vez para testar. Para obter a melhor estimativa de erro, é indicado usar dez partições [Witten et al., 2016].

4. Resultados

Após a realização das etapas da metodologia, os algoritmos de classificação e a análise dos resultados foram realizados. Assim, esta seção apresenta a avaliação dos algoritmos quanto à sua capacidade de classificar e diferenciar os sentimentos expressos nos *tweets*. Especificamente, são apresentadas as palavras identificadas como relevantes para identificar cada sentimento (Seção 4.1) e os resultados de uma análise comparativa dos algoritmos de classificação (Seção 4.2).

Tabela 2. Palavras consideradas relevantes para a análise de cada sentimento. As cores representam palavras semelhantes encontradas em classes diferentes.

#	Amor	Feliz	Triste	Chateado
1	ECoracao	ESorridente	ERostoFranzido	ERostoDesanimado
2	ESorridente	ECoracao	EChorando	ERostoFranzido
3	Amor	Boas	ERostoDesanimado	Tristeza
4	#Love	Alegria	Triste	ESemExpressao
5	Amamos	Agradecer	Acabando	Poxa
6	#Paixao	Especial	Coitado	Triste
7	Amado	Feliz	ESemExpressao	Raiva
8	Desejo	Sorriso	Infelizmente	Ruim
9	Flores	Sucesso	Acabaram	Problemas
10	Fofura	Excelente	Aff	Sozinho

4.1. Relevância das Palavras em Cada Sentimento

Uma das formas de entender os resultados dos algoritmos é analisar quais elementos foram importantes para a detecção dos sentimentos. A Tabela 2 apresenta as dez principais palavras que foram determinantes para a diferenciação dos sentimentos ao aplicar os algoritmos de classificação. As palavras estão ordenadas pela sua relevância e algumas delas se repetem em determinados sentimentos. Por exemplo, os *emojis*: *ERostoDesanimado* e *ERostoFranzido* são encontrados no sentimento Triste e Chateado, isso ocorre devido à proximidade dos dois sentimentos. Quando o sentimento chateado é expresso, logo, o sentimento de tristeza também pode ser expresso, pois são sentimentos relacionados. Essa proximidade também acontece nos sentimentos Amor e Feliz, com os *emojis ECoracao* e *ESorridente*, por exemplo.

Foram considerados quatro mil *tweets* para cada relação, sendo dois mil para cada sentimento da relação. Essa equiparação foi feita para que os algoritmos classificassem os sentimentos com a mesma quantidade de instâncias, evitando resultados influenciados pelo desbalanceamento das classes, ou seja, uma classe ter uma quantidade de instâncias excessivamente maior que a outra.

4.2. Análise Comparativa dos Algoritmos de Classificação

Com base na aplicação dos algoritmos e da divisão das diversas classes de sentimentos, alguns algoritmos se destacaram na classificação, já outros não apresentaram bons resultados. Em geral, os algoritmos baseados no modelo de Bayes foram os que apresentaram os melhores resultados.

Considerando o maior valor de acerto para combinação de classes que representam sentimentos semelhantes (positivo x positivo e negativo x negativo) e distintos (positivo x negativo), as taxas de acerto de classificação de sentimentos foram 85,54% para o Naive Bayes Multinominal Updateable, 85,41% para o Naive Bayes Multinominal e 85,64% para Complement Naive Bayes.

Tabela 3. Taxa de acerto dos melhores algoritmos para detectar sentimentos positivos x negativos.

Algoritmo	Amor x Triste	Feliz x Triste	Feliz x Chateado	Taxa de Acerto Média
Naive Bayes Multinominal Updateable	85,54%	81,35%	79,60%	82,16 %
Naive Bayes Multinominal	85,41%	81,02%	79,60%	82,01 %
Complement Naive Bayes	85,64%	80,34%	79,54%	81,84 %

As relações compostas por sentimentos distintos positivos x negativos apresentaram melhores resultados. Os três algoritmos responsáveis por tal desempenho são apresentados na Tabela 3. Em primeiro lugar, a relação Amor x Triste teve 85,64% de acerto na classificação com o método Complement Naive Bayes (a Tabela 4 apresenta resultados para todos os algoritmos de classificação aplicados para essa relação, o que permite observar as métricas F-Measure e ROC[2] com valores maiores também para Complement Naive Bayes em relação a outros algoritmos). Uma das razões para esse resultado é a quantidade de *emojis* encontrados na relação. Os *emojis Ecoracao* e *Esorridente* foram encontrados diversas vezes no sentimento amor. Já no sentimento triste, não foi detectado nenhuma ocorrência desses *emojis*, conforme mostra a Tabela 2. Em seguida, a relação Feliz x Triste obteve 81,35% de acerto com o método Naive Bayes Multinominal Updateable. Os *emojis ECoração* e *ESorridente* encontrados na classe feliz, também não são localizados na classe triste. Por fim, a relação Feliz x Chateado ficou em terceiro lugar, com 79,60% de acerto na classificação, com os algoritmos Naive Bayes Multinominal e Naive Bayes Multinominal Updateable. Vale ressaltar que as palavras relevantes para a classificação dos sentimentos positivos x negativos não se repetem. Consequentemente, essas relações são as que apresentam os melhores resultados comparado com as demais.

Em contrapartida, as relações compostas por sentimentos parecidos positivos x positivos ou negativos x negativos apresentam os piores resultados. Os algoritmos usados para a classificação não são capazes de distinguir ou encontrar alguma palavra que expresse um sentimento de maneira que diferencie um do outro. Como consequência, os resultados das relações representadas por sentimentos semelhantes são inferiores às relações compostas por sentimentos distintos. Ademais, nos sentimentos triste e chateado, quatro das dez palavras consideradas relevantes para a classificação se repetem, sendo elas: *ERostoFranzido*, *ESemExpressao*, *Triste* e *ERostoDesanimado*. Isso indica a semelhança entre os sentimentos, refletindo nos resultados obtidos.

A Tabela 5 mostra a taxa de acerto dos três algoritmos que apresentam os melhores resultados nas relações de sentimentos semelhantes. O método Naive Bayes Multinominal Updateable obteve melhor resultado dos algoritmos aplicados na relação

[2] F-Measure and ROC (*Receiver Operating Characteristic*) são métricas comumente utilizadas para avaliar a qualidade de algoritmos de aprendizagem de máquina.

Tabela 4. Resultado da aplicação dos algoritmos na relação Amor x Triste.

Algoritmo	Amor		Triste		Acerto (%)	Erro (%)
	F-Measure	ROC Area	F-Measure	ROC Area		
Naive Bayes	0,59	0,87	0,77	0,87	70,72	29,27
Naive Bayes Multinominal	0,84	**0,92**	**0,86**	**0,92**	85,41	4,58
Naive Bayes Multinominal Updateable	0,84	**0,92**	**0,86**	**0,92**	85,54	4,45
Sparge Generative Model	0,82	**0,92**	0,84	**0,92**	83,84	6,15
DMNB Text	0,82	**0,92**	**0,86**	**0,92**	84,50	5,49
Complement Naive Bayes	**0,84**	0,85	**0,86**	0,85	**85,64**	**4,35**
Bayesian Logistic Regression	0,79	0,80	0,80	0,80	80,38	9,61
IBK	0,76	0,83	0,69	0,83	73,59	6,40
Ramdom Forest	0,82	0,91	0,82	0,91	83,11	6,88
Ramdom Committee	0,80	0,88	0,80	0,88	81,03	18,96

Tabela 5.Taxa de acerto dos algoritmos para sentimentos negativos x negativos e positivos x positivos.

Algoritmo	Amor x Feliz	Triste x Chateado	Taxa de Acerto Média
Naive Bayes Multinominal Updateable	73,59%	66,48%	70,03 %
Naive Bayes Multinominal	73,25%	65,16%	69,20 %
Complement Naive Bayes	72,95%	64,14%	68,54 %

relação Chateado x Triste, com 66,48% de classificação correta. Já nos sentimentos feliz e amor, os *emojis*: *ECoracao* e *ESorridente* aparecem em ambos sentimentos, indicando a proximidade entre eles. Apesar da semelhança, a relação Amor x Feliz obteve um resultado de 73,59% de classificação correta com o método Naive Bayes Multinominal Updateable, sendo superior ao resultado obtido pela relação Triste x Chateado.

Os algoritmos IBK, Random Forest, Random Committee não apresentam resultados eficientes quando comparados com os resultados apresentados no trabalho feito por Garg e Khurana [2014]. Isso se dá pelo fato do formato dos dados serem diferentes e em outro idioma. Outro ponto importante que distingue os dois trabalhos é a quantidade de dados usados para as etapas de processamento e teste. Em Garg e Khurana (2014) foram considerados cerca de 142.000 para o conjunto de treino e teste.

Além das etapas de treino e de processamento dos dados, que foram importantes para a classificação dos sentimentos, os *emojis* se mostram essenciais para a classificação. Tais ideogramas representam as principais palavras para caracterizar um sentimento, diferenciando um sentimento positivo de outro negativo.

5. Conclusão

Neste trabalho, foram utilizados métodos de classificação baseados no modelo Naive Bayes e em outros modelos de classificação, como árvores de decisão, para classificar *tweets* em português brasileiro. Apesar de ser um tema popular, ainda existem vários desafios para a descoberta de sentimentos, especialmente, em textos na língua portuguesa. O estudo considerou sete sentimentos: triste, chateado, amor, feliz, raiva, inveja, ironia. Os dados foram obtidos do Twitter, no qual foram coletados 12.814 *tweets*. Ao final do processamento, inveja, ironia e raiva foram excluídos, restando 9.631 *tweets*. Após a coleta, os dados foram separados em relações de sentimentos: positivo x negativo, negativo x negativo e positivo x positivo, para, assim, serem aplicados os algoritmos de classificação. Os resultados mostraram que os algoritmos baseados no modelo Naive Bayes apresentaram melhor acurácia. As relações que possuem os sentimentos positivos x negativos foram as que obtiveram os melhores resultados, chegando a 85% de acerto na classificação com o Complement Naive Bayes. Apesar dos métodos produzirem resultados inferiores quando aplicados no português brasileiro, os resultados foram satisfatórios e de acordo com outras pesquisas na área.

Para os trabalhos futuros, planeja-se ampliar a coleta de dados, aplicar mais métodos de classificação e comparar com resultados obtidos em textos na língua inglesa. Ademais, métodos de análise de sentimentos podem ser utilizados a fim de realizar uma análise mais detalhada. Por exemplo, explorar a importância dos *emojis e emoticons* na identificação dos sentimentos expressos e detectar múltiplos sentimentos.

Agradecimentos. Este trabalho foi parcialmente financiado pelo CNPq.

7. Referências

Araújo, M.; Gonçalves, P.; Benevenuto, F. (2015). Métodos para Análise de Sentimentos em Mídias Sociais. In Procs. of *Simpósio Brasileiro de Multimídia e Web* (WEBMEDIA), p. 27 - 30, Manaus, Brasil.

Araújo, M.; Reis, J.; Pereira, A.; Benevenuto, F. (2016). An Evaluation of Machine Translation for Multilingual Sentence-level Sentiment Analysis. In Procs. of ACM Symposium on Applied Computing (SAC), p. 1140 - 1145, Pisa, Itália.

Bouazizi, M.; Ohtsuki, T. (2016). Sentiment Analysis in Twitter: From Classification to Quantification of Sentiments within Tweets. In Procs. of *Global Communications Conference* (GLOBECOM). p. 1 - 6, Washington, USA.

Caetano, J. A. C.; Lima, H. S. L.; dos Santos Santos, M. F.; Marques-Neto, H. T. M. N. (2017). Utilizando Análise de Sentimentos para Definição da Homofilia Política dos Usuários do Twitter durante a Eleição Presidencial Americana de 2016. In Procs. of *Brazilian Workshop on Social Network Analysis and Mining* (*BraSNAM)*, p. 480-491, São Paulo, Brasil.

Cavalcanti, D. C. et al. (2012). Análise de Sentimento em Citações Científicas para Definição de Fatores de Impacto Positivo. In Procs. of *International Workshop on Web and Text Intelligence* (WTI), 4. p. 1 - 10, Paraná, Brasil.

Dwivedi, S.; Kasliwal, P.; Soni, S. (2016). Comprehensive study of data analytics tools (RapidMiner, Weka, R tool, Knime). In Procs. of *Symposium On Colossal Data Analysis And Networking* (*CDAN*), p. 1 - 8, Indore, India.

França, Tiago C. de; Oliveira, Jonice. (2014). Análise de Sentimento de Tweets Relacionados aos Protestos que ocorreram no Brasil entre Junho e Agosto de 2013. In Procs. of *Brazilian Workshop on Social Network Analysis and Mining* (*BraSNAM)*, p. 128 - 139, Brasília, Brasil.

Garg, T.; Khurana, S. S. (2014). Comparison of classification techniques for intrusion detection dataset using WEKA. In Procs. of *International Conference on Recent Advances and Innovations in Engineering* (*ICRAIE*). p. 1 - 5, Jaipur, India.

Garimella, Kiran. et al. (2018). Quantifying Controversy in Social Media. *Journal ACM Transactions on Social Computing*, v. 1, n. 1, p. 3:1--3:27.

Matsuda, Y. (2017). Development of Emotion Teaching Interfaces using Emoticons and Emojis. In Procs of *International Conference On Intelligent Human-Machine Systems and Cybernetics* (*IHMSC*), 9, p. 253 - 258, Hangzhou, China.

Neri, F; Aliprandi, C; Capeci, F; Cuadros, M. (2012) Sentiment Analysis on Social Media. In Procs. of *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (*ASONAM*), p. 919 - 926, Istanbul, Turquia.

Pandey, A. K.; Rajpoot, D. S. (2016). A comparative study of classification techniques by utilizing WEKA. In Procs. of *International Conference On Signal Processing And Communication* (*ICSC*). p. 219 - 224, Bangalore, India.

Witten I. H.; Frank, E; Hall, M. A.; Pal, C. J. (2016). Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann Publishers, 4th edition. ISBN: 978-0128042915.

Yadollahi, A.; Shahraki, A. G.; Zaiane, O. R. (2017). Current State of Text Sentiment Analysis from Opinion to Emotion Mining. *ACM Computing Surveys*, v. 50, n.2, p. 1 - 33.

Caracterização e Análise das Redes de Colaboração Científica dos Bolsistas de Produtividade em Pesquisa do CNPq

Thiago M. R. Dias[1], Tales H. J. Moreira[1], Patrícia M. Dias[2]

[1]Centro Federal de Educação Tecnológica de Minas Gerais – CEFET-MG
[2]Universidade do Estado de Minas Gerais - UEMG

thiagomagela@cefetmg.br, tales.info@gmail.com, patrícia.dias@uemg.br

Abstract. *Studies of scientific collaboration networks have for some time been receiving attention from analysts in various fields of knowledge because of their potential to identify how groups of researchers have collaborated in their research. Such studies make it possible to identify with the aid of network analysis metrics, how networks are formed, how they evolve over time, and how they are structured. In this research, a characterization and analysis of the scientific collaboration networks of the CNPq Research Productivity Grantees is presented. As a result, it is presented as the main Brazilian researchers, who receive incentive for excellence in their research has collaborated with each other.*

Resumo. *Estudos sobre as redes de colaboração científica vêm a algum tempo recebendo atenção de analistas de diversas áreas do conhecimento devido seu potencial de identificar como grupos de pesquisadores têm colaborado em suas pesquisas. Tais estudos possibilitam identificar com o auxílio de métricas de análises de redes, como as redes são formadas, como evoluem ao longo do tempo e como são estruturadas. Neste trabalho é apresentada uma caracterização e análise das redes de colaboração científica dos Bolsistas de Produtividade em Pesquisa do CNPq. Como resultado, é apresentado como os principais pesquisadores brasileiros, que recebem incentivo pela excelência em suas pesquisas tem colaborado entre si.*

1. Introdução

Com o crescimento da divulgação científica impulsionada principalmente pela disponibilidade e acesso imediato a repositório de dados de publicações científicas, uma nova geração de serviços disponíveis principalmente na Web está mudando a forma de divulgar e disponibilizar a produção científica e tecnológica. Existe, atualmente, uma tendência que reforça a troca de informações e a colaboração entre as pessoas. A forte relação entre os domínios científico e socioeconômico tem gerado um interesse crescente pela compreensão dos mecanismos que norteiam as atividades científicas, sendo possível apontar diversos trabalhos que analisam aspectos específicos como as características da linguagem e dos discursos empregados na comunicação científica (HOFFNAGEL, 2009) ou, ainda, a relação de colaboração entre pesquisadores e grupos de pesquisa (DING, 2011; REVOREDO et al., 2012; STROELE; ZIMBRÃO; SOUZA, 2012).

Nesse cenário, a Bibliometria (ARAÚJO, 2006), que possui grande relação com a Cientometria (SILVA; HAYASHI; HAYASHI, 2011), se destaca como uma das principais ciências métricas de análise de conteúdo, podendo ser aplicada a fontes de dados científicos com o intuito de se obter informações quantitativas sobre publicações (PRITCHARD, 1969). Diante disso, é possível analisar dados de publicações científicas com o objetivo de identificar as tendências e o crescimento do conhecimento em diversas áreas, prever padrões de pesquisa, observar a dispersão do conhecimento científico, auxiliar políticas de auxílio à pesquisa e entender como ocorre a evolução científica de uma determinada área do conhecimento ou de grupos de pesquisadores.

A produção de indicadores bibliométricos considerados mais representativos tornou-se realidade concreta na s últimas décadas do século XX. Os principais motivos para a proposta de tais indicadores devem-se à construção, manutenção e informatização de repositórios de dados científicos (MUGNAINI; JANNUZZI; QUONIAM, 2004). Por sua vez, um dos fatores que têm motivado os estudos de avaliação da produção científica, principalmente com a adoção de métricas bibliométricas, tem sido a busca pela excelência em áreas de pesquisa e, também, a competição pelos recursos financeiros das agências de fomento (LOPES, 2012).

Nos últimos anos, além da análise bibliométrica de publicações científicas, diversos outros estudos têm procurado compreender como a ciência tem evoluído e como a colaboração científica ocorre. Diante disso, técnicas baseadas em análises de redes surgem como uma alternativa para verificar esse fenômeno. De modo geral, uma rede pode ser caracterizada como um grafo, que consiste de um conjunto de nós (vértices) e ligações (arestas) entre os nós (SZWARCFITER, 1986). Essas ligações podem ser direcionadas ou não e, opcionalmente, podem ainda ter um peso associado.

No domínio científico, um exemplo de uma rede social é a de colaboração científica, que pode ser observada como uma rede na qual os nós correspondem aos autores de publicações científicas e as arestas correspondem à relação de coautoria. Nesse tipo de rede, as arestas podem ou não ser ponderadas. A adição de um peso representa o número de trabalhos em que os autores relacionados pela aresta considerada participaram conjuntamente. Dessa forma, a intensidade dos relacionamentos presentes em uma rede de colaboração científica é medida pelo número de colaborações entre um par de autores. A presença do peso é útil para representar, por exemplo, a afinidade e os interesses comuns entre dois autores da rede (SONNENWALD, 2007).

Com a modelagem e caracterização das redes de colaboração científica, é possível aplicar diversas técnicas que permitem entender como essas redes são estruturadas, fornecendo assim subsídios para diversos estudos como predição de vínculos entre pesquisadores, recomendação de especialistas e identificação de grupos de pesquisa.

Neste contexto, este trabalho visa responder a seguinte questão: Como ocorre a colaboração científica entre os principais pesquisadores do Brasil ?

Atualmente no Brasil, pesquisadores com elevada capacidade de pesquisa são reconhecidos com o recebimento de uma bolsa de produtividade, paga pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pela excelência em suas pesquisas. A bolsa de produtividade em pesquisa do CNPq é destinada aos pesquisadores

que se destaquem entre seus pares na realização de pesquisas nas áreas científicas e tecnológica. Dessa forma, o CNPq oferece um estímulo constante aos pesquisadores de excelência no país, valorizando a qualidade, o aprofundamento e a possível aplicação de novos estudos (SANTOS, 2016).

Diante disso, este trabalho apresenta um estudo sobre as redes de colaboração científica dos Bolsistas de Produtividade em Pesquisa do CNPq, baseado em análises de redes sociais que possibilitam compreender como ocorre a colaboração científica do conjunto, e como estão colaborando os pesquisadores brasileiros de excelência nas diversas modalidades de bolsas. Para isso, são analisados dados dos currículos cadastrados na Plataforma Lattes do CNPq responsável por armazenar informações curriculares da comunidade científica nacional. Atualmente a Plataforma Lattes é considerada uma importante fonte de dados sobre a produção científica brasileira (LANE, 2010).

O restante deste trabalho está organizado da seguinte forma. A Seção 2 apresenta alguns trabalhos relacionados ao conteúdo do artigo, a Seção 3 descreve como foi construído e processado o conjunto de dados utilizados, a Seção 4 apresenta os resultados das análises realizadas e a Seção 5 apresenta as considerações finais e propostas para novos estudos.

2. Trabalhos Relacionados

Barabási et al. (2002) apresentam um estudo das redes de colaboração científica nas áreas de Neurociência e Matemática entre os anos de 1991 e 1998. Os autores concluíram que a rede evolui a partir do momento em que novos nós e vínculos entre os nós existentes são incluídos. E que esta evolução segue um processo denominado *preferential attachment*, em que novos nós tendem a criar seu primeiro vínculo com nós que possuem grande número de vínculos. Como resultado, autores experientes tendem a aumentar seu número de colaboradores, com maior frequência do que autores novatos.

Para Revoredo et al. (2012), redes como as de comunidades científicas de formação recente possuem poucos parâmetros para classificação de assuntos de interesse e pouco entendimento tanto da existência como o potencial de relações de colaboração. Nestas comunidades, a compreensão de sua composição e tendências de interesse se beneficia de técnicas de descoberta de conhecimento a partir de seus artefatos principais de produção, as publicações.

Segundo Stroele, Zimbrão e Souza (2012), a análise de redes de colaboração científica possibilita identificar como os grupos de pesquisadores e centros de estudos estão desenvolvendo seus trabalhos, verificar e entender qual o grau de envolvimento entre os pesquisadores de determinados grupos, de determinadas áreas do conhecimento, de instituições de ensino e pesquisa, e também permitir a indicação de padrões de colaboração que poderiam proporcionar um grande avanço na área, permitindo melhorias na comunicação e colaboração de toda a comunidade científica.

Mena-Chalco et al. (2014) utilizam dados dos currículos da Plataforma Lattes para identificar e caracterizar a rede de colaboração de pesquisadores brasileiros. O trabalho objetivou extrair os dados de currículos cadastrados na Plataforma Lattes, identificar automaticamente a colaboração baseada em informações bibliométricas, produzindo uma rede de colaboração, e aplicar métricas baseadas em análise topológica para compreender como ocorre a interação entre os pesquisadores.

No trabalho de Digiampietri et al. (2017), é apresentada uma análise da evolução, impacto e formação de redes nos cinco anos do BraSNAM (*Brazilian Workshop on Social Network Analysis and Mining*). Os autores analisaram baseado nas cinco primeiras edições do evento, a produção bibliográfica e a evolução da rede de coautoria. No estudo das redes de coautoria os trabalhos publicados em cada edição foram utilizados para a caracterização de cada uma das redes. Foi observado uma evolução do evento que anualmente tem atraído novos pesquisadores, contribuindo para a expansão e consolidação do BraSNAM.

Como diferencial deste trabalho em relação aos citados anteriormente, este estudo visa apresentar como ocorre a colaboração científica dos principais pesquisadores brasileiros considerando todo o conjunto de Bolsistas de Produtividade em Pesquisa do CNPq, bem como, por níveis das bolsas que recebem como auxílio.

3. Materiais e Métodos

Para a análise da colaboração científica do conjunto de Bolsistas de Produtividade em Pesquisa do CNPq, foram utilizados dados extraídos de seus currículos cadastrados na Plataforma Lattes.

Uma grande parte dos editais de financiamento de projetos de pesquisa realizados por agências de amparo à pesquisa, por instituições de ensino, bem como pelo próprio CNPq, utilizam dados registrados nos currículos cadastrados na Plataforma Lattes dos proponentes como uma das formas de avaliação das propostas. Isto passou a ser um grande incentivo para que os pesquisadores mantenham seus currículos com informações atualizadas, tornando a Plataforma Lattes uma fonte extremamente rica para análise da produção científica brasileira.

Para o processo de coleta e tratamento dos dados foi utilizado um *framework* denominado LattesDataXplorer (Dias, 2016). Ele é responsável por englobar todo um conjunto de técnicas e métodos para realizar a extração de todo o conjunto de currículos cadastrados na Plataforma Lattes, e realizar diversos processos para o tratamento, seleção e análises dos dados curriculares (Figura 1).

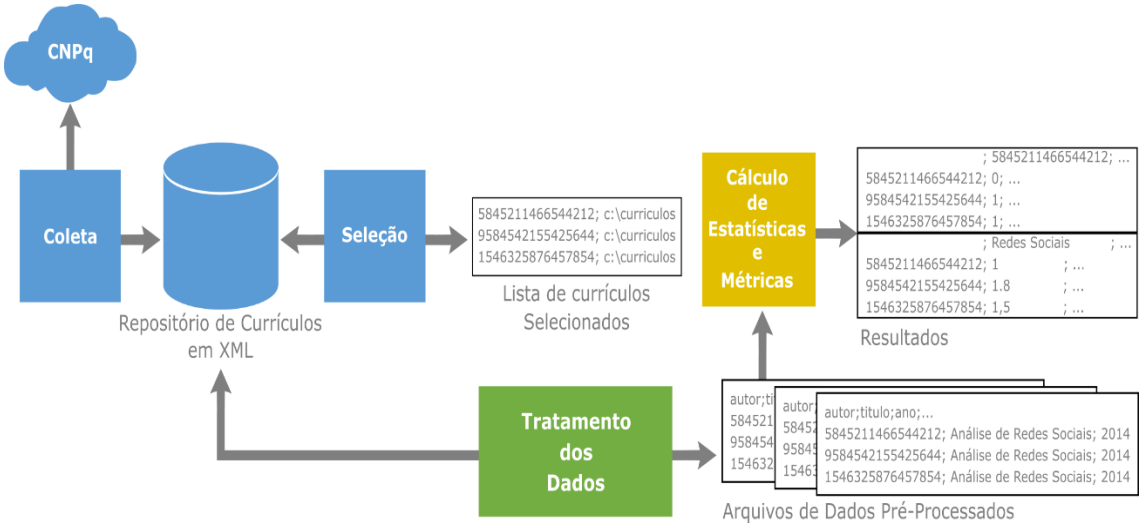


Figura 1. LattesDataXplorer, *framework* para extração e tratamento de dados curriculares da Plataforma Lattes (Dias, 2016).

O processo de coleta de todos os dados curriculares da Plataforma Lattes é dividido em três componentes que objetivam minimizar o custo computacional: 1) extração de URLs, que é responsável por extrair as referências únicas para todos os currículos cadastrados e desta forma possibilitar o acesso individual a cada currículo, 2) extração de Identificadores (Ids) e Data, que visa acessar cada currículo e extrair o seu identificador individual, bem como a data de última atualização, 3) extração de currículos, que é responsável por extrair e armazenar os currículos (em formato XML) cuja data de atualização na Plataforma Lattes seja divergente da data de atualização do currículo armazenado localmente.

Todas essas etapas se fazem necessárias, já que o ideal é manter os dados curriculares atualizados com a maior frequência possível, possibilitando a realização de análises com dados atualizados, tendo em vista que, com a estratégia adotada, não se faz necessário extrair todo o repositório de dados a cada nova extração. Além dos componentes responsáveis pelo processo de coleta dos dados, e considerando a necessidade de análises de grupos específicos, como por exemplo, dos Bolsistas de Produtividade em Pesquisa do CNPq, um componente denominado *Seleção*, caracteriza-se como importante mecanismo no contexto deste trabalho.

O componente de *Seleção* utiliza a linguagem de consulta XPath (*XML Path Language*) para pesquisa no repositório local de currículos e posterior geração do conjunto. A linguagem XPath permite construir expressões que vão processar e percorrer um documento XML de forma similar ao uso de expressões regulares. Logo, é possível agrupar um conjunto de currículos com base em parâmetros desejados. Diante disso, foi possível identificar no repositório local que dentre os 5.620.451 currículos extraídos em 02/2018, um conjunto de 14.475 indivíduos possuem registrado em seus currículos que eles são Bolsista de Produtividade em Pesquisa do CNPq em alguma modalidade de bolsa. Logo, os dados contidos nos currículos deste conjunto de indivíduos foram analisados, sendo apresentada uma visão geral deste conjunto, e posteriormente, caracterizadas as redes de colaboração científica para cada modalidade de bolsa, além da rede geral, contendo todos os bolsistas.

4. Resultados

Para análises da colaboração científica dos pesquisadores de excelência no Brasil, considerou-se o conjunto de Bolsistas de Produtividade em Pesquisa do CNPq. Este grupo de indivíduos que, em sua maioria, tem atuado em pesquisas, seja em instituições de ensino seja em institutos de ciência e tecnologia, ainda é responsável pela formação dos alunos nos principais programas de pós-graduação no Brasil. Com isso, ressalta-se que o conjunto de indivíduos analisado neste trabalho, apesar de englobar uma pequena quantidade de indivíduos, compreende grande parte dos principais pesquisadores em atuação no Brasil.

O sistema de bolsas Produtividade em Pesquisa do CNPq encontra se estruturado em modalidades, iniciando se pela modalidade 2, em que o bolsista deve ser portador de título de doutor há pelo menos três anos na ocasião da implementação da bolsa. Já na modalidade 1, ele deve possuir no mínimo, oito anos de doutorado na ocasião da implementação da bolsa. Essa modalidade encontra se dividida em quatro subníveis (A, B, C e D) estabelecidas em bases comparativas com a comunidade científica da área e

tendo por base os dados dos últimos dez anos, incluindo a capacidade de formação contínua de recursos humanos, especialmente orientações de pós-graduação stricto sensu. A distribuição do conjunto selecionado nas modalidades de bolsas pode ser visualizada na Figura 2.

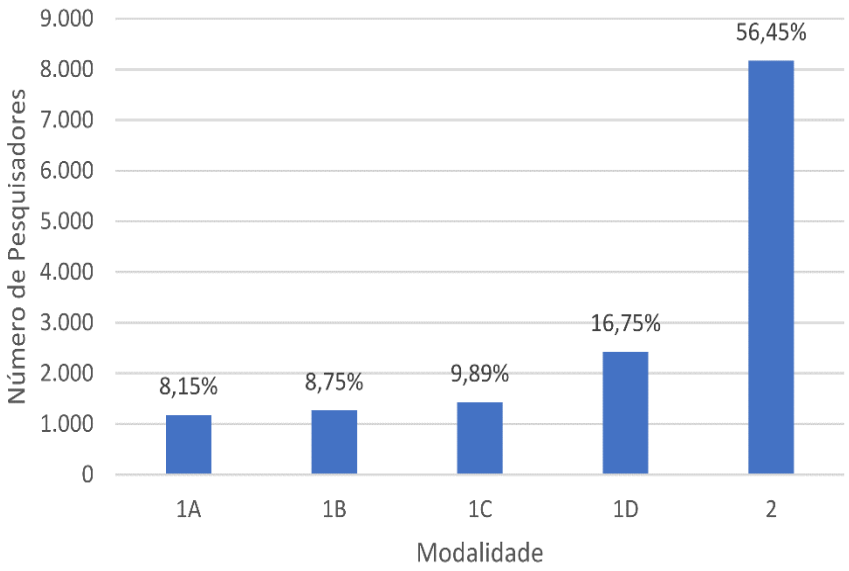


Figura 2. Distribuição dos bolsistas nas modalidades de bolsas.

Como pode ser observado, a modalidade 2 concentra a maior quantidade de bolsistas. Consequentemente, os outros pesquisadores que representam aproximadamente 44% do conjunto, são distribuídos nas modalidades superiores em que, a quantidade de indivíduos reduz em cada modalidade, à medida que o nível da bolsa aumenta.

Destaca-se que obrigatoriamente os bolsistas ingressam na modalidade 2, e posteriormente, podem progredir para as modalidades de níveis mais altos. Logo, a modalidade 2 tende naturalmente a concentrar uma maior quantidade de bolsistas, tendo em vista que é a modalidade de entrada para todos os pesquisadores.

No contexto deste trabalho, que visa caracterizar as redes de colaboração científica do conjunto de bolsistas, a produção científica é o principal elemento de estudo. Foi identificado que em ambas as modalidades de bolsas, os bolsistas tendem a divulgar sua produção científica preferencialmente em anais de congressos e em periódicos. A Figura 3 apresenta o total de publicações de todos os bolsistas, em ambos os meios de divulgação por modalidade de bolsas.

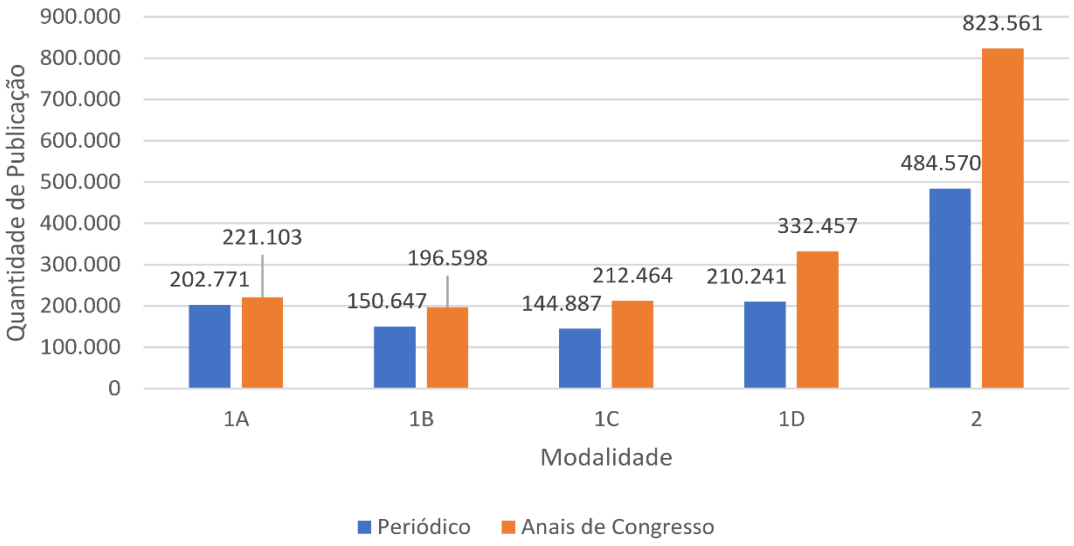


Figura 3. Total de publicações de artigos em anais de congressos e em periódicos

Ressalta que para a totalização das publicações em cada uma das modalidades, foram considerados todos os registros de publicações em cada um dos meios de divulgação encontrados nos currículos dos bolsistas, independentemente de quando o trabalho foi publicado.

Como pode ser observado, a modalidade 2 possui a maior média de publicações, tendo em vista que é a modalidade que contém a maior quantidade de bolsistas (56,45%). A quantidade de publicações nas modalidades mais altas, tendem a se reduzir, já que concentram uma menor quantidade de bolsistas. Destaca-se, a semelhança entre a quantidade de publicações da modalidade 1A com a 1C. E ainda, nota-se que, à medida que o nível das modalidades de bolsas vai aumentando, a diferença entre publicações de artigos em anais de congressos e em periódicos tendem a reduzir.

Um tipo de análise que comumente vem sendo realizada quando verificada a produção científica de um determinado conjunto de indivíduos é a análise sobre como a comunidade científica tem colaborado (MENA-CHALCO; DIGIAMPIETRI; CESAR-JUNIOR, 2012; BOAVENTURA et al., 2014; MENA-CHALCO et al. 2014; DIGIAMPIETRI, 2015). Neste trabalho, são consideradas apenas as publicações de artigos em anais de congressos e em periódicos dos bolsistas para a caracterização das redes de colaboração científica a serem analisadas, já que são os principais meios de divulgação do conjunto.

Tais redes possibilitam analisar, com a adoção de métricas específicas, como as colaborações têm evoluído ao longo dos anos ou sobre como elas estão estruturadas. Neste trabalho, foram caracterizadas redes por modalidades de bolsas e, para suas análises, algumas métricas de análise de redes sociais foram aplicadas.

Para caracterização das redes que contêm todos os bolsistas, foram adotados métodos que visam concentrar os nós mais conectados no centro da rede e, consequentemente, aqueles nós com menor quantidade de ligações ou isolados são deslocados para as extremidades das redes (Figura 4).

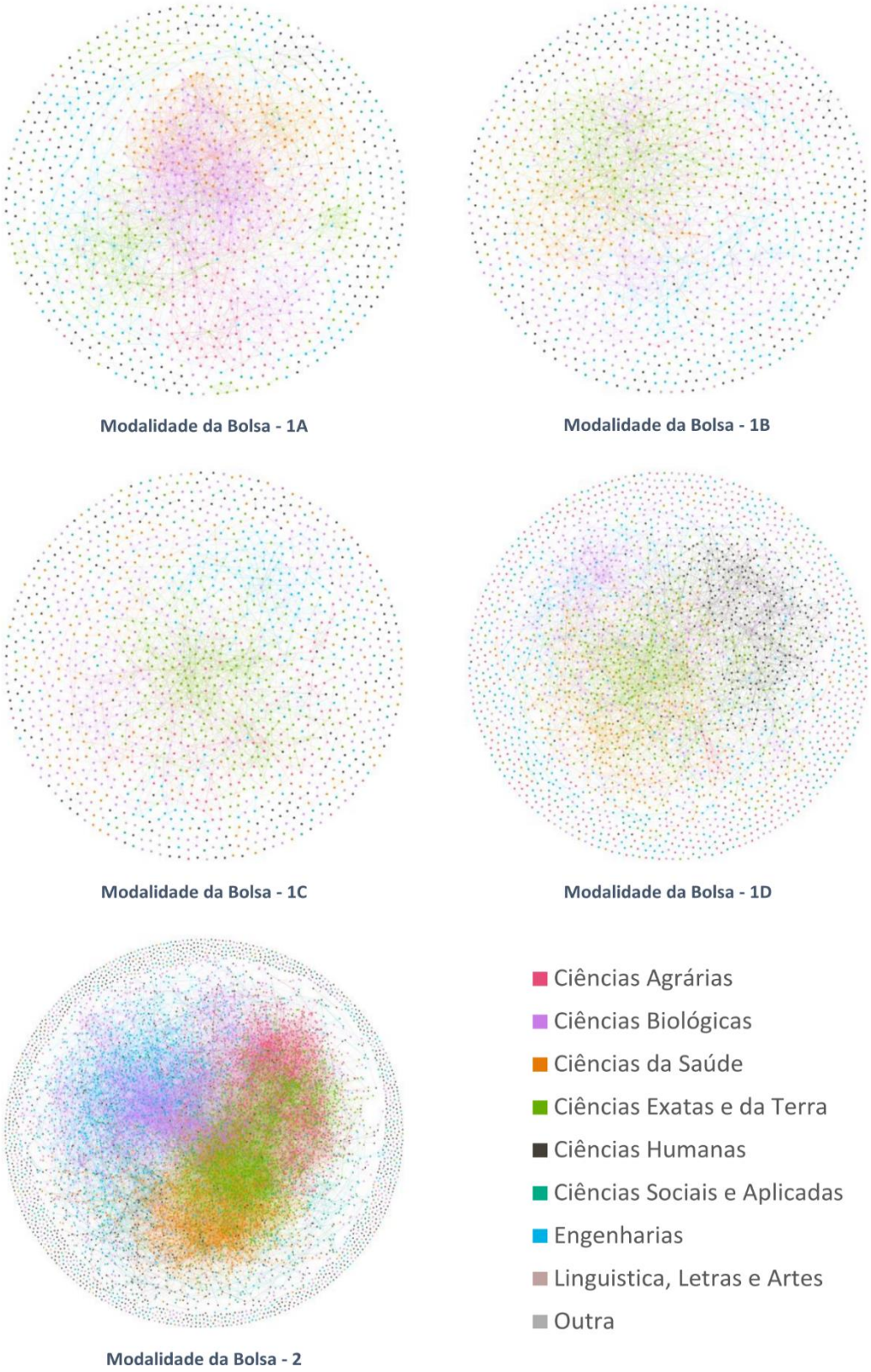


Figura 4. Redes de colaboração científica dos Bolsistas de Produtividade em Pesquisa do CNPq.

Ao analisar as redes de colaboração de cada modalidade, é possível observar grande similaridade entre as redes das modalidades 1A e 1D, a exemplo do que ocorre com a produção científica. Em que suas componentes gigantes são maiores, e existe uma maior centralidade dos bolsistas agrupados principalmente, pelas suas grandes áreas. Já as redes das modalidades 1B e 1C, também são semelhantes, possuindo a menor quantidade de bolsistas nas suas componentes gigantes, tornando as redes mais esparsas. A rede da modalidade 2, que possui a maior quantidade de bolsistas, dentre todas as modalidades, se destaca por possuir a maior componente gigante, e ainda, a maior quantidade de componentes isolados, concentrados nas extremidades da rede. Agrupando as redes de todas as modalidades, é possível caracterizar a rede global (Figura 5).

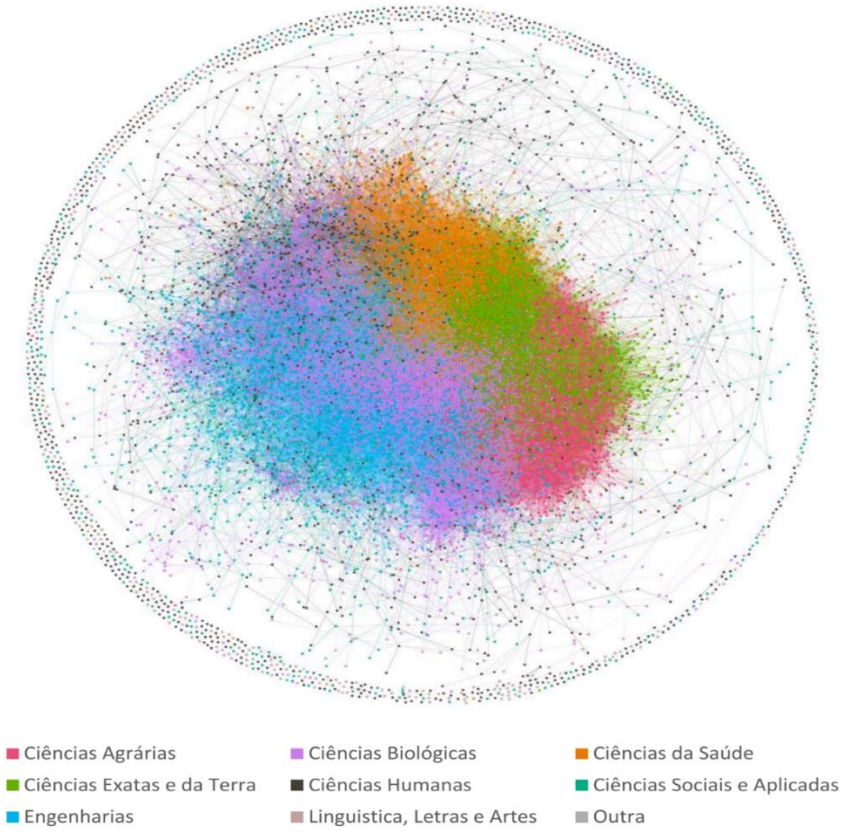


Figura 5. Rede global com todos os Bolsistas de Produtividade em Pesquisa do CNPq

A rede global, contempla todo o conjunto de bolsistas de Produtividade em Pesquisa do CNPq, possuindo a menor quantidade de nós isolados, que podem ser observados nas extremidades da rede, e sua componente gigante concentra conjuntos de bolsistas, agrupados principalmente pelas suas grandes áreas. No entanto, percebe-se considerável sobreposição de grandes áreas, o que representa colaboração entre bolsistas de grandes áreas distintas.

Após caracterizar as redes, algumas métricas foram aplicadas. As métricas adotadas (grau médio dos nós, total de nós no componente gigante, densidade da rede, diâmetro da rede e caminho mínimo médio) são métricas clássicas, usualmente adotadas por diversos trabalhos que analisam redes de colaboração (SZWARCFITER, 1986; NEWMAN, 2003; LEMIEUX; OUIMET, 2008; SCOTT, 2009; WASSERMAN; FAUST, 2009).

A Tabela 1 apresenta uma sumarização das redes caracterizadas e das métricas adotadas em todos as modalidades de bolsas. Pode-se afirmar que a análise das redes de colaboração científica do conjunto de Bolsistas de Produtividade em Pesquisa do CNPq possibilita obter uma visão de como ocorreu o processo de coautoria entre os principais pesquisadores em atuação no Brasil.

Tabela 1. Resultado das métricas adotadas

	Métrica									
Nível do Bolsista	Total de Nós	Total de Arestas	Grau Médio dos Nós	Total de Nós no Componente Gigante	% de Nós no Componente Gigante	Total de Arestas no Componente Gigante	Densidade da Rede	Diâmetro da Rede	Caminho Mínimo Médio	Total de Componentes Isolados
1A	1.180	5.402	9,15	884	74,92%	5.293	0,008	15	4,74	202
1B	1.267	3.156	4,98	835	65,90%	3.024	0,004	19	5,64	304
1C	1.432	3.170	4,42	957	66,83%	3.052	0,003	27	6,36	343
1D	2.425	6.399	5,27	1.707	70,39%	6.113	0,002	21	6,03	467
2	8.171	37.225	9,11	6.685	81,81%	36.839	0,001	19	5,69	1.079
Geral	14.608	218.788	14,95	13.375	91,56%	218.564	0,002	18	4,37	987

É possível observar a distinção entre as estruturas das redes, sendo que a rede global que comtempla todos os bolsistas, possui como era de se esperar, a maior componente conectada, concentrando 99,9% das arestas da rede e 91,56% dos bolsistas. No entanto, ela possui menor densidade que algumas redes específicas, como por exemplo das modalidades 1A, 1B e 1C. Logo, podemos afirmar que, apesar da rede geral possuir a maior componente conexa, as redes de algumas modalidades específicas, são bem mais densas, resultado que nestas redes, os bolsistas estão mais conectados, necessariamente, nos níveis mais alto de modalidade das bolsas.

Ressalta-se ainda, o diâmetro da rede da modalidade 1A, o menor dentre todas as outras, resultando em um menor esforço para conectar os dois bolsistas mais distantes da rede. Valor este, bem inferior ao de outras modalidades, como por exemplo o da modalidade 1C, em que existe uma distância de 26 bolsistas entre os dois mais distantes.

Destaca-se a rede geral, que possui o maior valor do grau médio (14,95), podendo ser considerada a rede de maior colaboração tendo em vista, que contempla todos os bolsistas. Foi possível ainda identificar que, apesar de as redes possuírem baixa densidade, as das modalidades de maior nível, possuem as menores quantidades de componentes isolados, que correspondem a autores que publicaram sem nenhuma colaboração ou que não publicaram com outros bolsistas de sua mesma modalidade.

Também foi possível verificar como alguns pesquisadores têm trabalhado em colaboração de forma muito intensa. Analisando as arestas mais densas das redes, alguns bolsistas se destacam com intensa quantidade de colaborações, em que para todas as

redes, as arestas com maior peso possuem centenas de colaborações, resultado de intensa frequência na produção de artigos científicos. Em geral, tais colaborações acontecem em sua maioria na publicação de artigos em anais de congressos, e são entre pesquisadores que atuam na mesma área ou em áreas correlatadas, que possuem a mesma formação acadêmica e, em determinadas situações, orientados pelos mesmos orientadores.

Logo, as análises apresentadas possibilitam compreender como ocorre a colaboração científica entre os pesquisadores de excelência no Brasil, analisando todo o histórico de suas publicações científicas registradas nos currículos cadastrados na Plataforma Lattes.

5. Considerações Finais

Como resultado deste estudo, foi possível verificar como ocorre a colaboração científica do conjunto de Bolsistas de Produtividade em Pesquisa do CNPq. Para isso, foram considerados os artigos publicados em anais de congresso e em periódicos contemplando todo o seu histórico de publicações. Identificou-se que a produção cientifica dos bolsistas da modalidade 1A é distribuída quase que igualitariamente entre os dois tipos de publicações, diferentemente dos níveis mais baixos de modalidade das bolsas, em que as publicações são realizadas, em sua grande maioria em anais de congressos.

Foi possível observar que nas modalidades mais altas de bolsas, a colaboração ocorre de forma mais intensa, e que nos níveis mais inferiores a quantidade de bolsistas que não colaboram com bolsistas da mesma modalidade é bem superior. Com isso concluímos que os pesquisadores de excelência do Brasil, têm colaborado entre si, e que esta colaboração ocorre de forma mais intensa entre os bolsistas das modalidades superiores. Além disso, analisando outros dados dos bolsistas, foi possível perceber também que existe colaboração entre bolsistas de diferentes níveis de modalidades de bolsas, e que a colaboração é fortemente influenciada pela localização geográfica dos bolsistas, por exemplo, entre bolsistas de uma mesma instituição ou estado.

Como trabalhos futuros pretende-se verificar como é estruturada a rede de colaboração científica do conjunto de doutores com currículos cadastrados na Plataforma Lattes que não são bolsistas, e dessa forma, verificar de forma comparativa a colaboração dos pesquisadores de excelência com os pesquisadores que não recebem bolsa.

Agradecimentos

Os autores agradecem ao CEFET-MG e a UEMG pelo auxílio na pesquisa.

Referencias

Araújo, C. A. (2006). Bibliometria: evolução histórica e questões atuais. *Em Questão, 12,* pp. 11-32.

Barabâsi, A. L., Jeong, H., Néda, Z., Ravasz, E., Schubert, A., & Vicsek, T. (2002). Evolution of the social network of scientific collaborations. *Physica A: Statistical mechanics and its applications, 311,* pp. 590-614.

Boaventura, M., Bonson, K., Silva, A. P., Veloso, A., & Meira Jr, W. (2014). Caracterização Temporal das Redes de Colaboração Científica nas Universidades Brasileiras: Anos 2000-2013. *In: Brazilian Workshop on Social Network Analysis and Mining.* Brasília.

Dias, T. M. R. (2016). *Um Estudo Sobre A Produção Científica Brasileira A Partir De Dados Da Plataforma Lattes.* Tese (Doutorado em Modelagem Matemática e Computacional) - Centro Federal de Educação Tecnológica de Minas Gerais, Belo Horizonte - MG.

Digiampietri, L. A. (2015). *Análise da Rede Social Acadêmica Brasileira.* (Livre Docência). Escola de Artes Ciências e Humanidades da Universidade de São Paulo, São Paulo.

Digiampietri, L., Mugnaini, R., Pérez-Alcázar, J., Delgado, K., Tuesta, E., & Mena-Chalco, J. (2017). Aná lise da evolução, impacto e formação de redes nos cinco anos do BraSNAM. *In Congresso da Sociedade Brasileira de Computação-CSBC.*

Ding, Y. (2011). Scientific collaboration and endorsement: Network analysis of coauthorship and citation networks. *Informetrics, 5*, pp. 187-203.

Hoffnagel, J. C. (2009). A prática de citação em trabalhos acadêmicos. *Cadernos de Linguagem e Sociedade, 10*, p. 71.

Lane, J. (2010). Let's make science metrics more scientific. *Nature, 464*, pp. 488-489.

Lemieux, V., & Ouimet, M. (2008). *Análise estrutural das redes sociais.* Lisboa: Instituto Piaget.

Lopes, G. R. (2012). *Avaliação e Recomendação de Colaborações em Redes Sociais Acadêmicas.* (Doutorado). Instituto de Informática UFRGS, Universidade Federal do Rio Grande do Sul, Porto Alegre.

Mena-Chalco, J. P., Digiampietri, L. A., & Cesar-Junior, R. M. (2012). Caracterizando as redes de coautoria de currículos Lattes. *In: Brazilian Workshop on Social Network Analysis and Mining.* Curitiba.

Mena- Chalco, J. P., Digiampietri, L. A., Lopes, F. M., & Cesar, R. M. (2014). Brazilian bibliometric coauthorship networks. *Journal of the Association for Information Science and Technology, 65*, pp. 1424-1445.

Mugnaini, R., Jannuzzi, P., Quoniam, L. (2004). Indicadores bibliométricos da produção científica brasileira: uma análise a partir da base Pascal. *CI, 33*, pp.123-131.

Newman, M. E. (2001). The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences, 98*, pp. 404-409.

Pritchard, A. (1969). Statistical bibliography or bibliometrics? *Journal of Documentation, 4*, pp. 348-349.

Revoredo, K., Araújo, R., Silveira, B., & Muramatsu, T. (2012). Minerando publicações científicas para análise da colaboração em comunidades de pesquisa. *Brazilian Workshop on Social Network Analysis and Mining (BraSNAM).* Curitiba - PR.

Santos, L. R. F. (2015). *Utilização de Dados da Plataforma Lattes para a Avaliação da Distribuição da Bolsa de Produtividade em Pesquisa do CNPq.* Dissertação (Mestrado em Modelagem Matemática e Computacional) - Centro Federal de Educação Tecnológica de Minas Gerais, Belo Horizonte.

Scott, J. (2009). *Social network analysis: a handbook* (2 ed.). London: SAGE.

Silva, M. R., Hayashi, C. R., & Hayashi, M. C. (2011). Análise bibliométrica e cientométrica: desafios para especialistas que atuam no campo. *InCID: Revista de Ciência da Informação e Documentação, 2*, pp. 110-129.

Sonnenwald, D. H. (2007). Scientific collaboration. *Annual Review of Information Science and Technology, 41*, pp. 643-681.

Ströele, V., Zimbrão, G., & Souza, J. M. (2012). Análise de redes sociais científicas: modelagem multi-relacional. *In: Brazilian Workshop on Social Network Analysis and Mining.* Curitiba.

Szwarcfiter, J. L. (1986). *Grafos e algoritmos computacionais* (2 ed.). Rio de Janeiro: Campus.

Wasserman, S., & Faust, K. (2009). *Social network analysis: methods and applications* (19 ed.).

Combinando Análise Bibliométrica e Análise de Redes Sociais para a Avaliação de Grupos Acadêmicos

Lucas Leal Caparelli[1], **Luciano Antonio Digiampietri**[1]

[1]Escola de Artes, Ciências e Humanidades – Universidade de São Paulo (USP)
São Paulo, SP – Brasil

lucas.caparelli@usp.br, digiampietri@usp.br

Abstract. *The characterization and evaluation of groups of researchers are relevant and complex activities. This paper aims to characterize Brazilian graduate programs in Computer Science according to different bibliometric and social networks analysis metrics. In order to do this, the four-year evaluation of the programs carried out by CAPES was taken as the object of study, trying to identify which classification method is the most appropriate for this task. Using machine learning algorithms, it was possible to produce a classification model of these programs, reaching an accuracy of 86.15%.*

Resumo. *A caracterização e a avaliação de grupos de pesquisadores são atividades relevantes e complexas. Este artigo visa a caracterizar programas brasileiros de pós-graduação em Ciência da Computação de acordo com diferentes medidas bibliométricas e oriundas da análise de redes sociais. Para tal, tomou-se como objeto de estudo a avaliação quadrienal dos programas feita pela CAPES, buscando identificar qual método de classificação é o mais indicado para esta tarefa. Utilizando algoritmos de aprendizado de máquina foi possível produzir um modelo de classificação destes programas alcançando acurácia de 86,15%.*

1. Introdução

A avaliação de grupos acadêmicos é extremamente importante para tarefas como a concessão de financiamentos, análise da viabilidade de projetos de pesquisa, entre outras. Essa avaliação combina diversas informações diferentes, muitas das quais podem estar distribuídas em fontes de dados distintas, sendo muitas vezes subjetivas ou difíceis de quantificar.

Na pós-graduação brasileira, a CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) realiza periodicamente a avaliação de todos os programas nacionais de pós-graduação e os resultados dessa avaliação indicam se um programa está apto ou não a oferecer turmas de mestrado e/ou doutorado, bem como indicam a quantidade de recursos federais (para ajudar no financiamento do programa) e a quantidade de bolsas que serão reservados a cada programa.

Nas avaliações, cada uma das áreas de conhecimento detalha os critérios e pesos utilizados. Tipicamente, estes critérios são divididos em cinco eixos/quesitos: 1) *Proposta do Programa* que inclui coerência, planejamento e infraestrutura; 2) *Corpo Docente* incluindo o perfil, titulação, distribuição na execução das atividades e dedicação/regime de

trabalho; 3) *Corpo Discente* que analisa a quantidade de teses e dissertações defendidas, distribuição das orientações entre os docentes, qualidade da produção dos discentes e eficiência do programa na formação de mestres e doutores, 4) *Produção Intelectual* que analisa a quantidade e qualidade das publicações produzidas pelos discentes e docentes do programa, a distribuição das publicações entre os docentes, a participação de discentes nas publicações, e outros tipos de produções intelectuais (por exemplo, patentes); 5) *Inserção Social* que avalia a inserção e o impacto do programa, integração com outros programas e a visibilidade do programa.

Ao longo dos últimos anos, diversos trabalhos tentaram caracterizar de maneira automática grupos acadêmicos (como programas de pós-graduação, grupos de pesquisa ou departamentos), ou inferir o resultado de avaliações (de ranqueamentos nacionais ou internacionais, ou de programas de pós-graduação) [Digiampietri et al. 2014, Mena-Chalco et al. 2014, Digiampietri et al. 2016, Silva et al. 2017, Linden et al. 2017].

O objetivo do presente trabalho é combinar análise bibliométrica com análise de redes sociais a fim de se analisar a importância de diferentes atributos (ou métricas) na avaliação dos programas brasileiros de pós-graduação em Ciência da Computação, bem como avaliar a capacidade desses atributos na inferência dos conceitos atribuídos pela CAPES. Este trabalho se diferencia dos trabalhos correlatos por analisar dados da última avaliação da CAPES (quadriênio 2013 a 2016) e por combinar informações de três fontes diferentes: CAPES (Plataforma Sucupira), CNPq (Plataforma Lattes) e Google Acadêmico (*Google Scholar*).

2. Materiais e métodos

Neste trabalho foram analisados os 66 programas acadêmicos de pós-graduação em Ciência da Computação que foram avaliados pela CAPES na última avaliação quadrienal (período de 2013 a 2016) e que iniciaram suas atividades anteriormente ao período de avaliação.

Dos 66 programas avaliados, dois não possuíam conceito atribuído na avaliação anterior, 45 mantiveram o mesmo conceito do triênio anterior, 16 tiveram sua avaliação elevada em um ponto e três tiveram seu conceito reduzido em um ponto.

As informações básicas dos programas (nome, conceito CAPES do último quadriênio, conceito CAPES do triênio anterior e lista de docentes) foi obtida manualmente a partir da Plataforma Sucupira[1] no dia 15 de outubro de 2017. Com base na lista dos orientadores, foram identificados os currículos Lattes de cada um dos orientadores por meio de um processo automático [Digiampietri et al. 2012, Digiampietri et al. 2014] que foi posteriormente verificado manualmente para assegurar a correta identificação dos currículos. Ao todo foram identificados 1.608 currículos de orientadores.

Adicionalmente, para cada orientador foi buscado o respectivo perfil do Google Acadêmico[2] utilizando um procedimento automático para identificação de perfis [Digiampietri et al. 2014, Digiampietri and Ferreira 2018] nos dias 15 e 16 de novembro de 2017. Este processo identificou o perfil de 1.093 orientadores.

[1]Plataforma Sucupira: https://sucupira.capes.gov.br/sucupira/ , acessado em 29/01/2018

[2]Google Acadêmico: https://scholar.google.com.br/ , acessado em 29/01/2018

Dois tipos de informações foram extraídas de cada programa: informações bibliométricas e métricas da análise de redes sociais. A tabela 1 contém a lista e uma breve descrição dos 29 atributos bibliométricos utilizados. Os três primeiros atributos, gerais de cada programa, foram obtidos com base nos dados da Plataforma Sucupira. Os 14 atributos relacionados a orientações e publicações foram obtidos a partir de dados dos currículos da Plataforma Lattes, destacando-se que os atributos relacionados a "pontuações" das publicações combinaram dados das publicações com os valores associados a cada conceito Qualis definidos pelo Comitê de Área da Ciência da Computação. Por fim, os 12 últimos atributos desta tabela, relacionados às citações, foram extraídos dos perfis do Google Acadêmico.

A tabela 2 contém a lista e breve descrição dos 14 atributos oriundos da análise de redes sociais. No presente trabalho dois tipos de redes sociais baseadas nas relações de coautoria foram construídas (utilizando as informações das publicações extraídas dos currículos Lattes dos orientadores de cada programa). O primeiro tipo, composto por apenas uma rede, possui como nós os programas de pós-graduação e como arestas as relações de coautoria entre docentes de diferentes programas. Esta rede é utilizada para o cálculo de quatro métricas de centralidade (ou importância) de cada programa em relação às ligações com os demais. Adicionalmente, foi construída uma rede por programa, na qual cada nó corresponde a um orientador e cada aresta corresponde a relações de coautoria entre orientadores de um mesmo programa. A partir da rede de cada programa foram extraídas 10 métricas globais de rede.

Com base nos atributos extraídos ou calculados, o objetivo deste trabalho foi analisar a importância desses atributos em relação ao Conceito CAPES obtido pelos programas na última avaliação quadrienal. Dois tipos de análise foram realizadas com este propósito: análise da importância dos atributos em relação ao Conceito CAPES (baseada na correlação de valores e em algoritmos de seleção de atributos) e a capacidade de se inferir o Conceito CAPES com base nos demais atributos utilizando algoritmos de classificação.

Tanto para a seleção de atributos quanto para a classificação dos programas de pós-graduação em Ciência Computação das universidades brasileiras utilizou-se o arcabouço Weka, o qual foi criado pela Universidade de Waikato, na Nova Zelândia [Witten et al. 2016]. Este arcabouço permite acesso a diversos métodos de classificação, dos quais fez-se uso em testes com o conjunto de dados a fim de se identificar o modelo de maior acurácia.

Para seleção de atributos fez-se uso de dois seletores de atributos: *ChiSquaredAttributeEval* e *CFSSubsetEval*. O seletor de atributos *ChiSquaredAttributeEval* avalia a importância de um atributo a partir da estatística de inferência qui-quadrado que serve para avaliar quantitativamente a relação entre um resultado (no caso o conceito CAPES atribuído aos programas) e a distribuição dos valores de um dado atributo. O método para a identificação dos atributos mais importantes utilizado foi o *Ranker* que, neste caso, simplesmente ordena os atributos de acordo com o respectivo valor de qui-quadrado.

O seletor *CFSSubsetEval* avalia a qualidade de um subconjunto de atributos considerando sua habilidade de predição individual juntamente do grau de redundância entre os atributos deste subconjunto. Subconjuntos que possuem alta correlação com o atributo

Tabela 1. Atributos bibliométricos utilizados

Sigla utilizada	Descrição
Conceito CAPES	Conceito CAPES do programa para o quadriênio 2013-2016.
Conceito CAPES Anterior	Conceito CAPES do programa para o triênio 2010-2012.
Pesquisadores	Número de pesquisadores em um dado programa.
Dissertações de mestrado	Número de dissertações defendidas orientadas pelos pesquisadores do programa.
Teses de doutorado	Número de teses defendidas orientadas pelos pesquisadores do programa.
Supervisões de pós-doutorado	Número de pós-doutoramentos supervisionados pelos pesquisadores do programa.
Dissertações de mestrado_PP	Número de dissertações por pesquisador do programa.
Teses de doutorado_PP	Número de teses por pesquisador do programa.
Supervisões de pós-doutorado_PP	Número de pós-doutoramentos por pesquisador do programa.
Publicações - Totais	Número total de artigos completos em periódicos ou anais publicados pelo pesquisadores do programa.
Publicações - Totais_PP	Número de artigos por pesquisador do programa.
Publicações - Pontuação	Total de pontos obtidos pelas publicações dos pesquisadores do programa.
Publicações - Pontuação_PP	Total de pontos por pesquisador do programa.
Publicações Índice Restrito - Totais	Número total de publicações dos pesquisadores do programa em veículos de índice restrito (A1, A2 ou B1).
Publicações Índice Restrito - Totais_PP	Publicações dos pesquisadores do programa em veículos e índice restrito por pesquisador.
Publicações Índice Restrito - Pontuação	Total de pontos obtidos pelas publicações dos pesquisadores do programa em veículos de índice restrito.
Publicações Índice Restrito - Pontuação_PP	Total de pontos das publicações em veículos e índice restrito por pesquisador.
Média das Citações	Média das citações totais do Google Acadêmico dos pesquisador do programa.
Média das Citações_5	Média das citações do Google Acadêmico dos pesquisador do programa nos últimos cinco anos.
Média do Índice H	Média do índice H do Google Acadêmico dos pesquisador do programa.
Média do Índice H_5	Média do índice H do Google Acadêmico dos pesquisador do programa nos últimos cinco anos.
Média do Índice I10	Média do índice H do Google Acadêmico dos pesquisador do programa.
Média do Índice I10_5	Média do índice I10 do Google Acadêmico dos pesquisador do programa nos últimos cinco anos.
Mediana das Citações	Mediana das citações totais do Google Acadêmico dos pesquisador do programa.
Mediana das Citações_5	Mediana das citações do Google Acadêmico dos pesquisador do programa nos últimos cinco anos.
Mediana do Índice H	Mediana do índice H do Google Acadêmico dos pesquisador do programa.
Mediana do Índice H_5	Mediana do índice H do Google Acadêmico dos pesquisador do programa nos últimos cinco anos.
Mediana do Índice I10	Mediana do índice H do Google Acadêmico dos pesquisador do programa.
Mediana do Índice I10_5	Mediana do índice I10 do Google Acadêmico dos pesquisador do programa nos últimos cinco anos.

classe e baixa intercorrelação são preferíveis [Hall 1998]. Este seletor trabalha associado a um método de busca. Fez-se uso do método *BestFirst*, sendo que a versão utilizada

Tabela 2. Atributos de rede utilizados

Métricas Locais - coautorias entre programas	
Sigla utilizada	**Descrição**
Centralidade de grau	Medida da importância de um programa com base em seu grau na rede de coautorias.
Centralidade de intermediação	Medida da importância de um programa com base na quantidade de vezes que se encontra nos caminhos mínimos entre programas na rede de coautorias.
Centralidade de proximidade	Medida da importância de um programa com base na média de seus caminhos mínimos com os demais programas na rede de coautorias.
Centralidade Page Rank	Medida da importância de um programa com base na medida Page Rank (medida que combina as relações de coautoria com a importância dos coautores).
Métricas Globais - coautorias inter-programas	
Sigla utilizada	**Descrição**
Arestas	Número de arestas na rede de coautorias de cada programa.
Média dos caminhos mínimos	Tamanho médio dos caminhos mínimos na rede de coautorias de cada programa.
Coeficiente de agrupamento	Medida de transitividade das relações de coautoria dentro de cada programa.
Assortatividade de grau	Medida da tendência de haver relacionamentos entre autores com o mesmo grau na rede.
Diâmetro	Diâmetro (maior caminho mínimo) da rede de coautorias de cada programa.
Densidade	Densidade (relação entre o número de arestas existentes e o número de arestas possíveis) na rede de coautorias de cada programa.
Componente Gigante	Porcentagem de nós no componente gigante (maior componente conexo) da rede de cada programa.
Centralização de grau	Medida da dependência da rede de coautorias de cada programa em relação ao seu nó mais importante (de acordo com o grau).
Centralização de intermediação	Medida da dependência da rede de coautorias de cada programa em relação ao seu nó mais importante (de acordo com a intermediação).
Centralização de proximidade	Medida da dependência da rede de coautorias de cada programa em relação ao seu nó mais importante (de acordo com a proximidade).

pelo arcabouço Weka também foi proposta pelo mesmo autor [Hall 1998]. Este consiste numa busca através do espaço dos subconjuntos de atributos fazendo uso de *hillclimbing* e *backtracking*.

Para a classificação, foram testados diferentes classificadores utilizando como estratégia de verificação dos resultados a validação cruzada em dez subconjuntos. Diferentes conjuntos de atributos foram utilizados (incluindo o uso de todos os atributos ou apenas daqueles selecionados pelos seletores de atributos). Os classificadores que atingiram os melhores resultados e compõe os modelos produzidos ao final dos testes foram: *BayesNet*, *NaiveBayes* e *RandomTree* e os melhores resultados foram aqueles utilizando os atributos selecionados por *CFSSubsetEval*.

Rede Bayesiana (*BayesNet*) corresponde à aplicação do Teorema de Bayes a modelos probabilísticos representados por grafos acíclicos direcionados. A estruturação da rede pode ser feita de maneira automática ao utilizar-se de um método de busca aliado a um sistema de avaliação [Korb and Nicholson 2010]. O algoritmo *NaiveBayes*, assim como o *BayesNet*, trata-se da aplicação do Teorema de Bayes a modelos probabilísticos. Entretanto, neste algoritmo aplica-se o teorema aos atributos do conjunto de dados assumindo forte independência entre estes [John and Langley 1995]. *RandomTree* trata-se de

um algoritmo de árvore de decisão no qual são escolhidos k atributos aleatórios a fim de realizar a classificação de uma instância [Hastie et al. 2001].

A próxima seção apresenta e discute os resultados da análise dos atributos e classificação dos programas.

3. Resultados

No Brasil, o oferecimento regular de programas de pós-graduação é condicionado à obtenção de conceitos CAPES entre 3 e 7. Na avaliação atual da CAPES, um programa na área de Ciência da Computação recebeu nota 2. Assim, dividiu-se o processo de classificação em duas principais vertentes: uma fazendo uso do conjunto de dados inteiro incluindo-se o programa de nota 2 e outra incluindo apenas programas com notas de 3 a 7.

A figura 1 apresenta os valores de correlação entre cada um dos atributos extraídos ou calculados e o Conceito CAPES atual dos programas. Observa-se que os valores das correlações considerando todos os programas ou apenas aqueles que possuem conceitos entre 3 e 7 são muito próximos. Nessa figura, os atributos estão ordenados de acordo com o valor de suas correlações (do maior para o menor). Os valores elevados de correlação indicam que quanto maior o valor de um dado atributo há maior chance do programa obter um conceito CAPES maior.

O maior valor de correlação ocorre entre o Conceito CAPES atual (quadriênio 2013 a 2016) e o Conceito CAPES anterior (triênio 2010 a 2012). As três correlações seguintes possuem valores muito próximos e estão ligadas às publicações ou pontuações obtidas via publicações, são elas: Publicações Índice Restrito - Totais, Publicações - Pontuação, Publicações Índice Restrito - Pontuação. A quarta maior correlação ocorre com o atributo Teses de doutorado. A sexta e a sétima maiores correlações ocorrem com duas medidas de centralidade ao se considerar a rede em que cada programa corresponde a um nó: Centralidade de grau e Centralidade Page Rank.

Apenas uma medida diretamente relacionada às citações dos artigos dos orientadores aparece entre as dez maiores correlações: Média do Índice H_5 (isto é, a média do índice H considerando-se as citações recebidas nos últimos cinco anos). A nona maior correlação ocorre com o atributo Publicações - Totais (número total de artigos publicados em anais de eventos ou periódicos). Por outro lado, a décima maior correlação ocorre com uma medida "por pesquisador": Publicações Índice Restrito - Pontuação_PP que é a mesma medida que apresentou a quarta maior correlação, porém ponderada pelo número de pesquisadores do programa. Destaca-se que todos os atributos avaliados "por pesquisador" apresentaram correlações inferiores do que aquelas obtidas pelo valor total (não ponderado) da respectiva medida.

Apenas quatro atributos apresentaram correlações negativas, sendo que nenhuma delas possui valor absoluto alto. Destaca-se apenas a correlação com valor mais alto, que ocorreu com o atributo Densidade. Esta correlação, apesar de não possuir valor muito alto, indica que há uma relação entre a rede de coautorias do programa ser mais densa e o programa possuir um conceito CAPES menor. Isto pode sugerir que programas cujas coautorias estão muito concentradas dentro do programa acabam tendo uma produção mais limitada do que aqueles cujas coautorias ocorrem mais frequentemente com colaboradores externos ao programa.

	Todos os programas	Programas com conceitos de 3 a 7
Conceito CAPES Anterior	0.920	0.924
Publicações Índice Restrito - Totais	0.887	0.889
Publicações - Pontuação	0.886	0.888
Publicações Índice Restrito - Pontuação	0.885	0.887
Teses de doutorado	0.850	0.854
Centralidade de grau	0.821	0.818
Centralidade Page Rank	0.819	0.816
Média do Índice H_5	0.803	0.807
Publicações - Totais	0.805	0.802
Publicações Índice Restrito - Pontuação_PP	0.803	0.800
Média do Índice H	0.796	0.801
Publicações Índice Restrito – Totais_PP	0.795	0.793
Média das Citações_5	0.792	0.795
Média do Índice I10_5	0.783	0.786
Média das Citações	0.778	0.781
Mediana do Índice H_5	0.776	0.775
Média do Índice I10	0.765	0.769
Teses de doutorado_PP	0.767	0.767
Publicações - Pontuação_PP	0.763	0.762
Dissertações de mestrado	0.760	0.758
Centralidade de proximidade	0.736	0.735
Pesquisadores	0.730	0.724
Centralidade de intermediação	0.714	0.714
Mediana do Índice H	0.674	0.665
Supervisões de pós-doutorado	0.641	0.645
Arestas	0.639	0.634
Mediana das Citações	0.628	0.629
Mediana do Índice I10	0.560	0.584
Diâmetro	0.559	0.548
Média dos caminhos mínimos	0.548	0.538
Supervisões de pós-doutorado_PP	0.419	0.425
Mediana das Citações_5	0.414	0.414
Mediana do Índice I10_5	0.415	0.413
Publicações - Totais_PP	0.370	0.361
Dissertações de mestrado_PP	0.297	0.295
Componente Gigante	0.282	0.273
Assortatividade de grau	0.219	0.199
Centralização de intermediação	0.198	0.182
Coeficiente de agrupamento	-0.053	-0.036
Centralização de proximidade	-0.122	-0.128
Centralização de grau	-0.228	-0.219
Densidade	-0.245	-0.234

Figura 1. Correlações entre os atributos utilizados e o conceito CAPES atribuídos aos programas

As demais análises realizadas nesta seção foram divididas em duas subseções, uma considerando todos os programas avaliados e outra apenas para os programas com conceitos de 3 a 7.

3.1. Análise para todos os programas avaliados

A tabela 3 apresenta os dez atributos melhor ranqueados de acordo com o seletor de atributos *ChiSquaredAttributeEval*. Observa-se que seis dessas medidas (ou atributos) são relacionadas às publicações ou citações, sendo duas relacionadas às pontuações baseadas nos extratos Qualis atribuídos aos veículos nos quais os artigos foram publicados

(Publicações Índice Restrito - Pontuação e Publicações - Pontuação), à quantidade de artigos publicados (Publicações Índice Restrito - Totais e Publicações - Totais) e às citações recebidas pelos artigos dos orientadores do programa (Média do Índice H e Média do Índice I10).

Tabela 3. Atributos melhor ranqueados - método baseado no valor qui-quadrado - programas notas 2 a 7

Atributo	Valor
Conceito CAPES Anterior	113,06
Dissertações de mestrado	111,52
Publicações Índice Restrito - Totais	100,46
Publicações Índice Restrito - Pontuação	99,80
Centralidade de Grau	97,46
Publicações - Pontuação	96,17
Média do Índice H	86,95
Publicações - Totais	79,99
Teses de doutorado	79,78
Média do Índice I10	79,67

Os demais atributos melhor ranqueados são o número total de Dissertações de mestrado e de Teses de doutorado defendidas no quadriênio e Centralidade de Grau que é a única métrica oriunda da análise de redes sociais entre os dez atributos listados, ocupando a quinta posição deste ranqueamento.

Destaca-se que o atributo melhor ranqueado foi Conceito CAPES Anterior, lembrando-se que mais de dois terços dos programas mantiveram a nota obtida no triênio anterior. Este atributo é seguido pelo número total de Dissertações de mestrado defendidas e Publicações Índice Restrito - Pontuação, isto é, a pontuação total obtida pelas publicações no quadriênio em veículos classificados com Qualis A1, A2 e B1.

O algoritmo *CFSSubsetEval* com método de busca *BestFirst* tem por objetivo selecionar o subconjunto de atributos mais representativos, produzindo uma lista não ordenada. Os atributos selecionados considerando os programas com conceito de 2 a 7 são: Publicações - Pontuação_PP, Publicações Índice Restrito - Totais, Publicações Índice Restrito - Totais_PP, Publicações Índice Restrito - Pontuação, Média do Índice H, Média do Índice H_5, Centralidade de Grau, Diâmetro, Conceito CAPES Anterior.

Dos nove atributos selecionados, cinco já haviam sido selecionados pelo algoritmo *ChiSquaredAttributeEval*: Centralidade de Grau, Conceito CAPES Anterior, Média do Índice H, Publicações Índice Restrito - Pontuação e Publicações Índice Restrito - Totais. Os atributos que não haviam sido selecionados são Diâmetro (da rede de coautorias interna de cada programa), Média do Índice H_5, Publicações - Pontuação_PP e Publicações Índice Restrito - Totais_PP. Destaca-se que este algoritmo visa a identificar um subconjunto de atributos que seja mais significativo em relação ao Conceito CAPES atual (analisados de maneira conjunta), o que é diferente de se observar os resultados anteriormente apresentados que são focados nos atributos de maneira individual.

Observa-se que nos resultados do algoritmo *CFSSubsetEval* houve preferência do uso das duas médias do índice H (considerando todas as citações e apenas as citações

dos últimos cinco anos), bem como da escolha de uma medida de rede que não havia sido selecionada anteriormente (Diâmetro) e também de duas medidas "por pesquisador": Publicações - Pontuação_PP e Publicações Índice Restrito - Totais_PP. Isto revela que estes atributos, apesar de não apresentarem as maiores correlações com o Conceito CAPES dos programas, são relevantes para a atribuição desses conceitos.

Ao aplicar ao subconjunto obtido pelo algoritmo *CFSSubsetEval* o algoritmo de classificação *NaiveBayes* foi possível classificar corretamente 81,82% das instâncias. É importante salientar também que o erro de classificação em todos os casos foi de apenas um ponto no conceito (por exemplo, todos programas de nota 4 foram classificados como programas de conceito 3, 4 ou 5), conforme pode ser observado na matriz de confusão (tabela 4), na qual a primeira linha representa as classificações do modelo e a última coluna apresenta o valor real do conceito daquele programa.

Tabela 4. Matriz de Confusão - classificação de todos os programas avaliados

a	b	c	d	e	f	
0	1	0	0	0	0	a = 2
0	21	3	0	0	0	b = 3
0	3	20	1	0	0	c = 4
0	0	2	5	0	0	d = 5
0	0	0	0	2	1	e = 6
0	0	0	0	1	6	f = 7

Não havendo conjunto de treinamento a parte para a criação do modelo, foi utilizado o método de validação cruzada em 10 *folds* para o treinamento e validação do modelo gerado. Vale salientar que, por haver apenas um programa de nota 2, quando este era utilizado no subconjunto de treinamento, nenhuma instância dessa classe seria utilizada como teste. Com isso, seria impossível para o algoritmo classificar corretamente este programa.

3.2. Análise para os programas avaliados com conceitos de 3 a 7

A tabela 5 apresenta os dez atributos melhor ranqueados de acordo com o seletor de atributos *ChiSquaredAttributeEval*. Observa-se que os atributos selecionados são os mesmos presentes na tabela 3, ocorrendo apenas mudanças na ordem de alguns atributos. Os quatro primeiros atributos aparecem em ordem diferente daqueles da tabela 3. Estas mudanças ocorreram porque, ao excluir o programa com conceito 2 da análise, o atributo Publicações Índice Restrito - Pontuação subiu da quarta posição para a primeira, deslocando os três atributos que haviam sido melhor ranqueados do que ele.

Destaca-se, assim, que neste conjunto de dados o atributo melhor ranqueado foi Publicações Índice Restrito - Pontuação, sendo melhor ranqueado do que o Conceito CAPES Anterior. Isto indica a importância desta medida que combina a publicação de artigos com a "qualidade" dos veículos nos quais os artigos foram publicados considerando os extratos Qualis A1, A2 e B1. Observa-se ainda que este atributo considera a pontuação total do programa e não aquela dividida pelo número de orientadores.

Fazendo uso do seletor de atributos *CFSSubsetEval* com o algoritmo de busca *BestFirst* foi possível reduzir o número de atributos a 10, sendo estes: Centralidade de

Tabela 5. Atributos melhor ranqueados - método baseado no valor qui-quadrado - todos os programas

Atributo	Valor
Publicações Índice Restrito - Pontuação	112,91
Conceito CAPES Anterior	110,65
Dissertações de mestrado	109,72
Publicações Índice Restrito - Totais	98,64
Centralidade de Grau	95,34
Publicações - Pontuação	94,38
Média do Índice H	84,32
Publicações - Totais	78,45
Teses de doutorado	78,36
Média do Índice I10	77,94

Grau, Conceito CAPES Anterior, Diâmetro, Dissertações de mestrado Média das Citações Média do Índice H, Publicações - Pontuação_PP, Publicações Índice Restrito - Pontuação, Publicações Índice Restrito - Totais, Publicações Índice Restrito - Totais_PP.

Destacam-se as diferenças ocorridas na seleção atual e naquele que envolvia todos os programas. Na seleção atual, não consta o atributo Média do Índice H_5, o qual havia sido selecionado ao se considerar todos os programas. Por outro lado, ao se considerar apenas os programas com conceitos de 3 a 7, dois novos atributos foram selecionados: Dissertações de mestrado (número de dissertações defendidas no quadriênio) e Média das Citações (média das citações recebidas pelos orientadores ao longo de suas carreiras).

O melhor resultado da classificação considerando-se os programas com conceitos de 3 a 7 ocorreu utilizando-se o conjunto de atributos selecionados pelo *CFSSubsetEval*. Foi realizada uma classificação em dois níveis. Inicialmente separou-se os programas de nota 6 e 7 dos demais e depois classificaram-se os demais programas. O classificador *RandomTree* foi capaz de classificar com 100% de acerto todos programas com conceitos 6 e 7, sem nenhum falso-positivo. Desta forma, é possível utilizar o resultado deste classificador para separar os programas com conceitos 6 e 7 e então usar uma nova abordagem para classificar os programas restantes. A tabela 6 apresenta a matriz de confusão referente a esta classificação.

Tabela 6. Matriz de Confusão - primeiro nível de classificação dos programas com conceitos de 3 a 7

a	b	c	d	e	
16	7	1	0	0	a = 3
6	14	4	0	0	b = 4
0	4	3	0	0	c = 5
0	0	0	3	0	d = 6
0	0	0	0	7	e = 7

Para os programas com conceitos entre 3 e 5 foi realizada uma nova seleção de atributos (utilizando a mesma técnica citada anteriormente) que resultou no seguinte sub-

conjunto: Centralidade de Grau, Conceito CAPES Anterior, Dissertações de mestrado Média das Citações_5 Média do Índice H, Média dos caminhos mínimos, Publicações - Pontuação_PP, Publicações Índice Restrito - Pontuação, Publicações Índice Restrito - Totais, Publicações Índice Restrito - Totais_PP.

Esta seleção de atributos difere da anterior em dois atributos. O atributo Média das Citações_5 está substituindo o atributo Média das Citações que havia sido selecionado anteriormente, indicando que para o conjunto de programas com conceitos de 3 a 5 a média das citações recebidas nos últimos cinco anos é mais importante do que a média de todas as citações recebidas. A outra diferença ocorre com o atributo Média dos caminhos mínimos que foi selecionado no lugar do Diâmetro.

Aplicando o classificador *BayesNet* a esse novo conjunto (considerando-se apenas os programas com conceitos de 3 a 5) obteve-se índice de acerto de 83,64%. A tabela 7 apresenta a matriz de confusão referente a esta classificação.

Tabela 7. Matriz de Confusão - programas com conceitos de 3 a 5

a	b	c	
7	0	0	a = 3
1	18	5	b = 4
0	3	21	c = 5

Com esta classificação em duas etapas, ao analisar os programas de nota 3 a 7, obteve-se 56 classificações corretas num total de 65 programas resultando num modelo com acurácia de 86,15%.

4. Conclusões

Neste trabalho foram medidos e analisados diferentes atributos ou características dos programas brasileiros de pós-graduação em Ciência da Computação do Brasil. Analisou-se a importância desses atributos em relação aos conceitos atribuídos pela CAPES em sua avaliação quadrienal, bem como a capacidade de se inferir o conceito com base nestes atributos.

Observando-se os resultados alcançados, é possível verificar que, dentre os métodos analisados, obteve-se melhores resultados ao se utilizar classificadores Bayesianos sobre subconjuntos de atributos selecionados tendo como objetivo a minimização da correlação entre estes e a maximização da correlação de cada atributo com o atributo classe, visando assim a eliminação de redundâncias.

Com o modelo obtido é possível, por exemplo, automatizar parte do processo de avaliação destes programas, bem como utilizá-lo para auto-avaliação por parte das universidades responsáveis pelos programas em janelas de tempo diferentes daquela em uso pela CAPES.

Como trabalhos futuros pretende-se explorar outras estratégias de seleção de atributos e classificação, bem como estender a análise realizada para programas avaliados por outros comitês.

Agradecimentos

Este trabalho foi parcialmente financiado pelo Programa de Educação Tutorial (PET) do Ministério da Educação e pelo CNPq.

Referências

Digiampietri, L., Linden, R., and Barbosa, L. (2016). Caracterizando departamentos e programas de computação utilizando análise de redes sociais e bibliometria. In *V Brazilian Workshop on Social Network Analysis and Mining (BraSNAM 2016)*.

Digiampietri, L. A. and Ferreira, J. E. (2018). Desambiguação de nomes de autores para a identificação automática de perfis acadêmicos. *Em Questão*, pages 1–12.

Digiampietri, L. A., Mena-Chalco, J., de Jesus Perez-Alcazar, J., Tuesta, E. F., Delgado, K., and Mugnaini, R. (2012). Minerando e caracterizando dados de currículos Lattes. In *Brazilian Workshop on Social Network Analysis and Mining (BraSNAM 2012)*.

Digiampietri, L. A., Mena-Chalco, J. P., Vaz de Melo, P. O. S., Malheiro, A. P. R., Meira, D. N. O., Franco, L. F., and Oliveira, L. B. (2014). Brax-ray: An x-ray of the brazilian computer science graduate programs. *PLOS ONE*, 9(4):1–12.

Hall, M. A. (1998). *Correlation-based Feature Subset Selection for Machine Learning*. PhD thesis, University of Waikato, Hamilton, New Zealand.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA.

John, G. H. and Langley, P. (1995). Estimating continuous distributions in bayesian classifiers. In *Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 338–345, San Mateo. Morgan Kaufmann.

Korb, K. B. and Nicholson, A. E. (2010). *Bayesian Artificial Intelligence, Second Edition*. CRC Press, Inc., Boca Raton, FL, USA, 2nd edition.

Linden, R., Barbosa, L. F., and Digiampietri, L. A. (2017). "Brazilian style science" – an analysis of the difference between Brazilian and international computer science departments and graduate programs using social networks analysis and bibliometrics. *Social Network Analysis and Mining*, 7(1):44.

Mena-Chalco, J. P., Digiampietri, L. A., Lopes, F. M., and Cesar, R. M. (2014). Brazilian bibliometric coauthorship networks. *Journal of the Association for Information Science and Technology*, 65:1424–1445.

Silva, T. H. P., Laender, A. H. F., Davis, C. A., da Silva, A. P. C., and Moro, M. M. (2017). A profile analysis of the top Brazilian computer science graduate programs. *Scientometrics*, 113(1):237–255.

Witten, I. H., Frank, E., Hall, M. A., and Pal, C. J. (2016). *Data Mining, Fourth Edition: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 4th edition.

Detecção Automática de Bolhas Sociais no Twitter em uma Rede de Usuários de Tecnologia

Bruno Evangelista, Gabriela Batista, Jaqueline Faria de Oliveira

[1]Centro Universitário de Belo Horizonte (UNIBH)
Belo Horizonte – MG – Brasil

{brunocarvalho107, gabi.viana11, jaquefari}@gmail.com

Abstract. *The phenomenon of Social Bubbles is known in several areas and research is being done to analyze its impact on society. It is characterized by the limitation of individuals to access information that has an affinity and lack of access to information that differs or differs from those of their interest. Technology professionals are not safe from social bubbles, and to measure the phenomenon for these professionals, this work seeks to detect social bubbles in Twitter using complex network metrics in a network of users related to Programming Languages most used in 2016. We collected 1,226,744 tweets, 896,556 profiles, and their followers. The Homophily by language presented that the profiles connected to the JavaScript programming language present greater Homophily than the other languages analyzed, followed by Python and Java. In this way, it is observed that these communities are more inserted in a social bubble.*

Resumo. *O fenômeno de Bolhas Sociais é conhecido em diversas áreas e pesquisas estão sendo feitas para analisar seu impacto na sociedade. Se caracteriza pela limitação dos indivíduos ao acesso a informações que tem afinidade e a falta de acesso a informações divergentes ou diferentes das de seu interesse. Os profissionais de tecnologia não estão a salvo das bolhas sociais, e com o objetivo de medir o fenômeno para esses profissionais, esse trabalho busca detectar bolhas sociais no Twitter utilizando-se de métricas de redes complexas em uma rede de usuários relacionados às 10 Linguagens de Programação mais utilizadas em 2016. Foram coletados 1.226.744 tweets, 896.556 perfis e seus seguidores. A Homofilia por linguagem apresentou que os perfis ligados à linguagem de programação JavaScript apresentam maior homofilia que as demais linguagens analisadas, seguida por Python e Java. Desta forma observa-se que essas comunidades estão mais inseridas em uma bolha social.*

1. Introdução

Nos últimos anos as redes sociais tem tido ritmo constante de crescimento [Chaffey 2017], sendo que o Facebook possui a maior fatia de mercado em todo o mundo, e no Brasil domina cerca de 76% do mercado [Newman et al. 2017]. As redes sociais apresentam uma larga escala de adesão dos usuários para comunicação e acesso à informação. No Brasil 57% da população recebe informações pelo Facebook, seguido de 46% pelo WhatsApp e 36% pelo Youtube. Segundo Guedes [Guedes 2014], por possuírem papel informativo, essas redes se tornaram uma forma importante de exercer o direito à informação, possibilitando o acompanhamento das ações do governo, ou o acesso a informações de interesse público em geral.

Para lidar com a grande quantidade de informações que as pessoas recebem no seu dia a dia, é comum o emprego de filtros em seu comportamento on-line, de forma consciente ou não [Nikolov et al. 2015]. A aplicação desses filtros tem por objetivo prover informações aos usuários conforme suas preferências, porém, a aplicação de filtros sociais e algorítmicos, pode criar e fortalecer a polarização do acesso a novas informações, estreitando a quantidade de pontos de vista a que as pessoas são expostas. Dessa maneira, as informações fornecidas são selecionadas de acordo com o círculo social no qual o indivíduo está inserido, minimizando a possibilidade do mesmo encontrar um ponto de vista contrário ao seu. A tecnologia que ajuda o usuário a encontrar conteúdo relevante na Internet está criando uma "bolha" em torno das pessoas, pois o usuário sempre recebe informações que reforçam seu ponto de vista [CAPELAS 2017].

As comunidades nas redes sociais vem sendo fonte de diversos estudos, inclusive para identificação de bolhas sociais. O Twitter é hoje uma grande fonte de informação para o estudo das comunidades e seu comportamento. Dentre os estudos, destacam-se os estudos para identificação de comunidades [Soares et al. 2014], Homofilia [Colleoni et al. 2014] política, estudos de Bolhas Sociais políticas [Caetano et al. 2016] e acesso a notícias [Nikolov et al. 2015].

Os profissionais de tecnologia também formam uma grande comunidade nas redes sociais e apresentam muitas vezes preferências semelhantes, por Sistemas Operacionais, Plataformas de Desenvolvimento, Áreas de Atuação e mesmo Linguagens de Programação. Haverá, também nessas comunidades, a presença do fenômeno de Bolha Social? Com o objetivo de responder a esta pergunta, desenvolveu-se nesse trabalho a análise de uma rede de usuários do Twitter que possuem comentários sobre linguagens de programação, a relação desses usuários e seus seguidores. Mediu-se a exposição destes a conteúdos diversos sobre linguagens de programação, com o objetivo final de medir a propensão de existência de Bolhas Sociais. Para isso foram investigados usuários que realizaram postagens sobre as 10 linguagens de programação mais populares em 2016 [Diakopoulos and Cass 2016] apresentadas na Figura 1.

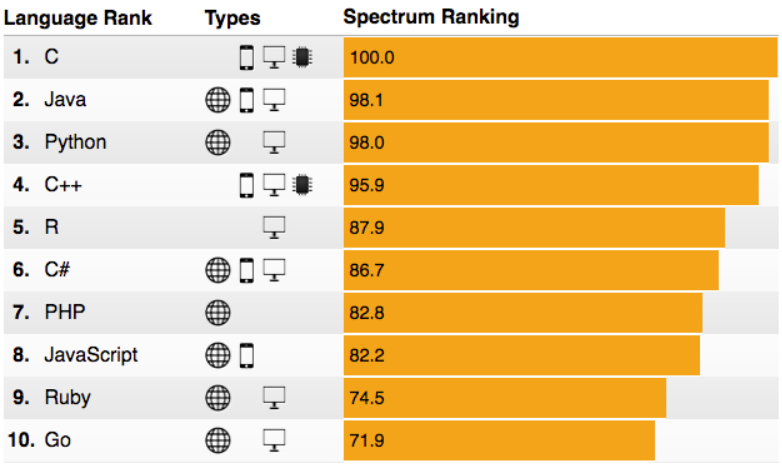


Figura 1. The Top Programming Languages 2016 [Diakopoulos and Cass 2016].

Foram coletados 1.226.744 tweets, 896.556 perfis e seus seguidores. A Homofilia por linguagem apresentou que os perfis ligados à linguagem de programação JavaScript

apresentam maior Homofilia que as demais linguagens analisadas, seguida por Python e Java.

As próximas sessões estão estruturadas da seguinte maneira. A seção 2 apresenta os trabalhos relacionados. A seção 3 traz informações sobre a as características do Twitter e da área de *Text Mining* para dar suporte ao entendimento do estudo. A seção 4 apresenta a metodologia desenvolvida para a extração dos dados e obtenção dos resultados. A seção 5 apresenta a análise e avaliação dos resultados dos dados coletados. Por fim, a seção 6 apresenta a Conclusão e as propostas para trabalhos futuros.

2. Trabalhos Relacionados

Nesta seção são apresentados os trabalhos relacionados com o tema desse artigo. São apresentados trabalhos sobre identificação de Comunidades e Bolhas Sociais nas Redes Sociais On-Line.

No trabalho "Measuring On-line Filter Bubbles" de 2015 [Nikolov et al. 2015], Nikolov e seus colaboradores analisaram um conjunto de dados da Universidade de Indiana contendo postagens e registros de pesquisa na rede social Twitter e conteúdos pesquisados no mecanismo de pesquisas da América Online (AOL) e concluíram que, de forma geral, as pessoas acessam informações de um conjunto de fontes significativamente mais "rasas" por meio das mídias sociais, comparado a uma pesquisa mais tradicional. Nesse artigo, o autor discute sobre a existência de bolhas sociais coletivas, que são grupos que se relacionam com outros e sobre a existência de bolhas sociais individuais, aquelas que não se relacionam com outros grupos.

Um estudo sobre encontrar comunidades em redes complexas realizado por Soares, Oliveira e Brito [Soares et al. 2014], descreve a detecção de comunidades em redes complexas em geral, através de um algoritmo denominado FastGreedy. Após detectar as comunidades, foi aplicado um algoritmo de agrupamento (K-means) em cada grafo e realizada a análise desses componentes. Após a realização dos estudos, foi possível aplicar o algoritmo de clustering, encontrando agrupamentos próximos aos detectados pelas outras métricas.

O trabalho publicado por Caetano (2016), aplica o cálculo da Homofilia a uma rede coletada no Twitter contendo perfis de pessoas politicamente liberais ou conservadoras. No trabalho, os autores comparam seus resultados com o esperado para redes "off-line", descobrindo que os valores de segregação ideológica encontrados são compatíveis com os presentes em relações interpessoais [Caetano et al. 2016].

3. Métricas de Redes complexas

Nesta seção são apresentados os conceitos teóricos necessários ao desenvolvimento desse projeto.

3.1. Redes Complexas

Uma rede pode ser definida genericamente como um conjunto de elementos que mantém conexões entre si [Brandão et al. 2007]. Na matemática, essas redes são reconhecidas como grafos, seus elementos são os vértices e suas conexões as arestas. Já na Ciência da Computação, segundo o autor, os elementos são conhecidos como nós e suas conexões como ligações.

Uma rede complexa é definida como uma estrutura que não possui um padrão regular, sendo compostas por estruturas complexas que possuem certo grau de imprecisão. Para solucionar este problema, ocorre a analogia entre a sua estrutura e a disposição de um grafo, com a finalidade de delimitar seu escopo [Barabási 2009].

Diversas métricas são utilizadas para caracterizar uma rede. Métricas de identificação de agrupamentos dentro de uma rede possibilitam a medida da aglomeração de conexões na rede [Blondel et al. 2008]). As medidas de centralidade indicam o grau de conectividade direta entre os nós de uma rede [Brandão and Parreiras 2010], dentre as métricas conhecidas na literatura, destacam-se a Centralidade de Grau e Betweenness [Vicente 2017].

Homofilia

O artigo de Mcpherson (2001), descreve o princípio da homofilia, que indica a tendência de indivíduos criarem laços com pessoas que possuam características e comportamentos sociais semelhantes ao utilizarem redes sociais [Mcpherson et al. 2001].

De acordo com Monge (2003), esse fenômeno já é conhecido e estudado também nos trabalhos "Similarity-Attraction hypotesis" de Byrne (1971) e "Theory of self-categorization" de Turner (1987), sendo esses fundamentais para dar suporte à teoria da homofilia [Monge and Contractor 2003].

Segundo [Colleoni et al. 2014], a homofilia de indivíduos de uma comunidade pode ser calculada através da Equação 1.

$$H_i = \frac{s_i}{s_i + d_i} \tag{1}$$

Na qual H_i é o valor da homofilia encontrado nesse grupo, s_i é o número de conexões que conectam os indivíduos dessa comunidade, ou seja, conexões homogêneas e d_i é o número de conexões que ligam os indivíduos desse conjunto, (grupo i), com indivíduos de outras comunidades (conexões heterogêneas).

Segundo trabalho publicado por Currarini et al. (2009), a fórmula mencionada não traduz a ideia de como um grupo tendencioso é comparado ao quão faccioso ele poderia ser potencialmente, apesar de fornecer informações interessantes [Currarini et al. 2009].

Para resolver esse problema, o autor sugere o uso da fórmula de Coleman (1958), "Inbreeding Homophily" (Equação 2), que realiza a normalização da homofilia pelo potencial de um grupo a ser tendencioso.

$$IH_i = \frac{H_i - w_i}{1 - w_i} \tag{2}$$

Nessa equação, IH_i é o indicador de Inbreeding Homophily, H_i é o valor calculado para a homofilia utilizando na Equação 1, e w_i é a probabilidade da ocorrência de indivíduos do grupo i, que consiste no total de indivíduos desse grupo dividido pelo total de indivíduos de uma rede T. Caso o valor de H_i seja maior que w_i, então existe homofilia. Quando ocorre o inverso, ou seja, o valor de w_i for maior, ocorre a heterofilia.

Ainda segundo o autor, nesse caso, a condição de homofilia e de heterofilia pode ser representada da seguinte forma:

$$IH_i > 0 \rightarrow homofilia$$

$$IH_i < 0 \rightarrow heterofilia$$

As medidas de Homofilia apresentadas nas Equações 1 e 2 serão utilizadas para medir a tendência de formação de Bolhas Sociais nas comunidades de Linguagens de Programação, que é a proposta deste trabalho.

4. Metodologia

Para realizar este trabalho foi feita a análise da rede formada pelos relacionamentos entre usuários a partir da coleta de tweets sobre 10 linguagens de programação distintas.

Foi adotado um processo metodológico que se dividiu em 6 passos. Esse processo inclui coleta de dados, tratamento dos dados, caracterização e análise de bolhas sociais. Os passos estão representados na Figura 2.

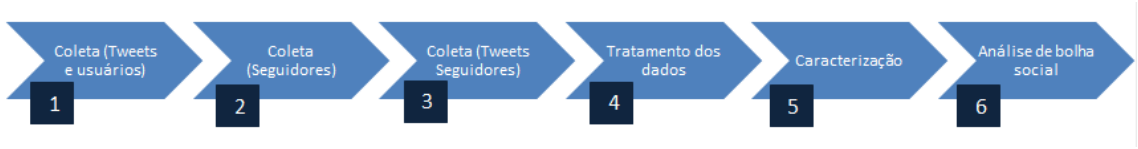


Figura 2. Processo adotado na metodologia.

A etapa 1 consiste na coleta de dados do Twitter. A Coleta foi realizada utilizando a biblioteca Tweepy[1], que implementa os métodos da API do Twitter na linguagem Python. Para busca dos tweets foi criada uma *bag of words* contendo termos relacionados às linguagens a serem pesquisadas. Esses termos são apresentados na Tabela 1. A coleta dos tweets ocorreu de 21 de maio a 30 de junho de 2017. Essa coleta foi realizada através de *streaming*, ou seja, os dados foram obtidos em tempo real através da API utilizada.

Na etapa 2 e 3 foram coletados, respectivamente, os seguidores dos usuários que realizaram as postagens e seus tweets. Essa coleta foi feita nos meses de julho a novembro de 2017. A diferença de período se justifica pela limitação da API do twitter, o que torna lenta a coleta de dados. Foram coletados os tweets dos usuários da base que postaram algum twitter referente à uma das linguagens de programação analisadas.

Na etapa 4, os dados foram tratados de forma a possibilitar a efetiva execução das etapas 5 e 6, respectivamente. Foram considerados somente os dados dos ”seguidores dos usuário” que continham a menção a alguma das linguagens de programação analisadas, os demais dados foram desconsiderados nas análises.

Na etapa 5 foi feita uma caracterização da rede gerada a partir dos usuários coletados. Os vértices representam os usuários coletados e as arestas as interações “seguindo / seguidor”, gerando um grafo direcionado, no qual as arestas saem do “seguidor” e incidem no “seguido”.

[1]http://www.tweepy.org/

Tabela 1. Termos utilizados para a busca dos tweets

Linguagens	Termos
C	clang; clanguage; cprogramming; ansic; learningansic; codingc; c programming; c language; coding c; learning c; ansi c
Java	javalang; javalanguage; java; learning java; coding java; java8; java7; java language; java lang
Python	csharp; csharp language; csharp programming; learning csharp; c#; c# programming; learning c#; c# language
C++	c++ language; c++ programming; cplusplusprogramming; learning cplusplus; learningc++; c plus plus; cpp; cpp programming; learning c++; learning cpp
R	rlang; rlanguage; rprogramming; learningr; codingr; rdevelopment; r lang; r language; r programming; learning r; coding r; r development
C#	csharp; csharp language; csharp programming; learning csharp; c#; c# programming; learning c#; c# language
PHP	php; php language; phplanguage; php programming; coding php; learning php; learningphp; phplang; php7; php5
JavaScript	javascript; javascript language; javascriptlang; jscode; jslanguage; codingjs; js programming; js; nodejs; angularjs; reactjs; vuejs; vue.js
Ruby	rubylang; ruby language; ruby programming; rubyprogramming; learnruby; codingruby; rubydevelopment; ruby; ruby lang
GO	goprogramming; go programming language; golang; golanguage; learngolang; godevelopment; learning go; learning golang

A partir da base inicial, definiu-se um conjunto de ”Comunidades de Linguagens de Programação” às quais cada usário pertence. Considera-se, no contexto desse trabalho, que um usuário pertence a uma Comunidade de Linguagem de Programação caso tenha mencionado em seus tweets alguma das linguagens analisadas. Esse fato possibilita que um usuário possa participar de mais de uma Comunidade. Métricas de Centralidade foram utilizadas para caracterização dos usuários da rede e Modularidade para a medida de agrupamentos de nós da rede.

Por fim, na etapa 6, foi medida a homofilia dos usuários das Comunidades de Linguagens de Programação com o intuito de identificar a similaridade entre os usuários. Na rede de usuários gerada, foram aplicadas as métricas de Homofilia (Equação 1 e Equação2) para verificar a tendência dos usuários das comunidades de conectarem-se com pessoas de interesses semelhantes.

5. Resultados

Esta seção fará a caracterização detalhada dos dados coletados, apresentando as informações relativas à quantidade de registros e como os mesmos se organizam. Serão apresentados ainda a caracterização da rede analisada e os resultados da homofilia da rede de usuários por Comunidades de Linguagens de Programação.

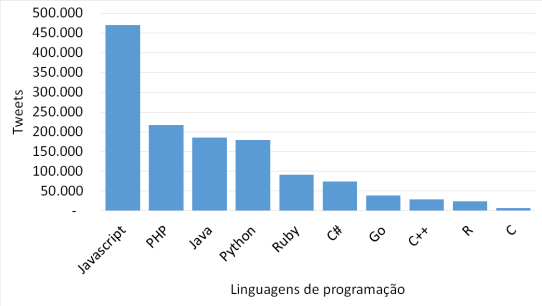
5.1. Conjunto de Dados

O processo de coleta consistiu nas etapas 1, 2 e 3 da metodologia. Foram obtidos 1.226.744 tweets e 896.556 usuários distintos que postaram sobre as linguagens de programação analisadas.

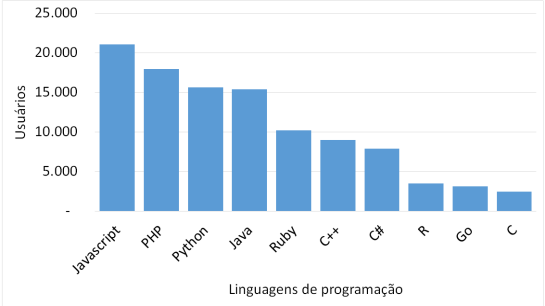
Com o objetivo de avaliar a popularidade das linguagens no Twitter, foram analisadas as distribuições de tweets por linguagem (Figura 3(a)) e usuários distintos por linguagem (Figura 3(b)).

Em comparação à popularidade das linguagens publicada pelo IEEE Spectrum em 2016 [Diakopoulos and Cass 2016], a quantidade de coleta dos tweets por linguagem demonstra um resultado diferente. A linguagem Javascript é a mais popular, enquanto o levantamento do IEEE mostra a linguagem C nesta posição. É importante notar que existe a possibilidade de um mesmo tweet falar a respeito de mais de uma linguagem, o que torna a soma das menções individuais, independente da quantidade total de tweets.

Ao se analisar a quantidade de usuários distintos que publicaram informações sobre cada linguagem, observa-se uma distribuição semelhantes ao total de tweets por linguagem nas duas primeiras posições. Há variação no *ranking* entre Java e Python, as demais linguagens apresentam distribuições distintas.



(a) Tweets coletados por linguagem.



(b) Total de usuários que realizaram publicações por linguagem.

Figura 3. Na Figura 3(a) é apresentado o total de tweets coletados por linguagem. Na Figura 3(b) são apresentados os totais de usuários que realizaram publicações para cada uma da linguagem.

5.2. Rede de usuários

A rede de usuários, formada por seguidos e seguidores que fizeram menção às linguagens de programação, foi analisada. A Figura 4 apresenta a distribuição de frequência de conexões dos usuários com outros usuários que também falam sobre linguagens de programação. Pode-se verificar que a grande maioria dos usuários possui poucas conexões com demais usuários que falam sobre o tema. Cerca de 65% tem somente uma conexão, as demais conexões são formadas por 35% da base, mas representam um valor considerável com mais de 480 mil usuários.

Foi gerado um grafo para representar a rede de usuários, onde os vértices são os usuários e as arestas as interações "seguindo / seguidor". O grafo é direcionado, no qual as arestas saem do "seguidor" e incidem no "seguido". A rede possui 896.556 vértices

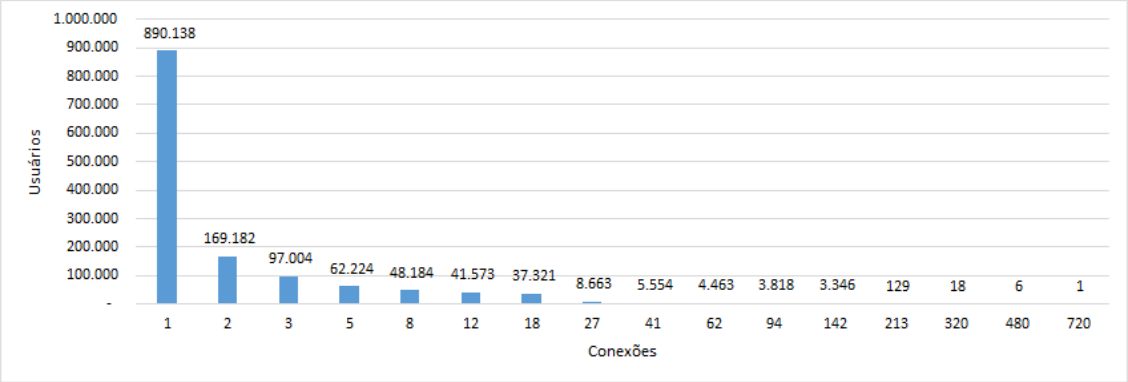
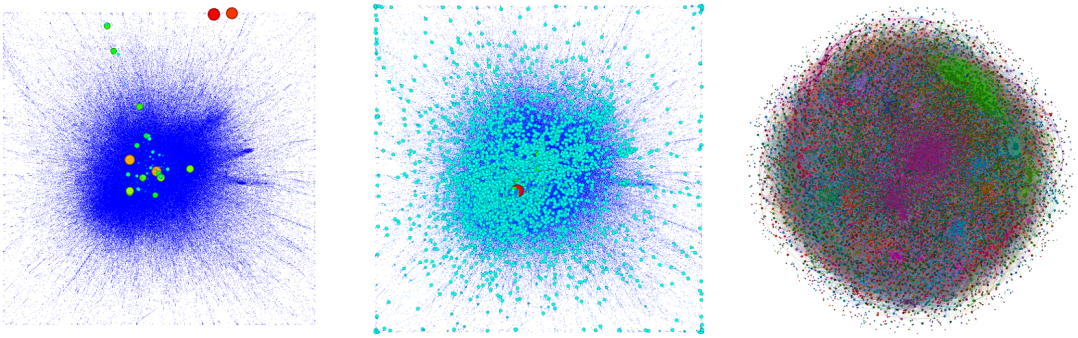


Figura 4. Frequência de conexões por usuário.

e 1.354.698 arestas e 538 componentes. Essa rede é analisada por meio de métricas de centralidade e de identificação de agrupamentos.

Betweenness

Foi calculada a métrica de Centralidade de Intermediação (Betweeness) ao grafo para verificar a importância dos nós na propagação de informação dentro da rede. Para destacar estes nós, foi configurada uma visualização em que o tamanho e a cor dos nós refletissem esta medida, como pode ser observado na Figura 5(a). Observa-se que a grande maioria dos vértices do grafo possuem um valor Betweeness muito baixo. Estes resultados indicam que uma quantidade muito pequena de usuários têm uma importância muito grande no fluxo de informações na rede coletada.



(a) Métrica Betweeness. (b) Distribuição de grau. (c) Comunidades da rede.

Figura 5. As Figuras apresentam os resultados de métricas de redes complexas aplicadas à rede em estudo. Na Figura 5(a) os vértices são coloridos e dimensionados de acordo com seu valor de Betweenness. Na Figura 5(b) também são apresentados os resultados através das cores e dimensão dos vértices, neste caso para representar a distribuição de grau. Na Figura 5(c) é apresentado o resultado da aplicação da métrica Modularidade para identificação de agrupamentos da rede (comunidades), sendo que as comunidades são representadas por diferentes cores.

Distribuição de Graus

A distribuição de graus foi calculada para verificar como as conexões do grafo se encontram distribuídas, para identificar quais os usuários mais conectados e quantas conexões eles possuem. Para obter essas informações foi calculado o grau médio do grafo, que representa a quantidade média de conexões por nós. Esse grau foi aferido em 3,039. Mais uma vez, a visualização do grafo foi configurada, desta vez para refletir o grau de cada vértice. A visualização é demonstrada na Figura 5(b) e mostra que muitos vértices possuem poucas conexões, o que justificaria o baixo valor do grau médio obtido para o grafo.

Modularidade

A métrica modularidade foi utilizada para detectar os agrupamentos (comunidades) existentes no grafo. O cálculo foi realizado para identificar esses grupos, sendo aplicado à toda rede, conforme representado na Figura 5(c).

A métrica de modularidade identificou 1.003 comunidades, obtendo um valor 0,735, que indica a existência de comunidades mais densamente conectadas internamente que externamente, ou seja, existem comunidades em que os indivíduos se conectam mais entre si do que com indivíduos externos. A maioria das comunidades possui poucos indivíduos, enquanto o número de comunidades com mais que 30 mil indivíduos é extremamente baixo.

5.3. Análise de Bolhas Sociais

Com o objetivo de analisar se há presença de Bolhas Sociais entre as Comunidades de usuários de Linguagens de Programação, foi aplicada a métrica de homofilia, procurando medir a similaridade das conexões entre os usuários.

Para cada Comunidade de Linguagens de Programação, formado por usuários que falaram das mesmas linguagens de programação, analisa-se se os usuários pertencentes às comunidades estão conectados a usuários que também estão em outras comunidades de linguagens. Para isso verifica-se as conexões homogêneas, conexões em que ambos os usuários falam sobre a linguagem, e conexões heterogêneas, conexões em que os usuários falam de linguagens distintas.

Foi feito o cálculo da homofilia dos usuários de cada Comunidade de Linguagem de Programação, representado pelo valor de H_i. O indicador *inbreeding homophily* (IH_i) também foi calculado para demonstrar um comparativo dos resultados. O valor de W_i é a probabilidade da ocorrência de indivíduos do tipo $_i$ e, quando H_i é maior que W_i, há homofilia na comunidade.

Pode-se verificar, conforme demonstrado na Figura 6, que os valores de H_i e IH_i estão muito próximos, embora os valores de IH_i sejam superiores aos valores de H_i para a maioria das Comunidades. Ambos os valores, H_i e IH_i, demonstram que todas as Comunidades de Linguagens de Programação apresentam homofilia e nenhuma apresenta heterofilia. Isso quer dizer que um grande número de usuários das comunidades se conectam a outros usuários da mesma comunidade.

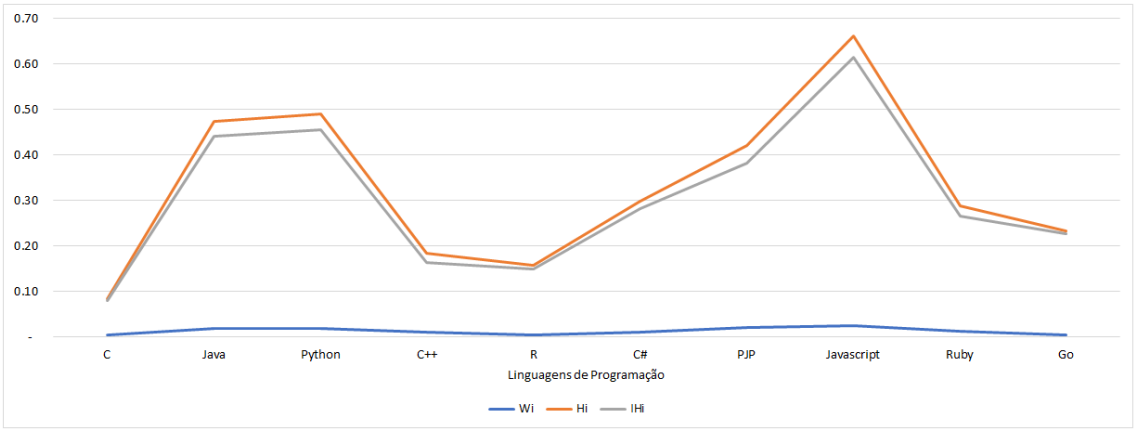
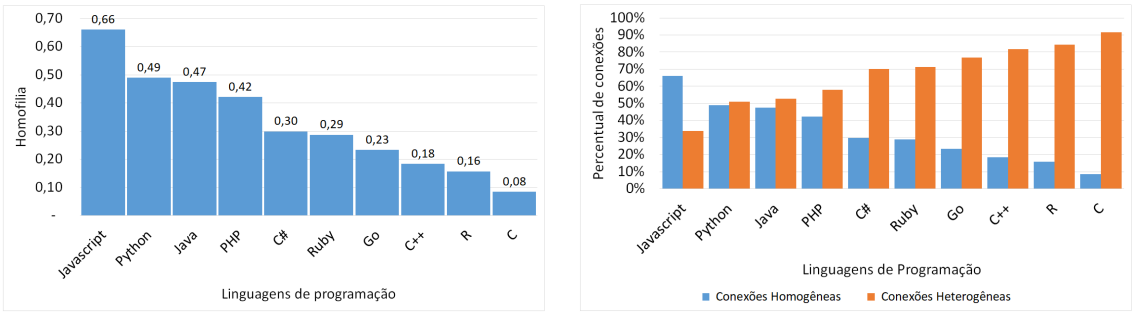


Figura 6. Demonstrativo dos resultados do cálculo da homofilia (H_i e IH_i) e a probabilidade de ocorrência de indivíduos da comunidade (W_i) por cada Comunidade de Linguagem de Programação.

Nota-se que o valor da homofilia para a comunidade de linguagem Javascript é o maior apresentado entre as comunidades, como demonstrado na Figura 7(a), seguida pelas comunidades Python, Java e PHP com valores superiores a 0,4 de homofilia. As demais apresentam homofilia, mas em valores mais baixos.



(a) Homofilia (b) Conexões homogêneas e heterogêneas

Figura 7. Na Figura 7(a) é apresentada a medida de homofilia por Comunidade de Linguagem de Programação. A Figura 7(b) mostra um comparativo de conexões homogêneas e heterogêneas dos usuários por Comunidade de Linguagem de Programação.

Pode-se perceber na Figura 7(b) que é proporcional o aumento das conexões heterogêneas e a diminuição da homofilia apresentada por comunidade (Figura 7(a)). E, apesar dos resultados apontarem que todas as comunidades contém conexões homogêneas, a linguagem Javascript é a única que possui mais conexões homogêneas que heterogêneas. As comunidades Python, Java e PHP apresentam um percentual de conexões heterogêneas e homogêneas balanceado. As demais linguagens apresentam valores discrepantes de conexões, sendo sua maioria conexões heterogêneas.

Pode-se presumir que a comunidade JavaScript está mais inserida em um cenário de Bolha Sociais que as demais comunidades. Das Comunidades de Linguagens de Programação analisadas, a Comunidade da Linguagem C foi a que menos demonstrou

conexões homogêneas, tendo consequentemente a menor probabilidade de estar envolvida numa Bolha Social.

5.4. Discussões

Os valores obtidos sobre a homofilia e sobre as conexões homogêneas e heterogêneas dos usuários das comunidades apontam na mesma direção, especialmente para a linguagem Javascript, pois apresenta a maior diferença entre conexões homogêneas e heterogêneas se comparada às de outras linguagens (Figura 7(b)). Por este motivo, os usuários da comunidade da linguagem Javascript tendem a estar mais inseridos em bolhas sociais que os das demais comunidades.

Os resultados são similares aos encontrados por [Halberstam and Knight 2013], embora o mesmo tenha efetuado a medição da homofilia dentro de dois grupos de eleitores americanos, encontrando valores numericamente próximos ao encontrados no presente trabalho.

6. Conclusão

Com a aplicação dos métodos expostos neste trabalho, foi possível fazer a análise da rede estudada culminando na medição da homofilia dos usuários das comunidades de linguagens de programação. A rede foi obtida a partir de usuários do Twitter que mencionaram alguma das linguagens de programação analisadas e seus seguidores que também postaram sobre essas linguagens. A rede demonstra um baixo valor de centralidade de grau e poucos usuários da rede possuem alto valor de Betweenness.

Os valores obtidos a partir da métrica de homofilia apontam para a possível existência de bolhas sociais, notadamente no grupo da linguagem Javascript, que possui um índice muito elevado de conexões homogêneas em relação às demais. Sua alta homofilia e a disparidade em suas conexões, se comparadas com as de outras linguagens, são os maiores indicativos da presença de bolhas sociais dentro do grupo.

Deve-se levar em consideração que os resultados apresentados foram obtidos a partir de uma amostra de tweets fornecidos pela plataforma do Twitter, e que pode não representar fielmente a realidade.

Apesar de o foco do trabalho ter sido a detecção de bolhas sociais nas Comunidades de Linguagens de Programação, a metodologia apresentada poderia estender-se aos mais variados assuntos, não limitando-se ao que foi tratado neste documento. O presente trabalho buscou detectar bolhas sociais dentro dos grupos de usuários do Twitter que falam sobre Linguagens de Programação, porém, ainda há a possibilidade do desenvolvimento de pesquisas buscando caracterizar tais bolhas, seus usuários e fluxo de informação.

7. Referências

Barabási, A.-L. (2009). Scale-free networks: a decade and beyond. *science*, 325(5939):412–413.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008.

Brandão, W. C. and Parreiras, F. S. (2010). Uma abordagem baseada em métricas de redes complexas para o estabelecimento do grau de influencia de termos em documentos. *ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO INOVAÇÃO E INCLUSÃO SOCIAL*, 11.

Brandão, W. C., Parreiras, F. S., and Silva, A. B. d. O. (2007). Redes em ciência da informação: evidências comportamentais dos pesquisadores e tendências evolutivas das redes de coautoria. *Informação & Informação*.

Caetano, A. J., Lima, H. S., and Santos, Mateus Freira, M.-N. H. T. (2016). Utilizando análise de sentimentos para definição da homofilia política dos usuários do twitter durante a eleição presidencial americana de 2016.

CAPELAS, B; MANS, M. (2017). Saiba como os algoritmos das redes sociais podem mudar a política.

Chaffey, D. (2017). Global social media research summary 2017.

Colleoni, E., Rozza, A., and Arvidsson, A. (2014). Echo chamber or public sphere? predicting political orientation and measuring political homophily in twitter using big data. *Journal of Communication*, 64(2):317–332.

Currarini, L., Jackson, M. O., and Pin, P. (2009). An economic model of friendship: Homophily,minorities,and segregation. *Journal of Communication*, 64(2):317–332.

Diakopoulos, N. and Cass, S. (2016). Interactive: The top programming languages 2016.

Guedes, T. M. (2014). As redes sociais-facebook e twitter-e suas influências nos movimentos sociais.

Halberstam, Y. and Knight, B. (2013). Are social media more social than media? measuring ideological homophily and segregation on twitter. Technical report, working paper.

Mcpherson, M., Smith-Lovin, L., and Cook, L. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27(1):415–444.

Monge, P. R. and Contractor, N. (2003). *Theories of Communication Networks*.

Newman, N., Fletcher, R., Kalogeropoulos, A., Levy, D. A., and Nielsen, R. K. (2017). Reuters institute digital news report 2017.

Nikolov, D., Oliveira, D. F., Flammini, A., and Menczer, F. (2015). Measuring online social bubbles. *PeerJ Computer Science*, 1:e38.

Soares, I., Oliveira, C. S., and de Moura Brito, J. A. (2014). Um estudo do problema de detecção de comunidades em redes. *Sistemas & Gestão*, 9(4):566–574.

Vicente, R. (2017). Redes complexas.

Detecção de Categorias de Aspectos Utilizando Redes Neurais Profundas em Avaliações Online

Bruno Á. Souza[1], **Alice A. F. Menezes**[1], **Carlos M. S. Figueiredo**[1,2], **Fabíola G. Nakamura**[1], **Eduardo F. Nakamura**[1]

[1]Universidade Federal do Amazonas (UFAM) – Manaus, AM – Brasil

[2]Universidade do Estado do Amazonas (UEA) – Manaus, AM – Brasil

{bruno.abia,alice.menezes,fabiola,nakamura}@icomp.ufam.edu.br

cfigueiredo@uea.edu.br

Abstract. *Virtual environments such as online stores (e.g. Amazon, Google Play and Booking) adopt a collaborative strategy of evaluation and reputation, where users classify products and services. User's opinion represents the satisfaction level of a rated item. The set of ratings of an item is a reference to its reputation/quality. Therefore, the automatic identification of a usersatisfaction related to an item, considering its textual evaluation, is a tool with singular economic potential. With deep learning researches evolution in sentiment analysis based in aspects, opportunities to apply several neural networks in this context arisen. However, the data representation models applied in these works focus only on Embeddings pre-trained networks as a way to perform feature extraction. In this way, this work aims to present a comparison between data representation techniques and deep networks approaches, to analyze which of them have better results in classifying categories of aspects. Thus, we can seethat TF-IDF with a Convolution Neural Network (CNN) had an F1 measure of 0.93%, being at least 0.02% higher than the others approaches applied in this work.*

Resumo. *Ambientes virtuais, como lojas online (e.g. Amazon, Google Play e Booking), adotam uma estratégia colaborativa de avaliação e reputação, onde os usuários classificam os produtos e serviços. A opinião do usuário representa o grau de satisfação em relação ao item avaliado. O conjunto de avaliações de um item é referencial de sua reputação/qualidade. Portanto, a identificação automática da satisfação do usuário em relação a um item, considerando sua avaliação textual, é uma ferramenta com potencial econômico singular. Com a evolução das pesquisas relacionadas a aprendizagem profunda na área de análise de sentimentos baseada em aspectos, têm surgido oportunidades de aplicar diversas redes neurais neste contexto. Porém, os modelos de representação de dados aplicados nesses pesquisas focam unicamente no uso de redes pré-treinadas de Embeddings como forma de realizar a extração de características dos dados. Desta forma, este trabalho tem como objetivo apresentar uma comparação entre técnicas de representação de dados e abordagens redes profundas, a fim de verificar qual apresenta melhor resultado na tarefa de classificar categorias de aspectos. Com isso, conseguimos observar que o uso de TF-IDF com uma Rede Neural Convolucional (CNN) apresentou uma me-*

dida F1 de 0,93%, sendo pelo menos 0,02% superior as demais aplicadas neste trabalho.

1. Introdução

Ambientes virtuais, como lojas online (e.g. Amazon, Google Play, Booking) e redes sociais, permitem que usuários avaliem produtos, serviços e compartilhem suas experiências. Esse compartilhamento de opiniões define de forma colaborativa a reputação tanto dos estabelecimentos quanto dos produtos/serviços disponibilizados. Essa realimentação dos clientes representa uma ferramenta importante para que as empresas possam identificar oportunidades de melhoria e crescimento. Portanto, identificar automaticamente a polaridade ou sentimento de uma avaliação realizada por um usuário (e.g. uma a cinco estrelas ou simplesmente positivo/negativo) é um problema com grande potencial econômico e estratégico para empresas [de Paula et al. 2017, Liu 2012, Ye et al. 2009].

A maioria dos trabalhos da literatura tem adotado técnicas de processamento de linguagem natural e algoritmos de aprendizagem de máquina (SVM, Naive Bayes e Max Entropy) para realizar as inferências de sentimentos e categorias expressas nos textos [Souza et al. 2016, Almeida et al. 2016, Araújo et al. 2013, Stiilpen Junior and Merschmann 2016, Santos and Moura 2016, Ye et al. 2009]. Porém, estas técnicas possuem limitações quanto ao retorno dos dados classificados, pois realizam a classificação apenas em positivo ou negativo.

Recentemente, técnicas de classificação de sentimento em nível de aspecto tem sido exploradas [Pavlopoulos 2014, Pontiki et al. 2015, Gulaty 2016], pois além de obterem o mesmo retorno dos algoritmos tradicionais também informam quais são os contextos que estão sendo tratados na sentença. Podemos citar como exemplo, a análise de uma entidade loja. Neste caso, podemos descobrir aspectos como atendimento e qualidade dos produtos e preços. Assim, dada uma sentença e um aspecto existente em um texto, o objetivo principal deste tipo de abordagem é inferir a polaridade/sentimento (positivo, negativo ou neutro) relacionado a este aspecto analisado. Em um exemplo prático, analisemos a revisão "*A comida é excelente, mas o atendimento é horrível*". O sentimento referente ao aspecto "*comida*" é positivo, enquanto o aspecto "*atendimento*" é negativo.

Segundo [Pontiki et al. 2016] esse problema de análise de sentimentos baseada em aspectos pode ser dividias em 4 subtarefas:

- Subtarefa 1: Dada uma sentença t, a abordagem ser capaz de reconhecer os aspectos existentes no texto;
- Subtarefa 2: Dados os aspectos extraídos do texto, ser capaz de classificar em positivo, negativo ou neutro, respectivamente, cada informação coletada;
- Subtarefa 3: Dado um determinado aspecto no texto em que existe uma categoria associada, como "A comida deste restaurante é ruim", o aspecto seria **comida** e a categoria seria **qualidade da comida**;
- Subtarefa 4: Dadas as categorias extraídas do texto, ser capaz de classificar em positivo, negativo ou neutro respectivamente cada informação coletada;

Como solução para estas subtarefas, Redes Neurais Convolucionais (CNN), Redes Neurais Recorrentes (RNN) e *Long Short-Term Memory* (LSTM) têm sido utilizadas

para resolver este problema no âmbito de Aprendizagem Profunda. Estas redes vem obtendo resultados representativos para este tipo de classificação de sentimento baseada em aspectos.

Neste trabalho, buscamos solucionar a subtarefa 3 (detecção de categorias). Assim, apresentamos as seguintes contribuições: (i) a análise de modelos de representação de dados para textos aplicados a tarefa de classificação; e (ii) a aplicação de 3 redes profundas (LSTM, CNN e RNN) para realizar a subtarefa de reconhecimento de categoria da análise de sentimentos baseada em aspectos.

O restante deste trabalho está dividido da seguinte forma: na Seção 2, são apresentados os trabalhos relacionados de análise de sentimentos baseada em aspectos; na Seção 3, é descrita a abordagem proposta para este trabalho; na Seção 4, são apresentados os experimentos e os resultados obtidos; por fim, a Seção 5 apresenta a conclusão e trabalhos futuros.

2. Trabalhos Relacionados

Atualmente, a análise de sentimentos baseada em aspectos tem sido assunto de muitos trabalhos na literatura dentro da área de mineração de textos [Kim 2014, Nguyen and Shirai 2015, Poria et al. 2016]. O objetivo principal desta área é a execução de duas tarefas principais: (i) reconhecimento de qual assunto (tópico) está sendo tratado em uma sentença s, onde $s \geq 1$, ou seja, onde pode existir um ou mais tópicos sendo abordados; (ii) a inferência de polaridade sob o texto que está sendo analisado (positivo, negativo ou neutro).

Abordagens anteriores realizam essa extração utilizando uma estratégia de classificação de múltiplas classes [Pontiki et al. 2016]. Este tipo de abordagem possui algumas limitações, pois em sua maioria apresentam dependências de domínio em relação a base de dados nas quais estavam sendo aplicadas. Já na análise de sentimentos eram utilizados diferentes classificadores com uma ampla variedade de recursos, como o uso de unigramas, *bag-of-words* com TF-IDF, *pos-tag* e sentenças léxicas.

Atualmente, trabalhos como o de [Xu et al. 2017] demonstram o uso de Redes Neurais Convolucionais (CNN) para a análise de sentimentos baseado em aspectos. Neste contexto, os autores demonstram a possibilidade de executar esta estratégia independente de domínio. Em seus experimentos, os autores comparam sua abordagem com o SVM e uma LSTM, alcançando uma acurácia de 76.90% nas avaliações relacionadas a computadores e 68.34% nas avaliações a respeito de restaurantes, ficando abaixo apenas do SVM, que obteve 80% na primeira base de dados e 72.1% na segunda. No trabalho de [Wang and Liu 2015], os autores também usaram CNN para este tipo de inferência. Em seus resultados, os autores conseguiram uma medida de F1 de 51% na detecção de aspectos, enquanto na análise de sentimentos obtiveram uma acurácia de 78%.

No trabalho de [Wang et al. 2016] é apresentado o uso de Redes Neurais Recursivas combinadas para a execução das duas tarefas. Em seus experimentos, os autores comparam seus resultados com outras técnicas como LSTM e CNN. Como resultado, os autores obtiveram melhor desempenho na mineração de opinião na base de dados de avaliações de restaurantes, atingindo uma medida F1 de 84.11%. Na base de avaliações de computadores, obtiveram melhor resultado que os demais métodos das duas tarefas, atingindo uma medida F1 acima de 78% nas classificações.

Diferente das pesquisas já realizadas, este trabalho utiliza outras representações de dados como entrada para as redes profundas, a fim de verificar se há melhora nos resultados da classificação das categorias dos aspectos no domínio de avaliações de restaurantes.

3. Abordagem Proposta

A abordagem proposta na elaboração deste trabalho consiste das seguintes etapas (ilustradas na Figura 1): (i) pré-processamento dos dados, a fim de retirar possíveis ruídos e termos não representativos para as avaliações; (ii) extração das características dos textos para entrada das redes profundas; (iii) treinamento das redes profundas para classificação das categorias existentes nos textos.

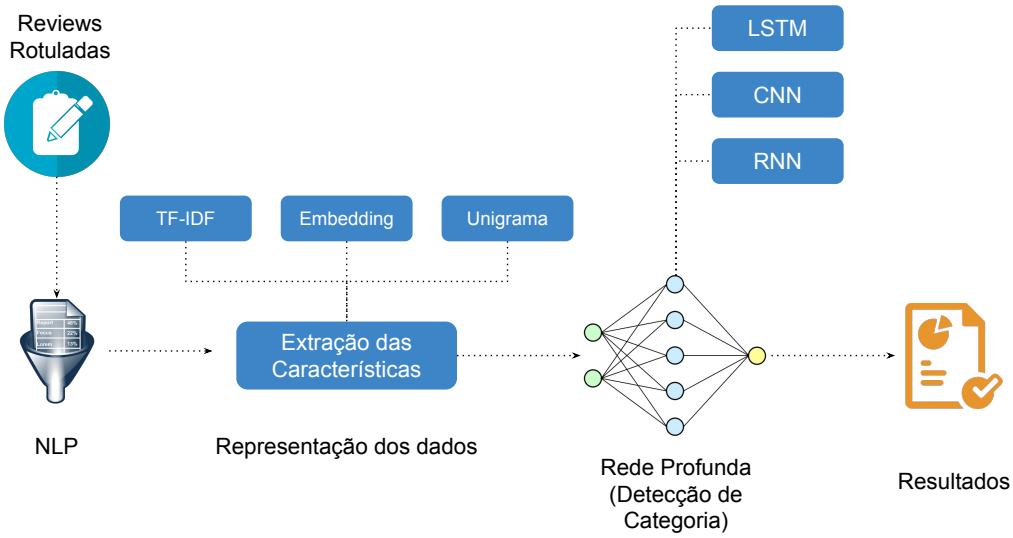


Figure 1. Arquitetura da abordagem utilizada.

3.1. Pré-processamento

Nesta etapa, os textos pertencentes a base de dados foram submetidos a uma filtragem, a fim de remover conteúdos com pouco valor semântico. Este processo consiste em separar os *tweets* em *tokens* (segmentos de sentenças), para logo em seguida, remover as menções, URLs e *emoticons* de cada um deles. Posteriormente, os *tokens* são normalizados, isto é, são submetidos a transformações (e.g., tratamento de pontuação, limpeza de caracteres especiais). Após a normalização, aqueles *tokens* que forem *stopwords* (palavras que podem ser considerados irrelevantes para o contexto estudado) são descartados e, por fim, os afixos dos *tokens* restantes são eliminados (*stemming*)

3.2. Extração de Características

Nesta etapa, nós selecionamos as principais técnicas de extração de características (TF-IDF, Unigramas e *Embedding*), a fim de comparar qual técnica fornecida como entrada para as redes neurais melhora o desempenho da classificação das categorias dos aspectos. Neste contexto, consideramos o total de dimensões extraídas por cada abordagem, de

forma que a rede pré-treinada utilizada possui 400 dimensões. A abordagem utilizando *bag-of-words* com TF-IDF construiu uma matriz com 2.480 dimensões e a abordagem com Unigramas obteve a mesma quantidade de dimensões, porém com valores em nível de atributo diferentes. Vale ressaltar que isso ocorre devido as abordagens terem formas diferentes de realizar o processo de extração de características.

Em relação as técnicas utilizadas, o TF-IDF (*Term Frequency and Inverse Document Frequency*) representa a distribuição ponderada dos termos, onde o TF representa a contagem de um termo t dentro de um documento d, e o IDF representa a distribuição de probabilidade observando o mesmo termo t, mas dessa vez observando a relação desse termo em toda base de dados. Os Unigramas, por sua vez, representam a contagem absoluta de um termo t dentro de toda a base de dados. Por fim, as redes de *Embeeding* assumem características distintas da estratégia de *bag-of-words*, pois enquanto o TF-IDF e os Unigramas assumem que os termos dentro de uma coleção são independentes, os *Embeedings* assumem que existe associações probabilísticas entre palavras e, com isso, estas redes conseguem construir processos de extração de características preservando a integridade do texto e com dimensões menores que a abordagem de *bag-od-words*.

3.3. Classificação

Após a construção da matriz, executamos a comparação de três redes profundas (CNN, RNN e LSTM), a fim de verificar qual apresenta melhor resultado na tarefa de classificação das categorias dos aspectos. Estas redes foram selecionadas porque observamos que nos trabalhos existentes no estado da arte, estas são as mais utilizadas para classificar tal análise semântica [Poria et al. 2016, Kim 2014, Nguyen and Shirai 2015]. Para obtenção e comparação de resultados, utilizamos os mesmos hiperparâmetros aplicados nas pesquisas mapeadas na seção dos trabalhos relacionados.

4. Experimentos e Resultados

A seguir, apresentamos uma descrição da base de dados utilizada neste trabalho, os experimentos realizados e os resultados obtidos após avaliação do método proposto.

4.1. Base de Dados

A base de dados utilizada neste artigo é disponibilizada no site SemEval (conferência focada em pesquisas de análise semântica). Neste contexto, esta base é composta por 2.290 avaliações de restaurantes[1] em inglês, onde os dados já se encontram particionados em amostras de treino e teste. Com isso, a base selecionada é composta por 1.708 amostras que serão utilizadas para treino e 582 amostras para teste. No total, foram reconhecidas 12 categorias referentes aos aspectos presentes, conforme ilustrado na Figura 2.

Conforme observado na Figura 2, as 3 categorias mais presentes na base de dados são FOOD#QUALITY, RESTAURANT#GENERAL e SERVICE#GENERAL, pois no contexto de restaurante é comum que os usuários comentem sobre qualidade de comida, nome de restaurantes e qualidade do serviço. Outra característica que pode ser destacada é que cada texto pode possuir mais de uma categoria presente em seu contexto. Com isso, adotamos a mesma estratégia de [Pontiki et al. 2016], onde o problema pode ser tratado com multi-rótulo, ou seja, a lista de rótulos será convertida em uma matriz de rótulos.

[1] `http://alt.qcri.org/semeval2014/task4/index.php?id=data-and-tools`

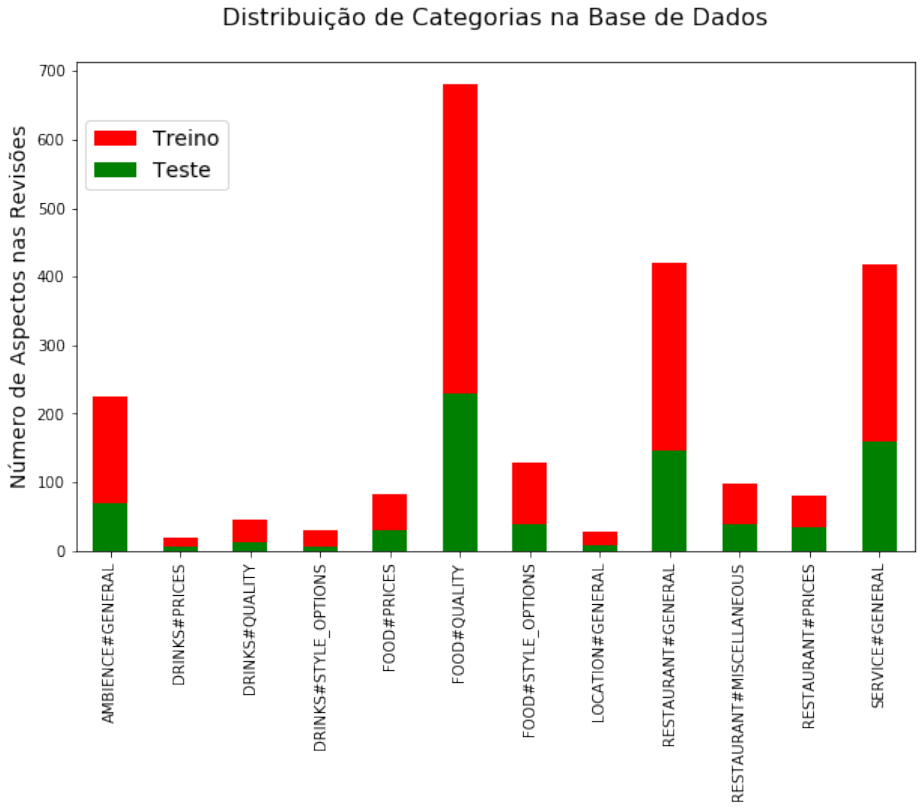


Figure 2. Distribuição dos dados dentro da coleção.

4.2. Classificação de Texto

Para a classificação de texto utilizamos a abordagem de organização dos rótulos de cada documento, adotando uma estratégia multi-classe. Assim, assumimos que cada documento pode possuir mais de uma categoria associada [Pontiki et al. 2016]. Para a tarefa treinamento utilizamos 3 redes profundas para classificação das categorias dos aspectos (LSTM, CNN, RNN) e, além disso, variamos os modelos de representação dados (TF-IDF, Unigramas e *Embeeding*), com o objetivo de verificar qual apresenta melhor desempenho estatístico segundo as métricas de avaliação. Durante o processo de classificação dos textos, utilizamos 100 épocas para medir o desempenho de aprendizado das redes profundas.

Os parâmetros utilizados nas redes podem ser observados na Tabela 1. Vale ressaltar que todos os parâmetros foram selecionados de forma empírica, onde alteramos esses parâmetros, a fim de verificar qual apresenta melhor resultado.

Table 1. Tabela de parâmetros usados nas redes profundas.

	Função de perda	Funções de ativação	Otimizador
LSTM	Categorical Crossentropy	Softmax	Rmsprop
CNN	Categorical Crossentropy	Relu/Softmax	Adam
RNN	Categorical Crossentropy	Softmax	Adam

4.3. Métricas de Avaliação

Para avaliar a eficácia dos classificadores na tarefa de identificação de categorias de aspectos, utilizamos as seguintes métricas: precisão, revocação, F1 e acurácia. Enquanto a precisão consiste na fração de documentos atribuídos a uma determinada classe que realmente pertencem a esta classe (segundo o conjunto de teste), a revocação representa a fração de todos os documentos que pertencem a uma determinada classe (segundo o conjunto de teste) e foram atribuídas corretamente a esta classe pelo classificador. Já a métrica F1 pode ser definida como uma medida que busca relacionar as métricas de precisão e revocação a fim de obter uma medida de qualidade que equilibre a importância relativa destas duas métricas. Esta medida pode ser atingida atraves da média harmônica entre a precisão e a revocação.

4.4. Resultados

Em nossos experimentos, medimos o desempenho estatístico das redes profundas alterando os modelos de representação de dados ao longo de 100 épocas, a fim de verificar se há melhora nos resultados da classificação dos textos. Para a avaliação do uso de *Embeeding*, utilizamos uma rede pré-treinada com 400 dimensões do Yelp, desenvolvida para realizar o processo de extração de características em avaliações. Para o uso do TF-IDF e dos Unigramas, utilizamos o processamento direto sobre os textos existentes em nossa base dados.

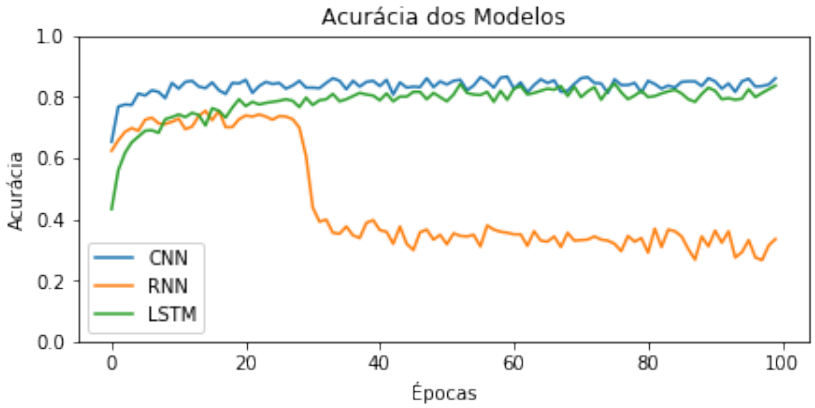
Como resultado, conseguimos observar que a rede profunda CNN apresentou melhor resultado em todas as variações de representação de dados (conforme ilustrado nas Figuras 3, 4 e 5), com destaque para o resultado com TF-IDF, onde obtivemos o melhor desempenho na classificação de categorias de aspectos (Figura 3(b)). Vale ressaltar que não eliminamos termos com baixa frequência e, com isso, a matriz de representação manteve 2.290 dimensões. Já o uso de Unigramas apresentou resultado similar ao do TF-IDF, porém é possível observar que a RNN não obteve resultados satisfatórios, pois nos três casos esta rede apresentou acurácia inferior a 50%.

Os experimentos que utilizam *Embeeding* em sua maioria apresentam resultados superiores ao uso de *bag-of-words* (TF-IDF ou Unigramas), pois essa abordagem tem uma característica de observar correlações entre os termos e, além disso, montar representações de dados com dimensões menores que as demais, preservando todas as características semânticas. Em nossos experimentos, foi possível observar que o uso de *Embeeding* apresentou uma medida F1 de 91%, enquanto o uso com TF-IDF apresentou 93%.

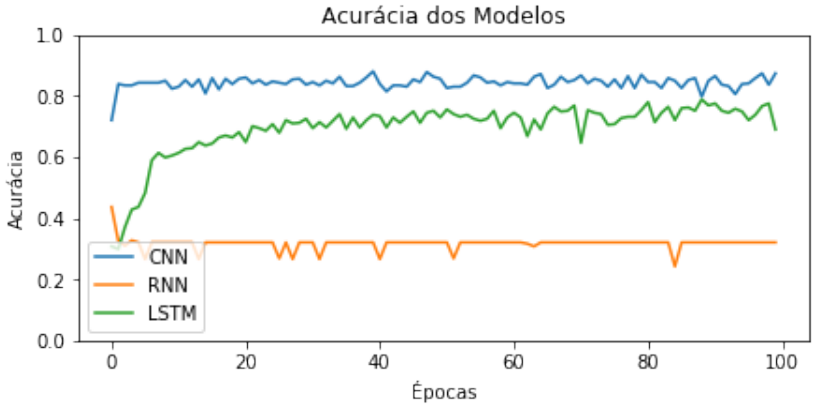
Vale ressaltar, que as métricas de precisão, revocação e F1 são apresentadas para a média das classes, pois estamos tratando esse problema com uma estratégia multi-classes. Assim, os valores representam todas as 12 classes detectadas. Outra característica importante que pode ser destacada é o aprendizado da rede profunda LSTM, pois com o passar das épocas é possível observar como sua acurácia aumenta. Com isso, é possível estimar que, algum ponto ao longo das épocas aconteça uma convergência de resultados com a rede que apresentou melhor resultado (CNN).

5. Conclusões e Diretrizes Futuras

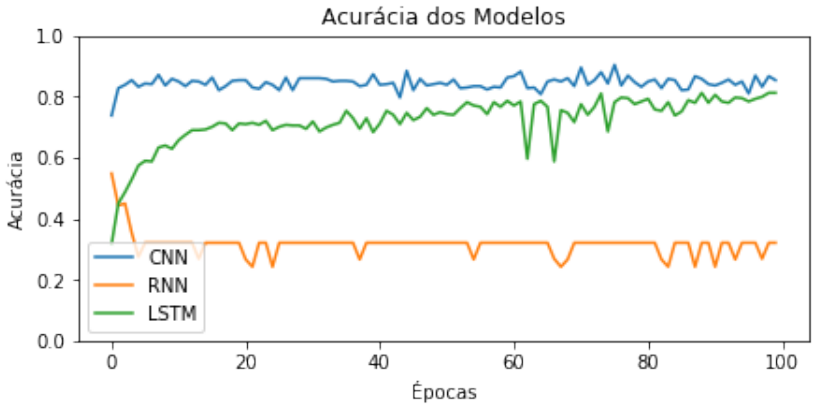
Neste trabalho, realizamos a classificação das categorias dos aspectos utilizando uma abordagem de aprendizagem profunda, onde comparamos 3 técnicas de representação



(a) Acurácia com uso de *Embeeding*.



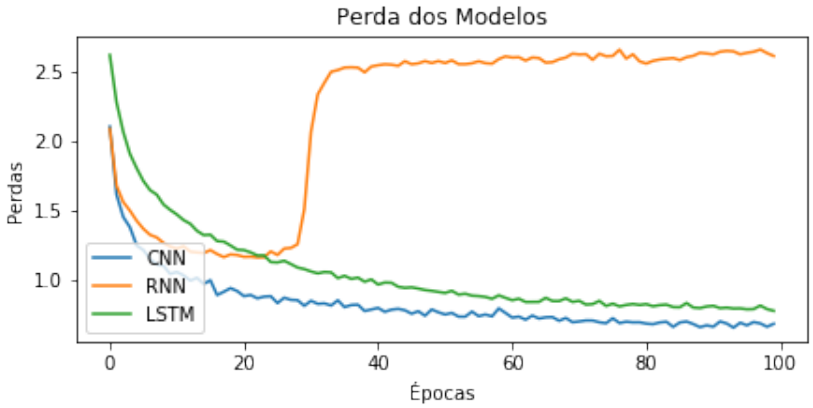
(b) Acurácia com uso de TF-IDF.



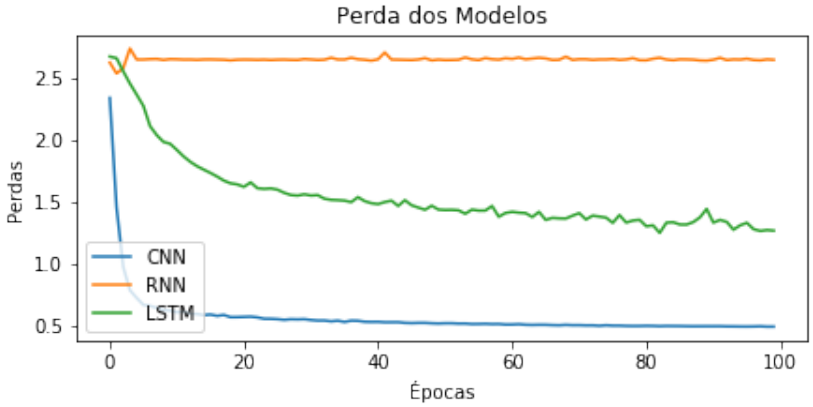
(c) Acurácia com uso de Unigramas.

Figure 3. Acurácia das redes variando os modelos de representação de dados.

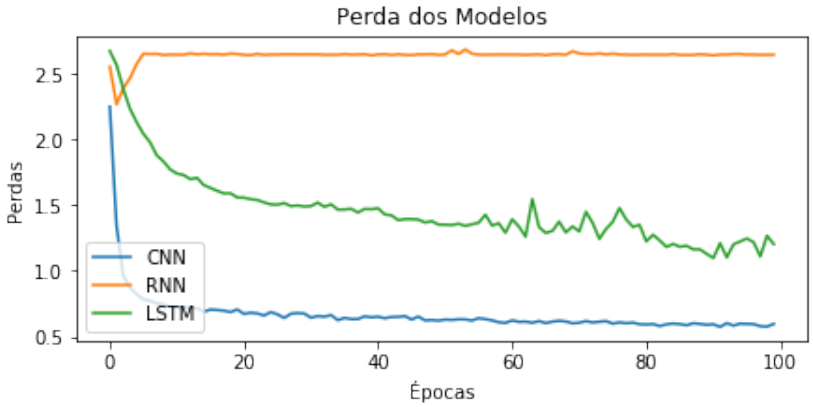
de dados (Unigramas, TF-IDF e *Embeedings*) combinadas as redes profundas, que tem sido utilizadas para realizar classificação na área de análise de sentimentos baseadas em aspectos no domínio de avaliação de restaurantes. Em nossos resultados, podemos observar que a abordagem utilizando TF-IDF combinada com a rede profunda CNN apresentou uma medida F1 de 0,93%, sendo superior ao uso de *Embeeding* com a mesma rede profunda. Porém, se observarmos em um contexto de todas as redes analisadas, o uso de



(a) Perdas com uso de *Embeeding*.



(b) Perdas com uso de TF-IDF.

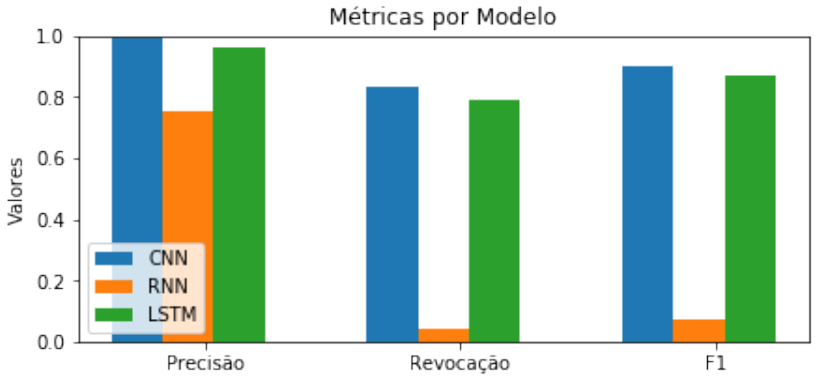


(c) Perdas com uso de Unigramas.

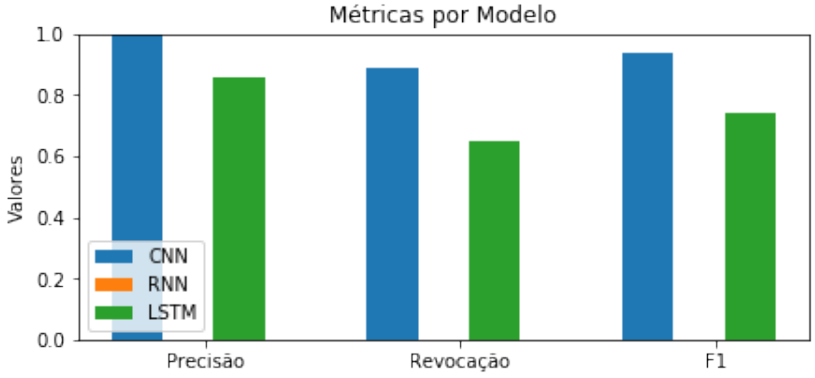
Figure 4. Perdas das redes variando os modelos de representação de dados.

Embeeding gerou uma capacidade de aprendizado mais eficiente nas 3 redes profundas que o uso de TF-IDF.

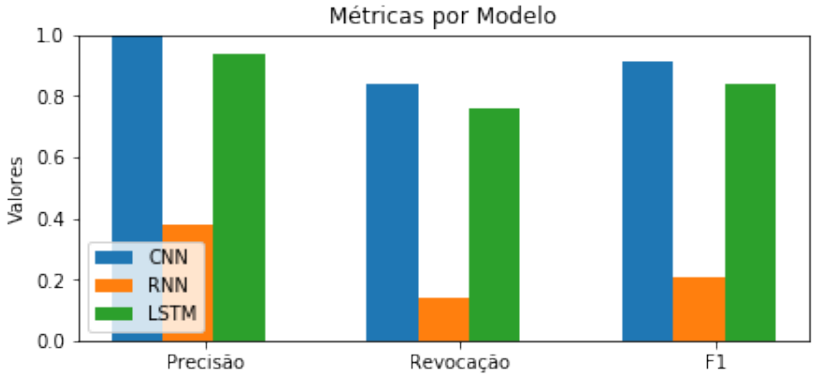
Vale ressaltar que tal classificação ainda apresenta muitas oportunidades de estudo, como o uso de outros modelos de representação de dados e a utilização de métodos estatísticos de extração de tópicos para reconhecimento de categorias. Com isso, como diretrizes futuras, podemos alterar as redes pré-treinadas de *Embedding*, a fim de veri-



(a) Métricas com uso de *Embeeding*.



(b) Métricas com uso de TF-IDF.



(c) Métricas com uso de Unigramas.

Figure 5. Métricas (Precisão, Revocação e F1) das redes variando os modelos de representação de dados.

ficar se a uma melhora na eficiência do modelo utilizando redes com maiores dimensões. Também podemos inserir a tarefa de análise de sentimentos sobre os aspectos detectados, com o objetivo de fornecer uma classificação com mais características sobre os textos analisados.

6. Referências

Almeida, T. G., Souza, B. A., Menezes, A. A., Figueiredo, C., and Nakamura, E. F. (2016). Sentiment analysis of portuguese comments from foursquare. In *Proceedings*

of the 22nd Brazilian Symposium on Multimedia and the Web, pages 355–358. ACM.

Araújo, M., Gonçalves, P., Benevenuto, F., and Cha, M. (2013). Métodos para análise de sentimentos no twitter. In *Proceedings of the 19th Brazilian symposium on Multimedia and the Web (WebMedia'13)*.

de Paula, H. L., Souza, B. A., Nakamura, F. G., and Nakamura, E. F. (2017). Quantificando a importância de emojis e emoticons para identificaç ao de polaridade em avaliaç oes online.

Gulaty, M. (2016). *Aspect-Based Sentiment Analysis*. PhD thesis, Dublin, National College of Ireland.

Kim, Y. (2014). Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167.

Nguyen, T. H. and Shirai, K. (2015). Phrasernn: Phrase recursive neural network for aspect-based sentiment analysis. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2509–2514.

Pavlopoulos, I. (2014). Aspect based sentiment analysis. *Athens University of Economics and Business*.

Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., Manandhar, S., Mohammad, A.-S., Al-Ayyoub, M., Zhao, Y., Qin, B., De Clercq, O., et al. (2016). Semeval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, pages 19–30.

Pontiki, M., Galanis, D., Papageorgiou, H., Manandhar, S., and Androutsopoulos, I. (2015). Semeval-2015 task 12: Aspect based sentiment analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 486–495.

Poria, S., Cambria, E., and Gelbukh, A. (2016). Aspect extraction for opinion mining with a deep convolutional neural network. *Knowledge-Based Systems*, 108:42–49.

Santos, R. L. d. S. and Moura, R. S. (2016). Extraç ao de métricas e analise de sentimentos em comentarios web no domınio de hotéis.

Souza, B. A., Almeida, T. G., Menezes, A. A., Nakamura, F. G., Figueiredo, C., and Nakamura, E. F. (2016). For or against?: Polarity analysis in tweets about impeachment process of brazil president. In *Proceedings of the 22nd Brazilian Symposium on Multimedia and the Web*, pages 335–338. ACM.

Stiilpen Junior, M. and Merschmann, L. H. C. (2016). A methodology to handle social media posts in brazilian portuguese for text mining applications. In *Proceedings of the 22nd Brazilian Symposium on Multimedia and the Web*, pages 239–246. ACM.

Wang, B. and Liu, M. (2015). Deep learning for aspect-based sentiment analysis.

Wang, W., Pan, S. J., Dahlmeier, D., and Xiao, X. (2016). Recursive neural conditional random fields for aspect-based sentiment analysis. *arXiv preprint arXiv:1603.06679*.

Xu, L., Lin, J., Wang, L., Yin, C., and Wang, J. (2017). Deep convolutional neural network based approach for aspect-based sentiment analysis. *Adv Sci Technol Lett*, 143:199–204.

Ye, Q., Zhang, Z., and Law, R. (2009). Sentiment classification of online reviews to travel destinations by supervised machine learning approaches. *Expert systems with applications*, 36(3):6527–6535.

Detecção de Posicionamento em *Tweets* sobre Política no Contexto Brasileiro

William Christhie, Julio C. S. Reis, Fabrício Benevenuto
Mirella M. Moro, Virgílio Almeida

[1]Universidade Federal de Minas Gerais (UFMG) – Brasil

{williamchristhie,julio.reis,fabricio,mirella,virgilio}@dcc.ufmg.br

Resumo. *Opiniões compartilhadas na Web constituem um grande volume de dados. Nelas, posicionamentos são expressos direta ou indiretamente, e sua detecção identifica qual a polaridade em relação a uma ideia alvo. Neste trabalho, apresentamos a caracterização de um amplo conjunto de tweets em português sobre a corrida presidencial brasileira de 2018. Tal conjunto serve como base para a detecção automática de posicionamento através de uma abordagem semi-supervisionada. Em nossa avaliação, encontramos indícios de bots na rede. Também avaliamos três classificadores com teste estatístico pareado, e nossos resultados apresentam F-Measure acima de 94%.*

Abstract. *Opinions shared over the Web constitute big volumes of data. Moreover, they may contain stances that are expressed directly or indirectly. Hence, stance detection may help to define the polarity related to a target idea. Here, we present the characterization of a broad set of tweets in Portuguese about the 2018 Brazilian presidential race. Such a set serves as the basis for automatic stance detection through a semi-supervised approach. In our evaluation, we find clues on the presence of bots in the network. We also evaluate three classifiers with paired statistical test, and our results present F-Measure above 94%.*

1. Introdução

As redes sociais online são um ambiente propício para o compartilhamento de vários assuntos, tais como visões políticas e opiniões pessoais sobre temas polêmicos como religião e sexualidade [Reis et al. 2015]. Dentre os milhões de usuários de redes sociais online, os brasileiros são especialmente ativos, e gastam em média 3 horas por dia no acesso a estes ambientes[1]. Com tamanha riqueza de conteúdo, várias aplicações têm surgido a fim de minerar as opiniões existentes. Uma das técnicas utilizadas na mineração de opiniões é a Detecção de Posicionamentos, que consiste em identificar como o autor de determinado texto se posiciona em relação a uma proposição alvo. Tal posicionamento pode ocorrer de forma neutra, favorável ou contrária sobre proposições que podem tratar de uma causa, pessoa, produto, organização ou política de governo, por exemplo.

A detecção de posicionamentos é abordada em trabalhos recentes sobre diversos assuntos, incluindo aborto, ateísmo, clima, feminismo e a corrida presidencial americana (Hillary e Trump) [Dias and Becker 2016a, Mohammad et al. 2017]. Porém , vários

[1]https://www.techtudo.com.br/noticias/2018/02/10-fatos-sobre-o-uso-de-redes-sociais-no-brasil-que-voce-precisa-saber.ghtml

fatores contribuem para que a tarefa não seja trivial. Por exemplo, embora existam registros favoráveis e contrários nos assuntos que envolvem controvérsia, o mesmo não ocorre em situações onde uma grande mobilização se manifesta em apenas uma das polaridades, como por exemplo os temas cura gay, aborto ou preconceito. Nesses casos, "posicionar-se abertamente" pode ter impacto nos relacionamentos internos e externos à rede, já que geralmente os usuários desses sistemas podem ser identificados no mundo real. Além deste fator, o uso de técnicas como Análise de Sentimentos não é uma solução viável, visto que os posicionamentos podem ocorrer indiretamente, através de opiniões sobre alvos relacionados [Mohammad et al. 2017]. Por exemplo, a manifestação de apoio a um determinado candidato que é contra a privatização de empresas públicas poderia ser expressa apenas com palavras negativas sobre a privatização, sem que o candidato fosse efetivamente citado no texto. Deste modo, a análise poderia classificar o texto como negativo, sendo que na verdade o texto é positivo em relação ao candidato em questão.

Embora existam trabalhos sobre Análise de Sentimentos no contexto brasileiro e para o idioma português (e.g., [Araújo et al. 2016] e [Bigonha et al. 2012]), o mesmo não ocorre com a Detecção de Posicionamentos, uma vez que não encontramos trabalhos abordando as especificidades do idioma e contexto brasileiros para o assunto (busca na DBLP em março/2018). Além disso, o desenvolvimento de técnicas e ferramentas capazes de detectar posicionamentos de usuários brasileiros é de grande importância. Por exemplo, a análise político-eleitoral do contexto brasileiro pode beneficiar-se da utilização de tais ferramentas, especialmente no monitoramento da aceitação de determinado candidato por parte da população, ou no apoio em pesquisas sobre comportamento político que são regularmente realizadas por institutos como Ibope[2].

Neste trabalho desenvolvemos um classificador que infere automaticamente o posicionamento expresso em *tweets* escritos em português. Especificamente, apresentamos uma metodologia que pode ser utilizada independentemente de contexto, e aplicamos ao cenário político-eleitoral brasileiro. A hipótese investigada é a de que o forte viés ideológico existente em *hashtags* pode ser utilizado para etiquetar de forma automática um conjunto de *tweets* que possa ser utilizado para treinar um classificador. Para isso, nós coletamos uma amostra do *Twitter* com mensagens favoráveis ou contrárias a quatro presidenciáveis, estabelecidos como alvo. Então, caracterizamos esse conteúdo e treinamos um classificador capaz de explorar diferenças psico-linguísticas expressas nos textos para a categorização das mensagens. Resultados mostram indícios de uma possível ação massiva de robôs (*bots*) atuando em polaridades e alvos específicos. Além disso, obtivemos F-Measure acima de 94% para as tarefas de classificação.

O restante desse artigo está organizado da seguinte forma. A Seção 2 descreve os trabalhos relacionados. A Seção 3 detalha a metodologia utilizada para a caracterização e classificação. A Seção 4 apresenta a análise e discussão dos resultados. A Seção 5 contém as conclusões e direções para trabalhos futuros.

2. Trabalhos Relacionados

A Detecção de Posicionamentos tem sido abordada por diferentes técnicas. Por exemplo, Shenoy et al [2017] empregam técnicas linguísticas em um processo supervisionado de aprendizado de máquina. Mourad et al. [2018] utilizam a técnica de maioria de votos

[2]http://www.ibope.com.br

entre os classificadores *Random Forest*, *Linear SVM* e *Naive Bayes*. Além disso, Dias e Becker [2016b] apresentam uma metodologia baseada em regras e análise de sentimentos para rotular os *tweets* que são utilizados para treinar um classificador SVM. Em outra abordagem, Xu et al. [2017] empregam fatores latentes para a extração de características dos textos e um *lexicon* de sentimentos para dividir o corpus em duas partes e treinar um classificador SVM. Por fim, Chen et al. [2017] abordam o problema com redes neurais convolucionais e *word embedding*.

De outro modo, Mohammad et al. [2017] demonstram que o sentimento expresso em um *tweet* é benéfico, mas não suficiente para a tarefa de classificação do posicionamento. Eles baseiam-se na ideia de que uma pessoa pode expressar a mesma posição em relação a um alvo usando linguagem negativa ou positiva. Além dessas, outra proposta para detecção de posicionamento em texto de *tweets* de forma não-supervisionada é apresentada em [Dias and Becker 2016a]. O algoritmo recebe como entrada apenas o alvo e um conjunto de *tweets* a rotular. Ele é baseado em uma abordagem híbrida composta pelas etapas de rotulação automática baseada em um conjunto de heurísticas e classificação complementar baseada em aprendizado supervisionado de máquina.

Entre a diversidade de estudos publicados sobre análise de sentimento, existem os que tratam especificamente do contexto político-eleitoral. Por exemplo, o cenário da eleição presidencial americana de 2016 é estudado em [Caetano et al. 2017], onde os autores utilizam análise de sentimentos para estudar a homofilia política entre os usuários do Twitter. Já no contexto específico brasileiro, Verona et al. [2017] estudam um ponto diferente: o de poder em redes sociais avaliado através das doações de campanha para o Senado Federal. Entretanto, não existem trabalhos que abordem a detecção de posicionamentos para textos em português no contexto político-eleitoral brasileiro.

3. Metodologia

A metodologia utilizada está dividida em duas partes, de acordo com as principais contribuições do trabalho: a primeira é a coleta e a descrição dos dados; e a segunda é a construção de *datasets* e o treinamento dos classificadores. A seguir, cada etapa é descrita individualmente.

3.1. Coleta e Descrição dos Dados

A API do Twitter[3] foi utilizada para coletar mensagens (*tweets*) sobre possíveis candidatos para as eleições presidenciais de 2018[4]. Neste estudo quatro presidenciáveis foram considerados em função do volume de mensagens: Alckmin, Bolsonaro, Dória e Lula.

Especificamente, definimos um conjunto de *hashtags* (Tabela 1) através de pesquisa manual para selecionar aquelas com forte viés ideológico por assunto, observando sua ocorrência em *tweets* favoráveis ou contrários à ideia alvo. Embora o número de etiquetas a favor e contra esteja desbalanceado, o total de registros coletados não é diretamente proporcional à essa quantidade. Por exemplo, o mesmo número de *hashtags* favoráveis e contrárias sobre o assunto Lula é utilizado para a coleta, porém o total de mensagens extraído é maior para o posicionamento contrário. O assunto Dória também apresenta situação semelhante, pois a maioria das mensagens coletadas está na polaridade

[3]https://dev.twitter.com

[4]http://politica.estadao.com.br/noticias/geral,veja-quem-quer-ser-presidente-em-2018,70002054149

Tabela 1. *Hashtags* utilizadas para a coleta

Candidato	Favor	Contra
Alckmin	#geraldoAlckmin, #alckmin2018, #geraldopresidente, #geraldoAlckminPresidente, #geraldoAlckmin2018	#ForaAlckmin, #alckminladrão
Bolsonaro	#BolsonaroMito, #BolsoMito, #BolsonaroPresidente	#ForaBolsonaro, #ForaBolsoLixo, #BolsonaroDitador
Dória	#doriaPresidente, #doria2018, #joaodoriapresidente	#foraDoria
Lula	#LulaPresidente, #TocomLula, #LulaPeloBrasil	#ForaLula, #LulaNaCadeia, #ForaPT

que possui menos etiquetas. Assim, o fator de maior influência no total de *tweets* coletados é a importância dada pelos usuários para cada *hashtags*, e não o número de etiquetas. Embora neste trabalho tenhamos utilizado esse conjunto de *hashtags*, a metodologia pode ser utilizada para outros conjuntos mais amplos.

Ao final da coleta, dois conjuntos de dados são obtidos para cada presidenciável pesquisado (*tweets* favoráveis, e *tweets* contrários). Em seguida, as *hashtags* são removidas do texto original para que elas não enviesem o processo de treinamento do classificador. Além do texto, cada *tweet* inclui informações adicionais sobre o usuário, como nome e identificador, além de informações sobre a mensagem, tais como *timestamp* e indicador de repostagem. Também é importante destacar que, embora a API restrinja o volume da coleta a 1% dos *tweets* públicos, isso representa um volume significativo de mensagens, visto que diariamente cerca de 500 milhões de *tweets*[5] são compartilhados. Por fim, utilizamos os conjuntos de dados extraídos e etiquetados como favoráveis ou contrários a cada assunto para efetuar uma análise estatística descritiva da coleta.

3.2. Treinamento do Classificador

Composição dos *Datasets* de treino. Criamos conjuntos de dados para treinar os classificadores utilizando os registros dos dois assuntos mais comentados após a filtragem efetuada durante a caracterização. O primeiro contém 41.498 *tweets* em português com o assunto Lula, sendo 27.717 contrários e 13.781 favoráveis. Porém, para o segundo *dataset*, o volume de *tweets* contrários coletado é muito baixo (159), inviabilizando a aplicação direta das técnicas de aprendizado de máquina. Para contornar tal situação, analisamos a viabilidade de utilização das mensagens de temas distintos na elaboração do *dataset*. Assim, a Figura 1 apresenta o cruzamento efetuado entre os usuários que postam mensagens em cada assunto, de modo a verificar a interseção entre os conjuntos. Os maiores valores encontrados são entre os usuários que postam: contrários a Dória e Alckmin (11,29%), além de contrários a Lula e favoráveis a Bolsonaro (5,72%), sendo que todas as demais apresentam valores baixos. Especificamente, a interseção entre os usuários que postam a favor de Lula e são favoráveis ao Bolsonaro é apenas 2,03%, e a interseção entre Lula Contra e Bolsonaro Contra é de somente 0,14%. Assim, os valores sugerem que os usuários realmente se posicionam em grupos contrários. Nota-se ainda que as interseções podem representar pessoas indecisas ou que mudaram de opinião. No contexto brasileiro tal cenário faz sentido, uma vez que esses presidenciáveis posicionam-se em extremos antagônicos de direita e esquerda. Com isso, utilizamos os dados favoráveis ao Lula como sendo contrários ao Bolsonaro para compor o segundo *dataset*, que contém 31.366 *tweets*, sendo 13.940 contrários e 17.426 favoráveis.

Extração de características. A estruturação dos dados em forma de texto é necessária

[5]http://www.internetlivestats.com/twitter-statistics/

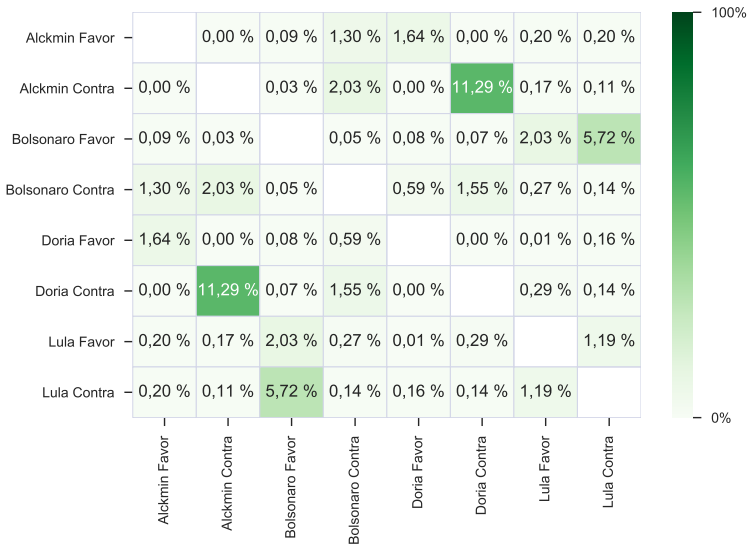


Figura 1. Cruzamento dos conjuntos de usuários que postam em cada classe

Tabela 2. Importância das características (ganho de informação)

Bolsonaro		Lula	
Importância	**Característica (*feature*)**	**Importância**	**Característica (*feature*)**
0,0149893	hora (hour)	0,13919312	todas pontuações (AllPunc)
0,0126505	dia da semana (dayOfWeek)	0,12782251	vírgulas (Colon)
0,0002846	lugar (place)	0,11818538	palavras de dicionário (Dic)
0,0000573	maiores que 6 letras (Sixltr)	0,10034218	outras pontuações (OtherP)
0,0000464	palavras com função gramatical (funct)	0,08829323	palavras com função gramatical (funct)

para possibilitar a utilização dos *tweets* por um classificador. Assim, utilizamos a ferramenta LIWC[6] para a extração de *features* dos textos em português de cada *dataset*. O software efetua a análise psicolinguística do texto e fornece um vetor com suas características quantitativas morfológicas, gramaticais e psicolinguísticas. O texto é então removido do *dataset* e as colunas com o processamento do LIWC são adicionadas.

Importância das características. A fim de investigar o poder discriminativo das *features* criadas e a possível redução de dimensionalidade dos dados, criamos um *ranking* com o ganho de informação de cada característica para a predição da classe. Desse modo, filtramos os atributos para manter somente aqueles com importância acima de zero, uma vez que os demais adicionam ruído no processo classificatório. Para exemplificar o processo, os cinco melhores atributos encontrados na análise são apresentados na Tabela 2. Ao final da etapa, restam 56 atributos para o *dataset* Bolsonaro e 82 para o *dataset* Lula.

Teste estatístico. Neste trabalho utilizamos os mesmos classificadores aplicados por Mourad et al. [2018] (*Naive Bayes*, *SVM* e *Random Forest*) para comprovar a hipótese inicial de que a técnica semi-supervisionada de etiquetamento funciona para o contexto brasileiro. Assim, executamos um teste-t pareado para comparar o desempenho de tais classificadores. Para tal, a ferramenta Weka Experimenter[7] é utilizada, uma vez que ela

[6]https://liwc.wpengine.com
[7]https://www.cs.waikato.ac.nz/ml/weka/

Tabela 3. Quantitativos de dados extraídos

Candidato	Extração			Idioma Português			% Português		Datas	
	Contra	Favor	Ambos	Contra	Favor	Ambos	Contra	Favor	Início	Fim
Alckmin	167	691	3	144	624	3	86%	90%	21/10/17	01/01/18
Bolsonaro	178	19.691	0	135	18.183	0	76%	92%	21/10/17	02/01/18
Dória	161	99	0	115	77	0	71%	78%	21/10/17	01/01/18
Lula	91.012	36.721	144	35.105	30.380	90	**39%**	83%	21/10/17	02/01/18

Tabela 4. Idiomas, maior número de postagens e de idiomas por usuário

Candidato	Posicionamento	# Idiomas	Maior # Postagens	Maior # Idiomas
Alckmin	Contra	3	72	2
Alckmin	Favor	5	441	2
Bolsonaro	Contra	3	21	3
Bolsonaro	Favor	19	152	4
Doria	Contra	4	26	2
Doria	Favor	4	20	2
Lula	Contra	35	11.882	33
Lula	Favor	16	1.229	6

permite fácil acesso as técnicas de aprendizado de máquina [Witten et al. 2016]. Realizamos 15 repetições entre os classificadores para cada *dataset*, utilizando a métrica F-Measure para *5-folds Cross Validation* em cada repetição, totalizando 75 medições para cada algoritmo. A possibilidade de rejeição da hipótese nula de que os classificadores são iguais é verificada com um nível de significância de 5%.

4. Resultados e Discussões

Nesta seção, apresentamos e discutimos os resultados, que estão divididos em caracterização dos dados coletados e classificação automática. Além disso, discutimos alguns pontos de ameaça à validade deste estudo.

4.1. Caracterização

Estatísticas Básicas. Caracterizamos os dados coletados para entender melhor sua composição e evitar que os classificadores deem maior importância para as características de *tweets* provenientes de postagens automatizadas e recorrentes. A Tabela 3 contém o volume coletado para cada tema, bem como o viés ideológico do conjunto de *hashtags*, visto que poucas mensagens aparecem com etiquetas dos dois posicionamentos (a favor e contra). Essas mensagens são descartadas, de modo a garantir a separação dos dados para treinamento dos classificadores. Além disso, também pode ser observado na Tabela 3 que o maior npumero de *tweets* aborda o assunto Lula, e a maioria dos presidenciáveis não apresenta grande diferença entre os totais de *tweets* em português e em outros idiomas. Por outro lado, somente 38% das mensagens abordando o assunto Lula possuem o idioma português, o que pode ser um indício da presença de *bots*.

Estatísticas de Idioma. Analisamos a divisão por idiomas envolvendo cada categoria (candidato e polaridade) para investigar os valores observados no assunto Lula. Então, selecionamos para cada categoria o maior número de mensagens postadas por um único usuário e o maior número de idiomas utilizados por um único usuário. Os resultados são apresentados na Tabela 4, que apresenta um grande número de idiomas distintos em algumas categorias, sendo 35 contrários ao Lula, 19 favoráveis ao Bolsonaro, e 16 favoráveis ao Lula. Investigamos manualmente alguns desses *tweets* e, aparentemente, o idioma

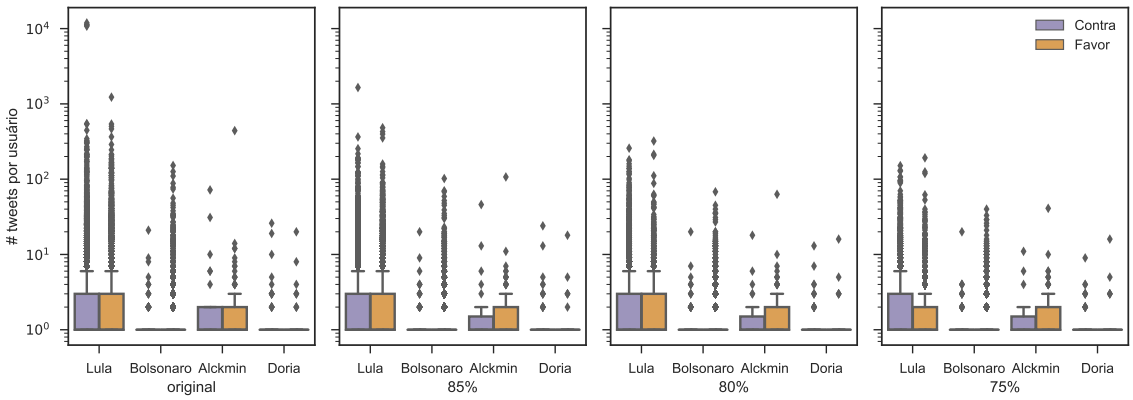


Figura 2. Distribuição do volume de *tweets* por usuário em cada classe com diferentes níveis de similaridade

de mensagens formadas por links, muitas *hashtags* e poucas palavras não é identificado corretamente. Um grande número desse tipo de mensagem foi encontrado para os temas citados, o que corrobora tal hipótese da existência de *bots* postando mensagens automaticamente. Também corrobora com a hipótese o volume de 33 idiomas distintos para um único usuário com *tweets* contrários ao Lula, e o alto volume de mensagens por usuário, sendo 11.882 *tweets* na categoria contrário ao Lula e 1.229 na favorável ao Lula.

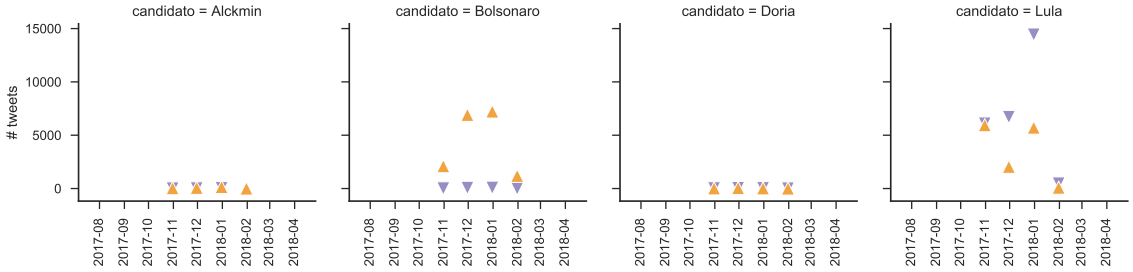
Volume por Usuário. A partir da possibilidade identificada da existência de *bots*, efetuamos uma análise mais detalhada da distribuição do volume de postagens por usuários para investigá-la. A Figura 2 apresenta as distribuições de volume de *tweets* por usuário em cada categoria. O primeiro *boxplot* representa a distribuição original dos dados coletados, no qual pode-se observar o grande npumero de mensagens postadas por usuários *outliers* em todos os temas. A exclusão de mensagens iguais postadas por um mesmo usuário não é suficiente para remover os *outliers*, uma vez que a maioria das mensagens possui pequenas alterações entre si.

Filtragem por Similaridade. Como a exclusão por igualdade não serve, adotamos a remoção de *tweets* através de funções de similaridade. Em [Christen 2006], várias dessas funções são testadas para a deduplicação de nomes. Além disso, essas funções são comparadas nas etapas de deduplicação de registros considerando 11 domínios distintos em [Silva et al. 2017]. Conforme os autores, a função *Jaro-winkler* está entre as que possuem o melhor desempenho. Por este motivo, neste estudo a função é utilizada para mensurar a semelhança entre dois textos.

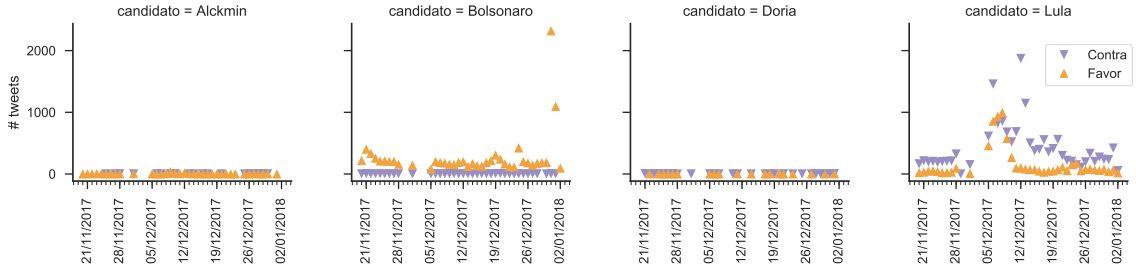
Especificamente, todas as mensagens pertencentes a um mesmo usuário e uma mesma classe são comparadas entre si utilizando-se três opções *threshold* para a definição de textos similares: 75%, 80% e 85%. Assim, os resultados das exclusões de mensagens identificadas como similares em cada um dos níveis podem ser observados nos três últimos *boxplots* da Figura 2. A remoção de *tweets* semelhantes nos três níveis de similaridade reduz o volume de mensagens dos usuários *outliers*. Entretanto, constatamos em inspeção manual que algumas mensagens similares continuam presentes nos níveis 85% e 80%. Além disso, a inspeção demonstra que muitas das mensagens semelhantes são removida no nível 75%, mas que o nível 70% remove mensagens legítimas. Deste modo, neste trabalho os *tweets* com 75% de semelhança são removidos para o restante

Tabela 5. Estatísticas do *dataset* antes e depois da remoção de *tweets* similares.

Candidato	Dataset original					Tweets duplicados removidos (75%)				
	Usuários	*tweets*	$\overline{x}$	75%	max	Usuários	*tweets*	$\overline{x}$	75%	max
Alckmin Contra	39	167	4 ± 12	2	72	39	64	2 ± 2	2	11
Alckmin Favor	122	691	5 ± 35	2	387	122	249	2 ± 4	2	41
Bolsonaro Contra	109	178	2 ± 2	1	21	109	159	1 ± 2	1	20
Bolsonaro Favor	14.832	19.691	1 ± 2	1	126	14.832	17.426	1 ± 1	1	40
Dória Contra	85	161	2 ± 3	1	26	85	114	1 ± 1	1	9
Dória Favor	61	99	2 ± 2	1	18	61	86	1 ± 2	1	16
Lula Contra	10.356	91.012	9 ± 222	3	11.877	10.356	27.717	3 ± 5	3	151
Lula Favor	7.051	36.721	5 ± 23	3	1.228	7.051	13.781	2 ± 4	2	192



(a) Distribuição mensal



(b) Distribuição diária destacando os meses de dezembro e janeiro

Figura 3. Volume de *tweets* em função do tempo

das caracterizações apresentadas.

A Tabela 5 apresenta estatísticas do conjunto de dados antes e depois da remoção das mensagens similares para analisar seu impacto. Aqui, 50% dos usuários de todas as classes possuem um único *tweet*, e por questões de simplicidade estas colunas não foram apresentadas na tabela. Observa-se uma pequena alteração no número de *tweets* da faixa de 75% dos usuários para a categoria a favor do Lula, que passa de quatro para três *tweets* com a remoção. Além disso, o total de usuários de cada classe não sofre alterações. Por outro lado, ocorre grande redução na média, desvio padrão e na quantidade de postagens do usuário que mais posta em cada categoria (coluna *max*) com a remoção de mensagens semelhantes. Como pode ser observado, o impacto da remoção de *tweets* semelhantes ocorre principalmente nos 25% dos usuários que mais postam, o que comprova o comportamento desejado de remoção de *outliers*.

Análise Temporal. Também analisamos a distribuição de *tweets* ao longo do tempo, de modo a complementar a caracterização. Assim, a Figura 3 apresenta: (a) o volume de postagens mensais em cada classe, onde é possível verificar um aumento de postagens

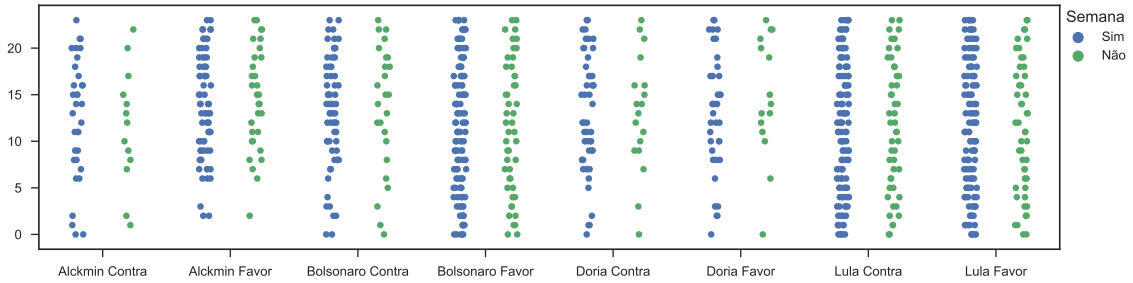


Figura 4. *Tweets* por horários da semana

sobre os presidenciáveis Bolsonaro e Lula nos meses dezembro de 2017 e janeiro de 2018; (b) mais detalhes sobre o período a fim de esclarecer tal comportamento, onde verifica-se um aumento dos *tweets* sobre o tema Lula em datas de eventos específicos. Como exemplo, por volta do dia 12/12/2017 houve a marcação do julgamento do recurso impetrado pela defesa de Lula na segunda instância para o dia 24/01/2018[8], e no dia 30/12/2017 houve a chegada dos militares no RN devido à paralisação de policiais[9].

A fim de refinar o estudo da distribuição de *tweets* no tempo, a Figura 4 apresenta a ocorrência de postagens nos horários do dia (eixo vertical). Para cada presidenciável a imagem apresenta duas colunas, uma para os dias da semana (à esquerda em azul) e outra para o final de semana (à direita em verde). Um comportamento mais intenso e bem distribuído pode ser notado para as categorias favorável e contrário ao Lula, bem como favorável ao Bolsonaro. Por outro lado, nota-se um comportamento mais disperso para as demais categorias. A frequente existência de postagens entre 3:00 e 5:00 para os três posicionamentos mais comentados pode significar que ainda existam postagens automatizadas (*bots*) no *dataset*. Entretanto, assumimos que essas mensagens não têm alto poder de enviesar o classificador, uma vez que pelo processo de remoção de mensagens semelhantes, a frequência de um mesmo *tweet* deve ser baixa. De qualquer modo, o desenvolvimento de ferramentas de detecção automática de *bots* em *tweets* escritos em português é necessária para validar os resultados.

Termos mais Frequentes. A Figura 5 apresenta uma nuvem de termos para cada posicionamento do *dataset*. Um padrão observado é a existência do termo "fora Temer" em todos os posicionamentos contrários, com exceção ao Lula, onde o termo não aparece. Outro padrão é a existência do termo "Minas Gerais" nos posicionamentos favorável e contrário a Lula, mas de modo mais intenso no posicionamento favorável. Também chama a atenção a falta de termos em destaque para o posicionamento a favor de Alckmin e a existência do termo "pais falavam" e "bullingDFuturo" para o posicionamento favorável ao Bolsonaro. Tais termos são utilizados de forma ironizada, demonstrando sarcasmo, sendo que existem inclusive comunidades sobre o assunto nas redes sociais.

[8]https://g1.globo.com/rs/rio-grande-do-sul/noticia/julgamento-de-lula-no-caso-triplex-em-segunda-instancia-e-marcado-para-24-de-janeiro-no-trf4.ghtml

[9]http://www1.folha.uol.com.br/cotidiano/2017/12/1947131-com-pm-parada-exercito-chega-ao-rn-e-forcas-armadas-assumem-a-operacao.shtml

Figura 5. Nuvem de termos

Tabela 6. Teste t pareado

Dataset	Naive Bayes	SVM	Random Forest
Bolsonaro	$0,03156 \pm 0,00365$	$\mathbf{0,99419 \pm 0,00103}$	$0,43404 \pm 0,01210$
Lula	$0,64877 \pm 0,00856$	$\mathbf{0,94185 \pm 0,00225}$	$0,93702 \pm 0,00220$

4.2. Avaliação do Classificador

O modelo criado para classificação dos dados é avaliado na Tabela 6, que apresenta o F-measure médio e o desvio padrão obtidos. O classificador SVM é estatisticamente superior aos demais classificadores para todos os *datasets* analisados, com um nível de significância de 5%. Porém, é importante observar a diferença encontrada para o classificador *Random Forest* nas duas bases de dados. Enquanto o classificador obteve um acerto de 93% para o *dataset* Lula, para o *dataset* Bolsonaro o acerto foi de somente 43%. Da mesma forma o classificador *Naive Bayes* encontrou maior dificuldade para classificar os registros do *dataset* Bolsonaro, conseguindo acertar somente 3%. Uma hipótese para o desempenho observado é o processo de elaboração do conjunto de dados Bolsonaro. Tendo em vista que os dados foram formados por características psico-linguísticas de usuários comentando sobre dois presidenciáveis distintos, o processo de inferência dos dois classificadores foi de alguma forma prejudicado.

4.3. Ameaças à Validade

Alguns pontos precisam ser observados quanto à validade dos resultados encontrados para os classificadores. O primeiro diz respeito às suposições de existência de *bots* nas postagens da rede social. Encontramos diversas evidências da existência de postagens automatizadas, embora este não seja um dos objetivos diretos deste trabalho. Porém, uma investigação mais detalhada sobre sua existência e extensão é necessária, uma vez que a frequente repetição de postagens pode influenciar no classificador. A segunda consideração diz respeito à presença de sarcasmo e ironia. Embora neste trabalho não tenhamos abordado diretamente o assunto, detectamos indícios de sua existência em termos frequentes para a classe a favor do Bolsonaro, o que também pode influenciar nos resultados apresentados. A última consideração refere-se à elaboração do *dataset* Bolsonaro. Reconhecemos que não necessariamente um *tweet* que seja favorável ao presidenciável Lula também demonstre posicionamento ideológico contrário ao presidenciável Bolsonaro. Embora nossos estudos evidenciem a separação entre os usuários defensores de

cada candidato, as mensagens podem conter posicionamentos exclusivamente sobre um assunto, sem indicar necessariamente oposição ao outro.

5. Conclusões e Direções Futuras

Este trabalho aborda a detecção automática de posicionamentos ideológicos em textos escritos em português. Para tal, coletamos *tweets* sobre quatro presidenciáveis para a corrida presidencial brasileira de 2018. Além disso, os dados são caracterizados e possíveis mensagens postadas por *bots* são identificadas e removidas. Utilizamos o viés ideológico das *hashtags* de coleta para construir dois *datasets* etiquetados. Enquanto um contém somente dados sobre o tema Lula, o outro contém dados de dois temas. Especificamente, devido ao baixo volume de mensagens contrárias ao Bolsonaro, utilizamos as mensagens favoráveis ao Lula como sendo contrárias ao Bolsonaro. Assim, os dois *datasets* foram utilizados para treinar dois modelos de classificação usando três algoritmos clássicos de aprendizado de máquina em cada, os quais foram testados com um teste estatístico.

Os resultados obtidos com a caracterização dos dados coletados apontam indícios de *bots* atuando principalmente em postagens dos dois presidenciáveis mais comentados. De fato, são removidas 59,93% das mensagens inicialmente coletadas por serem muito similares entre si e pertencerem a um mesmo usuário e categoria. Além disso, a grande maioria delas trata do tema Lula (57,98%), sendo que do total de cada categoria são removidas 69,55% contra e 62,47% a favor. Por fim, os resultados demonstram que as técnicas de aprendizado de máquinas exploram as diferenças linguísticas existentes entre textos favoráveis e contrários ao alvo definido, possibilitando a detecção de posicionamentos, mesmo no contexto multi polarizado político-eleitoral brasileiro. Neste sentido, conseguimos um F-Measure médio de 99% para o *dataset* Bolsonaro e 94% para o *dataset* Lula, sendo que em ambos o classificador SVM obteve o melhor desempenho.

Como trabalhos futuros, pretendemos utilizar um processo de expansão automática das *hashtags* de coleta através de um processamento de verificação do poder discriminativo das *hashtags* por classe. Essas *hashtags* poderiam retro-alimentar o processo, possibilitando um novo ciclo de coletas e análises. Além disso, pretendemos analisar outros alvos sobre assuntos em outros contextos. Sobre a acurácia dos classificadores, pretendemos utilizar um conjunto de dados aleatório como *baseline* para as comparações nas análises realizadas, o que possibilitaria mensurar o comportamento dos classificadores em relação a textos que não tratam do assunto alvo. Além disso, pretendemos realizar uma avaliação da melhoria dos classificadores com a adição de novas *features*, extraídas com outras técnicas como *bag-of-words* e *Word2Vec*, dentre outras.

Agradecimentos. Trabalho parcialmente financiado por CNPq, CAPES e FAPEMIG.

Referências

Araújo, M., Reis, J., Pereira, A., and Benevenuto, F. (2016). An Evaluation of Machine Translation for Multilingual Sentence-level Sentiment Analysis. In *Proceedings of the ACM Symposium on Applied Computing*, pages 1140–1145.

Bigonha, C., Cardoso, T. N. C., Moro, M. M., Gonçalves, M. A., and Almeida, V. A. F. (2012). Sentiment-based influence detection on twitter. *Journal of the Brazilian Computer Society*, 18(3):169–183.

Caetano, J. A. C., Lima, H. S. L., dos Santos Santos, M. F., and Marques-Neto, H. T. M.-N. (2017). Utilizando análise de sentimentos para definição da homofilia política dos usuários do Twitter durante a eleição presidencial americana de 2016. In *BraSNAM - Brazilian Workshop on Social Network Analysis and Mining*, pages 480–491.

Chen, Y.-C., Liu, Z.-Y., and Kao, H.-Y. (2017). Ikm at semeval-2017 task 8: Convolutional neural networks for stance detection and rumor verification. In *International Workshop on Semantic Evaluation*, pages 465–469.

Christen, P. (2006). A comparison of personal name matching: Techniques and practical issues. In *IEEE International Conference on Data Mining*, pages 290–294.

Dias, M. and Becker, K. (2016a). An Heuristics-Based, Weakly-Supervised Approach for Classification of Stance in Tweets. In *IEEE/WIC/ACM International Conference on Web Intelligence*, pages 73–80.

Dias, M. and Becker, K. (2016b). Detecção semi-supervisionada de posicionamento em tweets baseada em regras de sentimento. In *SBBD - Simpósio Brasileiro de Bancos de Dados*, pages 40–51.

Mohammad, S. M., Sobhani, P., and Kiritchenko, S. (2017). Stance and sentiment in tweets. *TOIT*, 17(3):26.

Mourad, S. S., Shawky, D. M., Fayed, H. A., and Badawi, A. H. (2018). Stance detection in tweets using a majority vote classifier. In *International Conference on Advanced Machine Learning Technologies and Applications*, pages 375–384.

Reis, J. C., Gonçalves, P., Araújo, M., Pereira, A. C., and Benevenuto, F. (2015). Uma abordagem multilıngue para análise de sentimentos. In *BraSNAM - Brazilian Workshop on Social Network Analysis and Mining*.

Shenoy, G. G., Dsouza, E. H., and Kübler, S. (2017). Performing stance detection on Twitter data using computational linguistics techniques. *arXiv preprint arXiv:1703.02019*.

Silva, L. S., do Amaral, D. C., and Moro, M. M. (2017). Uma avaliação de eficiência e eficácia da combinação de técnicas para deduplicação de dados. In *SBBD - Simpósio Brasileiro de Bancos de Dados*, pages 160–171.

Verona, L. V., Oliveira, J. O., and Campos, M. L. M. C. (2017). Métricas para aná lise de poder em redes sociais e sua aplicação nas doaçõ es de campanha para o senado federal brasileiro. In *BraSNAM - Brazilian Workshop on Social Network Analysis and Mining*, pages 544–554.

Witten, I. H., Frank, E., Hall, M. A., and Pal, C. J. (2016). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

Xu, X., Hu, F., Du, P., Wang, J., and Li, L. (2017). Efficient stance detection with latent feature. In *APWeb-WAIM - First International Joint Conference on Web and Big Data*, pages 21–30.

Estudo sobre Métricas para Definir Reputação do Autor de Comentários em Sites de Vendas de Produtos

Carlos Augusto de Sá[1], **Raimundo Santos Moura**[1]

[1] Universidade Federal do Piauí (UFPI)
Departamento de Computação – Teresina, PI – Brasil

{carlos.sa,rsm}@ufpi.edu.br

Abstract. *Knowing the author's reputation for opinionated texts is of utmost importance for evaluating a comment on the Web. This paper presents a study on measures used in the process of evaluating the author's reputation on product sales sites. Two experiments were carried out with neural networks Multilayer Perceptron (MLP) and Radial Basis Function (RBF), and the results show that the MLP gave slightly better performance, but not significantly so. In addition, an experiment was carried out to compare the TOP(X) approach, which is used to infer the best comments, with the new approach that uses MLP in the author's reputation dimension. The results showed that the new approach obtained a gain in the classification of the importance of the comments.*

Resumo. *Conhecer a reputação do autor de textos opinativos é de suma importância para avaliação de comentários na Web. Este artigo apresenta um estudo sobre medidas usadas no processo de avaliação da reputação do autor em sites de vendas de produtos. Realizou-se dois experimentos com as redes neurais Multilayer Perceptron (MLP) e Radial Basis Function (RBF), sendo que a rede MLP obteve melhor desempenho. Comparou-se também a abordagem TOP(X) original, usada para inferir os melhores comentários, com um novo modelo que utiliza rede MLP na dimensão da reputação do autor. Considerando os comentários excelentes e bons, a nova abordagem apresentou resultados significativamente superiores.*

1. Introdução

As Redes Sociais Online (RSO) Twitter e Facebook, e serviços de troca de mensagens pelo celular como o Whatsapp, permitem interação social entre seus os usuários. Além disso, sites de comércio eletrônico (*e-commerce*) permitem aos usuários deixar suas opiniões e comentários sobre o processo de aquisição de um produto ou realização de um serviço. Com o crescimento do volume de informações disponíveis e com o avanço da Computação, a área de Processamento de Linguagem Natural (PLN) ganhou bastante destaque por realizar análises de dados de maneira mais eficiente.

A Análise de Sentimentos ou Mineração de Opinião é uma subárea de PLN, que envolve Ciência da Computação, Linguística e Inteligência Artificial e tem atacado o problema de manipular grandes volumes de dados através de técnicas que analisam a linguagem escrita ou falada [Jackson and Moulinier 2007]. Um desafio da área se encontra na filtragem ou pré-processamento de comentários Web, já que existe uma tendência dos usuários de sistemas online a escrever com muitas gírias, o que dificulta o trabalho das

ferramentas tradicionais de PLN. Outro detalhe a ser considerado é a quantidade de *spam*, textos de baixa qualidade, erros ortográficos, *emoticons*, ”internetês”[1] e informações falsas [Liu 2011]. Além dessas considerações, algumas vezes os comentários apresentam sarcasmos e ironias, que são difíceis de serem captados pelas técnicas atuais de PLN. No entanto, existem esforços no sentido de resolver esses problemas, como os trabalhos de [Hartmann et al. 2014, Carvalho et al. 2009, Gonçalves et al. 2015].

Destaca-se que avaliações positivas a respeito de um produto ou serviço trazem ao novo consumidor mais segurança no processo de compra, porém, avaliações negativas também podem auxiliar na escolha, gerando um impacto positivo nas vendas [Hamilton et al. 2014]. Geralmente, comentários negativos são escritos de forma mais crítica e apresentam boa legibilidade, sendo, algumas vezes, melhores do que comentários positivos. A Figura 1 apresenta um comentário negativo retirado do site da Amazon[2]. Na parte inferior da figura, enfatiza-se um recurso bastante interessante para avaliar um comentário, conhecido como utilidade do review. Com este recurso, o usuário pode, ao terminar de ler, marcar a opção indicando se o comentário foi útil para ele, desta maneira, quanto mais votos ”Sim” um comentário possuir, melhor classificado ele será. Porém, uma desvantagem dessa medida é que comentários recentes e com alta significância ao consumidor, são ignorados por terem poucos votos [Li et al. 2013].

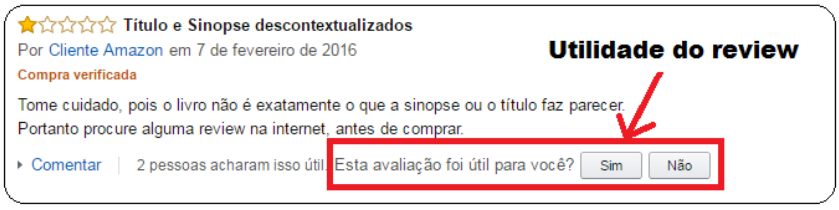


Figura 1. Review negativo no site Amazon

No contexto de reputação em RSOs, os sistemas apresentam diversos recursos para avaliação de um comentário, a saber: quantidade de comentários emitidos por um usuário, número de favoritos, quantidade de estrelas atribuídas, pontuação positiva e negativa, número de seguidores e amigos, entre outros. Um problema que os sites de *e-commerce* apresentam é o fato de possuírem cadastros independentes para os seus usuários e o acesso a esses dados não ser permitido, dificultando a coleta de informações. Desta forma, esta pesquisa sugere que os sites de *e-commerce* possam rever as suas políticas de privacidade no futuro. Uma tentativa de solucionar o entrave é ligar o perfil dos usuários com a suas contas em RSOs populares como Twitter e Facebook.

As RSOs levam vantagem sobre os sites de comentários no que se refere as informações sobre os autores das postagens. O Facebook, por exemplo, permite que os participantes possam curtir um conteúdo ou mesmo comentar. Destaca-se que as empresas tentam explorar ao máximo esse novo tipo *marketing*, o que se confirma com a grande quantidade de perfis nesta rede social.

De maneira geral, com o objetivo de identificar os comentários mais relevantes, em [de Sousa et al. 2015] os autores propuseram uma abordagem para inferir os melho-

[1]Neologismo (palavra: Internet + sufixo: ês) que designa a linguagem utilizada no meio virtual.
[2]http://amazon.com

res comentários sobre produtos ou serviços, denominada TOP(X), que utiliza um Sistema *Fuzzy* com três variáveis de entrada: reputação do autor, número de tuplas <*característica, palavra opinativa*> e riqueza de vocabulário; e uma variável de saída: grau de importância do comentário, representado pela variável "k" (ver Figura 2). No entanto, para definir a reputação do autor, os autores consideraram somente a quantidade de comentários publicados, ou seja, quanto mais comentários emitidos, melhor a reputação do autor. Destaca-se que essa hipótese é fraca e pode ser facilmente refutada pois um *spammer* [3] será considerado um bom autor.

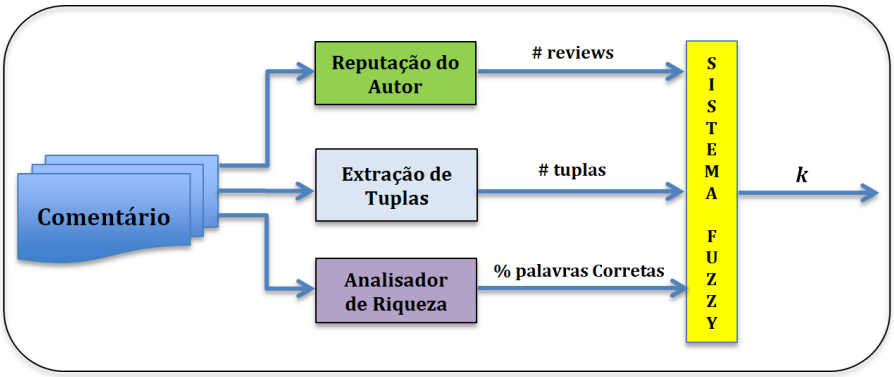


Figura 2. Abordagem TOP(X) proposta por [de Sousa et al. 2015]

Visando melhorar a análise da variável reputação do autor, este trabalho apresenta um estudo utilizando Rede Neural Artificial (RNA) para analisar um conjunto de medidas e definir quais são as mais relevantes no processo de avaliação. Outro aspecto mostrado é uma comparação entre a abordagem TOP(X) original com uma nova abordagem que utiliza RNA MLP na dimensão da reputação do autor. A hipótese é que o uso de uma RNA para inferir a reputação do autor no modelo original melhora o desempenho na classificação da importância dos comentários.

O restante deste trabalho está organizado de seguinte maneira: a Seção 2 apresenta alguns trabalhos relacionados com reputação de autor. A Seção 3 descreve a abordagem proposta para analisar o conjunto de medidas sobre reputação do autor, usando uma RNA. A Seção 4 explora a coleta e a preparação do *Córpus* utilizado na pesquisa. A Seção 5 apresenta e discute os resultados dos experimentos realizados. Por fim, a Seção 6 destaca as principais contribuições e trabalhos futuros.

2. Trabalhos Relacionados

Diversos autores têm investigado sobre avaliação de reputação de autor na Web e nas redes sociais, com destaque para os ambientes *Wiki* e para o *Twitter*. Os ambientes *Wiki* se caracterizam por permitir a colaboração mútua entre os usuários na produção de artigos dos mais variados temas. Um problema inerente desta liberdade é a possibilidade de se ter artigos de baixa qualidade, especialmente pela atuação de vândalos[4]. As principais formas de avaliar a reputação do autor nos ambientes *Wiki* são:

[3]Usuários que postam muitas propagandas sem a permissão dos demais usuários
[4]Usuários que editam os artigos com informações fora do contexto.

- **Histórico das edições**: os autores utilizam o histórico das páginas em busca de padrões de *edits* por parte dos usuários. [Wöhner et al. 2011] destacam que as contribuições persistentes de usuários na Wikipedia duram em média 14 dias sem sofrer modificações. Eles classificam os usuários como autores *vândalos* ou *regulares*. [Halfaker et al. 2009] definem um juiz para classificar um artigo como *aceito* ou *rejeitado* pela comunidade *Wiki* baseado em três características: qualidade dos colaboradores, experiência e no conteúdo postado. [Adler and de Alfaro 2007] indicam que autores *Wiki* ganham reputação quando seus *edits* são preservados por autores subsequentes e perdem reputação quando seus *edits* são desfeitos em um período curto de tempo. Adler e seus amigos definiram, então, o sistema de reputação *WikiTrust*[Adler et al. 2010] baseado em três características: qualidade do *edit*, reputação do autor e reputação do conteúdo.
- **Contexto social**: [Zhao et al. 2010] definiram a *SocialWiki*, um protótipo de sistema *Wiki*, que aproveita o poder das redes sociais para gerenciar automaticamente reputação e confiança para os usuários *Wiki*, baseado no conteúdo que eles contribuem e nas avaliações que eles recebem de outros usuários. Os autores consideram como colaboradores de um artigo, os usuários com interesses em comum, porém eles não descreveram a fórmula para calcular a reputação.
- **Mecanismos de recompensa**: [Hoisl et al. 2007] focaram sobre mecanismos de recompensa social, tais como aceitação, poder e *status*, para ranquear autores que mais colaboram com boas contribuições. Os autores concluíram que a abordagem de recompensas baseada em motivação pode produzir artigos de alta qualidade.

É importante destacar que a contribuição dos trabalhos que exploram o ambiente *Wiki* está na persistência das colaborações, ou seja, *quanto mais tempo um edit persistir, melhor a reputação do autor*.

Com relação a rede social *Twitter*, destacam-se os trabalhos para identificar os usuários mais influentes, usuários suspeitos e *spammers*. [Kwak et al. 2010], utilizam os dados coletados nos "*trending topics*" (assuntos do momento, em tradução livre) para criar *ranking* dos usuários de acordo com o número de seguidores e o algoritmo de *PageRank*. Eles notaram que esses dois *rankings* são similares. Os autores criaram um terceiro *ranking* baseado nos *retweets*, que é o processo de propagar na rede o *tweet* de outro usuário. Eles concluíram que um *retweet* possui alcance de, no mínimo, 1000 usuários, devido a forma de propagação instantânea e que mais de 85% dos tópicos classificados se referem a manchetes de provedores de conteúdo.

[Weitzel et al. 2014] definiram medidas baseadas nos *retweets* para calcular a reputação dentro do *Twitter*, abordando informações no domínio da medicina. Os autores concluíram que a maioria dos perfis no *Twitter* são individuais ou de *blogs* e que a aplicação das medidas baseadas em *retweets* conseguem identificar os usuários mais populares dentro da rede.

[Weng et al. 2010] propuseram a medida *TwitterRank*, baseada no número de seguidores e seguidos do usuário. De acordo com a abordagem dos autores, dados três usuários A, B e C, sendo que C segue A e B; se A e B publicam, respectivamente, 500 e 1.000 *tweets* sobre um dado tópico, então, a influência que B exerce sobre C é duas vezes maior que a influência de A. Ainda sobre medidas de ranqueamento, [Cappelletti and Sastry 2012] desenvolveram o algoritmo *IARank*, que observa o poten-

cial que um usuário possui de ampliar uma informação dentro do *Twitter*. Eles consideram dois fatores de entrada no algoritmo: a tendência de um usuário ser retuitado ou mencionado e o tamanho da audiência desses retuites ou menções. Destaca-se que essas duas medidas não substituem o algoritmo *PageRank*, utilizado pelo *Twitter*.

No trabalho desenvolvido por [Aggarwal and Kumaraguru 2014], os autores identificaram um ”mercado negro” que vende/compra contas fraudulentas, curtidas no *Facebook* e até mesmo seguidores no *Twitter* para, artificialmente, melhorarem a reputação social dos usuários. Os autores relatam uma precisão de 88,2% no mecanismo de aprendizagem de máquina supervisionado usado para predizer seguidores suspeitos.

No que se refere a detecção de *spammers*, [Wang 2010] definiu reputação do autor como sendo uma relação entre o número de amigos e o número de seguidores. Os resultados obtidos demonstram que o sistema de Wang consegue detectar comportamentos anormais de usuários.

Por fim, no contexto dos *sites* de *e-commerce*, existem soluções que criam *rankings* e filtros dos comentários sobre os produtos para auxiliar os consumidores no momento da compra. Os *rankings* podem ser ordenados por data ou número de estrelas. Adicionalmente, podem existir filtros para listar apenas os comentários positivos, negativos, de compradores verificados ou de produtos de uma determinada característica, por exemplo, produto da cor azul.

Este artigo investiga se a reputação do autor de comentários Web pode ser calculada a partir de seis medidas, abordadas com mais detalhes na Seção 3.

3. Abordagem Proposta

A abordagem proposta neste trabalho visa analisar um conjunto de medidas para definir a reputação do autor de comentários em sites de vendas de produtos. De forma geral, a proposta representa uma adaptação da abordagem Top(X) original, com ênfase na dimensão reputação do autor. O estudo foi conduzido através da aplicação de redes neurais artificiais para inferir a reputação dos autores dos comentários e descobrir a importância de cada medida da entrada.

A Figura 3 mostra a visão geral da abordagem proposta, considerando apenas a dimensão reputação do autor. Na figura, a variável ’a’ representa a saída da RNA e indica a reputação do autor normalizada para o intervalo de 0 a 10.

Considerando sites de *e-commerce*, definiu-se seis medidas para avaliar a reputação do autor dos comentários de produtos. Tais medidas foram extraídas levando em conta informações disponíveis em sites de lojas virtuais e comparadores de preços. Nossa pesquisa examinou também outras medidas utilizadas em ambientes Wiki, RSOs, Fóruns e Blogs. O estudo realizado sugere, então, que os sites de *e-commerce* devem expandir suas funcionalidades no sentido de reforçar a importância dos autores e seus relacionamentos. Uma maneira de realizar essa expansão é permitir a integração dos perfis de usuários dos sites com os seus respectivos perfis em redes sociais.

- **DataReview**: a data de escrita do comentário, convertida para dias em comparação com a data inicial de coleta do *Córpus*. Esta informação é importante pois quanto mais recente, mais atualizado o comentário e, hipoteticamente,

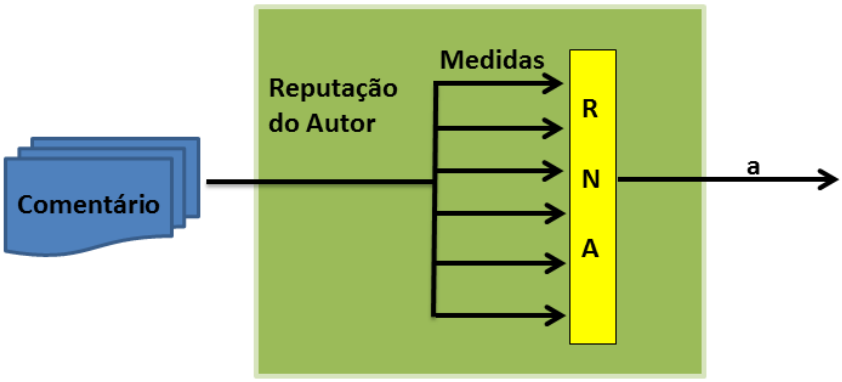


Figura 3. Abordagem proposta

deve ser melhor avaliado. No entanto, os comentários que são muito recentes podem ser prejudicados no processo de avaliação geral por não ter tempo hábil para leitura pelos consumidores;

- **DataCadastro**: a data em que o autor fez o seu cadastro no site, convertida para dias em comparação com a data inicial da coleta do *Córpus*. Esta informação é importante pois imagina-se que a reputação dos autores experientes seja melhor do que a autores novatos;
- **VotosPositivos**: quantidade de votos positivos atribuídos por outros usuários. A hipótese é que quanto mais votos positivos um autor receber de outros usuários, melhor será a sua reputação;
- **VotosNegativos**: quantidade de votos negativos atribuídos por outros usuários. A importância dos votos negativos é inversamente proporcional aos votos positivos, pois quanto mais votos negativos o autor receber em seus comentários, pior a sua reputação;
- **TotalVotos**: soma dos votos recebidos pelo comentário. De forma geral, imagina-se que quanto mais votos o usuário tenha em seus comentários, sejam positivos ou negativos, melhor a sua reputação pois o mesmo está sendo observado;
- **TotalReviewsAutor**: quantidade de comentários que o autor realizou no site. Esta informação é relevante pois indica a participação ativa do usuário dentro do ambiente.

4. Córpus: coleta e preparação[5]

Para avaliação da abordagem proposta, foi criado um *Córpus* com 2.433 comentários do site do Buscapé[6]. A decisão de trabalhar com comentários do site do Buscapé deu-se por três razões principais: i) ser o maior site comparador de preços da América Latina; ii) necessidade posterior de comparar o modelo proposto com a abordagem Top(x), que utiliza comentários sobre *smartphones* do referido site; e iii) os comentários serem disponíveis publicamente para coleta com rastreadores Web.

Os dados foram coletados no ano de 2016 e são referentes a comentários escritos em português sobre diversos *smartphones*. Após a exclusão de comentários duplicados e

[5]Córpus disponível em https://goo.gl/g5nrwJ
[6]http://www.buscape.com.br

vazios, definiu-se uma amostra de 2.000 comentários, sendo 1.000 de orientação positiva e 1.000 negativa. Destaca-se que no site do Buscapé, a orientação do comentário é definida pelo próprio autor e, são apresentados aos usuários em guias/abas separadas. No entanto, em uma análise mais detalhada, verificou-se que alguns comentários marcados como positivos eram, na verdade, negativos e vice-versa. Além disso, muitos comentários são considerados neutros.

Para solucionar esse problema, decidiu-se fazer uma análise manual do *Córpus* quanto à orientação semântica. Ao final do processo, o *Córpus* anotado ficou com 923 comentários positivos, 602 comentários negativos, 141 comentários neutros e 334 considerados ”lixo”, que são comentários de usuários que declaram ”não possuir o produto” e comentários totalmente sem sentido. Esses comentários foram desconsiderados nas nossas avaliações.

Em um segundo momento, criou-se um *Subcórpus* anotado com a reputação do autor, analisando uma amostra de 323 comentários (nível de confiança de 95% e margem de erro de 5%). Adicionou-se 33 comentários ao *Subcórpus* (10% da amostra), totalizando 356, sendo 132 positivos, 131 negativos e 93 neutros. A anotação foi realizada por três alunos da pós-graduação em Ciência da Computação da UFPI, considerando informações referentes ao autor, como a quantidade de votos positivos em seus comentários, quantidade de votos negativos, total de votos, entre outras medidas. Em seguida, aplicou-se uma nota de 0 a 10 para cada um dos autores dos comentários dentro da amostra definida, sendo guiado unicamente pelas variáveis de entrada propostas.

A Tabela 1 mostra o resultado da avaliação dos seres humanos para a reputação dos autores dos comentários. As 11 notas atribuídas aos autores foram generalizadas para o universo completo dos comentários através de uma RNA, como será descrito na próxima seção. Esta generalização se dá pela rede neural que infere, a partir das medidas de entrada, qual a reputação do autor para qualquer comentário dentro do *Córpus*.

Tabela 1. Resultado da anotação do *Subcórpus* por seres humanos

Reputação	#Total	Reputação	#Total	Reputação	#Total	Reputação	#Total
0	68	3	23	6	9	9	5
1	163	4	11	7	5	10	5
2	50	5	16	8	1	-	

É importante mencionar que a anotação manual de *Córpus* deu-se devido à ausência de recursos linguísticos para a língua portuguesa disponíveis publicamente. Sabe-se que o processo de anotação manual pode causar um viés no modelo proposto e, possivelmente, comprometer a viabilidade da solução. Porém, os riscos da anotação foram minimizados com o envolvimento de especialistas da área de linguística computacional.

5. Experimentos

Para avaliar o modelo de RNA proposto, realizou-se dois experimentos com as arquiteturas *Multilayer Perceptron* (MLP) e *Radial Basis Functions* (RBF), usando a ferramenta de análises estatísticas SPSS sobre o *Subcórpus* anotado. É importante mencionar que

nos dois experimentos foram usados 6 neurônios na camada de entrada da rede, correspondentes às medidas discutidas na seção anterior. Para comparar a abordagem TOP(X) original com a nova abordagem que usa RNA na dimensão da reputação do autor, fez-se um terceiro experimento discutido a seguir.

5.1. Experimento 1: RNA MLP

No primeiro experimento usamos uma RNA MLP e o melhor ajuste se deu com 8 neurônios na camada escondida e a função de ativação Tangente Hiperbólica. Na camada de saída utilizamos o atributo de supervisão ”ReputacaoManual” como variável dependente para testar a rede, classificando as 11 notas possíveis dos autores (0 a 10) e a função de ativação Softmax. É importante relatar que o ajuste é realizado pela ferramenta SPSS, através de diversos testes internos para atingir a melhor configuração possível para a performance geral da rede.

A Figura 4 mostra a disposição dos neurônios em cada camada na melhor topologia escolhida pela execução da RNA MLP. O processo de treinamento e teste foi aplicado sobre o *Subcórpus* utilizando o método de validação cruzada *10-fold cross validation*.

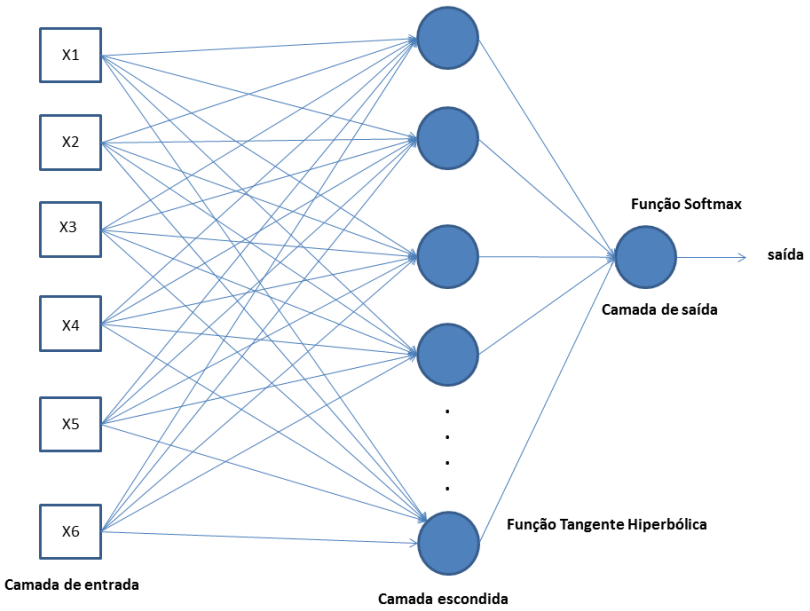


Figura 4. Topologia da RNA MLP

Em redes neurais, o valor da importância de uma variável de entrada é calculado levando em consideração o peso das conexões dos neurônios entre as camadas da rede. Já a importância normalizada é simplesmente os valores de importância divididos pelos maiores valores de importância e expressos em porcentagens.

A Tabela 2 apresenta a importância de cada variável de entrada na rede MLP. É possível observar que a variável mais importante para avaliar a reputação do autor foi ”VotosPositivos”, seguida de ”TotalVotos” e ”VotosNegativos”. Tal resultado se mostra interessante porque a variável ”VotosPositivos” é definida por outros usuários da rede, confirmando a boa reputação do autor daquele comentário que recebeu o voto positivo.

Por outro lado, a variável menos importante foi ”DataCadastro” que indica o tempo, em dias, do cadastro do usuário no site.

Tabela 2. Importância das variáveis de entrada na RNA MLP

Variável de Entrada	Importância	Importância Normalizada
DataReview	0,106	42,10%
DataCadastro	0,086	34,40%
VotosPositivos	0,252	100,0 %
VotosNegativos	0,205	81,70%
TotalVotos	0,245	97,60%
TotalReviewsAutor	0,105	41,80%

Com relação a precisão de inferência da RNA MLP dentro do conjunto de treinamento e teste, atingiu-se um valor de 62,08% no processo de classificação para os valores numéricos de 0 a 10. No entanto, considerando as faixas de valores: 0-3 para **baixo**, 4-7 para **médio** e 8-10 para **alto**, que são normalmente usados em sistemas de reputação, a precisão da rede atingiu o valor de 91,01%.

5.2. Experimento 2: RNA RBF

Visando apresentar uma alternativa para a arquitetura MLP, executou-se o segundo experimento com uma RNA com funções de base radial. O melhor ajuste se deu com 11 neurônios na camada escondida e a função de ativação Softmax. Na camada de saída utilizou-se também o atributo de supervisão ”ReputacaoManual” como variável dependente para testar a rede, classificando as 11 notas possíveis dos autores (0 a 10) e a função de ativação Identidade. Assim como na rede MLP, ressalta-se que o ajuste da rede RDF foi realizado usando a ferramenta SPSS. A topologia com a disposição dos neurônios é semelhante a topologia da rede MLP, com a diferença apenas da quantidade de neurônios na camada escondida e das funções de ativação usadas nas camadas escondidas e de saída. O processo de treinamento e teste utilizou o mesmo *Subcórpus* e o mesmo método de validação *10-fold cross validation*.

A Tabela 3 apresenta a importância de cada variável de entrada usando a rede RBF. Observa-se que os resultados são similares aos apresentados no experimento com a RNA MLP, porém existe uma mudança na ordem das variáveis mais importantes, sendo ”TotalVotos” a mais importante, seguida de ”VotosPositivos” e ”VotosNegativos”. Por outro lado, a variável menos importante foi ”DataCadastro”, assim como na rede MLP.

Tabela 3. Importância das variáveis de entrada na RNA RBF

Variável de Entrada	Importância	Importância Normalizada
DataReview	0,127	60,30%
DataCadastro	0,111	52,80%
VotosPositivos	0,210	99,60 %
VotosNegativos	0,210	99,50%
TotalVotos	0,211	100,0%
TotalReviewsAutor	0,132	62,60%

Com relação a precisão de inferência da rede dentro do conjunto de treinamento e teste, atingiu-se um valor de 52,25% no processo de classificação para os valores numéricos de 0 a 10 e 87,36% para as três faixas de valores baixo, médio e alto. Desta forma, devido a vantagem da rede MLP sobre a rede RBF, decidiu-se usar apenas a primeira arquitetura nos experimentos de comparação entre a abordagem TOP(X) original e a nova abordagem que utiliza uma RNA na dimensão da reputação do autor.

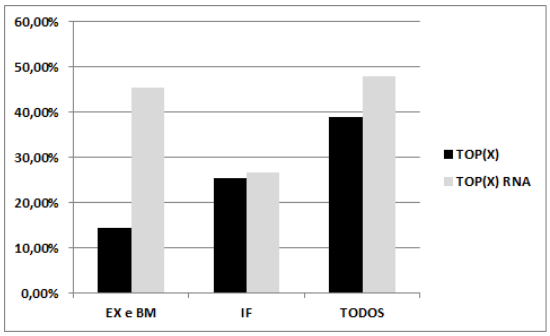
5.3. Experimento 3: Comparação entre as abordagens

Neste experimento 271 comentários foram selecionados aleatoriamente, sendo 100 positivos, 100 negativos e 71 neutros. Com relação a importância dos comentários, a amostra utilizada possui a seguinte anotação: 17 excelentes (EX), 24 bons (BM), 145 suficientes (SF) e 85 insuficientes (IF), conforme classificação proposta em [de Sousa et al. 2015].

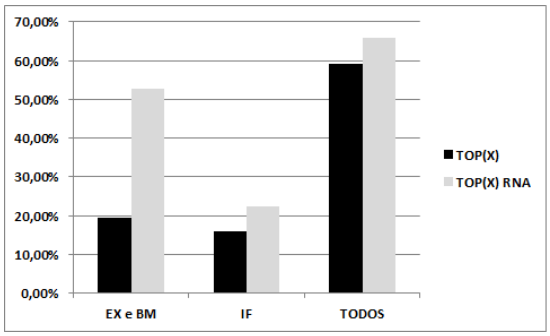
Para avaliar as abordagens, calculou-se as medidas de Precisão, Cobertura e a medida harmônica Medida-F para cada classe. Essas medidas são normalmente usadas em avaliação de abordagens na área de aprendizagem de máquina [Powers 2011].

A Figura 5(a) apresenta graficamente a comparação baseada na Medida-F entre as duas abordagens, relacionando os comentários positivos em termos de sua importância. É possível observar que abordagem TOP(X) com RNA supera com boa margem a abordagem TOP(X) original nos comentários excelentes e bons. Estes comentários são relevantes pois, normalmente, o usuário procura os melhores comentários para ler e decidir sobre a compra de um produto ou serviço. Desta forma, o usuário poderá focar em um pequeno grupo de comentários selecionados pela abordagem, gerando um ganho de tempo e esforço na pesquisa pelo produto que deseja adquirir.

A Figura 5(b) apresenta o gráfico comparativo com relação aos comentários negativos e sua respectiva importância, também baseada na Medida-F. Novamente é possível observar que abordagem TOP(X) com RNA também supera com boa margem a abordagem original nos comentários excelentes e bons.



(a) Comentários positivos



(b) Comentários negativos

6. Conclusão e Trabalhos Futuros

Neste artigo foi apresentado um estudo sobre métricas para definir a reputação do autor de comentários em sites de comparação de preços de produtos. De forma geral, o modelo proposto representa uma adaptação da abordagem TOP(X) [de Sousa et al. 2015], com ênfase na dimensão reputação do autor. O estudo foi conduzido através da aplicação

de redes neurais para inferir a reputação dos autores dos comentários e descobrir a importância de cada medida de entrada.

Realizou-se dois experimentos com a aplicação de RNAs MLP e RBF sobre um *Subcórpus* para realizar o treinamento da rede. Os resultados obtidos apresentaram similaridades entre as redes quanto a indicação da importância das variáveis de entrada. Observou-se que a quantidade de votos positivos que um autor recebe tem um peso significativo em sua reputação, sendo considerada a principal medida para avaliar a reputação do autor no contexto analisado. Com relação ao desempenho dos modelos, a rede RBF atingiu 87,36% de precisão, enquanto que a rede MLP atingiu 91,01%. Portanto, a rede MLP foi escolhida para a evolução de nossas pesquisas.

Em um terceiro experimento comparou-se as abordagens TOP(X) original e TOP(X) com RNA MLP, utilizando como base a média harmônica Medida-F. Com foco nos comentários excelente e bons, a nova abordagem apresentou resultados significantemente superiores. Conclui-se, então, que tal abordagem pode auxiliar os usuários na busca por produtos ou serviços, reduzindo o tempo e esforço gastos no processo. No entanto, considera-se ainda abaixo do esperado, pois os resultados na classificação ficaram em torno de 50%.

Como trabalhos futuros, pretende-se: i) aplicar a nova abordagem em um *Córpus* maior, realizando um processo mais extenso de anotação manual; e ii) investigar o impacto de reputação do autor em notícias falsas (*fake news*). Sabe-se que existem vários artifícios utilizados para potencializar o alcance de uma notícia ou comentário, bem como impulsionar a reputação de um autor.

Referências

[Adler and de Alfaro 2007] Adler, B. T. and de Alfaro, L. (2007). A content-driven reputation system for the wikipedia. In *Proc. of the Int. Conference on World Wide Web*, pages 261–270. ACM.

[Adler et al. 2010] Adler, B. T., de Alfaro, L., and Pye, I. (2010). Detecting wikipedia vandalism using wikitrust.

[Aggarwal and Kumaraguru 2014] Aggarwal, A. and Kumaraguru, P. (2014). Followers or phantoms? an anatomy of purchased twitter followers. *CoRR*.

[Cappelletti and Sastry 2012] Cappelletti, R. and Sastry, N. (2012). IARank: Ranking users on twitter in near real-time, based on their information amplification potential. In *SocialInformatics*, pages 70–77.

[Carvalho et al. 2009] Carvalho, P., Sarmento, L., Silva, M. J., and de Oliveira, E. (2009). Clues for detecting irony in user-generated contents: Oh...!! it's "so easy";-). In *Proc. of the Int. Workshop on Topic-sentiment Analysis for Mass Opinion*, pages 53–56.

[de Sousa et al. 2015] de Sousa, R. F., Rabelo, R. A. L., and Moura, R. S. (2015). A fuzzy system-based approach to estimate the importance of online customer reviews. In *International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–8.

[Gonçalves et al. 2015] Gonçalves, P., Dalip, D., Reis, J., Messias, J., Ribeiro, F., Melo, P., Araújo, L., Gonçalves, M., and Benevenuto, F. (2015). Bazinga! caracterizando e

detectando sarcasmo e ironia no twitter. In *Proc. of the Brazilian Workshop on Social Network Analysis and Mining (BraSNAM)*.

[Halfaker et al. 2009] Halfaker, A., Kittur, A., Kraut, R., and Riedl, J. (2009). A jury of your peers: Quality, experience and ownership in wikipedia. In *Proc. of the Int. Symposium on Wikis and Open Collaboration*, pages 15:1–15:10. ACM.

[Hamilton et al. 2014] Hamilton, R., Vohs, K. D., and McGill, A. L. (2014). We'll be honest, this won't be the best article you'll ever read: The use of dispreferred markers in word-of-mouth communication. *Journal of Consumer Research*, 41(1):197 – 212.

[Hartmann et al. 2014] Hartmann, N., Avanço, L., Balage, P., Duran, M., Nunes, M. D. G. V., Pardo, T., and Aluísio, S. (2014). A large corpus of product reviews in portuguese: Tackling out-of-vocabulary words. In *Proc. of the Int. Conference on Language Resources and Evaluation (LREC'14)*.

[Hoisl et al. 2007] Hoisl, B., Aigner, W., and Miksch, S. (2007). *Online Communities and Social Computing: Second International Conference, OCSC 2007*, pages 362–371. Springer Berlin Heidelberg.

[Jackson and Moulinier 2007] Jackson, P. and Moulinier, I. (2007). *Natural language processing for online applications: Text retrieval, extraction and categorization*. John Benjamins, Amsterdam.

[Kwak et al. 2010] Kwak, H., Lee, C., Park, H., and Moon, S. (2010). What is twitter, a social network or a news media? In *Proc. of the Int. Conference on World Wide Web*, pages 591–600. ACM.

[Li et al. 2013] Li, M., Huang, L., Tan, C., and Wei, K. (2013). Helpfulness of online product reviews as seen by consumers: Source and content features. *Int. J. Electronic Commerce*, 17(4):101–136.

[Liu 2011] Liu, B. (2011). *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*. Springer-Verlag New York, Inc., Secaucus, NJ, USA.

[Powers 2011] Powers, D. M. W. (2011). Evaluation: From precision, recall and f-measure to roc., informedness, markedness & correlation. *Journal of Machine Learning Technologies*, 2(1):37–63.

[Wang 2010] Wang, A. H. (2010). Don't follow me: Spam detection in twitter. In *Proc. of the Int. Conference on Security and Cryptography (SECRYPT)*, pages 1–10.

[Weitzel et al. 2014] Weitzel, L., de Oliveira, J. P. M., and Quaresma, P. (2014). Measuring the reputation in user-generated-content systems based on health information. *Procedia Computer Science*, 29:364 – 378.

[Weng et al. 2010] Weng, J., Lim, E.-P., Jiang, J., and He, Q. (2010). Twitterrank: Finding topic-sensitive influential twitterers. In *Proc. of the Int. Conference on Web Search and Data Mining*, pages 261–270. ACM.

[Wöhner et al. 2011] Wöhner, T., Köhler, S., and Peters, R. (2011). Automatic reputation assessment in wikipedia. In *Proc. of the Int. Conference on Information Systems*.

[Zhao et al. 2010] Zhao, H., Ye, S., Bhattacharyya, P., Rowe, J., Gribble, K., and Wu, S. F. (2010). Socialwiki: Bring order to wiki systems with social context. In *Social Informatics - Second International Conference, SocInfo*, pages 232–247.

Interdisciplinaridade e Teoria de Redes: rede semântica de cliques baseada em ementas

Júlia Carvalho Andrade[1], Renata Souza Freitas Dantas Barreto[2], Núbia Moura Ribeiro[1], Hernane Borges de Barros Pereira[1,2]

[1]Doutorado Multi-Institucional e Multidisciplinar em Difusão do Conhecimento – Universidade Federal da Bahia (UFBA) – Salvador – BA – Brasil

[2]Programa de Pós-Graduação em Modelagem Computacional e Tecnologia Industrial (SENAI-CIMATEC) – Salvador – BA – Brasil

juliacarvalhoandrade@yahoo.com.br,{nubiamouraribeiro, renatasouzabarreto, hernanebbpereira}@gmail.com

Abstract. *This work proposes to investigate the semantic networks of cliques based to the courses descriptions of a doctoral postgraduate program. A total of 46 courses descriptions were compiled. Using some indexes of social and complex networks as a starting point, it was observed that the centralities of degree, closeness and betweeness are adequate indexes to perceive the coherence and consistency of the proposal of a program with its set of courses descriptions. And the identification of the common topics in the courses descriptions helps in the perception of the interactions between the courses of a program.*

Resumo. *Este trabalho propõe investigar a rede semântica de cliques baseada em ementas de um programa de pós-graduação de doutorado. Foram coletadas 46 ementas de componentes curriculares. Usando alguns índices de redes sociais e complexas como ponto de partida, observou-se que as centralidades de grau, proximidade e intermediação são índices adequados para perceber a coerência e consistência da proposta de um programa com seu ementário. E a identificação dos temas em comum às ementas auxilia na percepção das interações entre os componentes curriculares de um programa.*

1. Introdução

Os componentes curriculares da matriz curricular de um programa de pós-graduação (PPG) têm como objetivo proporcionar os conhecimentos necessários ao aprendizado proposto pelo PPG. Cada componente curricular possui uma ementa, que consiste em uma descrição discursiva e resumida do seu conteúdo teórico (conceitual) ou teórico-metodológico (conceitual/procedimental). Neste sentido, o ementário de um PPG pode ser considerado um discurso escrito, o qual deve refletir sua proposta para o aprendizado. Ademais, cada ementa pode ser considerada uma sentença do discurso e o universo de palavras das ementas representa o vocabulário do discurso do PPG.

No Brasil, os programas de pós-graduação são avaliados e reconhecidos pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), e classificados em áreas de concentração, dentro de áreas de avaliação. Segundo a

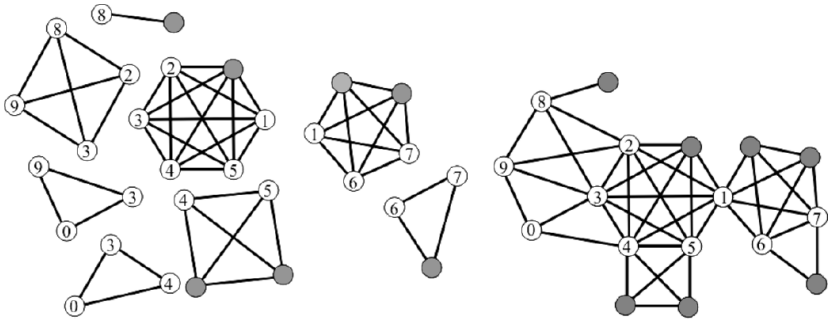
CAPES (2013), a estrutura curricular de um PPG deve ser adequada à formação de mestres e doutores, e constituída por um conjunto de componentes curriculares congruente com as áreas de concentração e as linhas de pesquisa. Desta forma, os textos descritivos das áreas de concentração e linhas de pesquisa de um PPG devem refletir sua proposta de aprendizado e espera-se coerência e consistência com seu ementário.

Para avaliar a coerência e a consistência entre a proposta de aprendizado baseada nas áreas de concentração e nas linhas de pesquisa de um PPG e o ementário relativo à sua matriz curricular, este estudo propõe a modelagem do ementário de um PPG em nível de doutorado, por meio de uma rede semântica de cliques baseada em ementas e o uso de temas-chave (i.e. conceitos representativos da proposta de aprendizado do PPG). Vale salientar que redes semânticas de cliques baseadas em ementas ainda não foram descritas na literatura científica até o presente momento.

2. Trabalhos Correlatos

Rede semântica é um sistema de representação do conhecimento de discursos. As redes semânticas de cliques é um tipo de rede semântica em que todas as palavras de cada sentença do discurso (e.g. texto) se conectam formando um subgrafo, chamado de clique. Assim, uma clique é um subgrafo completo de um grafo *G*. E as cliques se conectam formando a rede semântica do discurso [Fadigas e Pereira 2013].

As redes de cliques são formadas por justaposição e/ou sobreposição de cliques. O processo de justaposição significa conectar duas cliques com apenas um vértice em comum. E quando dois ou mais vértices conectam as cliques ocorre o processo de sobreposição [Fadigas e Pereira 2013], conforme ilustrado na Figura 1.



a) Estado inicial de cliques isoladas.

b) Uma "rede de cliques" conectada por justaposição e sobreposição.

Figura 1. Estado inicial de cliques isoladas e uma possível "rede de cliques". Fonte: Modificada de Fadigas e Pereira (2013, p. 2577).

Diversos trabalhos investigaram discursos por meio de redes semânticas de cliques. Caldeira et al. (2006) utilizaram redes de palavras baseadas em cliques para analisar a estrutura de conceitos significativos em textos (i.e. discursos) escritos. Redes semânticas baseadas em títulos de artigos de periódicos científicos (RST) foram estudadas por: Fadigas et al. (2009) que analisaram RST de divulgação em educação matemática; Pereira et al. (2011) estudaram a estrutura topológica de RST como um método de análise da eficiência da difusão da informação; e Henrique et al. (2014) utilizaram RST para comparar os títulos de artigos de periódicos de divulgação em

educação matemática, em inglês e português. Já Teixeira et al. (2010), Lopes et al. (2015) e Lima-Neto et al. (2018) utilizaram redes semânticas de cliques para analisar a relação entre as palavras que emergem em discursos orais.

Ademais, Fadigas e Pereira (2013) propuseram um conjunto de índices para capturar as propriedades de redes semânticas de cliques e um método para caracterizar o fenômeno mundo pequeno nestas redes, utilizando como fonte de dados RST. E Grilo et al. (2017) investigaram a robustez de redes semânticas de cliques.

3. Metodologia

O fluxograma do processo metodológico é apresentado na Figura 2.

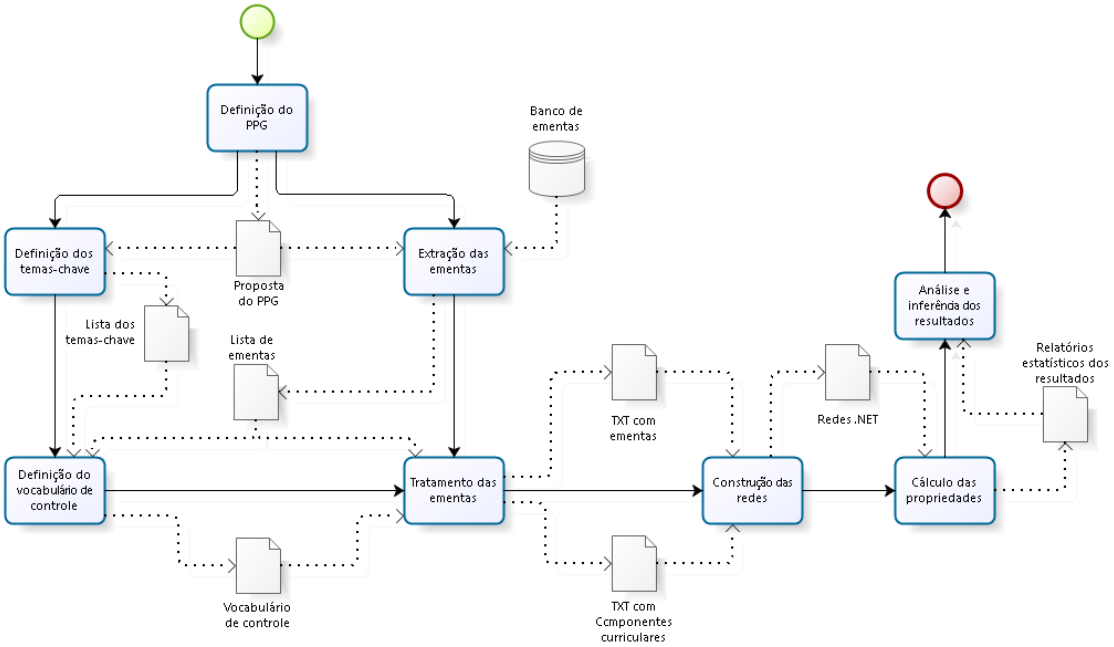


Figura 2: Fluxograma do processo metodológico. Fonte: Autores.

Para modelagem da rede de ementas foi definido como cenário um PPG, em nível de doutorado, da área de avaliação interdisciplinar da CAPES. O *corpus* de análise da sua proposta é constituído pelo texto descritivo da sua área de concentração – "Modelagem da Geração e Difusão do Conhecimento" e pelos títulos das três linhas de pesquisa – 1. Construção do Conhecimento: Cognição, Linguagens e Informação / 2. Difusão do Conhecimento: Informação, Comunicação e Gestão / 3. Cultura e Conhecimento: Transversalidade, Interseccionalidade e (In)formação. Estas informações foram obtidas na Plataforma Sucupira. E o ementário com 46 ementas foi disponibilizado pelo colegiado do PPG estudado, que autorizou a pesquisa por meio da assinatura de Carta de Anuência. Vale salientar que este trabalho foi apreciado e aprovado pelo Comitê de Ética em Pesquisa da Escola de Nutrição da Universidade Federal da Bahia (Número do parecer: 2.487.366).

Os processos "Definição dos temas-chave" e "Definição do vocabulário de controle" (Figura 2) foram realizados de forma manual, seguindo as etapas: (1) Pré-análise do *corpus* a partir da leitura para familiarização com o texto descritivo da área de concentração, títulos das linhas de pesquisa e ementas; (2) Identificação dos temas-

chave no texto descritivo da área de concentração e nos títulos das linhas de pesquisa que representassem o conteúdo teórico ou teórico-metodológico da proposta de aprendizagem do PPG investigado; (3) Identificação de palavras relacionadas semanticamente aos temas-chave nas ementas; (4) Quantificação da frequência de aparição das palavras relacionadas aos temas-chave nas ementas a fim de selecionar aquelas de maior ocorrência, que foram escolhidas para representar os temas-chave.

Após essas etapas, foram identificados 15 temas-chave e palavras relacionadas que compõem o vocabulário de controle usado no tratamento das ementas (Tabela 1). Cabe ressaltar que os temas-chave "conhecimento", "construção do conhecimento", "difusão do conhecimento" e "geração do conhecimento", apesar de estarem estreitamente relacionados entre si, possuem significados distintos.

Tabela 1: Vocabulário de controle. Entre parêntese, a frequência de aparição das palavras relacionadas aos temas-chave nas ementas. E em negrito, as palavras de maior ocorrência por tema-chave.

TEMAS-CHAVE (T)		PALAVRAS RELACIONADAS
T1	Metodologia	**Método** (7), metodologias (2), aspectos metodológicos (1), metodologias participativas (1), recorte teórico-metodológico (1)
T2	Cognição	**Análise cognitiva** (2), ciências cognitivas (2), analista cognitivo (1), cognição (1), cognólogo (1), processos cognitivos (1)
T3	Complexidade	**Complexidade** (5), redes complexas (1), sistemas complexos (1)
T4	Comunicação	**Comunicação** (2), canais de comunicação (1), cenário comunicativo (1), comunicação não verbal (1), relações comunicativas (1)
T5	Ciência	**Ciência** (7), desenvolvimento científico (1), difusão da ciência (1), disseminação da ciência (1), divulgação da ciência (1), invenção (1), científica (1), pesquisa científica (1), popularização da ciência (1), produção científica (1)
T6	Tecnologia	**Tecnologia** (4), desenvolvimento tecnológico (1), dimensão tecnológica (1), inovação tecnológica (1), tecnologias da comunicação (1), tecnologias da informação (2), transferência de tecnologia (1)
T7	Conhecimento	**Conhecimento** (11), gestão do conhecimento (3), padronização do conhecimento (2), bases de conhecimento (1), campos de conhecimento (1), controle do conhecimento (1), disseminação do conhecimento (1), divulgação do conhecimento (1), popularização do conhecimento (1)
T8	Construção do conhecimento	**Construção** do conhecimento (3)
T9	Cultura	**Cultura** (4), cibercultura (2), abordagens culturais (1), antropologia cultural (1), contextos culturais (1), cultura brasileira (1), culturalismo (1), diversidade cultural (1), indústria cultural (1), matrizes culturais (1), pluralismo cultural (1), politicas culturais (1), produção cultural (1)
T10	Difusão do conhecimento	**Difusão do conhecimento** (11)
T11	Epistemologia	**Epistemologia** (1), bases epistemológicas (1), correntes epistemológicas (1), escolhas epistemológicas (1)
T12	Geração do conhecimento	**Geração do conhecimento** (2)
T13	Informação	**Informação** (3), disseminação da informação (1)
T14	Linguagem	**Linguagem** (5)
T15	Modelagem	**Modelo** (9), modelagem (3), modelagem computacional (2)

Na construção de uma rede semântica de cliques baseada em ementas, cada ementa é considerada uma sentença de um discurso escrito, que, por sua vez, forma uma clique. As palavras de cada ementa são os vértices de uma clique e as arestas são as conexões entre as palavras que aparecem na mesma ementa. As palavras em comum às ementas são os vértices em comum às cliques, que fazem a justaposição e sobreposição das cliques, formando, assim, a rede de ementas (Figura 3).

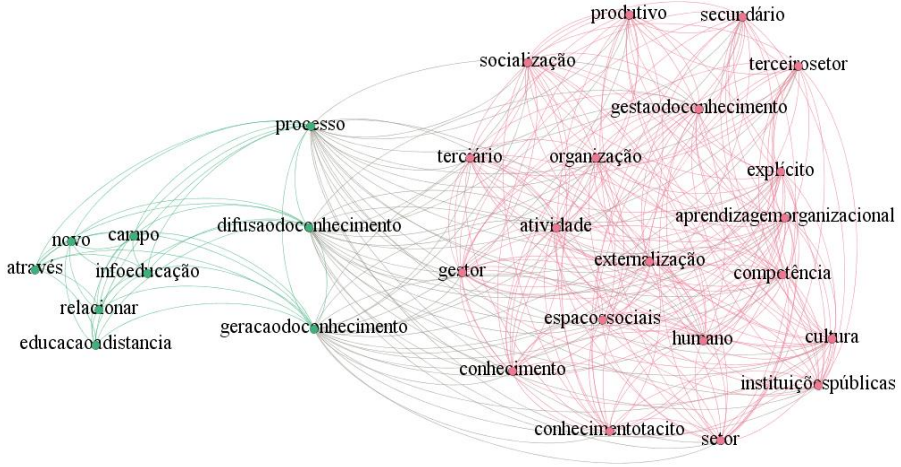


Figura 3: Excerto de uma rede de ementas. Fonte: Autores.

A Figura 3 mostra a rede semântica de cliques resultante das ementas das disciplinas “EDCA93 - Gestão do Conhecimento e Aprendizagem Colaborativa” e “EDCA94 - Infoeducação e Educação à Distância”. As palavras "processo", “difusão do conhecimento” e “geração do conhecimento” são as palavras em comuns às ementas, que atuam como vértices de conexão fazendo a sobreposição entre as duas cliques.

Os processos de “Tratamento das ementas” e “Construção da rede semântica de cliques baseada em ementas” (Figura 2) foram realizados conforme as etapas a seguir:

(1) União das ementas em um único discurso, sendo cada ementa inserida em uma linha em um arquivo de formato .txt.

(2) Tratamento manual das ementas, que consistiu na aplicação de 11 regras gerais propostas por Pereira et al. (2011) e aplicação de uma nova regra devido a uma especificidade deste estudo, a saber: Em alguns casos, foi necessária a introdução de palavras de maior ocorrência por temas-chave nas ementas que tratavam dos temas-chave, mas que não as continham. Isto propiciou a conexão entre as cliques que tratavam do mesmo tema-chave.

Por exemplo, havia ementas que explanavam sobre vários tipos de métodos (e.g. método analítico simplex, método de Householder e métodos estocásticos), e, seguindo o tratamento manual, estas palavras foram convertidas em “métodoanalíticosimplex”, “métododehouseholder” e “métodosestocásticos”, respectivamente. Como neste caso não aparecia a palavra “método” sozinha na ementa, a clique não iria se conectar a outras que tratassem do tema “T1: Metodologia” (Tabela 1), fazendo-se necessária a aplicação desta nova regra.

Outra situação para aplicação dessa regra ocorre quando a ementa apresenta pelo menos uma palavra relacionada ao tema-chave, segundo o vocabulário de controle (Tabela 1), mas não apresenta a palavra de maior ocorrência, sendo esta introduzida na ementa com o intuito de proporcionar a conexão das cliques por temas-chave.

(3) Após essa etapa, foi feito o tratamento computacional das ementas. As ementas passaram por um conjunto de programas da UNITEX [Paumier 2008] para classificação, modificação e eliminação, quando necessária, das palavras. Também foi utilizado o conjunto de ferramentas computacionais desenvolvido por Caldeira et al. (2005), que permite, dentre outras coisas, identificar a classe gramatical de cada palavra utilizada na criação da rede. Ao final do tratamento computacional foram gerados dois arquivos de formato .txt: um contendo o vocabulário e a classificação gramatical das palavras; e outro com os pares de palavras, frequência de aparição das palavras e dos pares nas ementas, o qual permite identificar os vértices em comum às cliques.

(4) Um arquivo no formato .net, também gerado na etapa anterior, foi utilizado para construção e análise da rede de ementas no software Gephi, versão 0.9.1. Para caracterizar a rede, foram empregados os índices: número de vértices ($n = |V|$), número de arestas ($m = |E|$), grau médio ($<k>$), diâmetro (D), coeficiente de aglomeração médio (C), caminho mínimo médio (L) e densidade (Δ). Estes índices foram calculados para a rede de ementas e para a rede aleatória correspondente, a qual foi construída para caracterização topológica da rede de ementas. A rede aleatória foi gerada com o mesmo número de vértices e grau médio da rede de ementas. E os índices empregados para medir a importância dos vértices foram: centralidade de grau (CG), centralidade de proximidade (CP) e centralidade de intermediação (CI).

4. Resultados e Discussão

Nesta rede de ementas, os vértices são as palavras empregadas nas ementas para descrever o conteúdo teórico (conceitual) ou teórico-metodológico (conceitual/ procedimental) dos componentes curriculares. Desta forma, os vértices representam conceitos e as arestas representam as conexões entre estes conceitos. Para reforçar essa afirmação, em revisão sistemática sobre redes semânticas, Rosa (2016) observou que, para a maioria dos estudos analisados (80%), um vértice de uma rede semântica representa conceitos ou objetos e as arestas não dirigidas de uma rede semântica representam relações entre estes conceitos ou objetos. Considerar os vértices da rede semântica de ementas como conceitos é importante na interpretação dos índices de centralidade de proximidade e intermediação, a seguir.

A rede de ementas do PPG investigado é constituída por 766 vértices e 17.201 arestas. É considerada esparsa ($\Delta = 0{,}059$) e as palavras estão conectadas em média a aproximadamente 45 outras palavras ($<k> = 44{,}91$). Isto sugere que as ementas possuem um vocabulário restrito para expressar os conhecimentos pretendidos.

A densidade é a razão entre o número de arestas existentes e o número máximo de arestas de uma dada rede. Em geral, percebe-se que a densidade da rede de ementas é maior que as observadas em RST (e.g. Pereira et al. 2016). Isto ocorre pelo fato de haver menor diversidade de palavras nas ementas do que entre as palavras dos títulos de artigos científicos. Quanto mais justaposições e sobreposições ocorram, maior densidade e maior coesão entre as ementas.

O coeficiente de aglomeração é um índice que mede o quanto os vértices vizinhos de um vértice estão conectados entre si. No caso da rede investigada, o coeficiente de aglomeração médio é considerado alto (C = 0,92), o que indica que os vizinhos dos vértices têm muitos vizinhos entre si, havendo muitas conexões entre os vértices. Isto pode ser explicado pelo próprio método de construção de redes semânticas de cliques e também pela grande quantidade de vértices em comum às ementas. Assim, o coeficiente de aglomeração médio alto pode ser um indicativo de menor diversidade de palavras e maior ocorrência de justaposição e sobreposição.

O diâmetro da rede investigada é igual a quatro (D = 4). O caminho mínimo médio informa que, em média, a distância entre as palavras é de 2,302. I, o que significa que as palavras são muito próximas. Isto acontece pela quantidade de justaposições e sobreposições entre as ementas, refletindo no maior compartilhamento de conceitos em comum, o que pode ser um indício de interdisciplinaridade.

Quanto à caracterização topológica, pode-se afirmar que a rede de ementas (*RE*) apresenta o fenômeno mundo pequeno, já que possui coeficiente de aglomeração médio elevado em relação ao da rede aleatória (*RA*) correspondente (C_{RE} = 0,92 >> C_{RA} = 0,03) e caminho mínimo médio similar ao da rede aleatória correspondente (L_{RE} = 2,302 ~ L_{RA} = 2,382), segundo o método de Watts e Strogatz (1998). Além disto, também foram observadas outras condições necessárias para que a rede investigada fosse considerada rede de mundo pequeno [Watts 1999], a saber: não dirigida, não ponderada, simples, esparsa e conectada (i.e. possui apenas um componente).

Considerando a análise de redes sociais, investigou-se também a importância dos vértices por meio dos índices de centralidade. As treze palavras da rede de ementas mais importantes em termo de centralidade de grau, em ordem decrescente, são: “conhecimento”, “teoria”, “abordagem”, “difusão do conhecimento”, “sociedade”, “análise”, “processo”, “complexidade”, “pesquisa”, “aplicação”, “rede”, “método” e “modelo”. Dentre estas, cinco representam temas-chave, e as outras, apesar de conceitualmente não representarem a proposta do PPG investigado, possuem significados relevantes para um curso de doutorado.

Tabela 2: Medidas de centralidades dos vértices que são as palavras de maior ocorrência por temas-chave.

TEMAS-CHAVE: VÉRTICES	*CG*	*CP*	*CI*
T7: Conhecimento	277	0,6105	0,0720
T10: Difusão do conhecimento	240	0,5930	0,0508
T3: Complexidade	210	0,5795	0,0306
T1: Método	190	0,5709	0,0585
T15: Modelo	179	0,5429	0,0353
T5: Ciência	150	0,5488	0,0172
T14: Linguagem	143	0,5422	0,0305
T9: Cultura	132	0,5387	0,0125
T11: Epistemologia	89	0,5222	0,0000
T13: Informação	76	0,4713	0,0054
T8: Construção do conhecimento	67	0,4926	0,0021
T4: Comunicação	60	0,4468	0,0022
T6: Tecnologia	48	0,4885	0,0031
T12: Geração do conhecimento	29	0,4279	0,0003
T2: Análise cognitiva	25	0,4540	0,0003

Do ponto de vista da importância dos vértices de maior ocorrência por temas-chave nas ementas, pode-se observar que o vértice de maior centralidade de grau é "conhecimento" (*CG* = 277), seguido pelo vértice "difusão do conhecimento" (*CG* = 240) (Tabela 2). Parece razoável afirmar que este fato é resultante da própria área de concentração em que o PPG está inserido, indicando assim que existe coerência e consistência entre a proposta do PPG e suas ementas em relação aos temas-chave "T7: Conhecimento" e "T10: Difusão do conhecimento".

Por outro lado, as palavras "análise cognitiva" (*CG* = 25) e "geração do conhecimento" (*CG* = 29) são os vértices que tiveram menor centralidade de grau e ambas aparecem em apenas duas ementas (Tabela 2). Esta constatação sugere que esses são os temas-chave menos abordados pelos componentes curriculares em comparação aos outros. Porém, deveriam ganhar maior destaque nas ementas, uma vez que "análise cognitiva" é o tema-chave central da linha de pesquisa três e a "geração do conhecimento" é almejada pela área de concentração do PPG investigado.

A centralidade de proximidade está relacionada com a distância total de um vértice em relação a todos os demais vértices do grafo [Freeman 1979]. Na rede de ementas, os vértices de maior centralidade de proximidade são os conceitos mais próximos, mais centrais aos outros no conjunto de ementas. Na rede investigada, os conceitos mais próximos aos outros são "conhecimento" (*CP* = 0,6105) e "difusão do conhecimento" (*CP* = 0,5930), respectivamente (Tabela 2), o que já era de se esperar uma vez que estes vértices representam os temas-chave centrais do PPG investigado.

Já a centralidade de intermediação mede quantos caminhos mais curtos entre todos os pares de vértices do grafo passam por um determinado vértice [Freeman 1979]. Este índice avalia a frequência de ocorrência de um determinado vértice entre pares de outros vértices em caminhos mais curtos que os conectam. Na rede de ementas, os vértices com maior centralidade de intermediação são "conhecimento" (*CI* = 0,0720) e "método" (*CI* = 0,0585). Estes vértices estabelecem pontes entre vários outros conceitos na rede de ementas (Tabela 2).

Na rede de ementas foram observadas 10 comunidades com o índice de modularidade igual a 0,619 (Figura 4). Newman e Girvan (2004) definem comunidades como grupos de vértices densamente conectados. Em redes semânticas, os vértices que compõem as comunidades formam grupos de palavras bem integrados ou semanticamente relacionados. Este fato também é observado em redes semânticas de cliques, mas, não necessariamente, as comunidades refletem as cliques, pois as justaposições e sobreposições entre as cliques aumentam a densidade das comunidades.

Observa-se na Figura 4 que a comunidade azul claro agrupa os seguintes vértices: "conhecimento", "difusão do conhecimento", "geração do conhecimento", "construção do conhecimento" e "cognição". Estas palavras possuem significados distintos, mas correlacionados entre si, e fazem parte do processo de "Modelagem da Geração e Difusão do Conhecimento", área de concentração do PPG investigado. Por isso, esta comunidade representa um importante núcleo semântico do PPG investigado. Também pode-se observar que os vértices que representam os temas-chave "difusão do conhecimento", "ciência", "complexidade" e "tecnologia" são vértices que atuam como "pontes" entre comunidades, ligando, assim, o que pode ser mais um indício da importância destes conceitos na promoção de interdisciplinaridade.

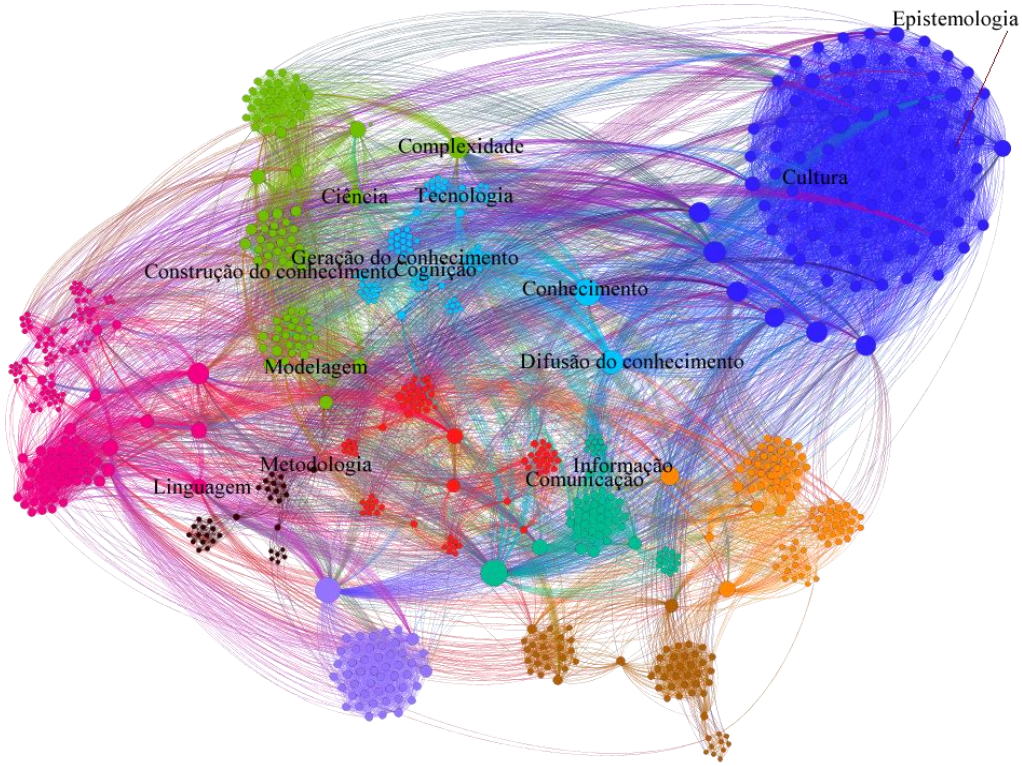


Figura 4. Rede de ementas, com destaque para os vértices de maior centralidade de grau e distribuição espacial por comunidades. Fonte: Autores.

A identificação dos temas-chave nas ementas permitiu agrupar os componentes curriculares por temas-chave (Tabela 3), e, a partir disto, foi construída a rede de componentes curriculares (Figura 5). Apesar da discussão sobre este tipo de rede estar fora do escopo deste artigo, far-se-á uma análise preliminar a seu respeito.

Tabela 3: Agrupamento dos componentes curriculares por temas-chave.

TEMAS-CHAVE (T)		COMPONENTES CURRICULARES
T1	Metodologia	EDCA91; EDCA86; EDCA98; EDCA99; EDCB04; EDCB09; EDCB76
T2	Cognição	EDCE31; EDCC42
T3	Complexidade	EDCA85; EDCA87; EDCA89; EDCB05; EDCB09
T4	Comunicação	EDCB07; EDCB10
T5	Ciência	EDCA85; EDCA89; EDCA92; EDCB05; EDCB06; EDCJ78; EDCE47
T6	Tecnologia	EDCA90; EDCA92; EDCB06; EDCH10
T7	Conhecimento	EDCA85; EDCA88; EDCA89; EDCE30; EDCA92; EDCA93; EDCB09; EDCC42; EDCB07; EDCJ78; EDCC42
T8	Construção do conhecimento	EDCA89; EDCH68; EDCA92
T9	Cultura	EDCE32; EDCB07; EDCB09; EDCA93
T10	Difusão do conhecimento	EDCA89; EDCA90; EDCA93; EDCA94; EDCA95; EDCB09; EDCC42; EDCC51; EDCE32; EDCH10; EDCJ78
T11	Epistemologia	EDCB09
T12	Geração do conhecimento	EDCA93; EDCA94
T13	Informação	EDCA90; EDCB07; EDCB10
T14	Linguagem	EDCA96; EDCA97; EDCB10; EDCC49; FISB39
T15	Modelagem	EDCA85; EDCA88; EDCA89; EDCA91; EDCA92; EDCA95; EDCB03; EDCB04; EDCB76

Na Tabela 3, observa-se o agrupamento dos componentes curriculares por temas-chave (i.e. principais conceitos extraídos a proposta de aprendizagem do PPG investigado). Isto pode auxiliar os discentes na identificação de componentes curriculares importantes para o desenvolvimento de suas teses.

Na rede de componentes curriculares apresentada na Figura 5, o tamanho dos vértices é proporcional ao seu grau. Este índice indica a quantidade de componentes curriculares adjacentes com que cada um deles se relaciona por meio de conceitos em comum. O grau também sofre influência do peso das arestas. Neste estudo, foram atribuídos pesos às arestas, equivalentes à quantidade de temas-chave em comum às disciplinas. Os pares de vértices com maior peso são aqueles que possuem mais temas-chave em comum. Por exemplo, os vértices "EDCA85 - Epistemologia e Construção do Conhecimento" e "EDCA89 - Processos de Construção do conhecimento" se conectam com peso quatro, uma vez que compartilham quatro temas-chave: "modelagem", "conhecimento", "complexidade" e "ciência" (Tabela 3).

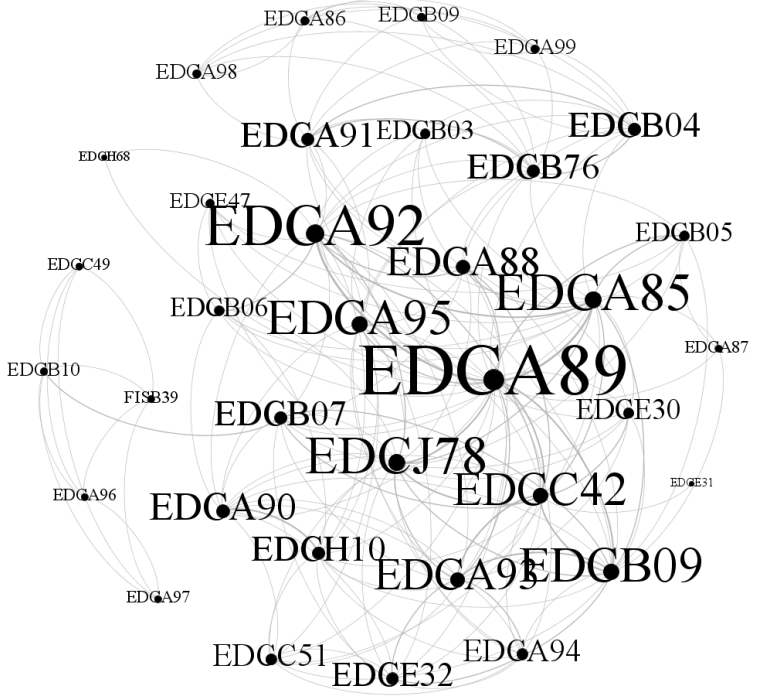


Figura 5. Rede de componentes curriculares do PPG investigado, com destaque para os vértices de maior centralidade de grau. Fonte: Autores.

Os vértices com as maiores centralidades de grau (EDCA89, EDCA85 e EDCA92) (Figura 5) são os componentes curriculares potenciais para a implementação de projetos de ensino interdisciplinaridade. Esta afirmação é corroborada pelo pressuposto que a interdisciplinaridade como uma abordagem de ensino propõe a interação entre duas ou mais disciplinas por meio de conceitos (e.g. temas-chave) compartilhados ou relacionados entre si. Segundo Japiassu e Marcondes (1993), a interdisciplinaridade consiste em um método de pesquisa e de ensino que possibilita que duas ou mais disciplinas interajam entre si, seja por meio da comunicação de ideias, integração mútua de conceitos, epistemologias, terminologias, metodologias, procedimentos e dados, ou pela forma de organização da pesquisa.

5. Considerações Finais

A rede semântica de cliques baseada em ementas permitiu avaliar a coerência e a consistência entre a proposta de aprendizado baseada nas áreas de concentração e nas linhas de pesquisa de um PPG e o ementário relativo à sua matriz curricular.

Neste estudo, a coerência e a consistência confirmam-se apenas em relação a alguns dos 15 temas-chave. Os vértices "conhecimento" e "difusão do conhecimento" apresentaram os maiores valores de centralidade de grau, proximidade e intermediação, denotando sua importância nas ementas. Esta relevância, porém, não foi observada em relação aos vértices "análise cognitiva" e "geração do conhecimento". Além disso, foi constatado que os processos de justaposição e sobreposição impactam sobremaneira nos índices da rede semântica de cliques baseada em ementas repercutindo em suas características topológicas e na preeminência dos vértices.

A partir da identificação dos temas-chave nas ementas, foi possível agrupar os componentes curriculares por temas-chave e, a partir disto, construiu-se a rede de componentes curriculares. Este tipo de rede demostra a integração entre os componentes curriculares por meio de conceitos relevantes (i.e. temas-chave).

As redes de ementas e de componentes curriculares sugerem que a teoria de redes pode ser empregada para análise de conexões entre palavras das ementas e entre componentes curriculares da matriz curricular. A análise das conexões entre os conceitos pode auxiliar na busca de conceitos relacionados ao objeto de pesquisa dos discentes de forma mais eficiente e na identificação do vocabulário relacionado a estes conceitos. Já a análise das conexões entre os componentes curriculares pode auxiliar no estabelecimento de critérios de pré-requisitos entre eles.

Desta forma, sendo a teoria de redes uma ferramenta que auxilia no estudo de sistemas compostos por elementos que se conectam, auxiliando na visão da complexidade do todo, pode auxilia na percepção da complexidade de um PPG. Ademais, espera-se que este trabalho contribua para elaboração de um modelo computacional baseado na teoria de redes capaz de apoiar o planejamento e avaliação de programas de pós-graduação.

Referências

Caldeira, S. M. G. (2005) "Caracterização da rede de signos linguísticos: Um modelo baseado no aparelho psíquico de Freud". Mestrado Interdisciplinar em Modelagem Computacional, Fundação Visconde de Cairu, Salvador.

Caldeira, S. M. G., Petit Lobão, T. C., Andrade, R. F. S, Neme, A. e Miranda, J. G. V. (2006). The network of concepts in written texts. In *The European Physical Journal B - Condensed Matter and Complex Systems*, 49(4):523–529.

Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES). (2013) Diretoria de Avaliação. "Avaliação Trienal 2013". In: Documento de Área. Área de Avaliação: Interdisciplinar.

Fadigas, I. S. e Pereira, H. B. B. (2013). A network approach based on cliques. In *Physica A: Statistical Mechanics and its Applications*, 392(10):2576 – 2587.

Fadigas, I. S., Henrique, T., Pereira, H. B. B., Senna, V. e Moret, M. (2009). Análise de redes semânticas baseada em títulos de artigos de periódicos científicos: o caso dos periódicos de divulgação em educação matemática. In *Educação Matemática Pesquisa*, 11(1):167–193.

Freeman, L. C. (1979). Centrality in Social Networks: Conceptual clarification. In *Social Networks*, 1(3):215-239.

Grilo, M., Fadigas, I. S., Miranda, J. G. V., Cunha, M. V., Monteiro, R. L. S. e Pereira, H. B. B. (2017). Robustness in semantic networks based on cliques. In *Physica. A (Print)*, 472:94-102.

Henrique, T., Fadigas, I. S., Rosa, M. G. e Pereira, H. B. B. (2014). Mathematics education semantic networks. In *Social Network Analysis and Mining*, 4:200.

Japiassu, H. e Marcondes, D. (1993) "Dicionário básico de filosofia". In: Zahar, Rio de Janeiro.

Lima-Neto, J. L. A., Cunha, M. V. e Pereira, H. B. B. (2018). Redes semânticas de discursos orais de membros de grupos de ajuda mútua. In *Obra Digital: Journal Of Communication And Technology*, 14: 51-66.

Lopes, C. R. S., Cardoso, J. P., Meira, S. S., Casotti, C. A., Vilela, A. B. A. e Pereira, H. B. B. (2015). Significado de corresidência na visão de idosos: uma estratégia para análise cognitiva com uso de redes semânticas. In *Revista Saúde.com*, 11:174-182.

Newman, M. E. J. e Girvan, M. (2004). Finding and evaluating community structure in networks. In *Phys. Rev. E.*, 69(2):026113.

Paumier, S. (2008) UNITEX 2.0. "User Manual", Université Paris-Est Marne-la-Vallée, http://unitexgramlab.org/releases/2.0/man/Unitex-GramLab-2.0-usermanual-en.pdf

Pereira, H. B. B, Fadigas, I., Senna, V. e Moret, M. (2011). Semantic networks based on titles of scientific papers. In *Physica A: Statistical Mechanics and its Applications*, 390(6):1192–1197.

Pereira, H. B. B, Fadigas, I. S., Monteiro, R. L. S., Cordeiro, A. J. A. e Moret, M. A. (2016). Density: A measure of the diversity of concepts addressed in semantic networks. In *Physica. A (Print)*, 441:81-84.

Rosa, M. G. (2016) "Modelo empírico para analisar a robustez de redes semânticas". Doutorado Multidisciplinar e Multi-institucional em Difusão do Conhecimento. Universidade Federal da Bahia, Salvador.

Teixeira, G., Aguiar, M., Pereira, H., Miranda, J., Cunha, M., Morais, J., Carvalho, C. e Dantas, D. (2010). Complex semantics networks. In *International Journal of Modern Physics C*, 21(3):333–347.

Watts, D. J. (1999) "Small Worlds: The dynamics of Networks between Order and Randomness", Princetown University Press, New Jersey.

Watts, D. J. e Strogatz, S. H. (1998). Collective dynamics of small-world networks. In *Nature*, 393(4):440-442.

O Que os Países Escutam: Analisando a Rede de Gêneros Musicais ao Redor do Mundo

Maria Luiza Botelho Mondelli, Luiz M. R. Gadelha Jr., Artur Ziviani

[1]Laboratório Nacional de Computação Científica (LNCC)
Petrópolis, RJ, Brasil

{mluiza,lgadelha,ziviani}@lncc.br

Abstract. *Music streaming platforms are increasingly popular, democratizing and facilitating the access to music content. This effect extends the reach and the penetration of different musical styles, increasing the diversity of listened genres in different countries around the world. In order to better understand this diversity and identify countries with common interests, in this paper we build and analyze a complex network of artists, musical genres, and countries using data from Spotify, one of the most widely used music streaming platforms today. As results, in addition to identifying communities of countries with similar musical styles, we show how the large amount and diversity of musical genres can influence the modeling and analysis of the considered network. We also classify the most commonly listened genres using different centrality metrics.*

Resumo. *Plataformas de streaming de música são cada vez mais populares, democratizando e facilitando o acesso ao conteúdo musical. Esse efeito amplia o alcance e a penetração de diferentes estilos musicais, incrementando a diversidade de gêneros escutados nos diferentes países do mundo. A fim de melhor entender essa diversidade e identificar países com interesses em comum, neste artigo foi construída e analisada uma rede complex de artistas, gêneros musicais e países utilizando dados do Spotify, uma das plataformas de streaming de música mais utilizadas atualmente. Como resultados, além de identificar comunidades de países com estilos musicais semelhantes, nós mostramos como a grande quantidade e diversidade de gêneros musicais pode influenciar a modelagem e análise da rede considerada. Nós também classificamos os gêneros mais comumente escutados utilizando diferentes métricas de centralidade.*

1. Introdução

A música é um tipo de arte cuja manifestação pode ser caracterizada como uma prática cultural. Seu papel está relacionado a aspectos como entretenimento e construção de identidade, memória e emoções em indivíduos [DeNora 2000]. Além disso, a música é comumente utilizada como uma forma de auto-expressão, fornecendo um meio de caracterizar e identificar, por exemplo, grupos de pessoas com interesses musicais em comum e também acontecimentos históricos [Rentfrow 2012].

Nas últimas décadas, o avanço da tecnologia tem exercido um papel importante na indústria da música, mudando a forma como ela é distribuída e consumida pelo público. Um exemplo disso consiste no surgimento das plataformas de *streaming* de música. Devido ao fato de permitirem o consumo em tempo real de músicas sem a necessidade de

fazer o *download* de arquivos [Trefzger et al. 2015], essas plataformas têm recebido cada vez mais destaque. A grande adesão a esse tipo de serviço por parte dos usuários pode ser entendida como um fator que tem democratizado o acesso ao conteúdo musical. Isso pode ser observado a partir de dados do IFPI (*International Federation of the Phonographic Industry*) [IFPI 2017], que mostram que 50% da receita da indústria fonográfica em 2016 foi proveniente do uso de plataformas de *streaming* de música e *downloads*.

Basicamente, através do serviço de *streaming*, usuários escutam músicas, que por sua vez são gravadas e disponibilizadas por artistas e bandas. Através dessa estrutura, é possível ainda registrar alguns metadados que incluem: gênero dos artistas, popularidade e quantidade de *streams* de determinada música, entre outros. A coleta e o armazenamento desse tipo de informação é um aspecto importante que surge com a popularização dessas plataformas e que abre espaço para um grande conjunto de possíveis estudos. Dentre eles, destaca-se a aplicação de técnicas de ciência de redes [Barabási 2016] para a análise da estrutura das plataformas, que podem ser facilmente caracterizadas como uma rede.

Neste contexto, o presente trabalho tem como objetivo analisar a rede de gêneros musicais ao redor do mundo. Para isto, utilizaremos o Spotify, uma plataforma de *streaming* de música bem consolidada e que disponibiliza os dados necessários para a construção da rede, que será composta por países, artistas e gêneros. Serão utilizados dados sobre as músicas mais tocadas em alguns dos países onde o Spotify está disponível. Esse tipo de estudo permitirá, por exemplo, entender quais são os gêneros mais tocados e identificar comunidades de países que compartilham gêneros musicais em comum.

Sendo assim, este artigo está organizado como segue. A Seção 2 descreve o processo de coleta de dados para a construção da rede. A Seção 3 apresenta o estudo e os resultados das análises da rede. A Seção 4 apresenta alguns trabalhos relacionados. Por fim, a Seção 5 conclui o trabalho e apresenta algumas oportunidades de trabalhos futuros.

2. Conjunto de dados e construção da rede

O Spotify é uma plataforma de *streaming* de música lançada em 2008 e que hoje reúne aproximadamente 30 milhões de músicas, 2 milhões de artistas e 140 milhões de usuários. Além de ser um serviço de *streaming* bem consolidado, o Spotify disponibiliza o acesso a parte de sua base de dados para consulta, possibilitando a construção da rede a ser analisada neste trabalho. Sendo assim, a seguir são descritas as duas fontes do Spotify utilizadas para a coleta de dados:

- Spotify Charts[1]: registra, diária e semanalmente, uma lista de até 200 músicas mais escutadas nos países onde o Spotify está disponível e seus respectivos artistas. Para este trabalho, foram coletadas duas listas: a primeira referente ao dia 01/07/2017, com 9231 registros, totalizando 1326 artistas e 2438 músicas diferentes; e a segunda referente ao dia 01/01/2018, com 9219 registros e um total de 1609 artistas e 2995 músicas diferentes. Vale ressaltar que para a primeira listagem foi possível obter registros para 57 países, enquanto que para segunda foram obtidos registros para 50 países.
- Spotify Web API[2]: disponibiliza metadados sobre artistas, álbuns e músicas do catálogo do Spotify. A API foi acessada a fim de recuperar os gêneros musicais

[1] `https://spotifycharts.com/`

[2] `https://developer.spotify.com/web-api`

de cada um dos artistas obtidos nas listagens do Spotify Charts descritas anteriormente. Vale destacar que cada artista pode possuir mais de um gênero e que, de acordo com a plataforma Every Noise [Noise 2017], o Spotify possui um registro de aproximadamente 1500 gêneros diferentes. A consulta aos metadados da API permitiu então associar cada artista ao seu respectivo gênero principal, resultando em um total de 367 diferentes gêneros para o conjunto de dados do dia 01/07/2017 e 425 gêneros para o conjunto do dia 01/01/2018. O impacto da grande quantidade de gêneros será discutido na Seção 3, onde serão feitas as análises das redes.

A estrutura dos dados permitiu então a construção de uma rede tripartida, não direcionada, para cada uma das listagens coletadas. Nessa rede, os conjuntos de vértices podem ser classificados como: (i) países, (ii) artistas e (iii) gêneros musicais. As ligações entre as classes de vértices são apresentadas na Figura 1 e acontecem da seguinte forma:

- Países se conectam a artistas, segundo o registro obtido no Spotify Charts que relaciona cada música e seu respectivo artista aos países onde ele é mais escutado.
- Artistas se conectam aos gêneros, de acordo com os metadados obtidos através da API Web.

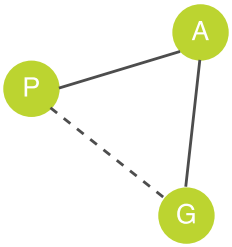


Figura 1. Exemplo da estrutura da rede de países, artistas e gêneros.

Com esse modelo, é possível ainda conectar cada país diretamente aos gêneros, por transitividade (linha tracejada na Figura 1). Essa conexão, que resulta na rede bipartida de país e gênero, será o objeto de estudo deste trabalho. Vale ainda ressaltar que, a construção dos conjuntos de dados e as análises apresentadas na seção a seguir, foram feitos utilizando o pacote estatístico R e estão disponíveis no GitHub [3].

3. Análise da rede de gêneros musicais

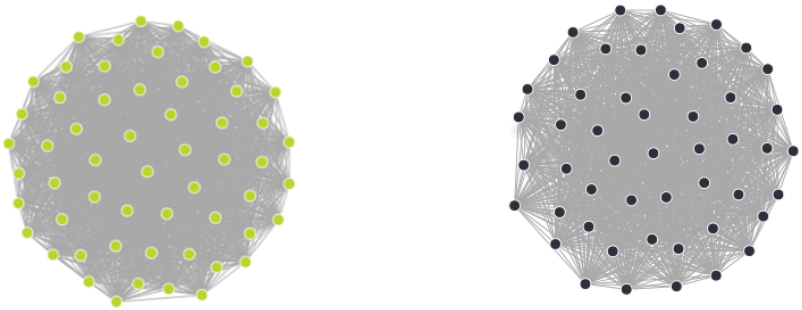
Esta seção analisa a rede de gêneros musicais ao redor do mundo para cada uma das datas coletadas. Primeiramente, consideramos a rede completa de gêneros musicais e, em seguida, analisamos a rede com somente os principais gêneros musicais escutados. A fim de diferenciar os resultados das análises, nomeamos como *Rede 1* e *Rede 2* as redes que utilizam os dados dos dias 01/07/2017 e 01/01/2018, respectivamente.

3.1. Análise das redes completas de gêneros musicais

Para esta análise, consideramos o modelo de rede bipartida que relaciona os conjuntos de vértices país e gênero. A Rede 1 possui 57 países e 367 gêneros, totalizando 424 vértices e 3579 arestas. A Rede 2, por sua vez, possui 50 países e 424 gêneros, totalizando 474 vértices e 3835 arestas. Com a finalidade de entender o perfil musical de cada um dos

[3] https://github.com/mmondelli/network-science

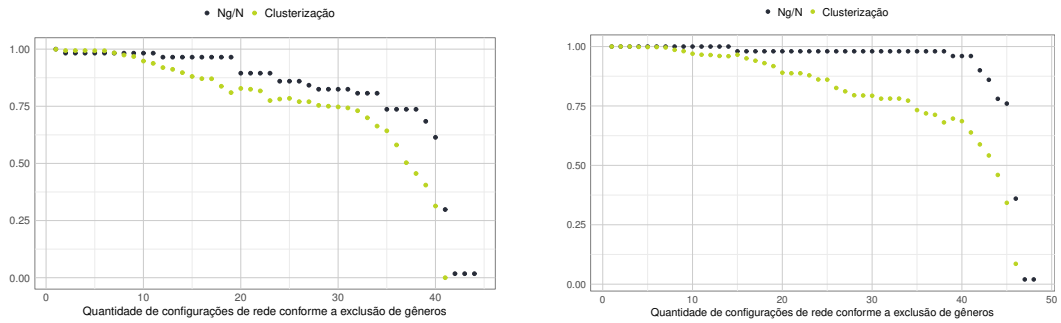
países e analisar as características das duas redes, foi realizada a projeção da rede de países conectados por gêneros em comum. Essa projeção resultou em uma rede completamente conectada, onde cada país está conectado aos demais, como mostra a Figura 2.



(a) Rede 1 (01/07/2017) (b) Rede 2 (01/01/2018)

Figura 2. Projeção de países conectados por gênero em comum.

Analisando os dados, observa-se que isso ocorre devido a existência de pelo menos um gênero que ocorre em todos os países. Esse tipo de gênero pode ser classificado então como sendo um gênero universal, que é o caso do gênero *pop*, por exemplo. Além da existência de gêneros universais, outro aspecto pode ser observado: o fator viral de uma determinada música ou artista pode fazer com que um gênero também ocorra em todos os países, mesmo que pontualmente. Um exemplo disso consiste no gênero *latin*, que aparece em todos os países da Rede 1 devido à popularidade da música "Despacito" do cantor Luis Fonsi, identificada como a música mais escutada ao redor do mundo. Os motivos pelos quais um artista ou música tornam-se virais é uma questão bastante discutida atualmente, mas que foge do escopo deste trabalho. No entanto, no caso da música "Despacito", é possível atribuir parte da popularidade ao aumento do uso de plataformas de *streaming* de música e vídeo[4].



(a) Rede 1 (01/07/2017) (b) Rede 2 (01/01/2018)

Figura 3. Comportamento da rede com a exclusão de gêneros.

Considerar a estrutura original dos conjuntos de dados deste trabalho impossibilita

[4]http://www.billboard.com/articles/columns/latin/7873798/luis-fonsi-daddy-yankee-despacito-streaming-popularity

o estudo das nuances acerca do perfil musical de cada um dos países, devido a existência de gêneros universais e músicas virais. Foi então realizada uma análise das Redes 1 e 2, a fim de entender seus comportamentos conforme uma determinada quantidade de gêneros fosse excluída. A exclusão foi feita iterativamente, de forma com que no primeiro passo fossem excluídos todos os gêneros que aparecem em todos os *n* países, onde *n* é o número total de países de cada rede. Depois foram excluídos os gêneros que aparecem em pelo menos *n - 1* países e assim por diante. Esse processo foi repetido até que não existissem mais gêneros a serem excluídos, totalizando 44 passos (ou redes diferentes) para a Rede 1 e 48 passos para a Rede 2. Foram contabilizados a fração de vértices do componente gigante (N_g) em relação ao número total de vértices de cada rede (N) e a clusterização em cada caso. O resultado é apresentado na Figura 3. É possível perceber que, em ambos os casos, a clusterização tende a permanecer alta. O componente gigante abrange a maioria dos nós da rede por muitos passos, diminuindo de forma mais significativa apenas quando restam poucos gêneros a serem excluídos. Sendo assim, escolher um limite de gêneros a ser considerado nesse caso para prosseguir com a caracterização da rede não é trivial. Neste trabalho, optou-se então por analisar apenas os gêneros mais frequentes em cada um dos países, conforme descreve a subseção a seguir.

3.2. Análise da rede para os 5 gêneros musicais mais escutados

Como observado na análise das redes com a exclusão de gêneros, existe uma sensibilidade no que diz respeito ao estado das redes que pode ser considerado para a continuação do estudo. Sendo assim, neste trabalho optou-se por fazer uma filtragem nos dados, de forma a considerar apenas os cinco gêneros mais frequentemente escutados em cada país. Vale ressaltar também que, para a construção destas redes, foram excluídos gêneros que aparecem em pelo menos 90% dos países, a fim de evitar as redes completamente conectadas e também de possibilitar o aparecimento de gêneros mais característicos de cada país.

A rede bipartida de países e gêneros, no caso da Rede 1, possui 157 vértices e 259 arestas. Para o caso da Rede 2, possui 155 vértices e 250 arestas. As redes podem ser analisadas a partir de suas projeções ponderadas. A Tabela 1 mostra as características das projeções, onde: (i) *Países* diz respeito à projeção de países conectados por gêneros em comum; e (ii) *Gêneros* é a projeção de gêneros conectados por países em comum. Em comparação ao que é esperado para redes aleatórias equivalentes às duas projeções das Redes 1 e 2, observa-se que ambas possuem distância e diâmetro baixos e alta clusterização. Além disso, é possível observar que as redes são bem conectadas, com grande parte dos vértices pertencendo ao maior componente conexo.

3.2.1. Ranqueamento dos gêneros musicais

A partir da projeção de gêneros conectados por países em comum, foi possível obter o ranqueamento dos gêneros mais centrais para as Redes 1 e 2. As Tabelas 2 e 3 apresentam uma listagem dos primeiros 5 gêneros classificados de acordo com três diferentes métricas de centralidade descritas a seguir:

- Grau: é considerada a medida mais simples de centralidade em redes e está relacionada à quantidade de conexões que um determinado vértice possui. Sendo assim, quanto mais conexões, maior a importância do vértice na rede. Para redes direcionadas, a centralidade de grau leva em consideração a direção da conexão.

Tabela 1. Características das projeções da rede bipartida de gêneros e países (*top* 5).

	Rede 1		Rede 2	
Características	**Países**	**Gêneros**	**Países**	**Gêneros**
Vértices	56	101	50	105
Arestas	459	310	357	327
Grau máximo	36	26	33	40
Grau mínimo	0	3	0	4
Grau médio	16.3	6.1	14.3	6.2
Distancia média	1.76	2.6	2.06	2.8
Diâmetro	5	7	9	9
Clusterização	0.79	0.43	0.8	0.39
Densidade	0.29	0.06	0.29	0.005
Componentes	6	6	3	3
Tamanho maior componente	91%	75%	96%	90%

- *Closeness*: calcula o comprimento médio dos caminhos mais curtos de um vértice para cada um dos demais vértices que compõem a rede. Quanto maior a centralidade de um determinado vértice, menor é a sua distância para os demais.
- *Betweenness*: estabelece a importância de um vértice com base na quantidade de caminhos mínimos pelos quais ele faz parte. Em outras palavras, esse tipo de centralidade quantifica o número de vezes em que um vértice atua como ponte no caminho mais curto entre dois outros vértices.

A respeito dos resultados de ranqueamento obtidos, vale ressaltar que a frequência com que esses gêneros aparecem em cada país não está sendo levada em consideração, mas sim se eles aparecem pelo menos uma vez, de acordo com a filtragem proposta nesta seção. O gênero *southern hip hop*, por exemplo, é mais frequente nos Estados Unidos e Canadá, enquanto que o *reggaetown* é frequente nos países da América Latina em geral. Em um trabalho futuro, uma oportunidade estaria em considerar a frequência dos gêneros nas análises.

Por se tratar de uma rede bipartida, a interpretação dos ranqueamentos apresentados nas Tabelas 2 e 3 deve levar em consideração o fato de que os gêneros estão conectados se possuem pelo menos um país em comum. Através dos dados da Rede 1 foi possível identificar que o gênero *southern hip hop* está presente em apenas 14 países, enquanto que o *reggaeton* aparece em 22 países. No entanto, comparando os dois casos, os países que escutam *southern hip hop* são mais distintos entre si, fazendo com que ele se conecte a mais gêneros e por isso possua maior centralidade de grau. Esse mesmo aspecto foi observado através de uma varredura nos dados da Rede 2, onde existem gêneros como *vegas indie* e *progressive house* que estão presentes em mais países do que os listados no ranqueamento, mas que não possuem alta centralidade de grau. Outro fator que pode ser observado é a similaridade dos resultados das métricas *closeness* e *betwenness*, principalmente para a Rede 1. Esses resultados para *closeness* e *betwenness* sugerem que os gêneros listados atuam tanto aproximando os países com estilos musicais em comum, quanto conectando países que possuem maior diversidade de gêneros musicais, respectivamente.

Além disso, é possível observar a classificação de gêneros como *west coast rap*, *latin hip hop*, *trap music* e *trap latino* dentre os gêneros mais centrais, mesmo com a exclusão dos gêneros *rap*, *hip hop* e *latin*. Isso mostra que, de alguma forma, esses últimos três gêneros ainda estão presentes através de suas variações, indicando sua importância na rede. Isso também ocorre com o *neo mellow* que, de acordo com mapa construído pelo Every Noise [Noise 2017], é um gênero muito similar ao *pop*, que também não está presente em nenhuma das redes.

Tabela 2. Ranqueamento dos gêneros musicais para a Rede 1.

	Grau	Closeness	Betweenness
1	southern hip hop	reggaeton	reggaeton
2	reggaeton	neo mellow	neo mellow
3	neo mellow	southern hip hop	southern hip hop
4	west coast rap	west coast rap	deep tropical house
5	tropical	deep tropical house	west coast rap

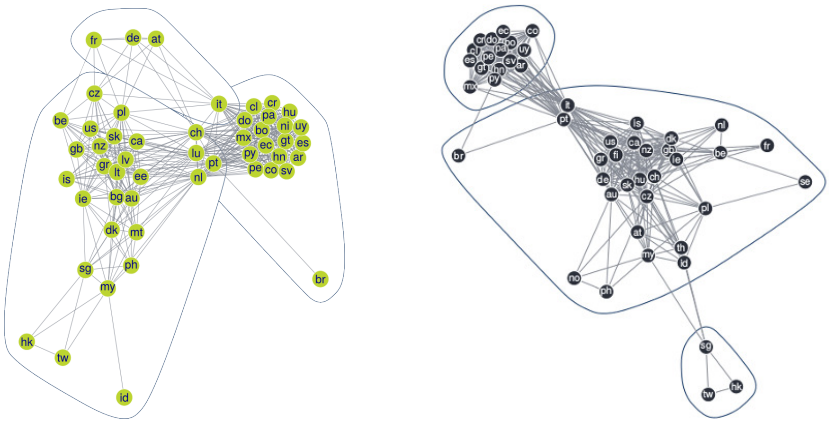
Tabela 3. Ranqueamento dos gêneros musicais para a Rede 2.

	Grau	Closeness	Betweenness
1	trap music	southern hip hop	trap music
2	big room	trap music	southern hip hop
3	tropical	rock	viral pop
4	southern hip hop	deep funk carioca	tropical
5	viral pop	viral pop	big room

3.2.2. Países com interesses musicais em comum

A identificação de comunidades em ciência de redes pode ser entendida como a busca por grupos de nós que possuem maior probabilidade de se conectarem entre si do que a nós de outras comunidades [Barabási 2016]. Sendo assim, comunidades reúnem um conjunto de nós que estão mais conectados internamente do que com outros nós em uma rede aleatória equivalente. Partindo dessa definição e da projeção de países conectados por gêneros, este trabalho buscou identificar os grupos de países que possuem interesses em comum. Para isso, foi aplicado o método de Louvain [Blondel et al. 2008], que possui uma abordagem baseada na otimização da modularidade e é considerado o estado da arte para detecção de comunidades em redes. Basicamente, em um primeiro momento, o método atribui cada vértice a uma comunidade. O método segue de forma iterativa, reatribuíndo os vértices às comunidades de forma com que eles se movam para as comunidades onde contribuem mais para a modularidade. Quando nenhum vértice pode ser mais reatribuído, o processo continua, mas considerando as comunidades como sendo os vértices. O processo é finalizado quando a modularidade global máxima é encontrada.

Como resultado, na Rede 1 foram encontrados 8 comunidades de países. Dessas comunidades, 5 são comunidades individuais de países que não estavam conectados ao maior componente da rede, sendo eles: Japão, Finlândia, Noruega, Turquia e Suécia. Com a filtragem dos 5 gêneros mais frequentes proposta na análise desta seção, esses



(a) Rede 1 (01/07/2017) (b) Rede 2 (01/01/2018)

Figura 4. Comunidades de países identificada pelo método de Louvain.

países passaram a não compartilhar nenhum gênero com os demais. No caso do Japão por exemplo, os gêneros mais comuns são o *anime*, juntamente com variações regionais do *pop*, *rap*, *r&b* e *rock*. As outras 3 comunidades são apresentadas na Figura 4(a) e descritas a seguir:

1. Argentina, Bolívia, Chile, Colômbia, Costa Rica, República Dominicana, Equador, Espanha, Guatemala, Honduras, México, Nicarágua, Panamá, Perú, Paraguai, São Salvador e Uruguai, Brasil, Portugal, Hungria e Suíça.
2. Áustria, Alemanha, Itália e França.
3. Estados Unidos, Reino Unido, Austrália, Bélgica, Bulgaria, Canadá, República Checa, Dinamarca, Estônia, Gŕecia, Holanda, Hong Kong, Indonésia, Irlanda, Islândia, Lituânia, Letônia, Luxemburgo, Malta, Malásia, Nova Zelândia, Filipinas, Polônia, Singapura, Eslováquia e Taiwan.

É interessante notar que, na comunidade 1 foram claramente agrupados a maioria dos países de línguas espanhola e castelhana. A comunidade 2 agrupou alguns países da Europa. A comunidade 3 por sua vez é a maior e mais diversificada, possuindo países da Europa, Ásia, América do Norte e Oceania.

Para a Rede 2 foi identificado um total de 5 comunidades, onde 2 são comunidades individuais dos países Japão e Turquia. No caso da Rede 1, esses dois países também não fazem parte de nenhuma outra comunidade, sugerindo que de fato eles possuem predileção por estilos musicais próprios. As demais comunidades, apresentadas na Figura 4(b), são listadas a seguir:

1. Argentina, Bolívia, Chile, Colômbia, Costa Rica, República Dominicana, Equador, Espanha, Guatemala, Honduras, México, Panamá, Peru, Paraguai, São Salvador e Uruguai.
2. Hong Kong, Singapura e Taiwan.
3. Estados Unidos, Reino Unido, Áustria, Austrália, Bélgica, Brasil, Canadá, Suíça, República Checa, Alemanha, Dinamarca, Finlândia, França, Grécia, Hungria, Indonésia, Irlanda, Islândia, Itália, Malásia, Holanda, Noruega, Nova Zelândia, Filipinas, Polônia, Portugal, Suécia, Eslováquia e Tailândia.

Observa-se que, diferentemente do resultado da Rede 1, a comunidade 1 não agrupa nenhum outro país que não seja de língua espanhola ou castelhana. A comunidade 2 passou

a agrupar países da Ásia e a comunidade 3 continua sendo a maior e mais diversificada.

Apesar dos agrupamentos, em ambos os casos percebe-se ainda que muitos países de uma comunidade se conectam a países de outras comunidades, podendo indicar que de fato eles são parecidos musicalmente ou ser consequência da grande quantidade de gêneros diferentes que existem mesmo com os filtros aplicados. No geral, observamos que o comportamento dos agrupamentos nas Redes 1 e 2 mantém um padrão. No entanto, as diferenças pontuais podem acontecer devido a filtragem de gêneros que foi proposta nesta seção e também ser consequência das músicas que são listadas como populares em cada um dos períodos. A identificação de comunidades nesse tipo de rede pode também sugerir a existência de zonas de influência de determinados gêneros em grupos de países, podendo ser objeto de estudo em análises mais detalhadas no futuro.

3.2.3. Quantos gêneros de um determinado país são de fato dele?

A plataforma Every Noise desenvolveu um mecanismo onde, além de identificar todos os gêneros musicais existentes a partir de características das músicas (incluindo aspectos sobre acústica, energia, entre outros), é possível atribuir os gêneros ao seu país considerado origem. Essa informação está também disponível através de seu site. Considerando então os 5 gêneros mais frequentes em cada um dos países, é possível determinar qual é a proporção desses gêneros que são de fato locais.

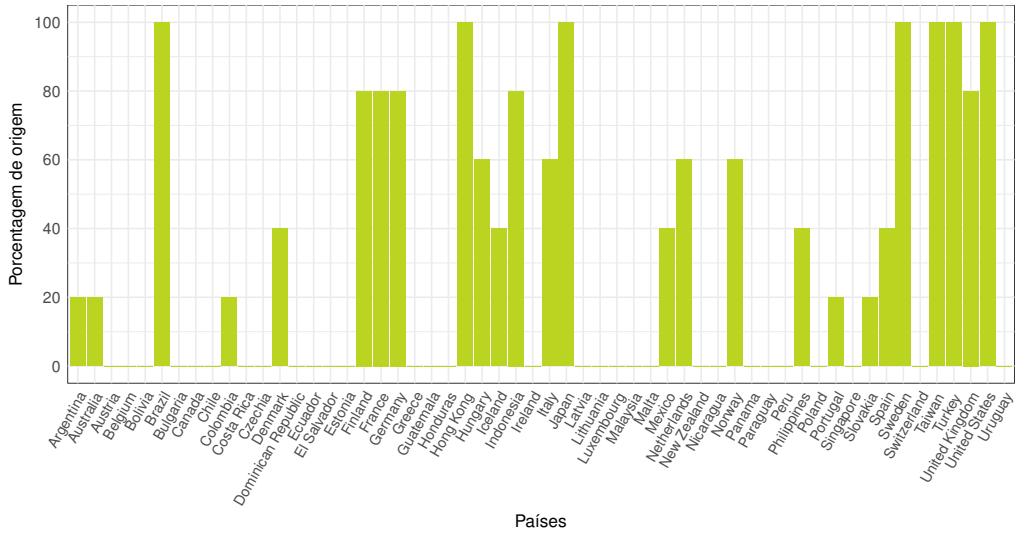


Figura 5. Proporção de gêneros que são originados em cada país.

Essa análise foi realizada para as Redes 1 e 2 mas, devido à similaridade dos resultados, optou-se por apresentar apenas um deles. Sendo assim, para cada país presente na Rede 1, os gêneros da rede foram comparados com a relação de gêneros e países da plataforma Every Noise. A Figura 5 mostra o resultado dessa comparação, onde 100% indica que os 5 gêneros de um país são de origem dele. Dos 57 países da Rede 1, apenas 12 possuem pelo menos 80% de gêneros considerados originais. Ou seja, pode-se dizer que, elimnando os gêneros universais existe pouca influência de outros países no estilo musical desse conjunto de 12 países. Por outro lado, um grande conjunto de países são

completamente influenciados ou importam gêneros musicais dos demais. O resultado desse tipo de análise pode também estar relacionado à produção musical própria de um determinado país. Em outras palavras, países que possuem uma indústria musical bem estabelecida podem estar produzindo um conteúdo mais voltado às suas questões culturais e costumes. Um exemplo disso é o Brasil, onde o forró, o *funk* carioca e o sertanejo e suas variações foram classificados como os gêneros mais escutados. Esse tipo de análise permite que estudos futuros busquem entender como ocorre a influência e o fluxo de estilos musicais entre países, identificar quais são os países mais influentes e explorar a estrutura da produção musical em cada país, por exemplo.

4. Trabalhos relacionados

O trabalho proposto por [Lee and Cunningham 2012] estuda o fluxo de música entre diferentes regiões do mundo, utilizando dados da plataforma de *streaming* de música Last.fm. O estudo busca entender quais são as preferências musicais de cada região e determinar os agrupamentos entre elas. São estabelecidas relações do tipo líder-seguidor entre os nós, permitindo entender qual é a dinâmica da preferência musical na rede. No entanto, como visto neste trabalho, a existência da grande quantidade de gêneros é um fator que influencia a estrutura de redes que visam estudar preferências músicais entre países. Em [Lee and Cunningham 2012] não fica claro qual foi a quantidade de gêneros considerada para o estudo e se houve ou não algum tipo de filtragem.

Em [Bryan and Wang 2011], a partir de dados da plataforma WhoSampled, são classificados os gêneros e artistas mais influentes no mercado musical de *sampling*. *Sampling* diz respeito à seleção de uma amostra de gravação de som a fim de reutilizá-la em uma outra gravação diferente. Sendo assim, um registro no conjunto de dados consiste em um artista que teve sua música amostrada e o artista que utilizou a amostra, incluindo metadados como o gênero. Esse tipo de dado possibilitou a construção de três redes direcionadas de músicas, artistas e gêneros. Em comparação com o presente trabalho, a abordagem em questão não aborda a grande diversidade de gêneros e difere no que diz respeito à estrutura da rede e ao tipo de dado que foi analisado.

A diversidade de gêneros é abordada em [Lambiotte and Ausloos 2006] com o foco no comportamento dos usuários em geral, e não nos países, como foi feito neste trabalho. São utilizados dados do Last.fm a fim de identificar padrões coletivos a partir das músicas escutadas. O trabalho identificou um total de 142 gêneros distintos, divididos em grupos musicais. A partir da correlação entre os gêneros, obteve-se um mapa onde foi possível observar que determinados gêneros musicais tendem a ser escutados pelos mesmos usuários. Em [Lambiotte and Ausloos 2005] a estrutura dessa mesma rede é explorada de forma aleatória a fim de mapear a estrutura interna da rede e as correlações em uma série temporal. A análise estatística dessa série permitiu identificar comportamentos não-triviais de usuários sugerindo, por exemplo, que eles se tornam mais ecléticos musicalmente.

A busca por comunidades em redes de música é também estudada em [Smith 2006] e [Gleiser and Danon 2003]. No entanto, ambos os trabalhos levam em consideração gêneros musicais específicos. Em [Smith 2006] é analisada a estrutura da rede de colaboração entre cantores de *rap* e em [Gleiser and Danon 2003] é analisada a rede de cantores de *jazz*. Neste último foi possível identificar a segregação racial entre as

comunidades de artistas encontradas.

Em contraste, o presente trabalho apresentou uma visão da rede de gêneros musicais escutados ao redor do mundo a partir dos dados do Spotify. O Spotify se destaca pela sua popularidade e pelo seu esforço em aprimorar a experiência do usuário no que diz respeito ao consumo de conteúdo musical. [Vlegels and Lievens 2017] destaca que estudos desse tipo de rede podem não refletir a possibilidade de mudanças nas características dos gêneros. No entanto, o Spotify classifica os gêneros musicais através de algoritmos que utilizam informações específicas das músicas, tais como força da batida e energia. Sendo assim, a estrutura da rede analisada reflete a grande quantidade de novos gêneros descobertos pela plataforma e não exclui as características intrínsicas a cada um deles.

5. Conclusões e trabalhos futuros

Neste trabalho, foi construída uma rede de países e gêneros com base no registro de músicas mais escutadas em cada país disponibilizado pelo Spotify. O modelo de rede foi estudado a partir de dois conjuntos de dados, coletados através da plataforma em duas datas diferentes. Dessa forma, demonstramos que a metodologia utilizada para as análises não se restringe à uma data específica. A análise da projeção de países conectados por gêneros em comum permitiu identificar o impacto que a quantidade de gêneros exerce sobre a estrutura da rede, com base nos conjuntos de dados utilizados. Gêneros considerados universais e virais foram responsáveis por conectar todos os países entre si.

Uma vez que o trabalho teve como objetivo entender o estilo musical dos países, foi feita uma filtragem que buscou excluir os gêneros universais e, a partir disso, considerar apenas os 5 gêneros mais frequentes em cada país. Com essa nova configuração de rede, foi possível classificar os gêneros não universais segundo métricas de centralidade de redes. A classificação mostra que, mesmo com a exclusão dos gêneros universais, esses mesmos gêneros continuam presentes através de suas variações (ou subgêneros), demonstrando sua importância no cenário musical como um todo. Além disso, com a aplicação do método de Louvain foram identificadas comunidades de países que compartilham interesses musicais em comum. Para os dois conjuntos de dados utilizados, foi interessante observar o surgimento de uma comunidade que agrupou países de língua espanhola e castelhana. Esse tipo de comunidade destaca que, apesar da língua falada nesses países não ter sido um atributo levado em consideração no conjunto de dados, esses países tendem a ter estilos musicais parecidos. Trabalhos futuros poderiam incluir as características específicas dos gêneros, como ocorre em sistemas de recomendação e classificação [Shakya et al. 2017], a fim de se obter uma análise mais detalhada sobre as relações musicais entre países.

Uma das análises do trabalho identificou qual a proporção de gêneros escutada por um país é de fato dele. Essa análise mostra que para um grande conjunto de países, os gêneros são provenientes de outros. Um trabalho futuro pode analisar como ocorre esse fluxo de gêneros entre países. Também como trabalho futuro, observa-se a possibilidade de analisar a rede com foco nos artistas. Essa perspectiva permitirá, por exemplo, identificar quais são os artistas responsáveis pelo surgimento de determinados gêneros nos países e classificar os artistas mais influentes tanto global quanto regionalmente. Pode-se ainda estudar a variação da rede deste trabalho no tempo, uma vez que o registro das músicas mais escutadas em cada país é feito diariamente. Com isso, seria possível entender o

comportamento de gêneros que, numa primeira análise não são universais, mas pelo fator viral de determinada música passam a conectar todos os países.

Agradecimentos

Agradecemos o suporte da CAPES, CNPq, FAPERJ e FAPESP.

Referências

Barabási, A.-L. (2016). *Network science*. Cambridge University Press.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008.

Bryan, N. J. and Wang, G. (2011). Musical influence network analysis and rank of sample-based music. In *ISMIR*, pages 329–334.

DeNora, T. (2000). *Music in everyday life*. Cambridge University Press.

Gleiser, P. M. and Danon, L. (2003). Community structure in jazz. *Advances in complex systems*, 6(04):565–573.

IFPI (2017). IFPI facts and stats. `http://www.ifpi.org/facts-and-stats.php`. Acessado em: 01/09/2017.

Lambiotte, R. and Ausloos, M. (2005). Uncovering collective listening habits and music genres in bipartite networks. *Physical Review E*, 72(6):066107.

Lambiotte, R. and Ausloos, M. (2006). On the genre-fication of music: a percolation approach. *The European Physical Journal B-Condensed Matter and Complex Systems*, 50(1):183–188.

Lee, C. and Cunningham, P. (2012). The geographic flow of music. In *Proceedings of the International Conference on Advances in Social Networks Analysis and Mining*, pages 691–695. IEEE Computer Society.

Noise, E. (2017). Every Noise at Once. `http://everynoise.com/engenremap.html`. Acessado em: 01/09/2017.

Rentfrow, P. J. (2012). The role of music in everyday life: Current directions in the social psychology of music. *Social and personality psychology compass*, 6(5):402–416.

Shakya, A., Gurung, B., Thapa, M. S., Rai, M., and Joshi, B. (2017). Music classification based on genre and mood. In *International Conference on Computational Intelligence, Communications, and Business Analytics*, pages 168–183. Springer.

Smith, R. D. (2006). The network of collaboration among rappers and its community structure. *Journal of Statistical Mechanics: Theory and Experiment*, 2006(02):P02006.

Trefzger, T., Rose, M., Baccarella, C., and Voigt, K.-I. (2015). Streaming killed the download star! how the business model of streaming services revolutionizes music distribution. *Journal of Organizational Advancement (Online), Strategie and lnstitutional Studies*, 7(1):29–39.

Vlegels, J. and Lievens, J. (2017). Music classification, genres, and taste patterns: A ground-up network analysis on the clustering of artist preferences. *Poetics*, 60:76–89.

That's my jam! Uma Análise Temporal sobre a Evolução das Preferências dos Usuários em uma Rede Social de Músicas

Fabíola S. F. Pereira[1], Cláudio D. G. Linhares[1], Jean R. Ponciano[1], João Gama[2], Sandra de Amo[1], Gina M. B. Oliveira[1]

[1]Faculdade de Computação – Universidade Federal de Uberlândia, Brasil

[2]INESC TEC – Universidade do Porto, Portugal

{fabiola.pereira,claudiodgl,jean,deamo,gina}@ufu.br, jgama@fep.up.pt

Abstract. *User preferences are fairly dynamic, since users tend to exploit a wide range of information and modify their tastes accordingly over time. Existing models that capture the evolution of preferences in the music domain are very restricted and do not take into account social influence. This work proposes an analysis on the evolution of musical preferences of the users of a social music network, using temporal networks of similarity. We found that few users significantly vary their preferences. The tendency is similar artists and users maintaining their similarities over time.*

Resumo. *As preferências dos usuários são muito dinâmicas, uma vez que eles tendem a explorar uma vasta gama de informações e modificar seus gostos ao longo do tempo. Modelos existentes que capturam a evolução de preferências no domínio de músicas são muito restritos e não levam em conta a influência social. Este trabalho propõe uma análise sobre a evolução das preferências musicais dos usuários de uma rede social de músicas, utilizando redes temporais de similaridade. Como resultado, foi detectado que poucos usuários variam significativamente suas preferências. A tendência é que artistas e usuários semelhantes mantenham suas similaridades ao longo do tempo.*

1. Introdução

O que direciona a evolução das preferências dos usuários? Entender como essas preferências evoluem é uma das tarefas de personalização mais importantes nos contextos de recuperação da informação e sistemas de recomendação. Quanto mais se conhece sobre o usuário, melhor será a qualidade dos itens recomendados a ele. De fato, a observação de como um usuário modifica seus gostos e costumes ao longo do tempo tem sido explorada em diferentes domínios, como filmes [Siddiqui et al. 2014], leitura de notícias [Pereira et al. 2016] e visões políticas [Macropol et al. 2013]. No domínio da música o cenário não é diferente. Estudos na área de psicologia indicam que o gosto musical das pessoas muda e evolui ao longo do tempo [Bonneville-Roussy et al. 2013].

Nesse contexto de evolução de preferências musicais, os trabalhos em geral propõem observações sobre as variações dos gêneros musicais e artistas com base em atributos como popularidade e periodicidade de audição [Kapoor et al. 2013]. Uma outra linha de pesquisa busca por modelos de aprendizado da evolução das preferências musicais levando em conta a sequência de músicas ouvidas em *playlists* [Moore et al. 2013]. Nenhum desses estudos, entretanto, levam em consideração a influência social que o usuário

sofre durante o processo de evolução das suas preferências musicais. De acordo com a psicologia, influência social é quando o comportamento de uma pessoa faz com que outra pessoa mude de opinião ou execute uma ação que, de outro modo, não executaria [Michener 2005]. As redes sociais são instrumentos propícios para a ocorrência de tal fenômeno, pois elas naturalmente facilitam a construção de relações entre pessoas que compartilham os mesmos interesses. Nelas, usuários conectados trocam informações sobre novas tendências, preferências de consumo e opiniões entre seus pares. Em especial, analisar a evolução dessas redes torna-se interessante para a compreensão temporal de tais relações sociais [Aggarwal and Subbian 2014].

Tendo como motivação o cenário descrito, neste trabalho é investigada a evolução do gosto musical dos usuários da rede social de músicas *This Is My Jam*[1] (TIMJ), utilizando técnicas de análise de redes sociais temporais [Holme 2014]. A proposta é modelar duas redes temporais de similaridade – uma entre artistas e outra entre usuários, e analisá-las separadamente. As redes temporais de similaridade são aquelas em que os nós estão ligados com base em alguma semelhança entre eles em um determinado momento. Similaridade entre artistas significa que dois artistas são semelhantes se há uma grande intersecção entre seus públicos. No contexto de usuários, dois usuários similares são aqueles que compartilham do mesmo gosto musical. Entender como esse grau de semelhança se altera ao longo do tempo nessas redes leva à compreensão da evolução das preferências dos indivíduos.

O objetivo principal do trabalho é verificar se as preferências musicais dos usuários realmente mudaram ao longo do tempo. Para tanto, as seguintes perguntas de pesquisa serão respondidas: (1) artistas semelhantes em relação ao seu público mantêm a semelhança ao longo do tempo? (2) Usuários com mesmo gosto musical evoluem de maneira similar compartilhando das mesmas preferências? A originalidade deste trabalho está na proposta do uso de redes temporais de similaridade para análise da evolução das preferências dos usuários sobre dados no domínio de música.

2. Trabalhos Correlatos

Evolução das preferências. A maioria dos trabalhos que tratam sobre a evolução das preferências do usuário têm como objetivo a recomendação de itens com base na variação dos perfis de cada usuário [Felício et al. 2016]. No domínio de músicas, o trabalho [Moore et al. 2013] desenvolveu um modelo probabilístico que incorpora o tempo para explorar como as preferências musicais de uma população evoluem. A ideia é representar usuários e músicas em um espaço Euclidiano, no qual suas posições mudam ao longo do tempo, demarcando a trajetória da evolução. Em [Summers and Popp 2015] os autores buscam por padrões de músicas ouvidas de acordo com diferentes contextos. O objetivo é identificar uma relação por exemplo, entre músicas natalinas e a época de Natal, com foco na predição da próxima música a ser tocada em uma *playlist*. São trabalhos de motivação similar a esta proposta, porém não utilizam técnicas de análise de redes temporais.

Análise de Redes Temporais. De acordo com [Aggarwal and Subbian 2014] existem duas maneiras de analisar uma rede que evolui ao longo do tempo: através de métodos de manutenção e através de métodos analíticos. No primeiro, é desejável manter os resultados de um processo de mineração de dados continuamente ao longo do tempo. Por

[1] https://www.thisismyjam.com/

exemplo, em tarefas como predição de *links* e detecção de comunidades (clusterização) deseja-se sempre manter os modelos atualizados e com boa acurácia à medida que a rede evolui. No segundo, a ideia é quantificar e entender as mudanças que ocorreram na estrutura da rede. Tais modelos estão focados em analisar a mudança, ao invés de apenas se ajustarem a ela. Este trabalho adéqua-se a essa segunda maneira de analisar uma rede em evolução – por meio do entendimento das mudanças.

Redes de Similaridade. Modelar a similaridade entre pessoas na forma de rede não é uma tarefa nova, tendo sido aplicada nos domínios de percepção visual em pinturas [Felício et al. 2016] e co-autoria de artigos científicos [Silva et al. 2015]. Em [Cano 2004] são analisadas redes de similaridade entre artistas no domínio da música, porém a evolução dessas redes não é levada em conta.

3. Modelagem das Redes

A análise foi conduzida sobre a rede social de músicas *This Is My Jam* (TIMJ). Nela, os usuários podem compartilhar suas músicas favoritas com seus seguidores. Apenas uma música pode ser compartilhada por vez - o *jam* atual, que dura no máximo uma semana no *status* dos usuários. Além disso, os usuários podem curtir o *jam* de outros usuários. A base de dados da TIMJ foi liberada por [Jansson et al. 2015] e contém 219.940 artistas, 132.299 usuários e 2.095.441 *jams*, referente ao período de 26/08/2011 a 26/09/2015. As análises apresentadas neste artigo levam em consideração períodos trimestrais, totalizando 17 instantes de tempo. Foram geradas duas redes de similaridade, descritas a seguir.

3.1. Rede de Similaridade entre Artistas (*art-art*)

Nessa rede os nós são os artistas e as arestas ocorrem entre artistas que interpretam as músicas compartilhadas como *jam* por um usuário em comum em determinado intervalo de tempo. O peso das arestas é definido pelo número de usuários em comum. Formalmente, tem-se: $\mathcal{A} = (V, E)$, onde $\mathcal{A}$ é uma rede não-dirigida, ponderada, sendo V o conjunto de artistas, e cada aresta $e = (u, v, t) \in E$ para $u, v \in V$, indica que u e v são intérpretes de *jams* compartilhados no instante t por um mesmo usuário. A função peso é definida por $w : E \rightarrow \mathbb{N}$, sendo $w(e)$ o número de usuários em comum entre u e v em t. Foi definido empiricamente um limiar de similaridade $\alpha = 20$ que atua como um filtro sobre a rede, onde $E = \{e|w(e) > \alpha\}$.

3.2. Rede de Similaridade entre Usuários (*usr-usr*)

Na rede não-dirigida ponderada *usr-usr* os nós são os usuários e as arestas indicam a similaridade entre eles com base nos artistas em comum que foram compartilhados como *jam* em um determinado instante de tempo. O peso das ligações representa a força da similaridade (*tie strength* [Brandao and Moro 2017]) entre os nós, calculado em função da quantidade de artistas em comum que dois usuários compartilharam ao mesmo tempo. Formalmente, tem-se $\mathcal{U} = (V, E)$, onde $\mathcal{U}$ é um grafo não-dirigido, ponderado, sendo V o conjunto de usuários, e cada aresta $e = (u, v, t) \in E$ indica que u e v são usuários que compartilharam o mesmo artista como *jam* no tempo t. A função peso é definida por $w : E \rightarrow \mathbb{N}$, sendo $w(e)$ o número de artistas compartilhados em comum entre u e v em t. Também foi definido um limiar de similaridade $\alpha = 100$ para filtro sobre a rede, onde $E = \{e|w(e) > \alpha\}$.

A Tabela 1 resume as características das redes implícitas analisadas neste trabalho, extraídas da base de dados TIMJ. Na Tabela 2 estão as propriedades que garantem que as redes geradas são redes do mundo real [Zafarani et al. 2014], com distribuição de graus seguindo a lei de potência, alto coeficiente de clusterização e baixo comprimento médio de caminho entre os nós.

Tabela 1. Descrição das redes implícitas geradas a partir da rede social de músicas *This Is My Jam*.

Rede	# nós	# arestas	Semântica	Limiar de Similaridade (α)
art-art	1670	3588	similaridade entre artistas	20
usr-usr	4388	9567	similaridade entre usuários	100

Tabela 2. Descrição das propriedades que garantem que as redes implícitas de similaridade geradas a partir da rede social de músicas são redes do mundo real.

Propriedade	*art-art*	*usr-usr*
distribuição dos graus	lei de potência	lei de potência
coeficiente de clusterização local médio	0,548	0,743
comprimento médio do caminho	4,57	3,849

3.3. Relação entre Redes de Similaridade e Preferências do Usuário

A ideia de investigar as preferências do usuário utilizando similaridade baseia-se no fato de que uma mudança na similaridade é um indício de que os usuários estão mudando seus gostos musicais. Do ponto de vista dos artistas, os nós densamente conectados indicam que aqueles artistas são ouvidos pelos mesmos usuários e, portanto, são similares em relação ao seu público. Com o passar do tempo, os artistas podem deixar de ser, tornarem-se ou manterem-se similares. Quanto mais mudanças nessa similaridade, mais os usuários estão mudando suas preferências. De fato, a rede *art-art* provê uma visão da evolução das preferências do usuário de uma maneira global. Os usuários podem, por exemplo, mudar suas preferências com base no lançamento de novos artistas ou músicas ou com base em acontecimentos como falecimento ou indicação para prêmios.

Em relação à similaridade entre usuários, a intuição da rede *usr-usr* é que os usuários estão conectados com base na semelhança entre suas preferências musicais. Uma mudança estrutural da rede pode indicar que os usuários estão deixando de ter gostos parecidos e tornando-se similares a outros. É uma visão do usuário em relação aos demais membros da comunidade musical, portanto uma evolução de maneira local. Uma pessoa pode mudar suas preferências com base em influências recebidas do meio externo no seu dia-a-dia, deixando de ser similar aos seus pares até então.

4. Metodologia

A análise temporal das redes de similaridade foi baseada em *snapshots* de cada instante de tempo (trimestre). Logo, não é uma abordagem incremental, pois cada métrica foi obtida baseada em apenas um instante de tempo, esquecendo o passado.

4.1. Métricas de Rede

Foram definidas as métricas ponte, influência e versatilidade para análise local dos nós das redes. Cada nó (artista ou usuário) tem o seu comportamento em relação a essas métricas. Também foi analisada a maneira como os nós evoluem em relação às suas vizinhanças ou comunidades. Tais definições são descritas a seguir.

1. **Comunidades**: foi utilizada a métrica *modularity* [Newman 2006] para detecção de comunidades em redes. Ela indica o quanto a estrutura de comunidade encontrada foi criada aleatoriamente. Assim, deseja-se dividir a rede em partições de modo que se a probabilidade de dois nós se conectarem aleatoriamente for baixa, eles devem estar na mesma partição e vice-versa [Zafarani et al. 2014]. No contexto de similaridade, tem-se que dentro de cada partição encontrada pela métrica estão os nós mais similares entre si.
2. **Ponte:** os "nós ponte" são aqueles que ligam duas ou mais comunidades, mantendo a rede conexa. A métrica utilizada foi a centralidade *betweenness* [Brandes 2001]. Essa medida indica o quão importante o nó é para conectar outros nós. Em redes de similaridade, um nó com alto valor de ponte é aquele que concentra características dos nós das comunidades que ele conecta.
3. **Influência:** um nó influente é aquele importante para os nós da sua vizinhança. A métrica *PageRank* [Brin and Page 1998] foi utilizada para detectar a influência de um nó nas redes de similaridade. Um nó com alto grau de influência na rede de similaridade tem interferência sobre seus pares, que por sua vez também têm alto grau de influência e assim por diante. Na similaridade entre artistas, por exemplo, um artista influencia demais artistas parecidos com ele.
4. **Versatilidade:** a versatilidade indica o quão próximo um nó está dos demais nós da rede. A métrica de centralidade *closeness* [Brandes 2001] foi utilizada. Quanto maior o *closeness*, mais similar é o nó em relação aos demais nós da rede, representando portanto um "nó eclético".

4.2. Análise e Avaliação da Evolução das Redes

Uma primeira análise da evolução foi conduzida baseada em *insights* obtidos por meio da visualização das redes. De fato, técnicas de visualização de redes temporais têm sido propostas [Beck et al. 2017] e auxiliam na compreensão do problema em análise. O *layout* temporal gerado pela ferramenta DyNetVis [Linhares et al. 2017] foi utilizado, bem como as visualizações estruturais geradas pela ferramenta Gephi [Bastian et al. 2009].

As comunidades foram avaliadas com base em nós selecionados. Dado um nó $v \in V$, para cada instante de tempo t, sua comunidade C_t^v foi detectada. Depois, a persistência de tal comunidade foi obtida por meio do coeficiente de Jaccard:

$$com(v) = \frac{|C_1^v \cap ... \cap C_n^v|}{|C_1^v \cup ... \cup C_n^v|} \tag{1}$$

onde $n = 17$ corresponde à quantidade de instantes de tempo analisada. Quanto mais próximo de 1, maior é a persistência da comunidade de v.

Por fim, a avaliação da evolução de todos os nós das redes ocorreu em função da variância $var_c(v)$ de cada nó $v \in V$ em relação a uma métrica de centralidade local c durante n instantes de tempo, definida como:

$$var_c(v) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \quad (2)$$

onde n é 17 representando a quantidade de instantes de tempo e x é a distribuição da centralidade c do nó v.

5. Resultados Obtidos

As redes de similaridade entre artistas e usuários foram analisadas de forma que a evolução de cada uma leva à conclusão sobre a evolução das preferências dos usuários de maneira global e local, respectivamente.

5.1. Como foi a evolução dos artistas?

Uma visão estrutural, estática, da rede *art-art* é dada na Figura 1(a). É possível observar que não é uma rede densa e que é composta por diversas comunidades, em sua maioria, pequenas.

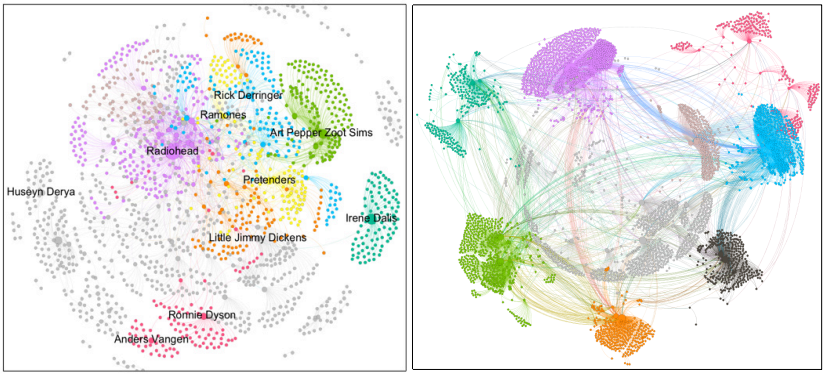
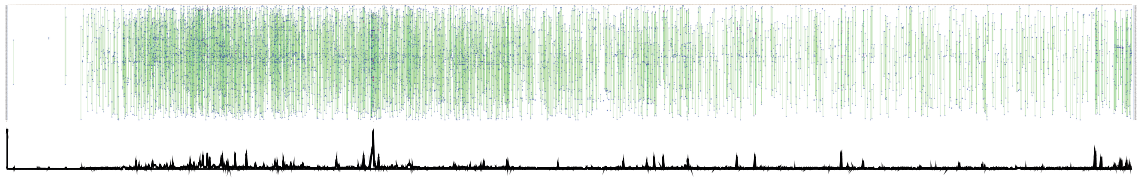


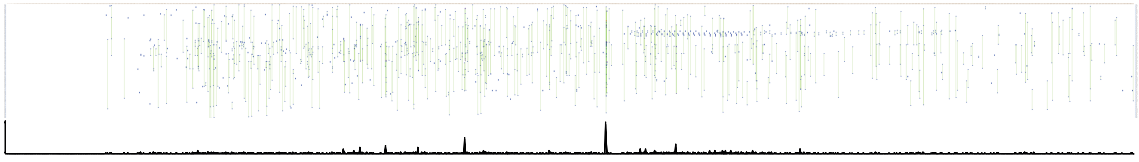
Figura 1. Visão geral das redes *art-art* (a) e *usr-usr* (b). As cores indicam a comunidade e o tamanho dos nós indica a influência. Alguns dos artistas mais influentes estão em destaque.

A visualização temporal da rede de similaridade entre artistas, mostrada na Figura 2(a), ilustra a intensidade de comunicação em relação aos 200 nós mais populares ao longo do tempo. No *layout* temporal, os nós são posicionados no eixo vertical e cada instante de tempo no eixo horizontal ilustra as arestas existentes entre os nós naquele momento [Linhares et al. 2017]. É uma perspectiva de atividade temporal na rede. Popularidade [Zafarani et al. 2014], no domínio dos artistas, indica aqueles que mais apareceram como *jam* nos perfis dos usuários considerando todo o período. São os artistas mais curtidos. Pode-se perceber uma interação intensa entre os trimestres 2 e 8 na rede de similaridade, ou seja, uma grande quantidade de artistas populares foram ouvidos pelos mesmos usuários.

Desconsiderando a evolução temporal, a Tabela 3 lista os 10 artistas que se destacam na rede estática, considerando todo o período de aproximadamente 4 anos. É uma visão geral da rede de similaridade *art-art*, que não provê informação sobre a evolução temporal de tais artistas. É interessante observar que os artistas *Radiohead*, *David Bowie*



(a) Rede *art-art*



(b) Rede *usr-usr*

Figura 2. Visualização temporal das redes de similaridade considerando os 200 nós mais populares de cada domínio, durante aproximadamente 4 anos com granularidade diária. Visualização gerada pelo software DyNetVis [Linhares et al. 2017].

e *R.E.M.* são aqueles mais centrais em relação a todas as métricas calculadas. Em termos de preferências, a análise estática leva à conclusão de que são os artistas mais preferidos da comunidade de músicas TIMJ.

Tabela 3. Top-10 artistas com maiores valores médios das métricas locais. Artistas que se repetiram em todas as métricas estão em destaque.

	Ponte	Influência	Versatilidade
1	**Radiohead**	**Radiohead**	**Radiohead**
2	Art Pepper n Zoot Sims	Irene Dalis	**David Bowie**
3	Ramones	**David Bowie**	The Beatles
4	Queen	**R.E.M.**	Queen
5	Pretenders	Little Jimmy Dickens	The Rolling Stones
6	Rick Derringer	Art Pepper n Zoot Sims	Bob Dylan
7	Pale Saints	Ramones	Nick Cave
8	**David Bowie**	Beastie Boys	**R.E.M.**
9	**R.E.M.**	Joe McPhee Quartet	The Kinks
10	Billie Holiday	Pretenders	Nina Simone

A Tabela 4 ilustra a evolução dos top-10 artistas com maior variância das métricas locais. A primeira conclusão é que as métricas são complementares e não necessariamente a variância em relação a uma, leva à variância em relação às demais. Os artistas que mais variaram em relação à versatilidade não variaram em relação à influência e ponte. Já entre ponte e influência houve uma sobreposição de 5 artistas dentre os 10 que mais variaram. A variância da centralidade para cada um dos artistas na Tabela 4 mostra que as preferências dos usuários que os escutaram oscilaram ao longo do tempo. Em determinados momentos, *Ramones* e *Pretenders*, por exemplo, foram muito influentes, em outros deixaram de ser.

As comunidades na rede de similaridade entre artistas representam os grupos de artistas que são ouvidos pelos mesmos usuários. Pelos *insights* visuais, dois padrões de evolução de comunidades foram detectados na rede. O primeiro corresponde às comunidades que existiram durante todo o tempo de observação ($\sim$ 4 anos), ou seja, uma quantidade representativa de nós que faziam parte de uma comunidade no primeiro instante de tempo permaneceu conectada durante os demais instantes. O segundo refere-

Tabela 4. Top-10 artistas com maior variância para cada uma das métricas locais. Artistas que se repetiram em pelo menos duas métricas estão em destaque.

	Ponte	Influência	Versatilidade
1	**Art Pepper n Zoot Sims**	Irene Dalis	Ronnie Dyson
2	**Ramones**	Little Jimmy Dickens	Dire Straits
3	**Pretenders**	**Art Pepper n Zoot Sims**	Kanye West
4	**Radiohead**	**Radiohead**	Roberta Flack
5	**Rick Derringer**	Beastie Boys	Aphex Twin
6	Pale Saints	Joe McPhee Quartet	Flower Travellin'band
7	Queen	**Pretenders**	Refused
8	Billie Holiday	Ricky Nelson	Self
9	The Damned	**Ramones**	Nicki Minaj
10	R.Seiliog	**Rick Derringer**	Rocket Ship

se às comunidades que surgiram e desapareceram ao longo da evolução. A Figura 3 ilustra a evolução de duas comunidades durante três trimestres. A primeira, cujo nó base é *Radiohead*, possui o comportamento persistente durante a evolução da rede, com $com(Radiohead) = 0,76$. A segunda, com *Ricky Nelson* como artista base, apenas se tornou volumosa no instante $t = 15$, sendo $com(RickyNelson) = 0,13$.

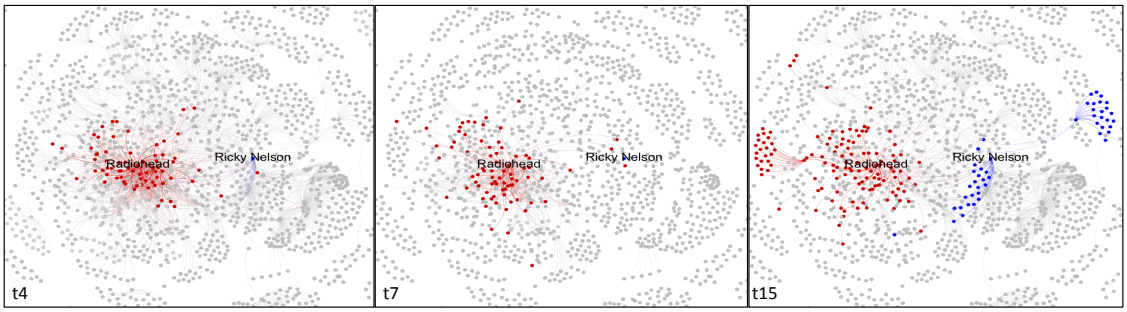


Figura 3. Evolução de duas comunidades a que pertencem *Radiohead* e *Ricky Nelson* em três instantes de tempo diferentes. *Radiohead* pertence a uma comunidade persistente. *Ricky Nelson* pertence a uma comunidade efêmera.

A Figura 4 ilustra a evolução da influência de determinados artistas. É um *heatmap* no qual quanto mais escura a cor, maior a influência. É possível observar que a influência de artistas como *Radiohead* e *Arctic Monkeys* se manteve alta na maior parte do tempo. Entretanto, para os demais artistas, o comportamento detectado é que em geral o pico de influência dura de 2 a 3 trimestres. É uma observação complementar àquela obtida pela análise da evolução das comunidades.

Por fim, o comportamento da variância de todos os nós da rede para cada uma das métricas é mostrado na Figura 5. A média das variâncias tende a zero, ou seja, a maioria dos artistas manteve seus valores de centralidade ao longo do tempo, com baixos índices de variação. A distribuição da quantidade de artistas e suas respectivas variâncias é mostrada nos histogramas. Os top-10 artistas na Tabela 4 estão entre os *outliers* dos *boxplots*. A conclusão é que as preferências dos usuários não variaram em relação à grande maioria dos artistas. Artistas que eram ouvidos por determinados usuários em comum no início, permaneceram sendo ouvidos pelos mesmos usuários em comum.

5.2. Como foi a evolução dos usuários?

A rede de similaridade entre usuários possui características de evolução diferentes da rede de similaridade entre artistas. É uma rede mais densa, porém com menos comunidades,

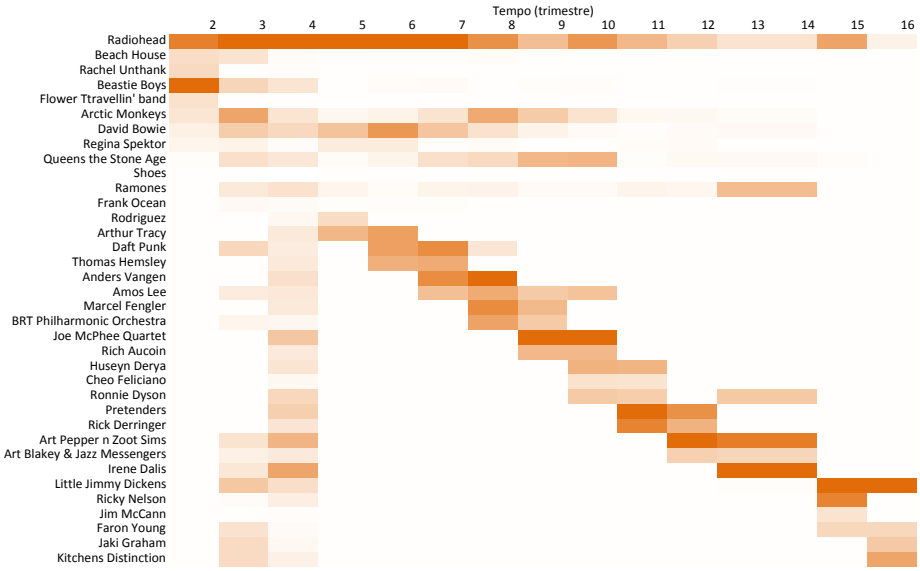


Figura 4. *Heatmap* baseado na influência dos artistas. Quanto mais escuro, maior a influência. Os artistas da figura são aqueles que em pelo menos um instante de tempo estiveram dentre os 5 mais influentes.

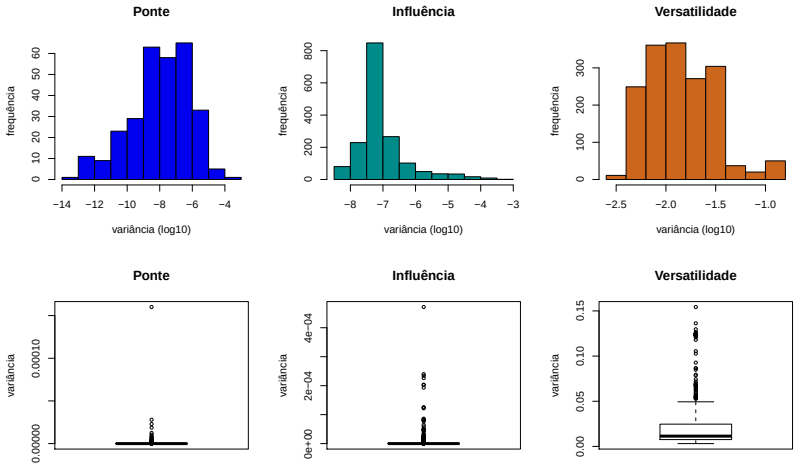


Figura 5. Comportamento da variância dos artistas durante o período observado em relação às métricas de ponte, influência e versatilidade.

sendo estas mais bem definidas de acordo com os *insights* de visualização (Figura 1(b)).

No contexto de usuários, popularidade foi definida em função do número de seguidores que os usuários possuem. Quanto mais seguidores, mais popular é aquele usuário. Observando a Figura 2(b), a rede *usr-usr* é esparsa quanto à interação entre os usuários populares, que, consequentemente, não são similares quanto às suas preferências musicais. A conclusão é que usuários populares formam comunidades de usuários similares a eles e que essas comunidades em geral não se misturam.

A evolução consiste em, a cada trimestre, uma comunidade se destaca com grande volume de interações entre os nós pertencentes a ela, com um nó central sempre em destaque quanto às três métricas investigadas. Tal comportamento pode ser observado

pelo *heatmap* na Figura 6, que explora a evolução da métrica de influência para alguns usuários. Os usuários destacados são aqueles que em pelo menos um trimestre estiveram dentre os 5 mais influentes. Diferentemente da rede de artistas, em todos os trimestres, cada um dos 5 usuários mais influentes pertencem a comunidades distintas. A conclusão é que as comunidades nessa rede são efêmeras, ou seja, não existe um grupo de usuários similares que perdura durante todo o período de observação. Além disso, os usuários não estiveram ativos o tempo todo na comunidade. Os picos de interação mostram que a participação na rede social, compartilhando artistas favoritos, é sazonal.

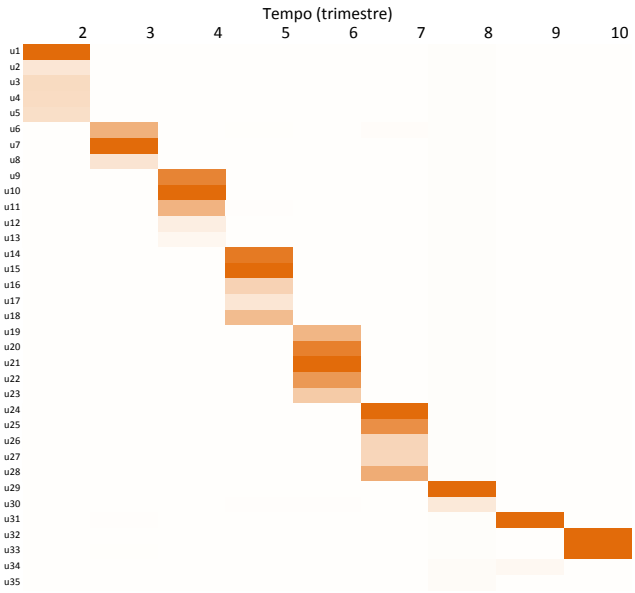


Figura 6. *Heatmap* baseado na influência dos usuários. Quanto mais escuro, maior a influência. Os usuários da figura são aqueles que em pelo menos um trimestre estiveram dentre os 5 mais influentes. Para todos os trimestres, os 5 mais influentes pertencem a comunidades diferentes.

O comportamento da variância entre as centralidades de toda a rede *usr-usr* é mostrado na Figura 7. Assim como na rede de artistas, na rede de usuários, a variância da maioria dos nós não foi significativa, tendendo a zero. Quando os usuários estiveram ativos na comunidade, eles se mantiveram similares. Ou seja, os mesmos usuários escutaram os mesmos artistas em comum durante o período de observação. A conclusão é que não houve variação nas preferências da maioria dos usuários da rede, uma vez que mantiveram suas similaridades.

5.3. Discussão

É importante ressaltar que os resultados obtidos possuem as seguintes limitações: (i) granularidade de tempo escolhida (trimestre). A variação da granularidade pode levar a conclusões diferentes. (ii) O limiar de similaridade (α) escolhido empiricamente. Quanto menor o limiar, mais conexões existirão na rede e vice-versa. Foi conduzida uma variação empírica de 10 valores diferentes para cada rede, na qual todas as redes geradas continuaram com propriedades do mundo real [Zafarani et al. 2014]. Como as análises não investigaram variações menores, focando apenas no comportamento das massas, o limiar não restringe as conclusões obtidas. (iii) Existência de muitos dados nulos, representando falta de atividade dos usuários. É característica da rede TIMJ a existência de usuários

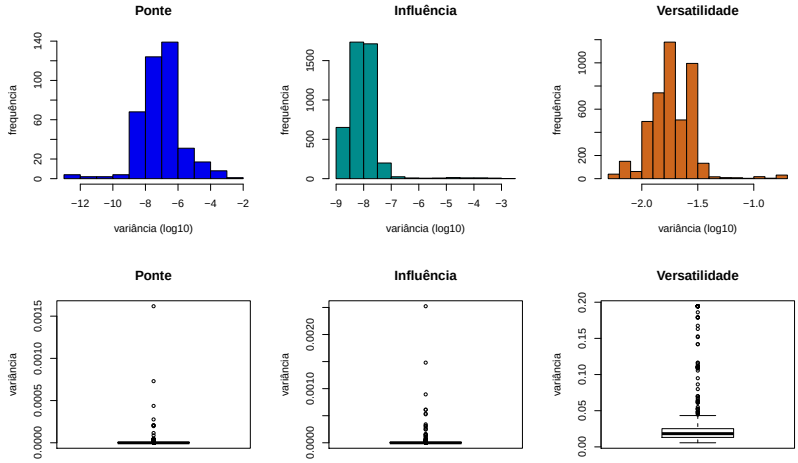


Figura 7. Comportamento da variância dos usuários durante o período observado em relação às métricas de ponte, influência e versatilidade.

sazonais, que não registraram atividades na rede durante todo o período de observação. Tais usuários podem ter mudado suas preferências musicais, porém deixaram de utilizar a TIMJ.

6. Conclusão

Este trabalho é uma contribuição no campo de ciência de dados aplicada: uma análise temporal sobre uma base de dados utilizando técnicas de mineração e análise de redes sociais. Foi conduzida uma análise sobre a evolução das preferências musicais dos usuários de uma rede social de músicas. A proposta foi modelar o domínio através de redes temporais de similaridade. Foi detectado que a maioria dos artistas e usuários semelhantes mantiveram suas similaridades ao longo do tempo, levando à conclusão de que as preferências musicais da maioria dos usuários não mudaram significativamente. Os usuários são ecléticos em relação à diversidade de artistas que apreciam, porém tal diversidade é mantida ao longo do tempo. Como trabalho futuro, pode-se utilizar o modelo de similaridade proposto aplicado à recomendação de músicas. Além disso, usar outras métricas de rede e atributos do domínio como gêneros musicais são extensões naturais deste trabalho.

Agradecimentos

Este trabalho é apoiado pela agências brasileiras CNPq, CAPES e Fapemig. Os autores agradecem à Microsoft Azure pelo financiamento à pesquisa (*research sponsorship* 65c28dfb-a346-455b-a644-c847ff5ac284).

Referências

Aggarwal, C. and Subbian, K. (2014). Evolutionary network analysis: a survey. *ACM Computing Surveys*, 47(1):10–36.

Bastian, M., Heymann, S., and Jacomy, M. (2009). Gephi: An open source software for exploring and manipulating networks.

Beck, F., Burch, M., Diehl, S., and Weiskopf, D. (2017). A taxonomy and survey of dynamic graph visualization. *Computer Graphics Forum*, 36(1):133–159.

Bonneville-Roussy, A., Rentfrow, P., Xu, M., and Potter, J. (2013). Music through the ages: Trends in musical engagement and preferences from adolescence through middle adulthood. *J Pers Soc Psychol.*, 4(105):703–717.

Brandao, M. and Moro, M. (2017). Tie strength analysis: New metrics and open problems. *VI Brazilian Workshop on Social Network Analysis and Mining*, pages 682–687.

Brandes, U. (2001). A faster algorithm for betweenness centrality. *Journal of Mathematical Sociology*, 25:163–177.

Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. In *World Wide Web*, pages 107–117.

Cano, P. (2004). The emergence of complex network patterns in music artist networks. In *Int. Symposium on Music Information Retrieval*, pages 466–469.

Felício, C. Z., de Almeida, C. M. M., Alves, G., Pereira, F. S. F., Paixão, K. V. R., and de Amo, S. (2016). *29th Canadian Conference on Artificial Intelligence*, chapter Visual Perception Similarities to Improve the Quality of User Cold Start Recommendations.

Holme, P. (2014). Analyzing temporal networks in social media. *Proceedings of the IEEE*, 102(12):1922–1933.

Jansson, A., Raffel, C., and Weyde, T. (2015). This is my jam – data dump. *16th Int. Society for Music Information Retrieval Conference*.

Kapoor, K., Srivastava, N., Srivastava, J., and Schrater, P. (2013). Measuring spontaneous devaluations in user preferences. In *ACM SIGKDD KDD*, pages 1061–1069.

Linhares, C. D. G., Travençolo, B. A. N., Paiva, J. G. S., and Rocha, L. E. C. (2017). Dynetvis: A system for visualization of dynamic networks. In *SAC*, pages 187–194.

Macropol, K., Bogdanov, P., Singh, A. K., Petzold, L., and Yan, X. (2013). I act, therefore i judge: Network sentiment dynamics based on user activity change. In *Int. Conf. on Advances in Social Networks Analysis and Mining*, ASONAM '13, pages 396–402.

Michener, H. A. (2005). *Psicologia Social*. Cengage Learning.

Moore, J., Chen, S., Turnbull, D., and Joachims, T. (2013). Taste over time: the temporal dynamics of user preferences. In *Int. Society for Music Information Retrieval Conf.*

Newman, M. E. (2006). Modularity and community structure in networks. *Proc Natl Acad Sci U S A*, 103(23):8577–8582.

Pereira, F. S. F., de Amo, S., and Gama, J. (2016). On using temporal networks to analyze user preferences dynamics. In *Int. Conf. Discovery Science*, 408–423.

Siddiqui, Z. F., Tiakas, E., Symeonidis, P., Spiliopoulou, M., and Manolopoulos, Y. (2014). xstreams: Recommending items to users with time-evolving preferences. In *Int. Conf. on Web Intelligence, Mining and Semantics*, pages 22:1–22:12.

Silva, V., Sampaio, F., and Oliveira, J. (2015). Temporal analysis of co-authorship networks: A study on the interactions of authors in the brazilian journal of computing in education. *Brazilian Journal of Computers in Education (RBIE)*, 23(2).

Summers, C. and Popp, P. (2015). Temporal music context identification with user listening data. *16th International Society for Music Information Retrieval Conference*.

Zafarani, R., Abbasi, M. A., and Liu, H. (2014). *Social Media Mining: An Introduction*. Cambridge University Press, New York, NY, USA.

Uma análise das seleções da copa utilizando uma rede de transferências de jogadores entre países

Lucas G. S. Félix[1] **, Carlos M. Barbosa**[1] **, Iago A. Carvalho**[2]**,**
Vinícius da F. Vieira[1] **, Carolina Ribeiro Xavier**[1]

[1]Departamento de Ciência da Computação - Universidade Federal de São João del-Rei
Av. Visconde do Rio Preto S/N - Colônia do Bengo

[2]Departamento de Ciência da Computação - Universidade Federal de Minas Gerais

{lucasgsfelix, cmagnobarbosa}@gmail.com, iagoac@dcc.ufmg.br

{carolinaxavier, vinicius}@ufsj.edu.br

Abstract. *Football is the most popular sport in the world. The growth in the number of transactions of purchase and sale, marketing, sponsorships, sale of tickets, TV contracts, among other forms of monetization of football makes the flow of values increasingly higher. The majority of works related to this sport is associated with sociological analysis. This work proposes a study focused on the transactions occurred among the football teams classified to the World Cup 2018 using complex networks techniques for an analysis of the transfer of players among these countries.*

Resumo. *O futebol é hoje o esporte mais popular do mundo. O crescimento no número de transações de compra e venda, marketing, patrocínios, venda de ingressos, contratos de TV, entre outras formas de monetização do futebol faz com que o fluxo de valores seja cada vez maior. Grande parte dos trabalhos relacionados a esse esporte são associados a análises sociológicas. Neste trabalho, é proposto um estudo focado nas transações feitas entre as seleções classificadas para a Copa do Mundo 2018 utilizando técnicas de redes complexas para uma análise da transferência de jogadores entre esses países.*

1. Introdução

O futebol é hoje o esporte mais popular do mundo [Liebig et al. 2012, Palacios-Huerta 2004]. Devido a sua grande popularidade, gera um enorme fluxo financeiro. Somente na Europa, o futebol movimentou aproximadamente 25 bilhões de euros na temporada 2016/2017 [Deloitte 2016].

Um time de futebol possui diversas formas de compensação financeira inclusive: vendas de camisas, patrocinadores, cotas pagas por canais de TV, além de uma grande movimentação de valores através de transações de jogadores. Um clube pode comprar um jogador permanentemente ou então adquirir seus direitos temporariamente, através de empréstimo. Ambos modelos de transação envolvem uma compensação financeira [Liu XF 2016].

A Copa do Mundo FIFA é um evento esportivo que ocorre de 4 em 4 anos. Criado em 1928, teve sua primeira edição realizada em 1930 e, desde então, foram realizadas 20 edições. A vigésima primeira edição da copa do mundo será realizada na Rússia em 2018.

O objetivo deste trabalho é estudar as propriedades da rede de transferências de jogadores de futebol entre os países participantes da Copa do Mundo de 2018 a fim de extrair informações para entender a dinâmica dessas transferências. Para isso, além das características básicas da rede, serão utilizados algoritmos de ranking e de detecção de comunidades em redes complexas.

Existem diversos estudos que abordam o futebol como tema principal, entretanto, apesar da importância e do tamanho do mercado de transferências de jogadores, existem poucos estudos sobre o mesmo, sendo que grande parte aborda somente o lado sociológico das transferências de jogadores, usando o futebol para ilustrar o movimento internacional de forças de trabalho [Maguire 1994, Maguire and Pearton 2000, Poli 2010, Roderick 2013].

No trabalho [Poli 2010], o autor mostra a globalização através das transações feitas no mercado do futebol. Nos trabalhos [Maguire and Pearton 2000, Maguire 1994], são feitas análises do impacto da migração no esporte, sendo o primeiro mais focado no desenvolvimento de jogadores europeus e o segundo na análise do fluxo de mão de obra não apenas no futebol, mas em todos esportes.

Em [Roderick 2013], os autores analisam o movimento de jogadores de futebol dentro de um mercado interno, entretanto, este é um trabalho qualitativo, onde não houve uma coleta de dados em quantidade, apenas uma entrevista feita com cerca de 49 jogadores. O trabalho de [Palacios-Huerta 2004] estuda o futebol através de análises estatísticas comportamentais temporais, investigando apenas ligas inglesas e dando uma visão econômica ao esporte. No artigo [Frick], é analisado o mercado de transferências de atletas na Europa dando um foco mais empírico ao assunto, avaliando aspectos não considerados em nossas análises como salário de jogadores e tempo de carreira de jogadores.

Um trabalho anterior que modelou uma rede de transferências de jogadores e a estudou foi [Liu XF 2016], apontando algumas propriedades da rede construída com o objetivo principal de analisar o sucesso de um time através de suas transferências.

2. Coleta e modelagem da rede

Para obtenção da rede de transferências, foi utilizado como base o transfermarkt[tra]. O transfermarkt é uma página na internet que possui uma grande base de dados, com diversas informações relacionadas ao futebol, como: estatísticas de resultados, tabelas de campeonatos e o principal objeto de estudo deste trabalho, os dados relacionados a transferências de jogadores. Para a coleta de dados, foram consideradas as transferências entre 1990 e 2017. Os dados relativos a anos anteriores ao ano de 1990 se mostraram pouco relevantes para o estudo, já que, além de estarem em pouca quantidade, os valores das transações se mostraram irrisórios quando comparado a valores dos anos seguintes.

Após a obtenção das páginas no formato *html*, foi montado um *parser* responsável por fazer um pré-processamento das páginas e permitir a estruturação dos dados para extração das informações. Após formatação padrão, foi montado um banco de dados para facilitar eventuais consultas, que poderia auxiliar também em diversas análises, principalmente para modelagem de diversas redes de interesse para análises específicas.

As redes foram modeladas da seguinte forma: cada país foi considerado um

vértice, e caso um país tenha realizado uma transferência para o outro país, uma aresta direcionada do país vendedor para o país comprador será adicionada, considerando como o peso das arestas a quantidade de transações feitas entre os países, levando em conta a direção.

3. Análise da rede

3.1. Propriedades da Rede de Transferências

A base de dados coletada possui transferências dos anos de 1990 a 2017, que totalizam aproximadamente 27 mil transferências nas quais estão 135 vértices, representando os países presentes na base e 597 arestas representando a conexão entre esses países quando houve transferência de jogadores entre eles. Contudo, como o estudo abrange apenas os países presentes na Copa do Mundo FIFA 2018, a rede foi reduzida para 32 vértices e 430 arestas, abrangendo um total de 13797 transações, o que representa um pouco mais 50% do total de transferências da base, mostrando que os poucos países participantes da copa, representam grande parte das transações mundiais.

Conforme mostram os dados da Tabela 3.1, a densidade da rede é 0,40, o diâmetro é 4 e o coeficiente de clusterização é 0,70, o que mostra que há uma grande movimentação de jogadores entre os países participantes da copa, e que essas transações, geralmente são recíprocas, já que a reciprocidade da rede é maior do que 70%.

Verifica-se que a rede é disassortativa, o que significa que os países que negociam com muitos países tendem a negociar com países que negociam com poucos países, apontando para um desbalanceamento entre os investimentos dos países em seus campeonatos internos com jogadores estrangeiros.

Propriedade	Valor
Densidade	0.40
Diâmetro	4
Coeficiente de clusterização	0.70
Reciprocidade	0.71
Assortatividade grau	-0.22
Grau máximo	46
Força máxima de saída	521
Força máxima de entrada	1034

Tabela 1. Propriedades da rede de países da Copa

O grau máximo da rede é 46, valor que está associado a dois países muito importantes no futebol mundial: França e Alemanha. Esse valor indica que esses países são os que possuem o maior número de parceiros para compra e venda de jogadores. A força máxima de saída é 521, referindo-se ao total de vendas internacionais na rede de um país, o Brasil. Ou seja, o Brasil é o país que mais forneceu jogadores para os clubes dos países que estão na Copa. Por outro lado, o país que mais comprou jogadores de outros países foi a Inglaterra, que possui a força máxima de entrada da rede, 1034 arestas. As Figuras 1 e 2 mostram as distribuições de grau de entrada e de saída da rede. Essas curvas mostram que a rede possui muitos vértices de grau alto, logo, é possível perceber que os países da copa são, de fato, bastante representativos no cenário do futebol mundial.

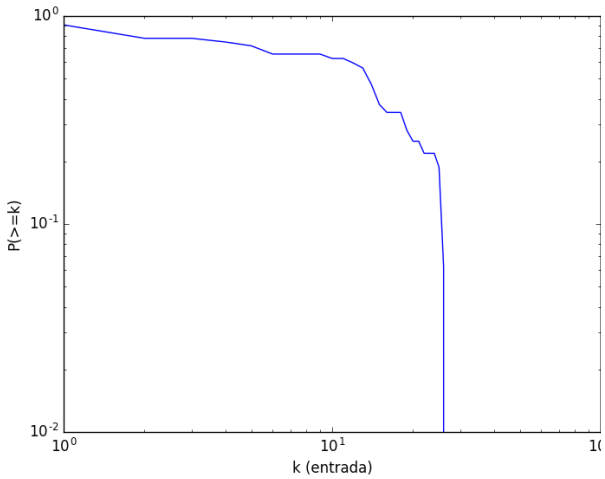


Figura 1. Distribuição de grau de entrada

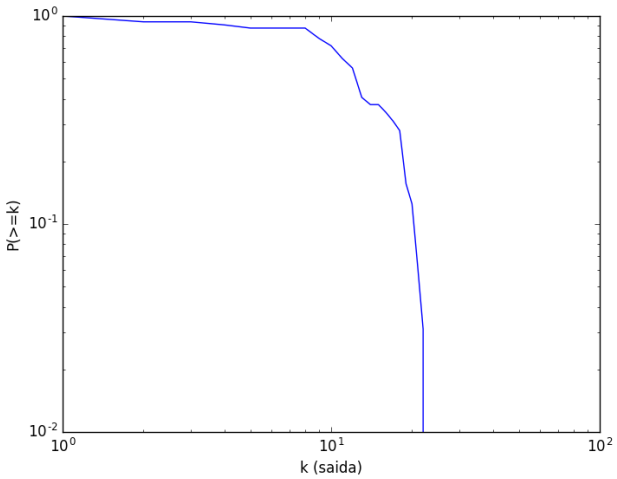


Figura 2. Distribuição de grau de saída

3.2. Análise de centralidade

Aplicando alguns algoritmos de centralidade de vértices, pode-se identificar os principais vértices da rede. Para isso foram consideradas três diferentes medidas de centralidade: *Closeness* [Freeman 1978], *Betweenness* [Newman and Girvan 2004] e *PageRank* [Page et al. 1999].

A centralidade de *Closeness* calcula o menor caminho entre um vértice e todos outros vértices do grafo, assim o vértice mais central é o que é mais próximo, em média, de todos os outros vértices.

A centralidade de *Betweenness* também é baseado em caminhos mínimos, contudo, este medida leva em consideração o número de caminhos mínimos que passam por um vértice. O vértice que é utilizado mais vezes como passagem nos caminhos mínimos será o vértice mais bem ordenado segundo esta medida.

Para aplicar medidas de *Closeness* e *Betweenness*, foi preciso inverter os valores dos pesos das arestas pois, como essas medidas utilizam como base os caminhos mínimos, a importância de um país, ligada ao número de transferências que ele faz, é mais apropriadamente representando os pesos de maneira inversa do original. Desta forma, para a aplicação dessas medidas, o peso das arestas foi calculado da seguinte forma: $PesoAresta = \sum 1/(QuantidadeTransacoes_a^b)$.

A medida de *PageRank*, por outro lado, considera que um vértice é importante se ele é apontado por outros vértices importantes. O *PageRank* leva em conta o número e a qualidade das ligações de um vértice para determinar o quão importante ele é, portanto o peso da aresta foi a quantidade de transações.

Após o cálculo das medidas de centralidade para a rede, foram obtidos três listas ordenadas distintas, das quais a Tabela 2 apresenta as 10 primeiras posições.

Como é possível observar a partir da Tabela 2, há uma grande consistência entre as medidades de centralidade consideradas e os *rankings* obtidos apresentam poucas variações. Cinco países aparecem entre as 10 primeiras posições nas três listas, e estão

Posição	Closeness	Betweenness	PageRank
1	**Inglaterra**	**Inglaterra**	**Inglaterra**
2	**Alemanha**	**Espanha**	**Alemanha**
3	**Espanha**	**Alemanha**	**Espanha**
4	**França**	**Brasil**	Rússia
5	Suíça	**França**	**França**
6	Rússia	Argentina	**Brasil**
7	Bélgica	Portugal	Bélgica
8	Argentina	Costa Rica	Portugal
9	**Brasil**	Dinamarca	Polônia
10	Dinamarca	Egito	México

Tabela 2. Top-10 países ordenados por cada uma das medidas de centralidade.

em negrito na Tabela 2, assim como na Tabela 3. As maiores surpresas ocorreram na lista ordenada pela medida *Betweenness*, mostrando Egito e Costa Rica, que são grandes parceiros da Inglaterra.

A Tabela 3 mostra a quantidade de transações realizadas pelos 10 países que mais realizaram transações internacionais dentro da rede. Percebe-se que todos países presentes nesta tabela também estão presentes em pelo menos um dos *rankings* apresentados pela Tabela 2, desta forma podemos considerar que a importância dos países na rede está diretamente ligada ao número de transações internacionais realizadas.

Posição	País	# de transações
1	**Inglaterra**	1384
2	**Alemanha**	1230
3	**Espanha**	1096
4	**França**	925
5	**Brasil**	719
6	Portugal	691
7	Argentina	513
8	Rússia	453
9	Bélgica	416
10	Suíça	337

Tabela 3. Quantidade de transações por país

3.3. Análise de comunidades

Uma investigação sobre a estrutura de comunidades na rede de transferências foi realizado com o objetivo de identificar grupos coesos de países dentro da rede e, para isso, utilizou-se alguns dos algoritmos para detecção de comunidades mais adotados na literatura. Apesar de se tratar de um mercado abrangente, onde qualquer país pode realizar o intercâmbio de jogadores com qualquer outro país, alguns "padrões de consumo"podem ser observados, como o fato de grandes países procurarem talentos estrangeiros para seus times locais em um mesmo grupo de países com mais frequência.

Os métodos para detecção de comunidades considerados neste trabalho foram: *Multilevel* [Blondel et al. 2008], *Eigenvector* [Newman 2006] e *Fastgre-*

edy [Clauset et al. 2004], todos eles baseados na otimização da modularidade [Newman and Girvan 2004], a medida mais utilizada na literatura para avaliação da qualidade da partição de uma rede. A Tabela 4 mostra os resultados encontrados por cada um dos métodos, em termos do valor de modularidade e do número de comunidades.

Algoritmo	Modularidade	# de comunidades
Eigenvector	0.53	8
Fastgreedy	0.14	7
Multilevel	0.54	6

Tabela 4. Comparação entre algoritmos de comunidade

Considerando que os métodos para identificação de comunidades visam otimizar a modularidade como uma medida a ser maximizada, a análise da Tabela 4 permite identificar que o método *multilevel* oferece uma partição de melhor modularidade entre os métodos analisados e, por isso, essa partição foi escolhida para as análises seguintes.

Como apresentado na Tabela 4, o método multilevel identificou 7 comunidades, descritas na Tabela 5.

Comunidade	Países
1	Rússia, Polônia, Sérvia e Panamá
2	Egito, Tunísia, Arábia Saudita, Nigéria e Marrocos
3	Irã, Bélgica, Croácia, Dinamarca, Costa Rica, Suécia e Austrália
4	Brasil, México, Espanha, Portugal, Argentina, Japão
	Coreia do Sul, Uruguai,Colômbia e Peru
5	Alemanha e Suíça
6	Inglaterra e Islândia
7	França e Senegal

Tabela 5. Comunidades formadas

A Figura3 apresenta uma representação gráfica da rede de transferências com as comunidades identificadas através de cores.

A comunidade 1, formada por Rússia (país anfitrião dessa copa do mundo), Polônia, Sérvia e Panamá, possui apenas um entre os países com maior número de transferências registradas na base (Tabela 3), sendo este a Rússia, com 882 transações. Polônia e Sérvia possuem 428 e 308 transferências respectivamente. Já o Panamá, possui apenas duas transações registradas com outros países participantes da copa, sendo uma com a Polônia, o que ajuda a explicar o porquê da presença do Panamá neste grupo.

Para auxiliar na análise dos dados e ilustrar o montante movimentado por um país, foi considerada a balança comercial do mesmo. Essa balança mostra o quanto um país ganhou ou perdeu com a compra e venda de atletas. Para o cálculo da balança comercial foi considerada a soma de todas vendas subtraída da soma de compras. Os países com balança comercial negativa podem ser considerados como mercado consumidor de atletas. Já os países com balança comercial positiva podem ser considerados países fazenda, que são países que se caracterizam pela venda de jogadores e são frequentemente procurados para a compra de jogadores jovens que se destacaram em campeonatos nacionais,

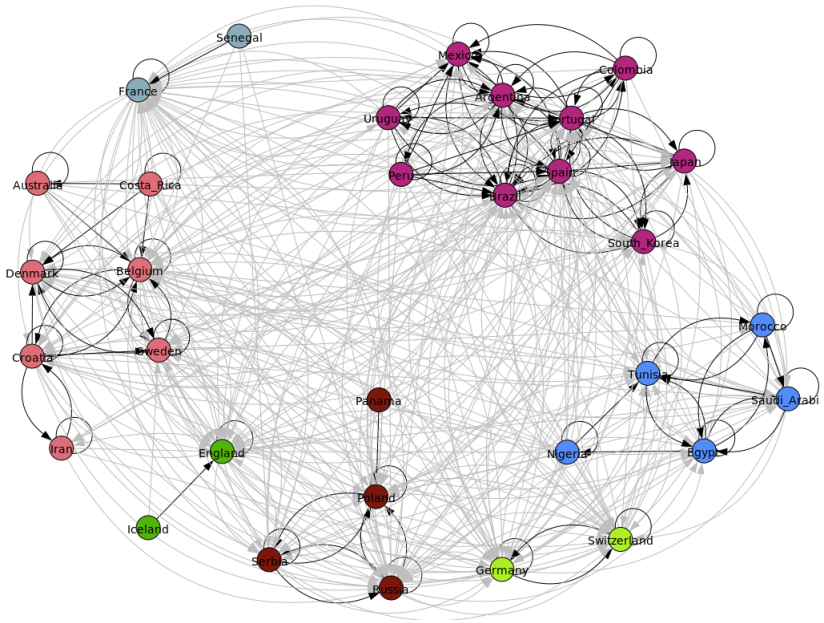


Figura 3. Rede de transferências de jogadores de futebol entre países da copa do mundo dividida em comunidades

País	Vendas	Compras	Balança	# de transferências
Rússia	Inglaterra,Alemanha,Espanha	Brasil,Portugal,Sérvia	-757416000 €	882
Polônia	Alemanha,Rússia,Inglaterra	Sérvia,Alemanha,Croácia	163645000 €	428
Sérvia	Rússia,Alemanha,Espanha	Brasil,Rússia,Portugal	466825000 €	308
Panamá	Espanha,Polônia	-	790000 €	2

Tabela 6. Dados sobre os países presentes na comunidade 1

regionais e às vezes até nos campeonatos juniores. O cálculo da balança comercial é da forma: $BalancaComercial = \sum Vendas \ - \sum Compras$.

A Tabela 6 analisa os países presentes na comunidade 1, apresentando os países que mais compram de cada país de referência, os países que mais vendem para aquele país, a balança comercial e a quantidade de transferências realizadas por aquele país.

A análise dos países com os quais os integrantes da comunidade 1 realizam transferências com maior frequência, permite observar o motivo da ligação entre os mesmos. De acordo com a balança comercial do país, é possível ver qual sua posição com relação aos outros países, se é um país consumidor ou um país fornecedor de atletas.

É possível perceber que, nessa comunidade, todos os países estão conectados pelos seus compradores e fornecedores, sendo a Rússia o principal país da comunidade, por possuir a maior quantidade de transações e estar entre os principais parceiros comerciais de cada um dos países, com exceção do Panamá. Desta forma observa-se que a movimentação financeira dos três primeiros países é consideravelmente grande, sendo a Rússia, o único país presente na comunidade com a balança comercial negativa, um país consumidor, que possui clubes ricos que buscam atletas em outros países, enquanto os outros países Polônia e Sérvia são países produtores de talentos por possuírem uma balança comercial positiva.

A comunidade 2 é formada por Egito, Tunísia, Arábia Saudita, Nigéria e Marrocos, sendo que apenas a Arábia Saudita não pertence ao continente africano. Nesta comu-

nidade, como pode-se observar na Tabela 7, os países não possuem uma grande quantidade de transferências e nenhum dos países presentes nessa comunidade se equipara, em número de transações realizadas, com grandes países da Europa e América. Além disso, assim como na comunidade 1, a maioria dos países possui a balança comercial positiva. Provavelmente esses países trocam frequentemente de jogadores entre si devido a proximidade geográfica deles, e essas transações foram suficientes para considerá-los como integrantes de uma comunidade.

País	Vendas	Compras	Balança	# de transferências
Tunísia	França,Arábia Saudita,Suíça	Nigéria,França,Marrocos	39491000 €	190
Egito	Arábia Saudita,Inglaterra,Suíça	Tunísia,Portugal,Inglaterra	48775000 €	164
Arábia Saudita	Brasil,Suíça,Rússia	Brasil,Egito,Tunísia	-166437000 €	123
Nigéria	Tunísia,França,Alemanha	Egito	22035000 €	35
Marrocos	França,Tunísia,Arábia Saudita	Egito,França,Tunísia	27789000€	68

Tabela 7. Dados sobre os países presentes na comunidade 2

A comunidade 3, formada por Irã, Bélgica, Croácia, Dinamarca, Costa Rica, Suécia e Austrália é a comunidade com maior diversidade de continentes, tendo países pertencentes a Europa, América, Ásia e Oceania. Essa comunidade é caracterizada por possuir somente países com uma balança comercial totalmente positiva, como pode ser observado de maneira detalhada pela Tabela 8.

País	Vendas	Compras	Balança	# de transferências
Irã	Alemanha,Rússia,Inglaterra	Brasil,Rússia,Espanha	7607000€	41
Bélgica	Inglaterra,Alemanha,França	França,Alemanha,Croácia	500960000€	627
Croácia	Alemanha,Bélgica,Rússia	Alemanha,Bélgica,Brasil	406378000€	286
Dinamarca	Alemanha,Inglaterra,Suécia	Alemanha,Suécia,Inglaterra	247644000€	354
Costa Rica	México,Bélgica,Inglaterra	Brasil	13645000€	13
Suécia	Inglaterra,Alemanha,Dinamarca	Dinamarca,Inglaterra,Alemanha	260493000€	188
Austrália	Inglaterra,Alemanha,Japão	Costa Rica	23305000€	37

Tabela 8. Dados sobre os países presentes na comunidade 3

A Tabela 8 mostra que os países europeus possuem uma quantidade maior de transações e, consequentemente, um maior valor na balança comercial. Estes países funcionam como um trampolim para futura venda de jogadores para outros países europeus, envolvendo times de maior prestígio.

Das poucas transações presentes na Oceania a maioria são transações de venda. O continente possui 62 transações de venda e 7 transações de compra registradas, fazendo desse continente o que possui menor quantidade de transações nesta comunidade.

Na Tabela 9, a distribuição de transações por continente é apresentada.

Continente	# compras	# vendas	# internas
Asia	834	281	539
América	435	1853	1296
Africa	60	388	384
Europa	2222	966	21057
Oceania	7	62	4

Tabela 9. Quantidade de transações de cada continente

A comunidade 4 é a maior comunidade presente na rede, com um total de 10 países. Essa comunidade é caracterizada pela presença de países de diversos continentes

Continente	Valor médio vendas
Asia	1.535.288 €
América	2.894.209 €
Africa	604.533 €
Europa	2.202.635 €
Oceania	415.322 €

Tabela 11. Valor médio pago em transações por continente

como Brasil, México, Argentina, Uruguai, Colômbia e Peru, representando a América, Japão e Coreia do Sul representando Ásia e, ainda, Espanha e Portugal representando a Europa.

A Tabela 10 apresenta os dados dos países presentes na comunidade 4. Nesta comunidade, está presente uma maior quantidade de países vencedores de copas do mundo, sendo eles Brasil, Argentina, Espanha e Uruguai.

País	Vendas	Compras	Balança comercial	# de transferências
Espanha	Inglaterra,Alemanha,Portugal	Argentina,Inglaterra,Portugal	-906437000 €	1704
Portugal	Espanha,Inglaterra,França	Brasil,Espanha,Argentina	1468658000 €	871
Brasil	Alemanha,Portugal,Espanha	Argentina,Portugal,Espanha	2257058000€	1133
Argentina	Espanha,México,Brasil	Espanha,Uruguai,México	1491868000€	794
México	Argentina,Brasil,Espanha	Argentina,Brasil,Colômbia	-72455000€	298
Japão	Alemanha,Brasil,Inglaterra	Brasil,Alemanha,Coreia do Sul	42627000€	144
Uruguai	Espanha,Argentina,Portugal	Brasil,Argentina,Espanha	328424000€	122
Colômbia	Argentina,Brasil,México	México,Argentina,Brasil	188658000€	89
Peru	Alemanha,México,Portugal	Argentina	22703000 €	27
Coreia do Sul	Alemanha,Japão,Inglaterra	Brasil,Sérvia,Croácia	58499000€	73

Tabela 10. Dados sobre os países presentes na comunidade 4

Assim como na comunidade 1, apenas uma pequena parte da comunidade possui uma balança comercial negativa, sendo estes Espanha e México.

Como esta comunidade possui um número bem grande países, optou-se por analisar a mesma por continente. Primeiramente, analisando os países da América, que são maioria nessa comunidade, observa-se que o Brasil possui o maior número de transações e o Peru o menor número de transações. Através dessa comunidade é possível visualizar um padrão que é compartilhado pela maioria dos países da América do Sul e América Central que é a balança comercial positiva, demonstrando a falta de poder de compra dos clubes pertencentes a esses países para trazer atletas de outros continentes ou mesmo manter atletas de origem nacional de qualidade dentro do próprio país.

A América hoje possui a maior média de venda de jogadores entre todos os continentes, como podemos ver na Tabela 11, sendo que considerando apenas transações externas de compra e venda, cerca de 81% das transações deste continente são transações de venda. Desta forma a Tabela 11 corrobora que este é um mercado produtor de talentos para clubes de outros países com poder aquisitivo maior.

Analisando os países asiáticos presentes na comunidade, tendo Japão e Coreia do Sul como representantes nesta comunidade, observa-se que ambos possuem uma balança comercial positiva, o que não representa o padrão de consumo asiático como um todo, pois a maioria das movimentações de compra superam as de venda. Esse padrão é alavancado pela China, que lidera como país com maior transações gerais dentro da Ásia movimentando grandes quantidade financeiras com grandes contratações de jogadores

estrangeiros [the]. A China, entretanto, mesmo sendo um país com fluxo considerável no mercado de transação de atletas não está presente entre os países classificados para copa de 2018, o que aponta que não são apenas investimentos em jogadores estrangeiros para fortalecimento das competições nacionais que influenciam no desempenho da seleção para classificação e participação em uma copa do mundo.

Os países do continente europeu nesta comunidade são Espanha e Portugal. Neste grupo fica claro o domínio de alguns países pertencentes a esse continente no mercado de transações. Cada um dos países presentes nesta comunidade funciona como um agente diferente na rede estudada e no mercado como um todo. A Espanha é um país comprador de talentos para fortalecimento do campeonato nacional, de forma que os clubes pertencentes a esses países adquirem tais atletas sem pensar em uma eventual venda do jogador. Portanto, a Espanha possui uma balança negativa, abrigando alguns dos clubes mais ricos do mundo [ric], que dispõem de montantes significativos para manter os atletas dentro de seu domínio e sempre comprar novos reforços para manutenção dos clubes.

Já Portugal tem servido como vitrine para os atletas, que passam pelos clubes portugueses para ter maior visibilidade no mercado europeu ao atuar neste país, servindo como um mercado impulsionador de atleta, aumentando atributos e o valores dos jogadores. Geralmente, times portugueses compram jogadores com o objetivo da venda por um preço maior que o comprado, obtendo assim um lucro na venda do jogador fazendo, assim, que Portugal possua uma balança comercial positiva.

Hoje a Europa é o centro do futebol mundial, onde há as principais ligas, com os melhores jogadores do mundo [fif]. Grande parte dos países europeus com balança comercial positiva funcionam como país trampolim para uma posterior venda desses jogadores. Entretanto, isso não ocorre com todos os países, já que com a consolidação de campeonatos mais fortes devido a compra de jogadores de maior habilidade faz com que o nível dessas competições aumente, fazendo com que haja uma evolução dos jogadores e dos preços de venda do atleta. Para uma melhor visualização disso, foi analisada a média de compra e venda nesses países.

País	Vendas	Compras
Rússia	1.681.381 €	1.830.670 €
Polônia	693.962 €	268.694 €
Sérvia	972.090 €	408.254 €
Espanha	3.128.515 €	2.907.998 €
Portugal	2.089.052 €	1.975.059 €
Bélgica	1.738.789 €	1.022.196 €
Croácia	1.090.413 €	646.089 €
Dinamarca	1.503.512 €	943.660 €
Suécia	1.124.419 €	418.479 €
Alemanha	1.529.849 €	1407114 €
Suíça	1.337.964 €	908.760 €
Inglaterra	2.886.619 €	3.013.590 €
França	2.098.838 €	2.145.160 €
Islândia	278.750 €	0 €

Tabela 12. Tabela mostra o valor médio pago em transações por continente

Pode-se ver na Tabela 12 a média de compra e vendas dos países pertencentes a Europa que estão presentes na copa do mundo 2018. De forma parcial pode-se ver países que compram jogadores por uma média mais baixa e vende a um preço mais alto. Se associarmos essa informação a dados de balança comercial e aos países ao qual esse país se associa para fazer compras, podemos definir se esse é um país consumidor ou fornecedor.

4. Conclusão

Este trabalho apresentou a análise de uma rede de transferências de jogadores de futebol entre países participantes da copa do mundo. Foram utilizadas técnicas de redes complexas como *ranking* por algumas medidas de centralidade e detecção de comunidades.

Observou-se que esses países são bastante representativos no futebol, abrangendo 50% das transferências mundiais as quais são bastante frequentes entre a maioria dos países da rede. Pode-se perceber alguns destaques de papéis de alguns países empregam no futebol mundial, por exemplo, se comparamos o Brasil, um país fazenda com a Inglaterra, um país de ligas ricas, pode-se ver dois países bastante diferentes em termos humanos e econômicos. A Inglaterra possui o 16^o IDH do mundo, uma expectativa de vida de $80,8$ anos e seus cidadãos estudam em média por $16,3$ anos. Enquanto o Brasil é o 79^o de acordo com o IDH, tem expectativa de vida de $74,7$ e seus cidadãos estudam em média $15,2$. A Inglaterra importa mais matéria prima e exporta produtos industrializados e medicamentos, enquanto o Brasil exporta matéria prima e produtos agrícola e importa produtos industrializados. No Brasil o futebol ainda é uma grande esperança de muitos jovens para mudança de realidade de vida.

As listas **top-10** geradas pelas diferentes medidas de centralidade mostraram que os países de ligas ricas são os países mais centrais da rede para qualquer medida de centralidade observada, apesar de países fazenda e países que tem tradição de comprar jogadores para valorização e futura venda e lucro também são encontrados nas listas.

Considerando a estrutura de comunidades, observa-se que os maiores parceiros comerciais estão unidos nos mesmos grupos. No entanto, os países mais centrais estão separados em comunidades isoladas com um parceiro eventual, possivelmente devido ao fato de estes países menores possuírem poucas transações com os demais países da rede.

Como trabalhos futuros, pretende-se realizar análises de comunidades com métodos que consigam identificar comunidades com sobreposição, pois acredita-se que este tipo de abordagem pode trazer mais informações interessantes. Pretende-se ainda utilizar uma rede com o montante financeiro como peso das arestas para comparação de resultados, além de analisar as redes separadas pelos anos que antecederam a copa e confrontar os resultados com o desempenho das seleções naquele período.

Agradecimentos

Os autores agradecem às agências de fomento: Capes, CNPq e FAPEMIG.

Referências

fifa-11. http://www.fifa.com/the-best-fifa-football-awards/fifa-fifpro-world11/index.html. Accessed: 2017-12-04.

Richest clubs. http://uk.businessinsider.com/the-20-richest-football-clubs-in-the-world-2017-1. Accessed: 2018-01-13.

The Guardian. www.theguardian.com/football/these-football-times/2017/jan/05/china-chinese-super-league-oscar-carlos-tevezWhy Chinese clubs are breaking transfer records – and why players are wise to go. Accessed: 2017-11-20.

Transfermarkt. transfermarkt.com/statistik/transferrekorde. Accessed: 2017-11-17.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008.

Clauset, A., Newman, M. E. J., and Moore, C. (2004). Finding community structure in very large networks. *Phys. Rev. E*, 70:066111.

Deloitte (June 2016). Annual review of football finance.

Freeman, L. C. (1978). Centrality in social networks conceptual clarification. *Social Networks*, page 215.

Frick, B. The football players' labor market: Empirical evidence from the major european leagues. *Scottish Journal of Political Economy*, 54(3):422–446.

Liebig, J., Rhein, A. V., Kastner, C., Apel, S., Dorre, J., and Lengauer, C. (2012). Large-scale variability-aware type checking and dataflow analysis.

Liu XF, Liu Y-L, L. X.-H. W. Q.-X. W. T.-X. (2016). The anatomy of the global football player transfer network: Club functionalities versus network properties. *PLoS ONE*, 11(6).

Maguire, J. (1994). Preliminary observations on globalisation and the migration of sport labour. *The Sociological Review*, 42(3):452–480.

Maguire, J. and Pearton, R. (2000). The impact of elite labour migration on the identification, selection and development of european soccer players. *Journal of Sports Sciences*, 18(9):759–769. PMID: 11043901.

Newman, M. E. J. (2006). Finding community structure in networks using the eigenvectors of matrices. *Phys. Rev. E*, 74:036104.

Newman, M. E. J. and Girvan, M. (2004). Finding and evaluating community structure in networks. *Phys. Rev. E*, 69:026113.

Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab. Previous number = SIDL-WP-1999-0120.

Palacios-Huerta, I. (2004). Structural changes during a century of the world's most popular sport. *Statistical Methods and Applications*, 13(2):241–258.

Poli, R. (2010). Understanding globalization through football: The new international division of labour, migratory channels and transnational trade circuits. *International Review for the Sociology of Sport*, 45(4):491–506.

Roderick, M. (2013). Domestic moves: An exploration of intra-national labour mobility in the working lives of professional footballers. *International Review for the Sociology of Sport*, 48(4):387–404.

Uma análise do fator cultural em tecnologias persuasivas: um estudo de caso da rede social *Facebook*

Mateus L. do Nascimento, Pedro H. B. Ruas, Otaviano Neves, Luis H. Zárate, Cristiane N. Nobre

[1] Instituto de Ciências Exatas e Informática
Pontifícia Universidade Católica de Minas Gerais
Av. Dom José Gaspar, 500 - Coração Eucarístico
Belo Horizonte - MG - CEP 30535-901

{mateus.nascimento,pedro.ruas,otaviano,zarate,nobre}@pucminas.br

Abstract. *Facebook is a social network used by more than one billion users, and is present in several countries trying to influence people so that they stay as long as possible on the network. This work was carried out in order to understand cultural influence in the interaction of users in the context of persuasive technologies (technologies that aim to influence behaviors). For this purpose, a case study was carried out on the use of the social network Facebook, comparing Brazil with other countries, using the data collected through questionnaires and analyzed by statistical methods. It was concluded that users of the analyzed cultures have similar behavior in the network.*

Resumo. *O Facebook é uma rede social usada por mais de um bilhão de usuários, e está presente em vários países tentando influenciar as pessoas a adotarem determinados comportamentos alvos. Este trabalho foi realizado visando entender a influência cultural na interação dos usuários no contexto das tecnologias persuasivas. Para isso foi realizado um estudo de caso sobre a utilização da rede social Facebook, comparando o Brasil com outros países, sendo os dados da utilização obtidos via questionários e analisados por métodos estatísticos. Concluiu-se que os usuários das culturas analisadas possuem comportamento semelhante na rede.*

1. Introdução

Uma *Tecnologia Persuasiva* é definida como qualquer produto interativo projetado para modificar hábitos ou comportamentos através de persuasão, de forma não coercitiva. Fogg (2003) cunhou o termo *captology* para se referir a *computers as persuasive technologies*, voltado para pesquisa e análise de produtos computacionais interativos criados para modificar o comportamento e/ou atitudes dos usuários. Assim, este conceito descreve uma área onde a tecnologia e a persuasão se intercedem. Um tipo de ferramenta computacional que utiliza estratégias persuasivas são, por exemplo, as redes sociais *online*.

As redes sociais *online* como conhecemos hoje, com perfil de usuários, rede de amigos, fotos e comentários surgiram no início dos anos 2000 com a popularização de redes como *MySpace*, *Linkedin*, *Orkut* e, posteriormente, com o *Facebook* e *Twitter*. Com o passar do tempo algumas dessas redes deixaram de existir, como por exemplo o *Orkut*, e outras se tornaram dominantes como é o caso do *Facebook* com mais de um bilhão de

usuários ativos por mês[1]. Essas redes são muitas vezes representações *online* do que as pessoas são ou pretendem parecer ser na sociedade. Por isso, os fatores culturais podem produzir efeitos na forma com que as pessoas utilizam essa ferramenta.

As culturas podem se diferenciar por serem mais individualistas ou mais coletivistas. Em Khaled et al. (2006b), os autores propõem um complemento à teoria sobre Tecnologias Persuasivas de Fogg (2003). Em uma análise preliminar, Khaled et al. (2005) afirmam que a maioria das estratégias persuasivas elaboradas por Fogg (2003) são voltadas para audiências individualistas, característica forte na cultura americana. Segundo o autor, isso se dá porque a maioria das tecnologias são criadas na cultura americana (ou para ela) e as análises das estratégias persuasivas são feitas a partir dessas tecnologias. Por esse motivo, Khaled et al. (2006b) propõem um conjunto de estratégias para culturas coletivistas, argumentando que, quando o aspecto cultural é considerado, a persuasão tem mais efeito. No contexto de redes sociais, o fator cultural pode ter uma grande relevância em como o usuário é persuadido pelas funcionalidades da ferramenta.

Ruas (2016) desenvolveram uma pesquisa para traçar o perfil dos usuários da rede social *Facebook* no Brasil. Os autores discutem como essa rede social funciona como uma tecnologia persuasiva e identificaram três perfis de usuários existentes nessa rede social: *espectadores*, usuário que tem como comportamento predominante visualizar o que é postado na rede social, *produtores*, usuários que criam e publicam conteúdo na rede, e *participantes*, usuários que têm como principal característica interagir na rede social.

Sendo uma rede social entendida como um grupo de pessoas que dividem os mesmos interesses, valores, objetivos em comum para interagir, se comunicar e compartilhar informações, este trabalho visa verificar a efetividade das estratégias persuasivas em diferentes culturas. A hipótese levantada neste trabalho é que as diferentes culturas podem resultar em diferentes estratégias de persuasão, já que elas podem dar importância a diferentes aspectos sociais ao utilizar uma rede social. Segundo Khaled et al. (2005), ainda não existem processos que orientem o desenvolvimento de tecnologias persuasivas destinadas a públicos de uma cultura particular. Os autores afirmam ainda que poucas pesquisas abordaram o tema de *design* de tecnologia persuasiva na perspectiva da integração da ideologia cultural. Assim, considerando-se esta lacuna, esse trabalho analisa como os grupos de usuários de diferentes culturas submetidos às mesmas estratégias reagem ao concordarem ou discordarem de determinado conteúdo postado na rede baseado nas funcionalidades disponível na ferramenta para a interação *online*, isto pode ajudar no desenvolvimento de tecnologias persuasivas mais eficientes.

Este trabalho está organizado da seguinte maneira: a Seção 2 trata da teoria necessária para entender os assuntos aqui tratados. A Seção 3 descreve os trabalhos relacionados à estratégias de persuasão aplicáveis a produtos computacionais em ambientes de redes sociais e o estudo de diferenças culturais. Na Seção 4 é apresentada a metodologia aplicada no desenvolvimento do trabalho. A Seção 5 contém os resultados obtidos, e por último, na Seção 6 as conclusões e as considerações finais deste trabalho são apresentadas.

2. Referencial Teórico

Nessa seção é apresentado o levantamento dos princípios persuasivos computacionais, bem como a caracterização de culturas e redes sociais.

[1]Disponível em http://br.newsroom.fb.com/company-info/

2.1. Princípios Persuasivos Computacionais

Elaborada por Fogg (2003), tecnologia persuasiva é definida como uma tecnologia desenvolvida para mudar atitudes ou comportamento dos usuários através de persuasão e influência social, sem o uso, porém, da coerção.

Para explicar como acontece um comportamento, Fogg (2009) apresenta o modelo comportamental denominado *Fogg Behavior Model* (FBM), o qual apresenta o comportamento do usuário como um produto de três fatores: habilidade, motivação e gatilho. O autor afirma que para um usuário realizar um comportamento alvo, ele deve estar suficientemente motivado, deve possuir a habilidade de executar aquela ação e receber um estímulo (gatilho) para realizar este comportamento. Para Fogg (2009), todos os três fatores devem estar presentes ao mesmo tempo para que o comportamento alvo ocorra.

O *Facebook* utiliza uma estratégia chamada de *redução* por Fogg (2003), que consiste em reduzir um comportamento complexo a etapas simples (no caso dos computadores, em poucos cliques), aumentando a relação custo/benefício e incentivando a execução do comportamento alvo. Essa estratégia pode ser identificada no *Facebook* nas funções "Curtir" e "Compartilhar". Clicando-se em compartilhar, o usuário tem a opção de redigir um texto para compartilhar o conteúdo desejado com seus contatos, não sendo obrigatório. Para executar uma destas funções no sistema, basta ao usuário clicar uma única vez em um botão para se completar a ação.

Já para o botão "Comentar" é necessário que o usuário primeiramente clique sobre o *link*, redija um pequeno texto para depois publicar seu comentário. Se utilizarmos o modelo de avaliação chamado *Keystroke Level* criado por Card et al. (1983) para medirmos o tempo gasto entre uma ação e outra, podemos melhor entender essa estratégia. Por exemplo, para curtir um *post* no *Facebook*, segundo o modelo *Keystroke*, o usuário levaria em média 0,35 segundos para realizar esta ação (tempo estimado para pressionar uma única tecla ou botão). Em contrapartida, para o usuário postar um comentário de 250 caracteres, por exemplo, sendo este um digitador mediano (40 palavras por minuto), seriam gastos 70 segundos. Sendo assim, sob a definição da estratégia de *redução*, pode implicar na desistência de comentar sua opinião contrária sobre algum conteúdo, pois para isso o esforço do usuário seria maior.

2.2. Caracterização de culturas e redes sociais

Segundo Fogg e Iizawa (2008), redes sociais podem causar efeitos diversos em diferentes culturas. Por isso, para conseguir estratégias persuasivas mais eficazes, é importante analisar o contexto cultural. Para conseguir caracterizar e categorizar as culturas, neste trabalho foi utilizado o modelo de Hofstede e Bond (1984), que define cultura como "o modo que a mente das pessoas é programada coletivamente, de forma a se distinguir de outro grupo ou categoria".

Os autores elencam seis dimensões culturais: 1) *Índice de distância do poder*, um país com baixo índice significa que as pessoas questionam mais as autoridades e existe uma tentativa de distribuir poder. Com alto índice, é aceita a inequalidade e a hierarquia é esperada; 2) *Aversão à incerteza*, culturas que tem pouca aversão à incerteza conseguem se estressar menos em situações de incertezas; 3) *Individualistas versus coletivistas*, culturas coletivistas pensam mais no grupo, colocam o relacionamento entre os pares acima

das tarefas e preferem cumprir obrigações impostas pelo grupo às suas e evitam confrontos diretos. Individualistas tendem a focar em cumprir as suas tarefas pessoais, obrigações e expressam os seus pensamentos de forma mais direta; 4) *Masculinidade versus feminilidade*, culturas femininas são mais focadas em qualidade de vida, trabalhar para viver, compaixão e resolução de conflitos através de compromisso e negociação. Culturas mais masculinas são centradas na ambição, admiração pelo sucesso e pelos mais fortes; 5) *Orientação de curto e longo prazo*, culturas voltadas para o curto prazo tendem a querer resultados imediatos, pressão para gastar mais. Já as de longo prazo tendem a enxergar o longo prazo, o esforço e a perseverança, busca economizar e ter cuidados com os recursos e adiar seus desejos por uma "boa causa". 6) *Prazer versus coibição*, culturas voltadas ao prazer tendem a ser mais felizes, buscam aproveitar a vida e se divertir. Em contraste, culturas mais coibidas tendem a controlar a felicidade e o prazer por normas sociais mais restritivas.

Khaled et al. (2006b) argumentam que uma deficiência no modelo de Fogg (2003) é o foco em culturas individualistas, por isso propõe um modelo para culturas coletivistas. A dimensão *individualistas versus coletivistas* é uma característica social e não individual, e diz respeito ao grau com que a sociedade está submetida a grupos. Na Tabela 1 é apresentado um comparativo das duas dimensões, segundo Hofstede (2011).

Tabela 1. Diferenças entre culturas coletivistas e individualistas

Individualistas	Coletivistas
Cada um deve tomar conta de si mesmo ou da sua família	Pessoas são extensões da sua família ou grupo social, devem protegê-los em troca de lealdade
Pensa em si mesmo	Pensa no grupo
Pensa em privacidade	Quer pertencer
O outro é visto como indivíduo	O outro é visto como pertencente ou não ao grupo
Opinião pessoal esperada, uma pessoa um voto	Opiniões e votos predeterminados pelo grupo
Na linguagem o "eu" é indispensável	Na linguagem o "eu" é evitado
Proposta da educação é aprender como aprender	Proposta da educação é aprender como fazer
Cumprir tarefas está acima do relacionamento	Relacionamento está acima das tarefas

Fonte: Adaptado de Hofstede (2011)

Em Hofstede et al. (2010) os autores fizeram uma listagem das dimensões individualista e coletivistas de 76 países para conseguir caracterizar a cultura destes indivíduos. Os autores comentam que individualismo tende a prevalecer em nações ocidentais; enquanto coletivismo prevalece em nações menos desenvolvidas e do leste Europeu, sendo o Japão inserido nessa dimensão.

O *Facebook* utiliza estratégias persuasivas para que o usuário fique o máximo de tempo na rede social (Fogg e Iizawa, 2008), e para que isso ocorra é interessante que os usuários interajam o máximo possível com a ferramenta. Para chegar a esse comportamento desejado, fazendo com que o usuário produza conteúdo e interações, a rede tenta ao máximo reduzir tarefas complexas a poucos cliques (estratégia de *redução*). As funcionalidades para interação, por exemplo, estão sempre visíveis sem precisar do usuário clicar em outra página antes de interagir nas publicações na rede social.

A questão principal a ser investigada é se estas estratégias surtem o mesmo efeito em pessoas de diferentes culturas, pois culturas individualistas tendem a expressar opiniões mais diretamente. Em contrapartida culturas coletivistas pensam mais na opinião do grupo antes de opinar, ainda que estejam submetidos às mesmas estratégias dentro da rede. Esta diferença cultural pode moldar a forma como o indivíduo interage na

rede. Por mais que as interações sejam simplificadas para influenciar um comportamento, a própria cultura poderá inibir o usuário de realizar tal comportamento.

3. Trabalhos relacionados

Ruas et al. (2015), por meio de técnicas de clusterização, caracterizaram os usuários do *Facebook* em três perfis de interação: *participante*, *espectador* e *produtor de conteúdo*. Os autores descreveram cada perfil da seguinte forma: o *produtor de conteúdo* é aquele que possui como principal característica a criação de conteúdo na rede social e, com isso, é esperado que os usuários na rede que possuem conexão com este produtor interajam com este conteúdo através das ações disponíveis (curtir, comentar, compartilhar, dentre outras) na ferramenta. O usuário com perfil de *participante* interage com conteúdo já produzido por meio das funções curtir comentar ou compartilhar, e o *espectador*, preferencialmente observa o que acontece na rede, sem interagir ou interagindo muito pouco com o conteúdo disponível.

Para identificar as estratégias persuasivas utilizadas pelo *Facebook*, Ruas et al. (2014) realizaram uma pesquisa para tentar descrever, com base nos perfis (*participante, espectador e produtor de conteúdo*), como cada perfil de usuário utiliza a rede. Para tal, os autores disponibilizaram um questionário *online* em que os usuários dessa rede social responderam questões demográficas e sobre a utilização do *Facebook*. Além disso, os usuários foram questionados também sobre as frequências com que eles curtiam, comentavam e compartilhavam informações quando concordavam com o conteúdo e a frequência que eles comentavam, ocultavam ou deixavam de seguir um usuário quando discordavam do conteúdo.

A pesquisa de Ruas (2016) foi feita com 686 usuários brasileiros. Em relação aos perfis dos usuários da rede social, as pessoas se declararam da seguinte forma: 76,06% são espectadores, 12,63% são participantes e apenas 11,31% são produtores de conteúdo. Além disso, foram elaborados, por meio de ACP (Análise de Componentes Principais), dois índices de utilização: o primeiro indicando o grau de utilização geral, ou seja, o uso das funcionalidades disponíveis no *Facebook* para manifestar tanto a aprovação (“curtir”, “comentar”, e “compartilhar”) quanto a reprovação (“comentar”, “deixar de seguir usuário”, e “ocultar publicação”). Assim, se determinado usuário obtém um valor numérico alto para o primeiro índice, significa que ele possui um alto grau de utilização para todas as funções.O segundo índice aponta uma utilização denominada “radical”, representando os usuários que, quando desaprovam determinada publicação, não manifestam sua opinião contrária, mas deixam de seguir o autor da publicação e/ou ocultam o conteúdo. Observou-se que aqueles usuários que se declararam *espectadores*, possuíam tendências de comportamento *online* mais radicais.

Pesquisas interculturais afirmam que para que a persuasão seja mais eficaz é necessário conhecer fatores culturais para elaborar as melhores estratégias persuasivas. Neste sentido, Khaled et al. (2006b) propõe um conjunto de estratégias para culturas coletivistas, sendo elas: 1) *Opiniões de grupo*: Pessoas de culturas coletivistas tendem a se importar mais com o que os outros membros do grupo vão pensar delas; se sentem desmotivadas quando acham que estão fazendo algo isoladamente. Levando em conta esse aspecto, é proposto um sistema de recomendação separado por grupo, que reúna pessoas com interesses parecidos com foco na opinião desse mesmo grupo e partir do

perfil desse grupo são feitas as recomendações, reduzindo a ideia de isolamento; 2) *Vigilância em grupo*: culturas coletivistas tendem a usar a vergonha e exposição perante ao grupo como uma forma de persuasão. Dessa forma, os indivíduos de um mesmo grupo são compelidos a ajudarem e se sentirem úteis neste grupo e são mais efetivos quando pertencem a este grupo. A estratégia proposta é tentar utilizar formas de penalizar o grupo ao invés do indivíduo, quando um comportamento indesejado acontece; 3) *Condicionamento por desaprovação*: enquanto para indivíduos seria antiético usar incentivos negativos, culturas coletivistas tendem ser mais acostumados com esse tipo de incentivo e mais reativos a eles. Uma forma de usar incentivo negativo seria reforçar a desaprovação de um certo comportamento perante ao grupo; 4) *Monitoramento de desvios*: coletivistas tentam se encaixar no padrão do grupo, uma estratégia seria notificar quando o indivíduo está saindo desse padrão; 5) *Personalização do grupo*: geralmente coletivistas se comportam de acordo com o contexto do grupo que está inserido. Ainda, que as pessoas tenham personalidades individuais, a capacidade de personalizar em prol do grupo pode ajudar as pessoas a se adaptarem dentro do grupo.

4. Metodologia

Esta pesquisa busca comparar como diferentes culturas reagem aos mesmos estímulos do *Facebook* comparando usuários brasileiros e não brasileiros. Tendo como base os tipos de perfis identificados por Ruas (2016) e o padrão de uso do *Facebook* que o autor identificou nos usuários brasileiros, foi realizada essa pesquisa aplicando um questionário nos mesmos moldes para usuários não brasileiros. Para discutir as diferenças culturais foi utilizada as dimensões culturais propostas por Hofstede (2011) e para descrever a correlação das diferentes culturas com as tecnologias persuasivas foi utilizado os estudos de Khaled et al. (2006a).

Para isso, foi aplicado um questionário contendo as mesmas perguntas utilizadas por Ruas (2016) para reproduzir a pesquisa com usuários não brasileiros. O questionário foi aplicado de forma inteiramente virtual, disponibilizado no próprio *Facebook*, via *e-mail* e no site *Reddit*[2] de Agosto de 2017 a Outubro de 2017. No *Facebook* o questionário foi divulgado em comunidades acadêmicas e entre pessoas que fizeram intercâmbio em outros países. No *reddit*, onde foi obtido o maior número de respondentes, o questionário foi divulgado dentro da comunidade de cada país. O questionário foi disponibilizado em inglês sendo dividido em duas partes: a primeira é relativa a questões demográficas (nacionalidade, idade, gênero, escolaridade) e a segunda refere-se a como o respondente utiliza as ferramentas da rede para interagir no *Facebook*, os usuários responderam de forma anônima. Nessa etapa, os usuários tiveram que responder em qual perfil ele melhor se encaixava: se *Espectador* (o usuário predominantemente vê o que se passa na rede), *Participante* (o usuário curte, comenta e/ou compartilha conteúdo) e *Produtor de conteúdo* (se ele compartilha conteúdo original na rede). Além disso, foram realizadas perguntas em relação à maneira como o usuário reage ao se deparar com conteúdo com o qual ele concorda ou discorda na rede. As perguntas eram em relação à frequência que o usuário usa as opções "Curtir", "Comentar" e "Compartilhar" ao concordar com um conteúdo e "Comentar", "Ocultar postagem" e "Deixar de seguir" ao discordar do conteúdo.

A análise dos dados foi realizada utilizando-se técnicas de estatísticas descritiva

[2]Reddit é um site de mídia social no qual os usuários podem divulgar ligações para conteúdo na Web.

para analisar como os dados se correlacionam. Utilizou-se a Análise de Componentes Principais para a partir das seis variáveis sobre as interações na rede social (curtir, comentar, compartilhar quando concorda; e comentar, ocultar, deixar de seguir quando discorda) identificar as componentes principais que explicam o modelo.

Para validar as hipóteses a respeito dos componentes principais encontrados, foi aplicado o teste de análise de variância (ANOVA), que testa a hipótese de que as médias de duas ou mais populações são iguais avaliando a importância de um ou mais fatores, comparando-se as médias de variáveis de resposta nos diferentes níveis de fator. A hipótese nula afirma que todas as médias de população (médias de nível de fator) são iguais, enquanto a hipótese alternativa afirma que pelo menos uma é diferente. No caso desse trabalho, o teste ANOVA foi utilizado para verificar se existe diferença significativa entre as médias e se os fatores exercem influência nas variável dependentes (Espectador, Participante e Produtor de conteúdo).

5. Resultados e discussões

Nesta seção são descritos os resultados do questionário para os usuários não brasileiros, além de uma comparação com as mesmas análises a partir dos brasileiros.

5.1. Análise do questionário

O questionário elaborado para usuários não brasileiros ficou disponível de Agosto a Outubro de 2017. Foram coletadas 378 respostas, porém foram excluídos 30 respostas provenientes de respondentes que afirmaram não utilizar a rede social *Facebook* e/ou serem usuários brasileiros. As nacionalidades e o número de respondentes ficaram assim distribuídos: indiano (21), australiano (32), americano (56), mexicano (123), Portugueses (53), chinês, canadense, Britânico, Dinamarquês dentre outros (63). O questionário voltado para usuários brasileiros ficou disponível entre Outubro a Dezembro de 2015 e Ruas (2016) obtive 686 respostas no total.

No caso dos Brasileiros, a maioria dos respondentes são do sexo feminino, totalizando 53,4% (367). 38,92% dos respondentes possuem entre 18 e 23 anos. Quanto ao grau de escolaridade, 52,33% (359) dos respondentes afirmaram possuir superior incompleto. Em relação ao perfil de uso da rede, eles se declaram da seguinte forma: 4,51% são criadores de conteúdo, 46,20% são participantes e 49,27% são espectadores. Quanto aos Não Brasileiros, percebe-se que a maioria dos respondentes são do sexo masculino (65,8%), com idade entre 24 e 35 anos, de nacionalidade mexicana. Quanto à escolaridade, 34,4% afirmaram possuir curso superior completo. Dentre os perfis observados, 4,9% são criadores de conteúdo, 22,4% são participantes e 72,2% são espectadores.

Comparando-se os dois resultados demográficos, em porcentagem, é possível notar algumas diferenças. A base de dados estrangeira possui mais respondentes do sexo masculino (65,8%), enquanto a brasileira possui mais usuários do sexo feminino (46,6%). No caso dos estrangeiros, a maioria dos respondentes são mais velhos que os brasileiros na faixa entre 24 e 35 anos. Além disso, enquanto a maioria dos respondentes brasileiros possuem curso superior incompleto, a maioria dos estrangeiros possuem curso superior completo.

5.2. Princípio da redução

Como já definido, o princípio de redução visa reduzir um comportamento complexo em tarefas simples, aumenta a relação custo/benefício do comportamento e influencia os usuários a executar o comportamento alvo (Fogg, 2003).

As três principais funcionalidades encontradas no *Facebook* para que o usuário manifeste concordância quando vê um *post* que concorda, foi "Curtir, Compartilhar e Comentar". Sendo "Curtir" e "Compartilhar" as ações mais fáceis a serem executadas, sendo possível sua execução com apenas um clique, encaixando na estratégia de redução de Fogg (2003). Em contrapartida, para a ação de "comentar", o usuário precisa redigir um texto, o que é mais trabalhoso que as funções acima. Para manifestar discordância de um *post*, o usuário pode usar o comentário manifestando na escrita a discordância, ocultar a postagem fazendo com que ela não apareça mais na sua tela e/ou deixar de seguir o autor da postagem, fazendo com que todas as postagens daquele autor não apareçam mais na sua *timeline*. Considerando que as duas últimas funcionalidades, além de serem executadas facilmente, sendo concluídas com poucos cliques, é visível apenas para quem as realizam, ou seja, os outros usuários da rede não são informados que determinada publicação foi ocultada .

A Tabela 2 apresenta a frequência com que usuários brasileiros e não brasileiros utilizam as funções ao se depararem com um conteúdo com o qual concordam. Nesta tabela é apresentada também a média ponderada[3] para cada funcionalidade utilizada.

Tabela 2. Funções utilizadas para manifestar *aprovação* de conteúdo

	Não Brasileiros			Brasileiros		
Frequência	**Curtir**	**Comentar**	**Compartilhar**	**Curtir**	**Comentar**	**Compartilhar**
Nunca	23 (6,61%)	72 (20,69%)	121 (34,77%)	10 (1,46%)	64 (9,33%)	26 (3,79%)
Raramente	66 (18,97%)	160 (45,98%)	136 (39,08%)	37 (5,39%)	256 (37,32%)	216 (31,49%)
Às vezes	142 (40,8%)	106 (30,46%)	77 (22,13%)	158 (23,03%)	246 (35,86%)	329 (47,96%)
Quase sempre	78 (22,41%)	6 (1,72%)	5 (1,44%)	254 (37,03%)	88 (12,83%)	88 (12,83%)
Sempre	39 (11,21%)	4 (1,15%)	9 (2,59%)	227 (33,09%)	32 (4,66%)	27 (3,94%)
Média	3,13	2,17	1,98	3,95	2,66	2,82

Percebe-se que a função mais utilizada pelos usuários brasileiros para manifestar aprovação de um conteúdo é a função "Curtir" (com média de 3,95), que exige apenas um clique. De modo contrário, a função menos utilizada é de "Comentar" (com média de 2,66). No caso dos não brasileiros, a ação preferencial também é "Curtir" (média = 3,13), seguida pelas ações de "comentar" e "compartilhar". Observado que a grande maioria dos usuários estrangeiros se consideraram espectadores (72,7%), a baixa utilização das funções pode ser explicada devido a esse fato. A diferença entre a utilização dos estrangeiros é que eles tendem a utilizar a ação de "Comentar" com mais frequência do que a de "Compartilhar".

A Tabela 3 mostra a frequência com que usuários brasileiros e não brasileiros utilizam as funções "Comentar", "Deixar de seguir"e "Ocultar" postagem ao se deparar com um conteúdo com o qual discorda.

Nos dois casos foi observado que a opção de "Comentar" é a menos utilizada em

[3]Esta média é calculada da seguinte forma: soma-se os produtos do número de respondentes que selecionaram cada resposta e o peso de cada resposta, e dividi-se pelo número total de respondentes. Foi atribuído o valor 1 para "Nunca utilizo", 2 para "Raramente utilizo", 3 para "Às vezes utilizo", 4 para "Quase sempre utilizo" e 5 para "Sempre utilizo".

Tabela 3. Funções utilizadas para manifestar *desaprovação* de conteúdo

	Não Brasileiros			Brasileiros		
Frequência	**Comentar**	**Deixar de seguir**	**Ocultar**	**Comentar**	**Deixar de seguir**	**Ocultar**
Nunca	170 (48,85%)	17 (4,89%)	139 (39,94%)	306 (44,61%)	208 (30,32%)	209 (30,47)
Raramente	115 (33,05%)	130 (37,36%)	93 (26,72%)	216 (31,49%)	191 (27,84%)	164 (23,91%)
Às vezes	54 (15,52%)	96 (27,59%)	81 (23,28%)	123 (17,93%)	179 (26,09%)	152 (22,16%)
Quase sempre	2 (0,57%)	10 (2,87%)	22 (6,32%)	27 (3,94%)	67 (9,77%)	85 (12,39%)
Sempre	7 (2,01%)	95 (27,30%)	13 (3,74%)	14 (2,04%)	41 (5,98%)	76 (11,08%)
Média ponderada	1,74	3,10	2,07	1,87	2,33	2,50

casos de discordância. As funções de "ocultar publicação" e "deixar de seguir", mesmo sendo realizadas com apenas dois cliques, ainda assim são pouco utilizadas pelos usuários brasileiros. No caso dos não brasileiros, esse padrão é similar, mas é possível observar que eles tendem a deixar de seguir usuários com uma frequência maior do que os usuários brasileiros quando discordam de determinada publicação.

5.3. Relação da utilização das funcionalidades e o perfil dos usuários

Para analisar como cada perfil de usuário (espectador, participante e produtor) reage ao concordar ou discordar de um conteúdo, aplicou-se a Análise de Componentes Principais. Dessa forma, é possível transformar as seis variáveis relacionadas à concordância de conteúdo (*"curtir", "compartilhar" e "comentar"*) e discordância (*"comentar", "ocultar publicação" e "deixar de seguir"*) em componentes e, a partir dos componentes principais, extrair as informações mais relevantes que explicam a maior parte dos dados.

Conjuntamente, estas seis variáveis explicam o comportamento dos usuários em relação à frequência de utilização das funcionalidades, de acordo com a sua concordância ou discordância a um determinado conteúdo. A frequência com que os usuários interagem na rede social foi medida por meio de uma escala Likert, em que foi atribuído o valor 1 para "Nunca utilizo", 2 para "Raramente utilizo", 3 para "Às vezes utilizo", 4 para "Quase sempre utilizo" e 5 para "Sempre utilizo".

Para ser possível a comparação, foi adotado o mesmo critério de análise para usuários brasileiros e não brasileiros, ou seja, foram mantidas as componentes com autovalores > 1, mantendo-se as combinações lineares que conseguem explicar pelo menos a mesma quantidade de variância de uma variável original padronizada (Mingoti, 2005). As Figuras 1(a) e 1(b) apresentam os valores obtidos para cada componente, para os não brasileiros e brasileiros, respectivamente.

Dessa maneira, analisando os autovalores (*Eigenvalues*) apresentados nas Figuras 1(a) e 1(b), que indicam a quantidade de variância nos dados originais, apenas as duas primeiras componentes principais, PC1 e PC2, foram utilizadas para ambos os casos. Estas componentes principais explicam, conjuntamente, 65,2% da variabilidade total dos dados no caso brasileiro e 69,9% no caso dos não brasileiros.

As componentes resultantes tem comportamentos parecidos tanto para usuários brasileiros quanto não brasileiros. A PC1 pode ser interpretada como um índice de utilização geral das funções para se interagir com os conteúdos no *Facebook*. Se determinado usuário obtém um valor numérico alto para esta componente, significa que ele possui um alto grau de utilização para todas as funções, ou seja, este usuário utiliza frequentemente, tanto as funções para manifestar sua concordância em relação a determinado conteúdo, quanto para demonstrar sua discordância.

Figura 1. Análise de componentes principais de usuários não brasileiros e brasileiros

Eigenanalysis of the Covariance Matrix

Eigenvalue	2,475	1,443	0,581	0,525	0,354	0,225
Proportion	0,442	0,257	0,104	0,094	0,063	0,040
Cumulative	0,442	0,699	0,803	0,897	0,960	1,000

Variable	PC1	PC2	PC3	PC4	PC5	PC6
Curtir (c)	0,498	0,330	0,738	0,075	0,124	-0,278
Comentar (c)	0,380	0,250	-0,063	-0,294	0,190	0,817
Compartilhar (c)	0,359	0,324	-0,389	0,707	-0,337	0,037
Comentar (d)	0,349	0,291	-0,502	-0,545	0,033	-0,494
Deixar de seguir (d)	0,392	-0,505	0,131	-0,240	-0,716	0,063
Ocultar (d)	0,451	-0,619	-0,178	0,232	0,566	-0,083

(a) Não Brasileiros

Eigenanalysis of the Covariance Matrix

Eigenvalue	2,677	1,671	0,753	0,678	0,544	0,328
Proportion	0,402	0,251	0,113	0,102	0,082	0,049
Cumulative	0,402	0,654	0,767	0,869	0,951	1,000

Variable	PC1	PC2	PC3	PC4	PC5	PC6
Curtir (c)	0,210	-0,439	-0,432	-0,505	-0,390	0,412
Comentar (c)	0,216	-0,454	-0,034	-0,012	-0,206	-0,839
Compartilhar (c)	0,222	-0,507	-0,205	0,249	0,753	0,151
Comentar (d)	0,227	-0,393	0,773	0,212	-0,251	0,298
Deixar de seguir (d)	0,562	0,292	0,301	-0,620	0,338	-0,100
Ocultar (d)	0,702	0,325	-0,287	0,503	-0,247	0,070

(b) Brasileiros

A PC2, por outro lado, pode ser entendida como um índice de utilização radical. Apesar das diferenças nos sinais das componentes, a PC2 mantém o mesmo comportamento entre os brasileiros e não brasileiros. No caso dos brasileiros, um usuário que obtém um valor numérico mais elevado para esta componente, tem relativamente uma frequência maior na utilização das funções de discordância "ocultar publicação" e "deixar de seguir o usuário" em comparação com as demais, já que elas representam uma mudança de direção e são usadas em casos contrários a outras variáveis. Usuários que possuem um valor numérico alto para essas componentes são considerados usuários mais "radicais" se considerarmos essas funcionalidades mais extremas, no caso de discordância a determinado conteúdo. No caso dos usuários não brasileiros, a PC2 também representa um índice de utilização radical, mas de forma inversa, o que significa que quanto menor o valor obtido por um usuário nessa componente, maior a frequência na utilização das funções de discordância "ocultar publicação" e "deixar de seguir o usuário".

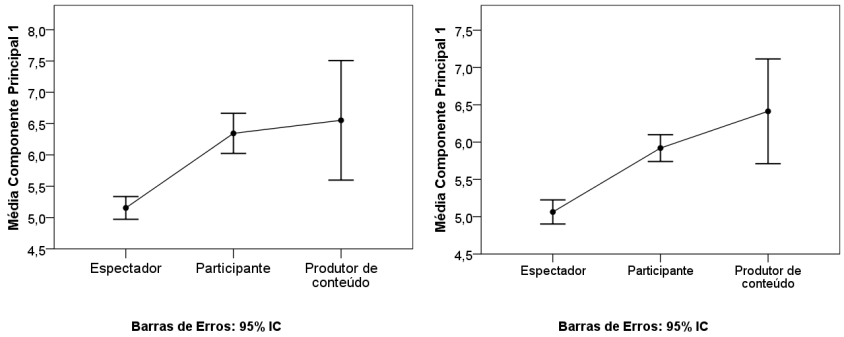
Encontradas as componentes principais, aplicou-se o teste ANOVA (Análise de variância) para ajudar a entender o significado de cada componente principal em relação ao perfil do usuário. Para isso foi feito o caminho inverso, sendo este cálculo feito a partir do *score* das componentes multiplicado pelo valor das respectivas variáveis, e com o resultado é possível reconstruir a base a partir dos Componentes Principais. Assim, com os valores da componente principal é aplicado a ANOVA para verificar a relação entre os valores das componentes e o perfil do usuário.

Observa-se pela Figura 2 (PC1) que o perfil "espectador" possui média de utilização geral menor que os demais perfis. Esta afirmação possui um grau de confiança estatístico de 95%. Os perfis de usuários "participantes" e "produtor de conteúdo" não possuem diferenças significativas ao nível de 5% de significância.

Analisando as componentes principais, a PC1 (grau de utilização geral) teve resultados parecidos entre usuários brasileiros e não brasileiros, de forma que aqueles que se declaravam espectadores acabaram tendo um baixo índice de utilização geral, como esperado e descrito por Ruas (2016).

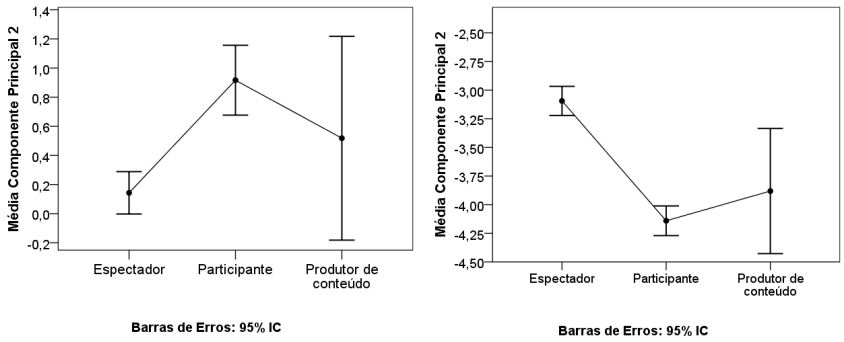
No caso da PC2, grau de utilização radical (Figura 3), também não houve diferença entre as diferentes culturas. Assim, os brasileiros e não brasileiros demonstraram comportamentos parecidos ao se utilizar a rede social. Os usuários radicais, representados pelos espectadores, geralmente optam pela opção mais simples, que é o de

Figura 2. Relação entre o grau de utilização geral (PC1) e o perfil do usuário para não brasileiros e brasileiros



(a) Não Brasileiros (b) Brasileiros

Figura 3. Relação entre o grau de utilização radical (PC2) e o perfil do usuário para não brasileiros e brasileiros



(a) Não Brasileiros (b) Brasileiros

"ocultar a postagem" ou "deixar de seguir" ao discordar de um conteúdo.

Assim, nos dois casos avaliados, brasileiros e não brasileiros, os usuários espectadores tendem a esconder mais a discordância a respeito de uma publicação com a qual não aprova, indicando talvez uma preocupação com a opinião do grupo. Ressalta-se que mesmo isolando os dados e comparando apenas países de culturas individualistas (EUA e Austrália) com a Brasileira os resultados foram os mesmos (dados não apresentados).

6. Considerações finais e proposta de trabalhos futuros

Neste trabalho levantou-se a hipótese de que diferentes culturas podem requerer estratégias diferentes de persuasão. Para isso, analisamos como usuários brasileiros e não brasileiros reagem quando aprovam ou desaprovam um conteúdo postado na rede social *Facebook*, por meio das funções "curtir", "comentar", "compartilhar", "ocultar a postagem" e "deixar de seguir".

Para as funcionalidades analisadas, não foi possível observar diferenças na maneira de utilizar a rede social entre brasileiros e não brasileiros, ainda que a maioria dos respondentes sejam de culturas similares. Isso indica que nas funcionalidades analisadas as estratégias utilizadas pelo *Facebook* tiveram o mesmo efeito persuasivo nessas culturas.

Como propostas de trabalhos futuros, sugere-se: 1) aumentar a base de dados

para as nacionalidades consideradas, além de considerar usuários de outras nacionalidades não previstas (com uma cultura mais coletivista, por exemplo); 2) realizar uma análise separada por dimensões culturais do uso do *Facebook*; 3) caracterizar, por meio de regras, o perfil dos usuários de cada nacionalidade.

Referências

S. K. Card, A. Newell, e T. P. Moran. *The Psychology of Human-Computer Interaction*. L. Erlbaum Associates Inc., Hillsdale, NJ, USA, 1983.

G. Hofstede e M. H. Bond. Hofstede's culture dimensions: An independent validation using rokeach's value survey. *Journal of Cross-Cultural Psychology*, 15(4):417–433, 1984.

R. Khaled, J. Noble, e R. Biddle. An analysis of persuasive technology tool strategies. In *7th International Workshop on Internationalisation of Products and Systems*, pages 167–173, 2005.

R. Khaled, P. Barr, J. Noble, R. Fischer, e R. Biddle. Our place or mine? exploration into collectivism-focused persuasive technology design. In *Proceedings of the First International Conference on Persuasive Technology for Human Well-being*, PERSUASIVE'06, pages 72–83, Berlin, Heidelberg, 2006a. Springer-Verlag.

R. Khaled, R. Biddle, J. Noble, P. Barr, e R. Fischer. Persuasive interaction for collectivist cultures. In *Proceedings of the 7th Australasian User Interface Conference*, AUIC '06, pages 73–80, Darlinghurst, Australia, 2006b. Australian Computer Society.

B. J. Fogg e D. Iizawa. Online persuasion in facebook and mixi: A cross-cultural comparison. In *Proceedings of the 3rd International Conference on Persuasive Technology*, PERSUASIVE '08, pages 35–46, Berlin, Heidelberg, 2008. Springer-Verlag.

G. Hofstede, G. J. Hofstede, e M. Minkov. *Cultures and Organizations: Software of the Mind, Third Edition*. McGraw-Hill Education, 2010.

P. H. B. Ruas, C. N. Nobre, e A. M. P. Cardoso. A influência das estratégias persuasivas no comportamento dos usuários no facebook. In *Proceedings of the 13th Brazilian Symposium on Human Factors in Computing Systems*, IHC '14, pages 255–264, Porto Alegre, Brazil, 2014. Sociedade Brasileira de Computação.

P. H. B. Ruas, A. M. P. Cardoso, L. E. Zarate, e C. N. Nobre. Caracterização do comportamento dos usuários da rede social facebook utilizando métricas de redes complexas e algoritmos de clusterização. In *Proceedings of Satellite Events of the 30th Brazilian Symposium on Databases*, volume 1, pages 39–44, 2015.

B. Fogg. A behavior model for persuasive design. In *Proceedings of the 4th International Conference on Persuasive Technology*, Persuasive '09, pages 40:1–40:7, New York, NY, USA, 2009. ACM.

B. J. Fogg. *Persuasive Technology: Using Computers to Change What We Think and Do*. Morgan Kaufmann, San Francisco, CA, USA, 2003.

G. Hofstede. Dimensionalizing cultures: The hofstede model in context. *Online Readings in Psychology and Culture,*, 2(1):8, 2011.

S. A. Mingoti. *Análise de dados através de métodos de estatística multivariada: uma abordagem aplicada*. Editora UFMG, 2005.

P. H. B. Ruas. Influência de princípios persuasivos no comportamento de usuários de redes sociais: uma análise no Facebook. Master's thesis, Pontifícia Universidade Católica de Minas Gerais, Belo Horizonte, 2016.

Uma Análise do Mercado de Ações Baseada na Correlação entre Ativos no StockTwits

**Gabriela B. Alves, João Paulo S. R. Bastos,
Michele A. Brandão, Adriano C. M. Pereira**

Universidade Federal de Minas Gerais (UFMG) – Belo Horizonte, MG – Brasil

{gabrielabrant,joaopaulosr,micheleabrandao,adrianoc}@dcc.ufmg.br

Abstract. *StockTwits is a social microblog for the financial and investing community which is becoming very popular. In this paper, we investigate why companies are mentioned together in StockTwits. Furthermore, we analyze how this information can be used to help on decision making in the stock market. In particular, we propose a new graph-based model in which stocks are nodes and edges are formed when stocks are mentioned together in the same post. Then, we analyze the main features and isolated pairs in the network. The results show that stocks cited together are correlated with financial results on stock market.*

Resumo. *O StockTwits é um microblog social cada vez mais popular e voltado para o público interessado em mercado financeiro. Neste trabalho, o porquê de empresas distintas serem citadas juntas no StockTwits é investigado. Além disso, analisamos como utilizar essa informação para auxiliar na tomada de decisão no mercado de ações. Especificamente, é proposta uma modelagem em grafo na qual os ativos são os nós e as arestas se formam quando os ativos são citados juntos em uma postagem. Em seguida, é feita uma análise das principais características e dos pares isolados da rede. Os resultados mostram que ativos citados juntos estão correlacionados com as variações no mercado financeiro.*

1. Introdução

As redes sociais são cada vez mais importantes e não param de surgir variações delas para as mais diversas finalidades (e.g., interação social, troca de fotos, compartilhamento de música, etc.). Entre as mais conhecidas e exploradas por pesquisas acadêmicas está o Twitter [Ciotec et al. 2014, Santos et al. 2015], que com seu formato dinâmico e postagens de no máximo 140 caracteres, atrai muitos usuários a compartilharem opiniões e informações. Na mesma linha de fluxo rápido de informações e postagens curtas, o StockTwits[1] é uma rede social criada para a disseminação de conteúdo sobre finanças. Diferente de outras redes sociais, o StockTwits possui como público-alvo pessoas interessadas no mercado financeiro, dando possibilidade aos usuários amadores de interagirem livremente com profissionais.

Até o momento, o mercado de ações associado a redes sociais ainda é pouco explorado pela área de Ciência da Computação [Sprenger et al. 2014]. Atualmente, é possível encontrar pesquisas que utilizam análise de sentimentos para prever movimentações no

[1]StockTwits: https://stocktwits.com/

mercado de ações [Oh and Sheng 2011]. Outros se envolvem em esforços para encontrar os usuários mais experientes e que podem oferecer twits[2] de maior relevância. Esses twits podem ser utilizados como entrada para a modelagem de aprendizado de máquina objetivando prever movimentos no mercado de ações [Bar-Haim et al. 2011].

Diferente dos trabalhos citados, o objetivo principal aqui é analisar a correlação entre ativos, os quais identificam empresas que estão na bolsa de valores e são representados através de $TICKERS. Por exemplo, $GOOG e $APPL representam os ativos da Google e da Apple, respectivamente. Inicialmente, a rede de ativos é modelada como um grafo em que os nós são os ativos e as arestas se formam quando os ativos são citados em um mesmo twit. Assim, são considerados apenas os twits que possuem dois ou mais ativos citados. Em seguida, são realizadas análises topológicas para entender as características da rede e análise dos pares isolados (componentes desconectados) para estudar a interação entre pares de ativos. Note que os pares isolados são considerados por não se repetirem com a mesma frequência que os pares pertencentes ao componente conectado.

O estudo da correlação entre ativos aqui apresentado é inédito, portanto, este trabalho mostra o valor dessa linha de análise para auxiliar à tomada de decisão no mercado financeiro. Além disso, é investigado como a quantidade de citações está relacionada com a variação no preço das ações dos pares mais citados. Também é brevemente analisado o impacto que notícias podem causar na variação dos resultados financeiros dos ativos. Especificamente, duas perguntas principais de pesquisa motivam este trabalho: **P1:** Quais as características de uma rede de ativos? e **P2:** Por que ativos são citados juntos?

O restante deste artigo está organizado da seguinte forma: a Seção 2 discute os trabalhos relacionados. A Seção 3 descreve os materiais e métodos utilizados para desenvolvimento da pesquisa. Em seguida, a Seção 4 discorre sobre os resultados encontrados e suas análises. Finalmente, a Seção 5 apresenta as principais conclusões.

2. Trabalhos Relacionados

Existem trabalhos que exploram conteúdo de redes sociais e revelam conclusões promissoras sobre a importância de analisar esses tipos de dados para prever o mercado de ações. Grande parte dos trabalhos realizam análise de sentimentos em sua metodologia. Por exemplo, Oliveira et al. (2013) utilizam indicadores de sentimento associados a um modelo de regressão para prever três variáveis do mercado: retorno, volatilidade e volume negociado. Tais autores chegam a conclusão de que prever o andamento do mercado é uma tarefa muito complexa e que são necessários modelos muito bem embasados. Estudos mais recentes, como o realizado por Li et al. (2017) , consideram métodos mais específicos de análise de sentimentos, através de indicadores léxicos associados à redes neurais, e obtém resultados melhores na predição de sentimento das postagens (twits).

Ademais, Atkins et al. (2018) estudaram como as notícias impactam na volatilidade do mercado de ações. Nessa pesquisa, utilizaram métodos de aprendizado de máquina aplicado a notícias de jornais e concluíram que a predição do mercado de ações apresenta maior acurácia quando notícias midiáticas são consideradas nos modelos. Similarmente, neste trabalho, são utilizadas fontes de notícias que envolvem pares de ações selecionados, buscando entender a correlação entre eles.

[2] *Twits*: termo utilizado no StockTwits que funciona de forma similar aos *tweets* do Twitter.

Figura 1. Principais etapas para construir e investigar a rede de ativos.

Existem linhas de pesquisa que investigam a influência dos usuários em relação ao mercado financeiro. Por exemplo, Wang et al. (2014) utilizam dados do StockTwits e do SeekingAlpha[3] para determinar quem são os usuários especialistas e se suas postagens têm utilidade para investidores individuais. Similarmente, Tu et al. (2016) investigam os especialistas no StockTwits por meio de técnicas de aprendizado de máquina e utilizam as postagens de usuários *experts* para gerar recomendações de portfólio.

Nesse contexto, a modelagem em grafos aqui proposta é pouco explorada, sendo que foi encontrada apenas uma pesquisa que modela os dados do StockTwits em grafo. Diferentemente de nossa abordagem, Cortez et al. (2016) modelam a interação entre usuários em grafo, considerando os *retweets*, compartilhamentos e *replies*. Assim, o objetivo é medir a influência dos usuários no StockTwits, partindo do pressuposto que usuários mais influentes são mais úteis para criar modelos de previsão no mercado financeiro.

Portanto, a principal contribuição deste trabalho é um estudo da correlação entre ativos no StockTwits através de uma modelagem em grafos em que os ativos são os nós. O objetivo é identificar as características (e.g., nível de agrupamento, conexão entre vizinhos) da rede e entender o motivo dos ativos serem citados juntos. Além disso, a análise dos pares considerando notícias fornece indícios do porquê deles serem citados juntos repetidas vezes. Assim, essas formas de análise têm potencial para contribuir para a tomada de decisão no mercado financeiro.

3. Materiais e Métodos

Esta seção apresenta os materiais e métodos utilizados para investigar a rede de ativos. Uma visão geral das etapas para realização deste trabalho é apresentada na Figura 1 com suas cinco principais etapas: (1) construção da rede de ativos na qual os nós são ações e a aresta se forma quando um par é citado no mesmo twit; (2) cálculo das métricas topológicas para cada mês com o objetivo de analisar a variação das características durante o ano; (3) remoção dos índices da rede; (4) seleção dos vinte pares mais citados de cada mês; e (5) análise de um par isolado para cada trimestre do ano para entender como a variação no número de citações se relaciona com os dados financeiros daquelas ações.

Descrição do conjunto de dados. Os dados utilizados foram gentilmente cedidos pela plataforma StockTwits, portanto, não houve fase de coleta ou tratamento dos dados. A base dos dados contém 22.434.953 twits postados de 1º de janeiro a 31 de dezembro de 2015, sendo que desses, 37,2% citam algum tipo de ativo. Visto que o principal objetivo é entender o motivo de ativos serem citados juntos, são considerados apenas twits com dois ou mais ativos. No total, foram processados um montante de 991.772 twits.

[3]Seeking Alpha: https://seekingalpha.com/

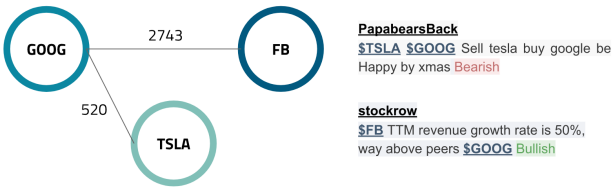


Figura 2. Modelo da rede de ativos

Tabela 1. Propriedades topológicas

Métrica	Definição	Interpretação
Coeficiente de Clusterização (CC)	Sendo $T(u)$ o número total de triângulos aos quais u pertence e $deg(u)$ o grau (número de arestas conectadas) de u, então CC é: $CC(u) = \frac{2T(u)}{deg(u)(deg(u)-1)}$.	Representa a tendência de um nó formar uma comunidade, ou seja, a tendência dos anos em agruparem-se.
Neighborhood Overlap (NO)	Sendo $\mathcal{N}(u)$ e $\mathcal{N}(v)$ o conjunto de nós u e v vizinhos, respectivamente, então $NO(u,v) = \frac{\lvert\mathcal{N}(u)\rvert \cap \lvert\mathcal{N}(v)\rvert}{\lvert\mathcal{N}(u)\rvert \cup \lvert\mathcal{N}(v)\rvert}$	Mede a similaridade entre dois pares de nós vizinhos e pode gerar bons insumos a respeito de como os ativos interagem entre si.
Preferential Attachment (PA)	Dado $\mathcal{N}(u)$ e $\mathcal{N}$ como conjuntos de vizinhos de u e v, respectivamente, $PA(u,v) = \lvert\mathcal{N}(u)\rvert\lvert\mathcal{N}(v)\rvert$	Assume o potencial de receber novas arestas através do grau do nó, sendo que quanto maior o número de vizinhos, maior o valor do preferential attachment.
Coeficiente Adamic-Adar (AA)	No contexto de rede social, as características são os vizinhos em comum e a métrica é formalizada da seguinte forma: $AA(u,v) = \frac{\sum \forall z \in \lvert\mathcal{N}(u)\rvert \cap \lvert\mathcal{N}(v)\rvert}{log\lvert\mathcal{N}(z)\rvert}$	Dado um conjunto de características, essa métrica foi originalmente criada para computar a similaridade entre duas páginas Web.

Construção da rede social de ativos. A rede foi modelada como um grafo ponderado $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, sendo $\mathcal{V}$ o conjunto de nós (representando os ativos), e $\mathcal{E}$ o conjunto de arestas não direcionadas (formadas quando um par de ativos é citado no mesmo twit). O peso das arestas é o total de vezes que um par de ativos é citado junto. A Figura 2 mostra um exemplo da formação da rede de ativos, baseado em dois twits aleatórios. Os pares de ativos ($TSLA, $GOOG) e ($FB, $GOOG) se conectam, pois foram citados juntos na mesma postagem. O peso é um valor fictício utilizado apenas para exemplificar.

Para entender o impacto da quantidade de vezes em que um par de ativos é citado junto, as análises são realizadas considerando os pares de ativos mensalmente. Dessa forma, a escolha mensal permite analisar dados mais específicos e melhor correlacionar com o mercado e suas variações.

Análise topológica. Para responder à pergunta de pesquisa **P1**, foram calculadas métricas topológicas para cada mês do ano de 2015. As métricas topológicas caracterizam a estrutura da rede, considerando como os nós e arestas interagem entre si. Em particular, o objetivo é comparar as variações destas características ao longo dos meses. Em nosso estudo, as métricas de análise de redes complexas utilizadas foram coeficiente de clusterização, *neighborhood overlap*, *preferential attachment* e Adamic Adar, calculadas na forma de média e detalhadas na Tabela 1.

Análise dos pares isolados. Além da análise topológica, é necessária uma análise mais aprofundada para entender como os pares de ativos citados juntos podem ser utilizados para auxiliar na tomada de decisão no mercado de ações. Assim, são feitas análises em pares de ativos com o objetivo de responder a pergunta **P2** deste trabalho.

Nessa etapa, consideramos apenas os vinte pares de ativos mais citados de cada mês (rede TOP20), pois o objetivo é fazer uma análise detalhada. Em média, o par mais citado possui aproximadamente 1098 citações e o vigésimo par mais citado 364 citações. A partir dos grafos TOP20 de cada mês, questiona-se o porquê de alguns pares aparecerem isolados do componente principal. Em geral, os componentes conectados principais são formados pelos mesmos ativos, como AAPL, FB, TWTR, GOOGL, TSLA, NFLX, etc. Portanto, os pares isolados são mais interessantes dado o seu comportamento atípico, em que aparecem uma vez ou poucas vezes entre os vinte pares mais citados nos meses. Os índices, *Exchange Traded Funds* (ETFs)[4], não são considerados nas análises de pares isolados, visto que os índices representam diversos ativos e o escopo do trabalho é entender porque ativos são citados juntos.

Na rede TOP20 de cada mês, existem diversos pares isolados, portanto, é necessário uma abordagem para selecionar os pares a serem estudados, descrita a seguir. Inicialmente, o ano de 2015 foi dividido em trimestres, uma vez que três meses é o intervalo em que as empresas liberam seus balanços financeiros. Em seguida, analisamos os dados de cada trimestre e selecionamos o par isolado com maior número de citações. Após selecionado o par mais citado de cada trimestre, o gráfico de preço anual de cada par é comparado com a quantidade de citações mês a mês. Após a análise anual, os ativos são estudados em uma granularidade de tempo menor, ou seja, o mês em que o par mais citado do trimestre aparece é analisado em termos diários. No gráfico diário, são considerados apenas dias úteis, pois é quando o mercado de ações está aberto. Finalmente, a variação dos preços e números de citações é estudada utilizando notícias midiáticas com o objetivo de apresentar explicações para a correlação entre ativos.

4. Resultados

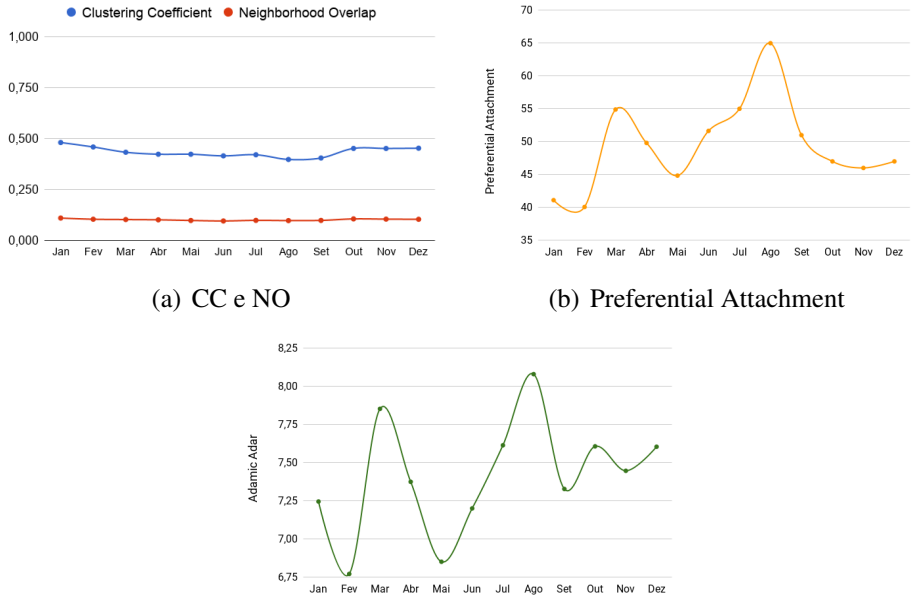
Esta seção apresenta resultados da análise topológica (Seção 4.1) para responder a pergunta de pesquisa P1 e da análise de pares isolados (Seção 4.2) para responder P2.

4.1. Análise Topológica

Para investigar as características da rede de ativos, as métricas coeficiente de clusterização, *neighborhood overlap*, *preferential attachment* e *Adamic Adar* são consideradas e apresentadas na Figura 3. Os valores das métricas são referentes à rede completa de cada mês, ou seja, consideram todos os ativos que apareceram naquele mês, incluindo índices. Note que essa análise não é feita para a rede TOP20, pois o relevante aqui é entender as características da rede como um todo e sua variação ao longo do ano.

A Figura 3(a) mostra que as características são homogêneas durante o ano. O coeficiente de clusterização se manteve estável ao longo de 2015, variando apenas entre 0,4 e 0,5. Isso significa que os pares de ativos tendem em média a formar triângulos [David and Jon 2010]. Em outras palavras, a citação de dois pares de ativos tende a ocasionar a citação conjunta de um terceiro par com ativos já envolvidos em citações anteriores. Para mercado financeiro, isso pode indicar, por exemplo, que ativos de mesmo setor tendem a se influenciar mutualmente. Os resultados para neighborhood overlap são constantes ao longo do ano, se mantendo em torno de 0,1. Em geral, valores próximos de

[4]Um índice de ação é uma carteira teórica (i.e., não existe de fato, é apenas calculada) criada com objetivo de que os investidores tenham um indicador de comparação de um conjunto de ações.



(a) CC e NO

(b) Preferential Attachment

(c) Adamic Adar

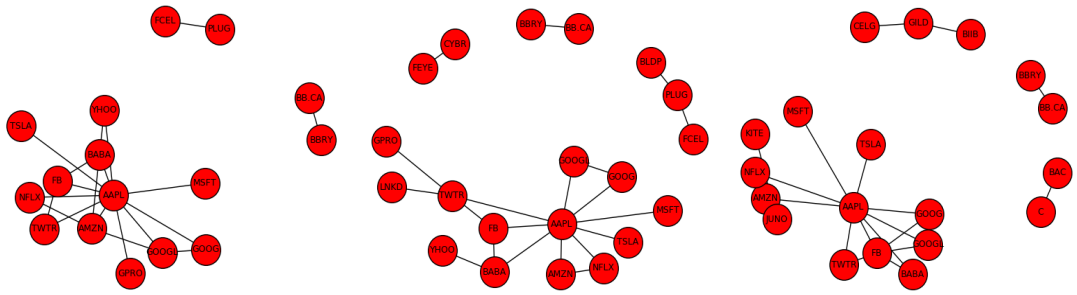
Figura 3. Valor médio das métricas topológicas para o grafo completo de citações de ativos de cada mês.

0 indicam que existem muitas pontes locais na rede e aqui representa que existem diversos ativos formando uma ponte entre componentes conectados [David and Jon 2010].

Por outro lado, a Figura 3(b) mostra picos de variação em alguns meses, como março e agosto que apresentam valores elevados comparados a seus adjacentes. *Preferential attachment* refere-se a ideia de que há uma tendência de novos nós se conectarem a nós que possuem alto grau, causando o efeito rico fica mais rico. Assim, considerando o contexto financeiro, um alto PA pode indicar que existem ativos que se destacam na rede em meses específicos. Em relação à Figura 3(c), existe uma variação grande entre os meses. A métrica *Adamic Adar* calcula a similaridade entre dois nós e a intuição por trás da métrica é que características raras são mais importantes do que as triviais. Dessa forma, pares com alto valor de AA tendem a ser novidade na rede e não são citados aleatoriamente ou influenciados por ativos sempre citados. A análise da Figura 3(c) mostra que nos meses de fevereiro e maio não houveram pares de ativos muito raros, enquanto que em março e agosto observa-se a maior presença de novos pares. Por exemplo, nesses quatro meses, os ativos AAPL e SPY são sempre citados em conjunto com outros pares. Assim, um ativo citado em conjunto com esses pode não indicar alguma mudança no mercado financeiro, enquanto que o surgimento de outra citação como BNI-UNP pode trazer mais informação relevante.

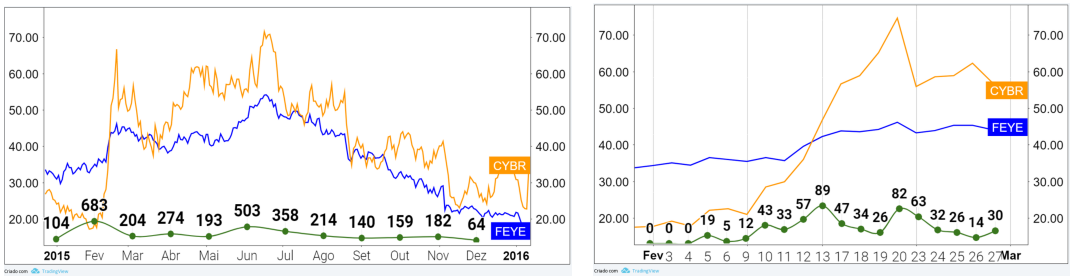
4.2. Análise dos Pares Isolados

Esta seção apresenta a análise dos pares isolados da rede dos pares TOP20 de ativos mais citados a cada trimestre conforme os grafos apresentados nas Figuras 4, 6, 8 e 10. Observa-se que em todos os meses existe um componente conectado predominante e outros isolados. Neste artigo, são realizadas análises sobre pares os isolados, como explicado na Seção 3.



(a) Janeiro (b) Fevereiro (c) Março

Figura 4. Primeiro trimestre - Rede de ativos TOP20



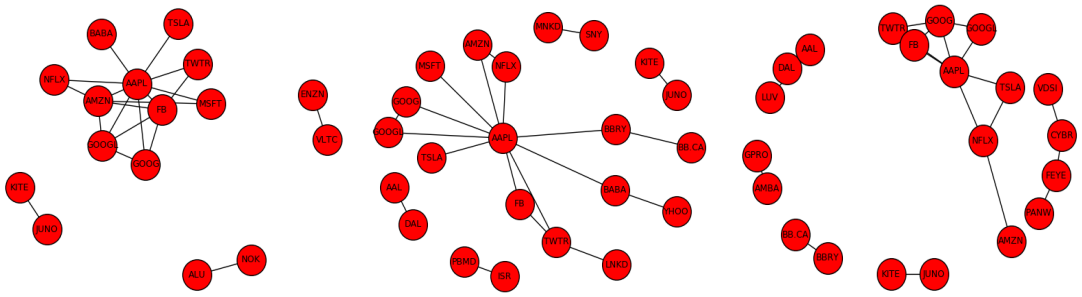
(a) Janeiro a Dezembro de 2015 (b) Fevereiro de 2015

Figura 5. CYBR (laranja), FEYE (azul) - Variação de preço em contraste com número de citações

Primeiro trimestre. Começando a análise pelo primeiro trimestre de 2015, vemos na Figura 4 que existem diversos componentes nos meses de janeiro, fevereiro e março. No entanto, consideramos elegíveis apenas os componentes isolados e entre eles, analisaremos o mais citado. O par mais citado ocorre em fevereiro com 683 citações, sendo ele a aresta CYBR-FEYE. O fato desse par não aparecer nos outros meses nos grafos de TOP20 dá indícios de que houve algum acontecimento relevante naquele mês, levando essas duas empresas a serem citadas juntas muitas vezes.

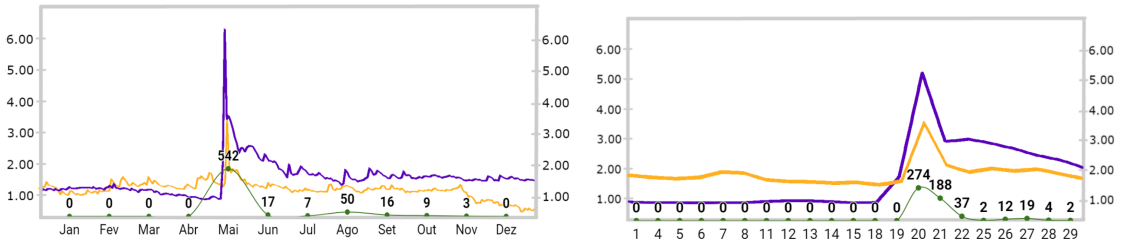
Ambas empresas são do setor de Tecnologia da Informação, sendo que a CyberArk Software Ltd, representada pelo *ticker* CYBR, é uma empresa Israelense que oferece soluções em segurança da informação voltadas para proteção contra ataques cibernéticos. Em comparação, o *ticker* FEYE representa a FireEye Inc. que oferece soluções baseadas em inteligência artificial, também voltada para segurança cibernética.

A Figura 5(a) mostra o gráfico da variação do preço das duas ações justapostos no período de janeiro a dezembro de 2015. A linha laranja refere-se à ação CYBR e a linha azul à ação FEYE. A linha verde na parte inferior do gráfico mostra a quantidade de citações por mês desse par de ativos. É possível observar que o gráfico das duas ações em questão tem picos e comportamentos incomuns nos meses em que os ativos são mais citados. Além disso, analisando especificamente o mês de fevereiro de 2015, vemos na Figura 5(b) a variação diária no preço e número de citações do par CYBR-FEYE. Novamente, o número de citações é representada pela linha verde no canto inferior do gráfico. Vemos que no dia 12 de fevereiro de 2015 ocorreu a inversão das curvas das ações, colocando a FEYE acima da CYBR. Ao mesmo tempo, o número de citações cresceu até alcançar o



(a) Abril (b) Maio (c) Junho

Figura 6. Segundo trimestre - Rede de ativos TOP20.



(a) Janeiro a Dezembro de 2015 (b) Maio de 2015

Figura 7. ISR (laranja), PBMD (azul) - Variação de preço em contraste com número de citações

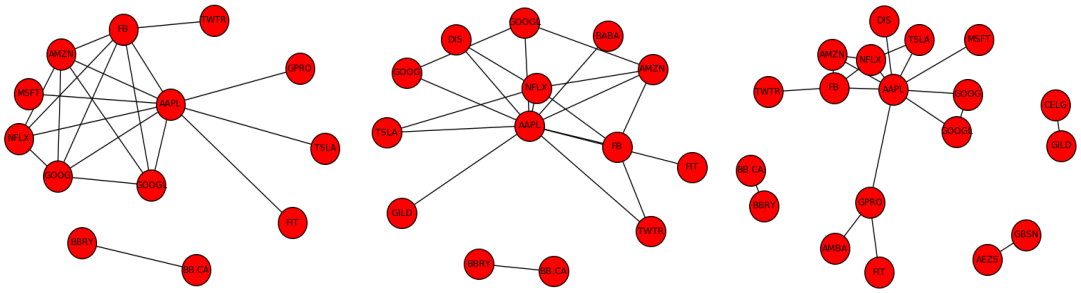
pico no dia 13 de fevereiro. Essa subida abrupta no número de citações e a inversão de posições na linha de preços entre CYBR e FEYE se deu pela divulgação de resultados financeiros da CYBR no mês de fevereiro[5]. Em geral, os resultados superaram as expectativas de analistas de mercado e fizeram com que as ações da empresa tivessem seu valor elevado em aproximadamente 15%, sendo essa mesma subida nos preços apontada como uma possível bolha por analistas na semana seguinte.[6]

Segundo trimestre. No segundo trimestre, também vemos alguns componentes isolados na Figura 6 e o par ISR-PBMD é escolhido entre eles, uma vez que possui a maior quantidade de citações daqueles três meses (542 citações). O par em questão foi o segundo mais citado de maio, ficando atrás apenas de APPL-FB que possui 572 citações. Essa posição reforça a relevância dessa dupla no segundo trimestre.

A empresa IsoRay Inc. é do ramo de biotecnologia. Através de sua subsidiária IsoRay Medical Inc., a empresa desenvolve, manufatura e vende medicamentos a base de isótopos e aparelhos para o tratamento de câncer e outras doenças malignas. Por outro lado, a Prima BioMed Ltd, também voltada para biotecnologia, tem foco em pesquisa, desenvolvimento e comercialização de produtos licenciados no segmento imunoterapêutico.

Ao analisar o gráfico de preço anual das ações na Figura 7(a), o pico de preço da Prima BioMed chama atenção. Tal comportamento atípico está relacionado ao número de citações do par ISR-PBMD em maio, que foi o mais alto do ano inteiro. No mesmo mês, a

[5]The Street: `http://bit.ly/subida-preco-CYBR` - Acesso em 12/2017.
[6]SeekingAlpha: `http://bit.ly/possivel-bolha-CYBR` - Acesso em 12/2017.



(a) Julho (b) Agosto (c) Setembro

Figura 8. Terceiro trimestre - Rede de ativos TOP20.

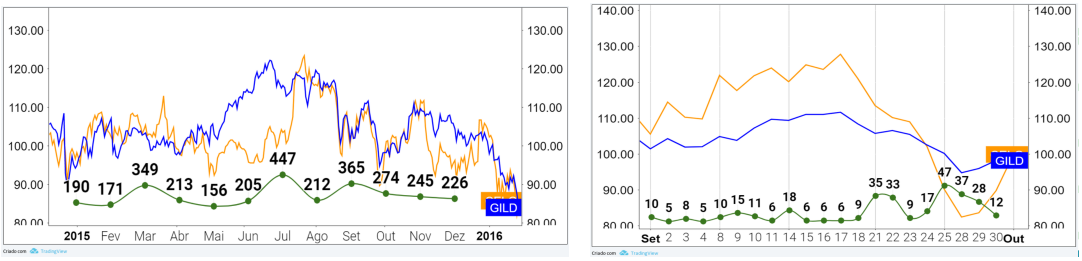
IsoRay Inc. também tem um pico, porém menos significativo que a Prima BioMed. Além do gráfico de preços durante todo o ano, temos também o gráfico que mostra a variação de preço e do número de citações durante o mês de maio, exibidos na Figura 7(b). Em uma reportagem de 20 de maio de 2015, explica-se a subida de 270% no preço das ações da Prima Biomed após o anúncio de resultados promissores nos testes de uma nova droga para o tratamento de câncer de ovário[7]. Na Figura 7(b), vemos que esta é exatamente a data na qual o par de ações tem maior número de citações no mês de maio. Além disso, é interessante ver que o número de citações era baixo e constante nos dias anteriores à essa notícia, tendo um aumento abrupto com a novidade. Por fim, considerando-se que ambas empresas são da mesma área, a descoberta de novos tratamentos em uma delas, gera expectativa de novidades na principal concorrente da área.

Terceiro trimestre. Neste trimestre, o par de citações mais relevante foi CELG-GILD, que aparece no mês de setembro. Essa é a amostra menos significativa entre os trimestres, pois como vemos na Figura 8 os grafos de julho, agosto e setembro são altamente conectados. Assim, temos apenas dois candidatos à análise aprofundada, sendo eles CELG-GILD (365 citações) e AEZS-GBSN (346 citações). O par BBRY-BB.CA representa a mesma empresa, Blackberry Inc.[8], portanto não é um par elegível para a análise.

Por ser o par mais citado do terceiro trimestre, vamos estudar as ações CELG-GILD. A empresa Celgene Corporation representada pelo *ticker* CELG, é do ramo biofarmacêutico e em conjunto com suas subsidiárias, a companhia tem como objetivo a descoberta, desenvolvimento e comercialização de terapias para o tratamento de câncer e doenças inflamatórias. Em comparação, o *ticker* GILD representa a empresa Gilead Sciences Inc., também da área biofarmacêutica voltada para pesquisa e que desenvolve e comercializa medicamentos nas áreas de necessidades médicas ainda sem cura, como Vírus da Imunodeficiência Humana/Síndrome da Imunodeficiência Adquirida (HIV/AIDS), doença hepática, câncer, doenças respiratórias e cardiovasculares.

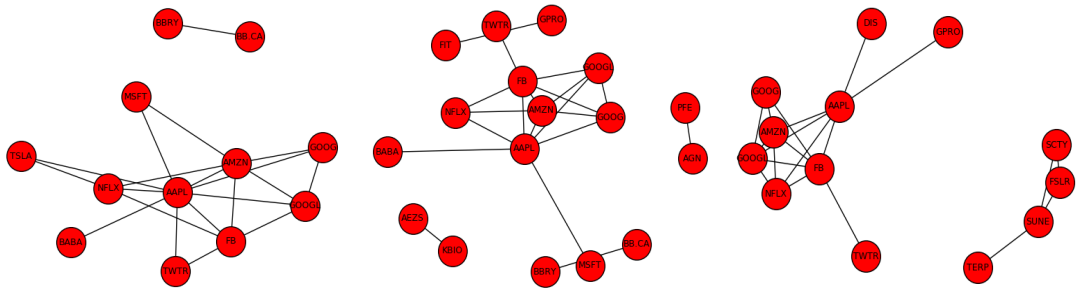
Em 2015, o par selecionado CELG-GILD não tem uma variação muito expressiva na quantidade de citações por mês, como pode ser visto na Figura 9(a). Além disso, o mês de setembro não é o que possui maior número de citações como nos outros pares que estudamos. Ainda assim, é possível observar comportamentos interessantes no gráfico de

[7]Sacks: `http://bit.ly/pbmd-decolou-270-porcento` - Acesso em 01/2018
[8]BB.CA:TSE (Toronto Stock Exchange); BBRY:NYSE (New York Stock Exchange)



(a) Janeiro a Dezembro de 2015 (b) Setembro de 2015

Figura 9. CELG (laranja), GILD (azul) - Variação de preço em contraste com número de citações



(a) Outubro (b) Novembro (c) Dezembro

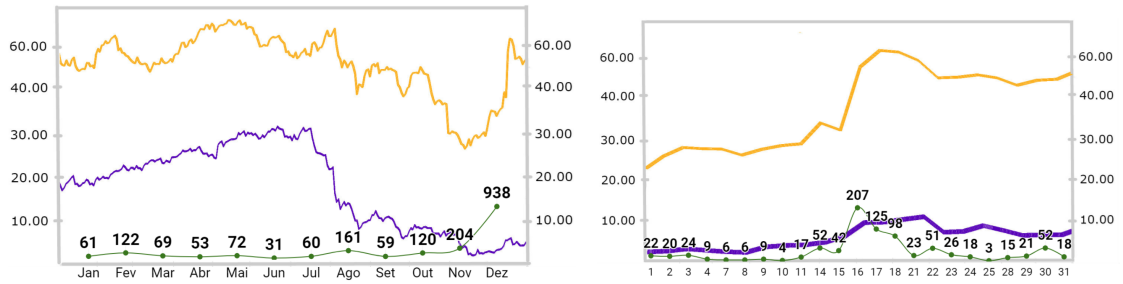
Figura 10. Quarto trimestre - Rede de ativos TOP20.

ações das empresas CELG e GILD, já que o mês de setembro apresentou a queda mais expressiva no preço das duas ações. Esse é um resultado diferente daqueles vistos no primeiro e segundo trimestre, pois está associado à diminuição do valor da ação. Em detalhes, conseguimos observar na Figura 9(b) a varição do preço e número de citações do mês de setembro. No dia 21 de setembro de 2015 houve um pico no número de citações e queda abrupta no valor das ações do par CELG-GILD. Esse foi exatamente o dia em que Hillary Clinton postou um *tweet* prometendo tomar alguma atitude em relação aos altíssimos preços praticados pela indústria farmacêutica[9]. Sendo assim, podemos observar que houve grande impacto do comentário nas ações CELG e GILD, que são da indústria farmacêutica

Quarto trimestre. No último trimestre, o par de ações com o maior número de citações é o SCTY-SUNE, que representam as empresas SolarCity Corporation e SunEdison Inc., respectivamente. Esse par teve 938 citações no mês de dezembro e ambas são empresas do setor energético norte-americano. Entre as análises realizadas nos trimestres anteriores, o mês de dezembro é o que tem quantidade mais expressiva em número de citações.

Sobre as ações aqui analisadas, apresentamos a SolarCity Corporation que oferece serviços relacionados à energia limpa. Entre eles, energia solar, eficiência energética e desenho de veículos elétricos, monitoramento e manutenção em ambiente reseidencial, escolar e governamental nos Estados Unidos. Paralelamente, a SunEdison Inc. opera em dois segmentos: materiais semicondutores e energia solar. A companhia produz e vende

[9]CNN Money: `http://bit.ly/tweet-hillary-clinton` - Acesso em 03/2018



(a) Janeiro a Dezembro de 2015 (b) Dezembro de 2015

Figura 11. SCTY (laranja), SUNE (azul) - Variação de preço em contraste com número de citações

wafers e produtos relacionados à semicondutores e à indústria de energia solar.

As notícias publicadas no mês de dezembro tiveram impacto claro na variação de preço e citações do par de SCTY-SUNE no mês de dezembro. Primeiro, ocorreu, no início de dezembro, a Conferência do Clima Paris 2015 (COP21) que reuniu quase 200 governos mundiais os quais concordaram em limitar a emissão de gases estufa e investir no desenvolvimento de energias renováveis.[10]. Em seguida, no dia 15 de dezembro os rumores sobre a possível aprovação de incentivos fiscais para indústrias do setor energético se espalha[11] e no dia 18 de Dezembro as expectativas se concretizam. O governo norte-americano aprova o incentivo milionário[12]. Essa sequência de notícias comparadas à Figura 11(b) mostra que a expectativa de mercado fez o preço das ações aumentarem. Em seguida, a quantidade de citações estabilizou novamente, após a sequência de novidades.

Após analisar todos os trimestres de 2015, aprofundando em quatro pares de ações, é possível responder à pergunta de pesquisa P2. Os ativos são citados juntos quando são do mesmo setor, ou seja, concorrentes. Em todos os exemplos, essa é uma constante e mostra que os usuários do StockTwits estão atentos às empresas que competem entre si.

5. Conclusão

Este trabalho apresentou indícios de que pares de ativos podem ser utilizados no auxílio à tomada de decisão no mercado financeiro. Até o momento, não identificamos pesquisas nesse sentindo, com foco na análise da correlação entre ativos citados juntos através de uma modelagem em grafos. Os resultados mostraram que os pares de ativos mais citados refletem o comportamento de mercado dos mesmos, em relação a preços e notícias financeiras. Em resposta a pergunta P1, as características da rede de ativos revelaram que há uma tendência na formação de triângulos e pontes na citação de ativos, bem como mostraram a presença de ativos que são citados com muita frequência. Em geral, tais ativos não trazem muita informação nova para análise de mercado. Sobre a questão P2, observou-se que o principal motivo para ativos serem citados juntos repetidas vezes em uma mesma postagem é a divulgação de notícias que impactam as ações em questão. Em todos os casos, as ações são impactadas por mudanças políticas ou econômicas. Com essa informação, conclui-se que ações que atuam no mesmo setor são as mais citadas juntas,

[10]Yahoo Finance: `http://bit.ly/porque-alta-setor-energetico` - Acesso 02/2018
[11]Forbes: `http://bit.ly/investimentos-industria-energitica` - Acesso em 03/2018
[12]Forbes: `http://bit.ly/novidades-congresso-US` - Acesso em 03/2018

mostrando que os investidores no StockTwits estão atentos à competição entre empresas.

Finalmente, ainda há muito a ser explorado e existem diversos trabalhos futuros. Por exemplo, gerar uma nuvem de palavras para cada par de ativos mais citados e entender o que está sendo dito. Também analisar nós que aparecem e saem do componente conectado a cada mês, não apenas os pares isolados. Outra possibilidade é estudar a criação e recomendação de portfólios de investimentos através da análise dos pares de ações.

Agradecimentos. Este trabalho foi parcialmente financiado pelo Instituto Nacional de Ciência e Tecnologia para a Web (grant no. 573871/2008-6), MASWeb (grant FAPEMIG/PRONEX APQ-01400-14), CAPES, CNPq e Fapemig.

Referências

Atkins, A., Niranjan, M., and Gerding, E. (2018). Financial news predicts stock market volatility better than close price. *The Journal of Finance and Data Science*.

Bar-Haim, R., Dinur, E., Feldman, R., Fresko, M., and Goldstein, G. (2011). Identifying and following expert investors in stock microblogs. In *Procs. of EMNLP*, pages 1310–1319, Edinburgh, United Kingdom.

Ciotec, S., Dascalu, M., and Trausan-Matu, S. (2014). A comprehensive study of twitter social networks. In *RoEduNet Conference 13th Edition*, pages 1–7. IEEE.

Cortez, P., Oliveira, N., and Ferreira, J. a. P. (2016). Measuring user influence in financial microblogs: Experiments using stocktwits data. In *Procs. of WIMS*, pages 23:1–23:10, Nîmes, France.

David, E. and Jon, K. (2010). *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge University Press, New York, USA.

Li, Q. and Shah, S. (2017). Learning stock market sentiment lexicon and sentiment-oriented word vector from stocktwits. In *CoNLL*, pages 301–310, Vancouver, Canada.

Oh, C. and Sheng, O. (2011). Investigating predictive power of stock micro blog sentiment in forecasting future stock price directional movement. In *Procs. of ICIS*, page 57–58, Shangai, China.

Oliveira, N., Cortez, P., and Areal, N. (2013). On the predictability of stock market behavior using stocktwits sentiment and posting volume. volume 854, page 355–365, Berlin, Germany.

Santos, H. S., Laender, A. H., and Pereira, A. C. (2015). Uma visão do mercado brasileiro de ações a partir de dados do twitter. In *BraSNAM*, Recife, Brazil.

Sprenger, T. O., Tumasjan, A., Sandner, P. G., and Welpe, I. M. (2014). Tweets and trades: The information content of stock microblogs. *European Financial Management*, 20(5):926–957.

Tu, W., Cheung, D. W., Mamoulis, N., Yang, M., and Lu, Z. (2016). Investment recommendation using investor opinions in social media. In *SIGIR*, pages 881–884, Pisa, Italy.

Wang, G., Wang, T., Wang, B., Sambasivan, D., Zhang, Z., Zheng, H., and Zhao, B. Y. (2015). Crowds on wall street: Extracting value from collaborative investing platforms. In *CSCW*, pages 17–30, Vancouver, Canada.

Visibilidade no Facebook: Modelos, Medições e Implicações

Eduardo Hargreaves[1],
Daniel Menasché[1], **Giovanni Neglia**[2], **Claudio Agosti**[3]

[1]Dept. Ciência da Computação, UFRJ, Rio de Janeiro, Brasil

[2]INRIA, Sophia Antipolis, França

[3]Universidade de Amsterdã, Amsterdã, Holanda

eduardo@hargreaves.tech, sadoc@dcc.ufrj.br,giovanni.neglia@inria.fr

claudio.agosti@logioshermes.org

Resumo. *O Facebook tem um impacto significativo na vida de milhões de usuários da Internet, todos os dias. Entretanto, os mecanismos usados pelo Facebook para filtrar mensagens apresentadas aos usuários não são de domínio público, motivando uma engenharia reversa dos mesmos. Neste trabalho, propomos modelos e medições para melhor entender o comportamento de* timelines. *Em particular, reportamos resultados sobre medições de visibilidade de fontes das eleições italianas de 2018, que motivam um modelo analítico para caracterizar a visibilidade de* posts. *Dentre as implicações práticas de nossos estudos, indicamos seu potencial para inferir diferentes métricas de visibilidade a partir de medidas colhidas do sistema.*

Abstract. *Facebook news feed personalization algorithm has a significant impact, on a daily basis, on the lifestyle, mood and opinion of millions of Internet users. Nonetheless, such algorithms lack transparency challenging researchers to improve their fairness and accountability. In this paper, we propose a model to capture the dynamics of contents over a timeline (also known as news feed). The input to our model is a fundamental quantity associated to timelines, which we show that can be easily parameterized using real world data: the arrival rate of posts of a given publisher followed by the user. Using real world Facebook traces from the latest elections in Italy, we validate the accuracy of the proposed model and use the model for conterfactual what-if analysis.*

1. Introdução

O Facebook tem um impacto significativo na vida de milhões de usuários da Internet, todos os dias [Tsaparas 2017]. Entretanto, os mecanismos usados pelo Facebook para filtrar mensagens apresentadas aos usuários não são de domínio público, motivando pesquisas que envolvem desde estatísticas básicas até engenharia reversa de algoritmos. Tais estudos são fundamentais para garantir certo nível de transparência aos usuários do sistema.

Compreender um sistema complexo como o Facebook não é trivial. A visibilidade de uma certa fonte, por exemplo, depende da taxa de publicação de postagens por parte da fonte e dos interesses dos usuários. No artigo publicado em [TechCruch 2016] é dito que o Facebook utiliza aproximadamente 100.000 variáveis para escolher a publicações exibidas na suas News Feeds. Uma abordagem para avaliar o impacto de diferentes fatores

na visibilidade consiste na criação de modelos analíticos que permitam derivar métricas de interesse a partir de dados colhidos das redes, mas tais dados em geral não estão disponíveis de forma aberta.

Neste trabalho, propomos medições e um modelo para melhor entender o comportamento de *timelines*. O modelo proposto é baseado em modelos de filas e de caches, para os quais existem uma vasta literatura. Em particular, parametrizamos e validamos o modelo proposto usando dados do Facebook, indicando o poder expressivo do mesmo.

Contribuições Dentre as principais contribuições deste trabalho, destacamos as seguintes.

Medições do Facebook Usando usuários virtuais, colhemos visões distintas sobre as postagens no Facebook nas eleições da Itália em 2018. Tais medições motivam a criação de modelos para capturar a essência da dinâmica das publicações em *timelines*.

Modelo de visibilidade Propomos um modelo para estimar a visibilidade das publicações em função das taxas de criação das fontes e do algoritmo de filtragem da rede social. Usando dados reais do *Facebook*, validamos os modelos propostos.

Estudo contrafactual de caso Usando as medições e os modelos propostos, avaliamos qual teria sido a visibilidade de diferentes publicações, sob diferentes critérios de interesse.

O restante deste artigo está organizado da seguinte forma. Na Seção 2 apresentamos a metodologia de medições adotada neste trabalho. Em seguida, a Seção 3 traz os resultados empíricos obtidos nas eleições italianas. A Seção 4 traz uma visão geral sobre o modelo proposto, seguida pela Seção 5 que apresenta o modelo analíticos proposto. A Seção 6 faz uma análise contrafactual dos dados usando o modelo proposto, seguida da Seção 7 que cobre trabalhos relacionados. A Seção 8 conclui.

2. Medições

Obtenção das métricas de interesse Usando a infraestrutura do Facebook Tracking Exposed,[1] criamos seis usuários virtuais no Facebook. Todos os usuários seguem as mesmas fontes. Entretanto, os usuários possuem diferentes perfis (e.g., um curte páginas de direita e outro de esquerda). A cada hora, colhemos as publicações apresentadas na *timeline* de cada usuários. Tais fotografias constituem nossa base de dados.

Cabe destacar que a API do Facebook desde 2015 não oferece os dados acima [Facebook 2018b]. Mesmo quando a API era aberta, o Facebook dizia que a informação fornecida pela API não era fidedigna [Facebook 2018a]. Utilizamos esta API somente para colhermos informações sobre o número de publicações por fonte.

Métricas de interesse Dentre as métricas de interesse, destacamos as seguintes:

Repercussão (probabilidade de acerto) de uma fonte é a probabilidade de um usuário efetivamente ler (e possivelmente clicar) em uma publicação. Nesse trabalho, não distinguimos entre probabilidade de acerto e probabilidade de *click*. A probabilidadde de acerto de uma fonte pode ser dada pela visibilidade ou pela ocupação das postagens desta fonte.

[1]Facebook tracking exposed: `https://facebook.tracking.exposed/`

Visibilidade é a probabilidade de existir uma publicação de uma determinada fonte na *timeline*.

Ocupação é o número de publicações de uma determinada fonte na *timeline*.

Denotamos por π_{ij} a visibilidade da fonte j na *timeline* do usuário i. Denotamos por N_{ij} o número médio de publicações do *publisher* j na *timeline* de i. Neste trabalho, assumimos que a repercussão de uma fonte pode ser dada em função da visibilidade ou da ocupação. A repercussão igual a visibilidade condiz com usuários que eventualmente irão ler uma das publicações de cada uma das fontes presentes nas K primeiras posições de suas *timelines*. A repercussão igual a ocupação, por outro lado, captura o comportamento de usuários que irão influenciar-se mais por fontes que ocupam mais posições em suas *timelines*, ou seja, quanto mais posições ocupadas por uma fonte, maior a repercussão da mesma.

3. Descobertas empíricas sobre a eleição italiana

3.1. Coleta dos dados

As eleições italianas foram no dia 04 de março de 2018 e o experimento ocorreu entre os dias 10 de janeiro de 2018 e 06 de março de 2018 de forma que procuramos analisar o período antecedente às eleições e os dois dias seguintes para capturarmos as reações ao resultado das eleições.

Foram selecionadas 30 fontes italianas que posteriormente foram classificadas de acordo com uma das cinco orientações políticas: esquerda, centro-esquerda, direita, ultra-direita e movimento 5 estrelas. Também foram criados 6 usuários fictícios. Todos os usuários seguem as mesmas 30 fontes. No entanto, os usuários foram polarizados de forma que cada usuário curtia publicações de apenas uma orientação política. O sexto usuário foi caraterizado como indeciso uma vez que não curtiu publicações de numa página.

As "fotografias" das *timelines* foram "tiradas"através de uma extensão dos navegadores Chrome e Firefox chamada *facebook.tracking.exposed*. Essa extensão coleta os dados públicos, e retorna a data de criação da publicação, o momento da visualização, o usuário que visualizou, a fonte, o conteúdo, a quantidade de reações, o número de compartilhamentos, e a ordem de aparição da publicação. Em paralelo, a API do Facebook foi utilizada para a obtenção de todas as publicações das fontes selecionadas. Chamamos de S_i o número de fotografias tiradas no i-ésimo usuário. Nos nossos experimentos, os usuários foram indexados de 1 a 6, denotando as orientações de centro-esquerda, ultra-direita, esquerda, M5S, direita e a indecisa. Os valores de S_i obtidos foram: 577, 504, 623, 674, 655, 576, com $i = 1, \ldots, 6$. A diferença entre o número de fotografias advém de falhas de medição, e.g., máquinas travarem, falta de luz e queda de conexão. Embora existam tais falhas, acreditamos que estatisticamente elas não afetam os resultados, tendo em vista o grande volume de fotografias colhidas.

3.2. Achados empíricos

A seguir, apresentamos uma visão geral dos dados colhidos. A Figura 1 (a) ilustra o número de publicações por fonte. Esta informação foi colhida diretamente da API do Facebook. Algumas poucas fontes geraram milhares de publicações durante o período considerado, enquanto que a maioria gerou dezenas de publicações.

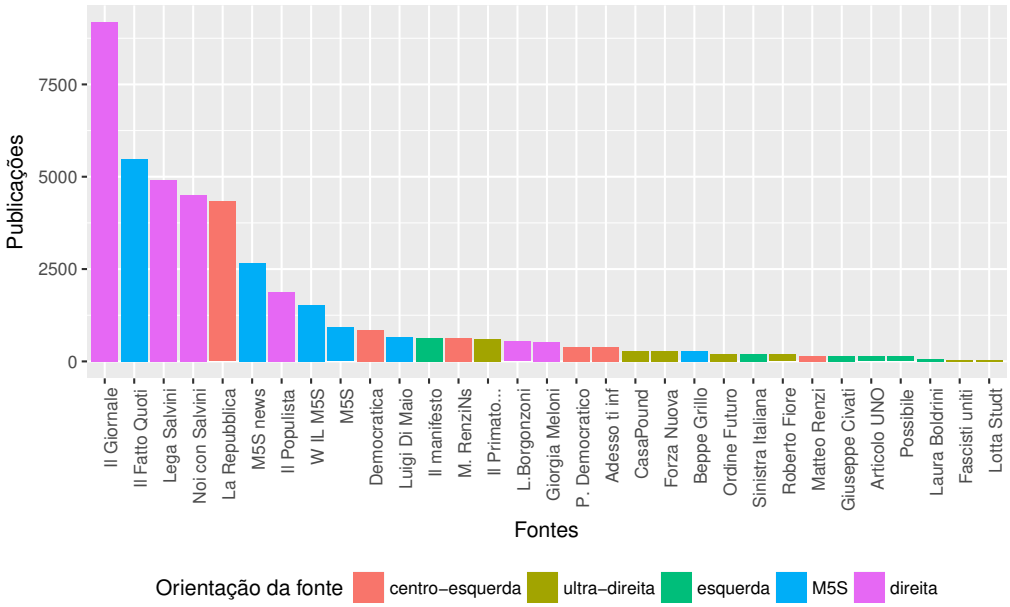


Figura 1. Total de publicações por fonte

Chamamos cada publicação visualizada de impressão, e chamamos de I_{ij} o número total de impressões da fonte j no usuário i. Se uma publicação é visualizada duas vezes, ela conta com duas impressões. Esta informação foi colhida a partir de nosso aplicativo (Facebook Tracking Exposed). A Figura 2 mostra o total de impressões I_{ij} por usuário. Em nenhum usuário a distribuição das publicações vistas foi semelhante a distribuição das publicações criadas. Em todos os usuários, as fontes mais vistas refletiram a polarização dos mesmos. Por exemplo, o usuário que curtiu fontes de esquerda (no topo a direita da Figura 2) visualizou mais postagens de fontes de esquerda do que os demais. Cabe relembrar que todos os usuários seguiram todas as fontes (e apenas distinguiram-se pelas curtidas) e que o viés fica claro independente da taxa com quem as fontes geraram conteúdos (ilustrada na Figura 1).

A Figura 3 (a) mostra o número de impressões em cada um dos seis usuários, agrupadas por orientação política, e a Figura 3 (b) mostra o número total de publicações por orientação política. Essa Figura corrobora as observações anteriores: o viés das publicações reflete o viés dos usuários.

Cabe destacar que a grande diferença entre a ordem de aparição das fontes nas Figuras 1 e 1 é fruto da filtragem realizada pelo Facebook. Um dos objetivos do presente trabalho é propor um modelo analítico que nos permita compreender os efeitos de tal filtragem nas métricas de visibilidade e ocupação das fontes nas *timelines*.

Embora as observações apresentadas acima em parte sejam esperadas, cabe destacar que analisamos também usuários neutros. Para usuários neutros, que não curtem nenhuma fonte, seria de esperar que a presença de publicações fosse semelhante a Figura 1, ou então que existisse uma uniformidade entre as fontes. A Figura 2 mostra que esse não é o caso. Podemos constatar o alto número de impressões da fonte M5S no eleitor indeciso. É importante observar que o partido M5S foi o partido que recebeu mais votos nas eleições italianas. O modelo analítico apresentado a seguir nos permite realizar

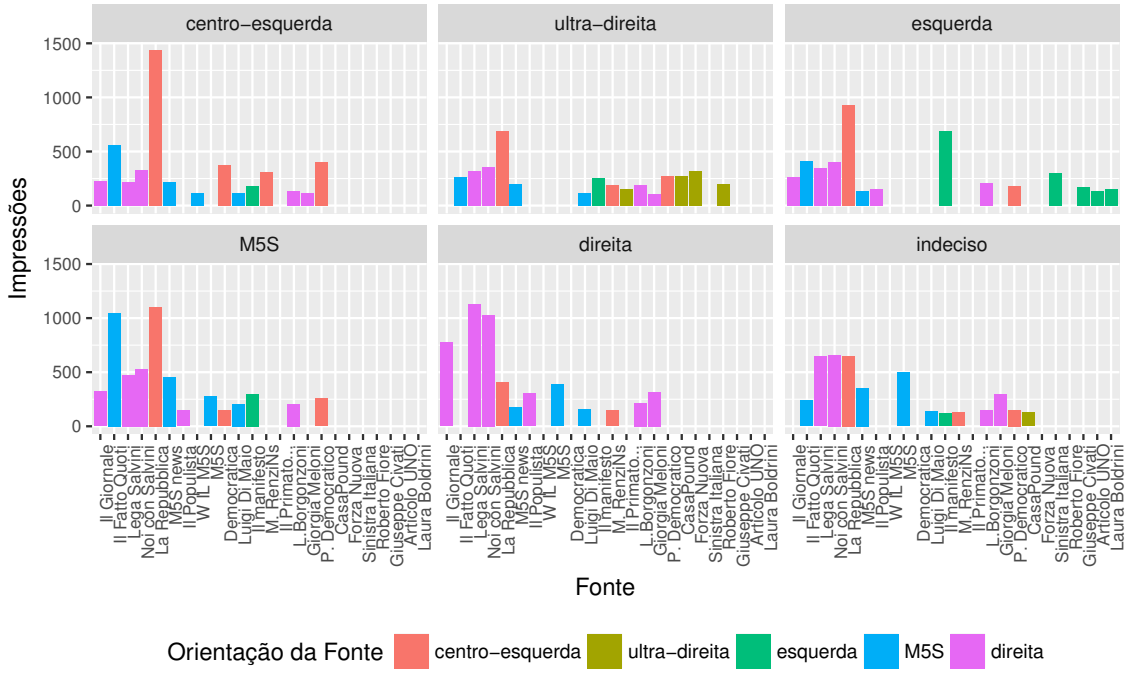


Figura 2. Total de impressões no topo da timeline por fonte em cada usuário (Apenas ocupações superiores a 0.2 estão representadas)

estudos contrafactuais, para averiguar o que ocorre com a visibilidade das fontes caso diferentes filtros sejam adotados.

4. Uma visão geral sobre a modelagem de *timelines*

A seguir, apresentamos intuitivamente as ideias que suportam o modelo analítico introduzido na seção seguinte.

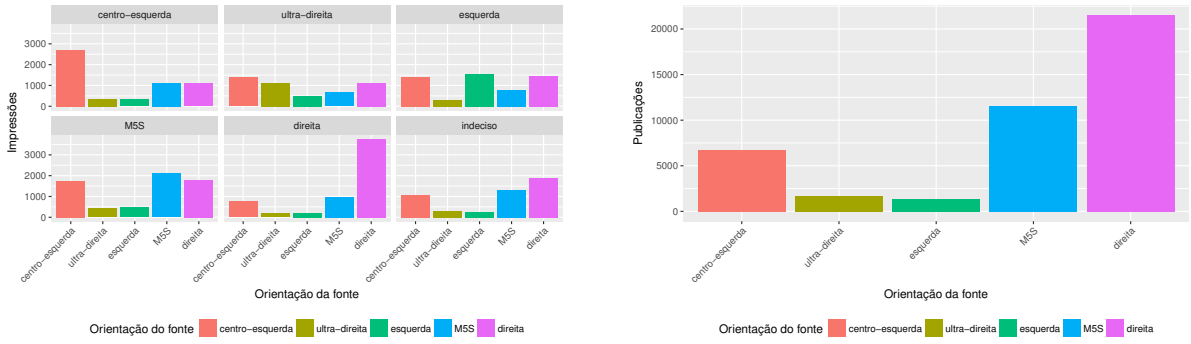
4.1. Analogias entre filas, *caches* e *timelines*

Em redes sociais virtuais, as publicações criadas pelas fontes fluem através das conexões entre os seus respectivos membros e alcançam as *timelines* dos usuários interessados. Desta forma, *timelines* são um canal de comunicação entre fontes e usuários.

Na sua forma mais simples, as publicações são organizadas por ordem de chegada, de forma que estas entram e saem das timelines de acordo com uma ordem cronológica. Os algoritmos de personalização somente filtram as publicações da fonte j na timeline do usuário i, transformando uma taxa de criação de conteúdo Λ_j em uma taxa de exibição de mensagens λ_{ij}.

O comportamento recém descrito, é o comportamento de uma fila do tipo primeiro a entrar, primeiro a sair, ou, como é mais conhecida, como uma fila FIFO (*first-in, first-out*). Desta forma, acreditamos ser natural utilizar a teoria de filas para modelar as métricas de interesses das *timelines*.

Também argumentamos que existem inúmeras similaridades entre *timelines* e *caches* de conteúdos. Tanto *timelines* quanto *caches* são utilizadas para armazenar conteúdos de interesse a usuários. Ambas podem ser encaradas como filtros, tendo em



(a) (b)

Figura 3. (a) Total de impressões no topo da timeline (por orientação política), (b) Total de publicações por orientação política

vista que tem tamanho limitado. Além disso, ambas em geral armazenam conteúdos mais recentes, e removem conteúdos que não são mais de interesse dos usuários. E, principalmente, ambas aumentam a eficiência da obtenção de informação do que está guardado nelas.

Existe uma ampla literatura sobre *caches* e teoria de filas. Estabelecendo a relação entre filas, *caches* e *timelines*, podemos nos aproveitar desta literatura para estudar *timelines* que, por serem mais recentes, receberam menos atenção da comunidade científica em comparação com *caches* e filas, que são melhor entendidas. No caminho oposto, acreditamos que resultados obtidos através dos estudos de timelines podem ser resultar em novas políticas de *caching* ou em novas formas de distribuição de conteúdo.

4.2. *Timelines* são *caches*?

Neste trabalho, um de nossos objetivos é compreender como projetar *timelines* (tendo em vista suas similaridades com *caches*) para melhor prover conteúdo de interesse dos usuários. Cabe destacar, entretanto, que existem importantes diferenças entre *timelines* e *caches*. Usuários de timelines estão tipicamente interessados numa *classe* de itens relacionados a uma certa categoria ou usuários. Esta é a principal diferença entre os dois. Por exemplo, um usuário deseja seguir as últimas notícias de seu jornal favorito. Esta trata-se de uma consulta flexível, em comparação com uma busca mais específica, por exemplo, a busca por um determinado episódio de uma série de televisão. A Tabela 1 ilustra algumas das diferenças entre *caches* e *timelines*.

Pelas razões acima, embora *timelines* e *caches* tenham muitas similaridades, elas necessitam de políticas de inserção e remoção de conteúdos distintas. Algoritmos de *caching* clássicos prestam-se a servir requisições por itens específicos. Algoritmos para *timelines*, em contrapartida, precisam lidar com a distribuição de conteúdo baseada em classes, como por exemplo, um tópico especifico ou uma fonte preferida. Por estes motivos, consideramos que *timelines* são *caches* orientados a fontes ou classes e divergindo dos *caches* tradicionais que são orientados à requisição.

5. Um modelo analítico para *timelines*

A seguir, apresentamos o modelo analítico proposto, seguido por sua validação usando dados das eleições italianas de 2018.

Tabela 1. Comparação entre timelines e caches

	Timelines	Caches
Evento de interesse	publicação de conteúdo	chegada de requisições
Decisões de inserção e remoção	tomadas após a publicação	tomadas depois de um *miss*
Requisições	para classe de conteúdos	para conteúdos específicos
Controle de ocupação	de itens por classe	de itens específicos

5.1. Descrição do modelo

Nesta seção, descrevemos um modelo analítico para capturar a dinâmica das publicações em uma *timeline*.

Tabela 2. Tabela de notação

Variável	descrição
K	número de posições de interesse no topo da *timeline* (*top* K)
j	j-ésima fonte
i	i-ésimo cliente (usuário)
S_i	número de fotografias tiradas no usuário i
I_{ij}	número de impressões da fonte j no usuário i
D_{ij}	número de publicações distintas da fonte j visualizadas pelo usuário i
Λ_j	taxa de criação de publicações da fonte j
λ_{ij}	taxa efetiva de chegada de publicações de j na *timeline* de i
$\lambda_{i,-j}$	taxa efetiva de chegada de publicações de outras fontes (que não j) na *timeline* de i
λ_i	taxa total de chegada na *timeline* de i
T_{ij}	valor esperado para tempo de permanência das publicações da fonte j na *timeline* de i
$\hat{\pi}_{ij}$	visibilidade de j medida na *timeline* de i
π_{ij}	visibilidade segundo modelo
$\hat{N}_{ij}$	ocupação média das publicações da fonte j medida na *timeline* de i
N_{ij}	ocupação média das publicações de j medida na *timeline* de i segundo modelo proposto

Dividimos os membros das redes sociais virtuais entre fontes e usuários. Membros que geram conteúdos são fontes, e membros que consomem tais conteúdos são os usuários. Ao administrar sua *timeline*, um usuário ou grupo de usuários pode subscrever a fontes, e um subconjunto das publicações destas fontes será exibida na *timeline*. Note que numa rede em que a maior parte dos conteúdos é gerada pelos próprios usuários (ex., Facebook), o papel de cada usuário dinamicamente muda entre fonte e usuário.

Seja $\mathcal{I} := \{1, \ldots, i, \ldots I\}$ o conjunto de usuários em estudo, cada usuário associado a uma *timeline*, e seja $\mathcal{J} := \{1, \ldots, j, \ldots J\}$ o conjunto de fontes. Seja i o usuário de interesse, cuja *timeline* desejamos modelar (a notação é sumarizada na Tabela 2).

A fonte j cria publicações segundo um processo Poisson com taxa Λ_j. Seja $\lambda_{ij} \leq \Lambda_j$ a taxa efetiva com que a fonte j alimenta a *timeline* i. A taxa agregada de publicações chegando na *timeline* do usuário i é dada por

$$\lambda_i = \sum_{j=1}^{J} \lambda_{ij} \tag{1}$$

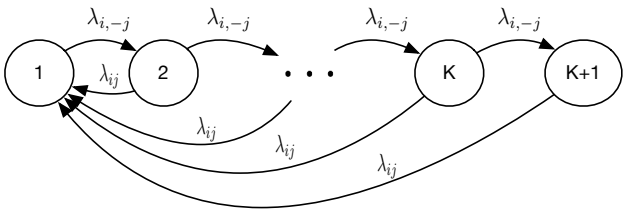


Figura 4. CTMC caracterizando a publicação da fonte j na posição mais no topo na *timeline* do usuário i.

A taxa com que publicações de fontes, diferentes de j, chegam ao usuário i é

$$\lambda_{i,-j} = \lambda_i - \lambda_{ij} \tag{2}$$

Dinâmica FIFO Assumimos que cada publicação entra no topo da *timeline*, na posição 1. Com taxa λ_i, as publicações tem sua posição incrementada em uma unidade. Estamos interessados nas primeiras K posições (slots) da *timeline*. De forma equivalente, assumimos que a *timeline* tem tamanho K. Uma publicação é removida da *timeline* quando ela passa da posição K para a posição fictícia $K + 1$.

Cadeia de Markov A seguir, consideramos uma cadeia de Markov para capturar a evolução das publicações da fonte j, na *timeline* do usuário i. O objetivo é calcular a visibilidade e a ocupação de j na *timeline* do usuário i. Por isso, não levamos em conta a posição de cada publicação de j na *timeline* de i, mas apenas a posição da publicação mais no topo. Enquanto houver um *post* de j na *timeline* de i, a fonte j estará visível.

A Figura 4 mostra uma cadeia de Markov tempo contínuo (CMTC) cuja variável de estado X representa a posição mais no topo ocupada por uma publicação da fonte j. Seja $\tilde{\pi}_{ij}(x) = P(X = x)$. Com taxa λ_{ij}, uma publicação de j chega no topo *timeline*. Com taxa $\lambda_{i,-j}$, uma publicação de outra fonte faz com que as publicações de j sejam movidas para a posição k+1. A fonte j estará fora da *timeline* se a sua publicação mais no topo encontrar-se na posição fictícia $K + 1$. Logo, a visibilidade π_{ij} é dada por

$$\pi_{ij} = 1 - \tilde{\pi}_{ij}(K + 1) \tag{3}$$

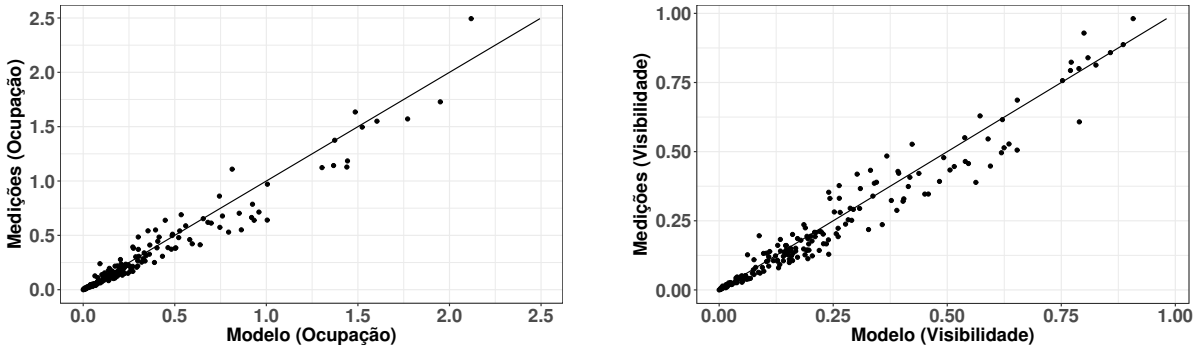
Proposição 1. *Em uma FIFO* timeline*, a visibilidade da fonte j na* timeline *do usuário i é*

$$\pi_{ij} = 1 - \left(\frac{\lambda_{i,-j}}{\lambda_i}\right)^K$$

Demonstração. O resultado segue imediatamente a partir das equações de balanço do sistema, que podem ser derivadas diretamente a partir da Figura 4. □

Proposição 2. *Em uma FIFO* timeline*, a ocupação média da fonte j na* timeline *do usuário i é*

$$N_{ij} = \frac{\lambda_{ij} K}{\lambda_i} \tag{4}$$



(a) (b)

Figura 5. Validação do modelo: (a) Ocupação, (b) Visibilidade

Demonstração. Quando a m-ésima publicação entra na *timeline* ela fica visível até a chegada do $(m + K + 1)$-ésima publicação. Já que o tempo médio entre chegadas é $1/\lambda_i$, o tempo de residência da m-ésima publicação é K/λ_i. Em estado estacionário, seja T_{ij} o tempo médio que publicações de j ficam na *timeline* de i, $T_{ij} = K/\lambda_i$. Pela lei de Little, $N_{ij} = \lambda_{ij} T_{ij} = \frac{\lambda_{ij} K}{\lambda_i}$. □

Resumo modelo FIFO O modelo FIFO é um dos modelos analiticamente tratáveis mais simples que se possa conceber para uma *timeline*. Nesta seção, apresentamos o modelo FIFO e indicamos como ele pode ser usado para derivar a visibilidade e ocupação média de cada fonte em uma *timeline*. Em trabalhos futuros, pretendemos considerar modelos alternativos, mais flexíveis, como aqueles baseados em *caches* do tipo TTL.

5.2. Validação

A seguir, apresentamos a validação do modelo proposto. Nosso objetivo é indicar que o modelo tem capacidade expressiva para capturar os dados colhidos em ambiente real (eleições da Itália), e que ele é de fácil parametrização (requerendo apenas a taxa efetiva de publicações por fonte por usuário).

Chamamos de D_{ij} o número de publicações distantas da fonte j vistas pelo usuário i. Cabe ressaltar que esse número é menor ou igual a I_{ij}. A taxa λ_{ij} é dada pela razão D_{ij}/S_j. Cada taxa obtida λ_{ij} é substituída em (1) e em (2) para a obtenção da visibilidade, através da Proposição (1), e da ocupação, através da Proposição (4).

Cada ponto nas Figura 5 corresponde a um usuário e uma fonte. Um ponto (x, y) indica que, para o usuário e a fonte em questão, o modelo prediz uma ocupação x (respectivamente, visibilidade), e empiricamente observamos ocupação y (resp., visibilidade). Os erros advém do fato, por exemplo, de nosso modelo assumir que as publicações nunca são reordenadas. Ainda assim, o fato de a maioria dos pontos estar próxima à reta $x = y$ ilustra o poder preditivo do modelo. O erro médio quadrático da ocupação foi igual a 0.0971 e o da visibilidade foi igual a 0.0527.

6. Análise contrafactual (*what-if analysis*)

A seguir, combinamos o modelo proposto com os dados colhidos para realizar análise contrafactual de ocupações. Para tal, calculamos o viés entre as ocupações medidas em-

piricamente e as ocupações preditas pelo modelo, usando a quantidade de publicações criadas por cada fonte (API do Facebook). O objetivo é comparar a ocupação após a filtragem do Facebook (ocupação medida empiricamente) com aquela que nosso modelo prediz como sendo a ocupação condizente com um sistema sem filtros (ocupação analítica obtida com o modelo proposto).

O viés é definido como a diferença $\hat{N}_{ij} - N_j$, onde $\hat{N}_{ij}$ é a ocupação média empírica da fonte j na *timeline* i, e N_j é dado por (4) e $\hat{N}_{ij}$ é dado por I_{ij}/S_i. Observe que o índice i referente aos usuários é suprimido na expressão de N_j pois na análise de ocupação ignorando filtros assumimos que a ocupação é igual para todos os usuários.

As Figuras 6 e 7 mostram que, de um modo geral, o viés positivo reflete a orientação dos usuários. No entanto, a fonte que mais produziu publicações, Il Giornale, foi penalizada em todos os usuários, e a segunda fonte que mais produziu, o Il Fatto Quotidiano só não foi penalizado no usuário que curtiu a sua página. O usuário neutro sofreu viés tanto positivo quanto negativo. A fonte M5S teve um forte viés positivo no usuário neutro. No entanto, o forte viés negativo do Il Fatto Quotidiano anulou esse efeito de forma que o viés total do M5S no usuário neutro foi negativo.

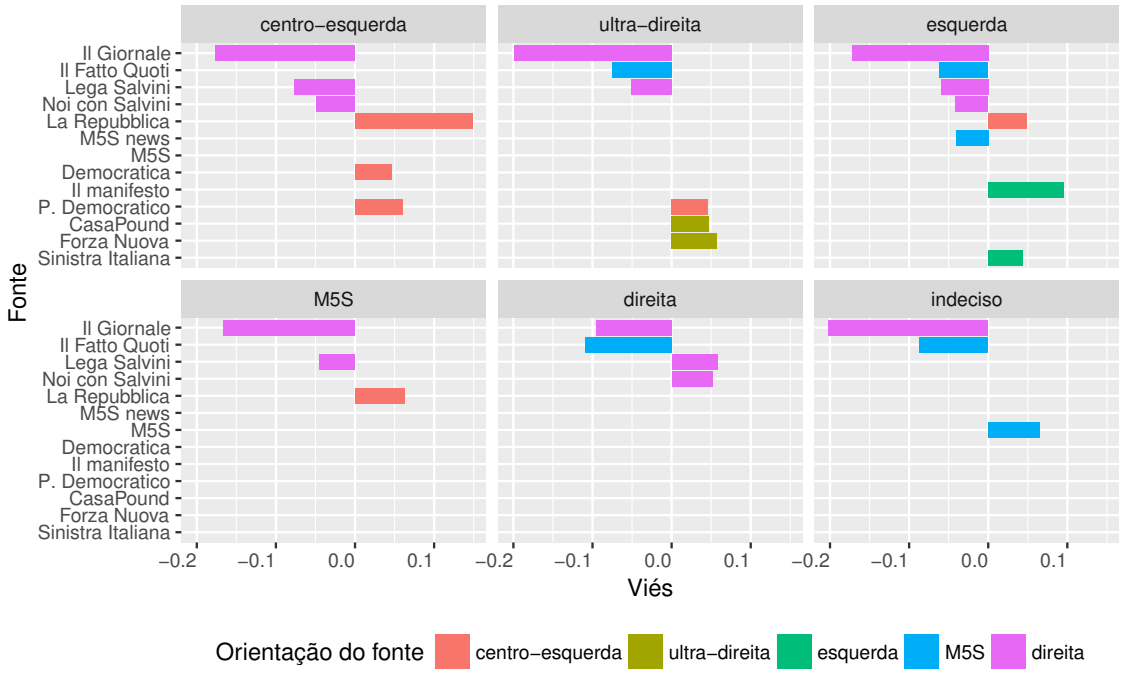


Figura 6. Viés por usuário e por fonte (estão representados apenas os casos em que o valor absoluto do viés foi maior do que 0.4)

7. Trabalhos relacionados

O livro de [O'Neil 2016] mostrou diversas situações nas quais algoritmos podem reforçar preconceitos e tomar decisões que podem influenciar a sociedade como um todo. No contexto de mídias sociais e política, em [Epstein and Robertson 2015] foi demonstrado que manipulações em mecanismos de buscas são capazes de influenciar eleitores indecisos. Robôs fazendo propaganda política tentando influenciar eleições foram estudados no mundo todo em [Woolley and Howard 2017] e particularmente no Brasil em [Arnaudo 2017] . Um estudo realizado por [Eslami et al. 2015] mostrou que mais de

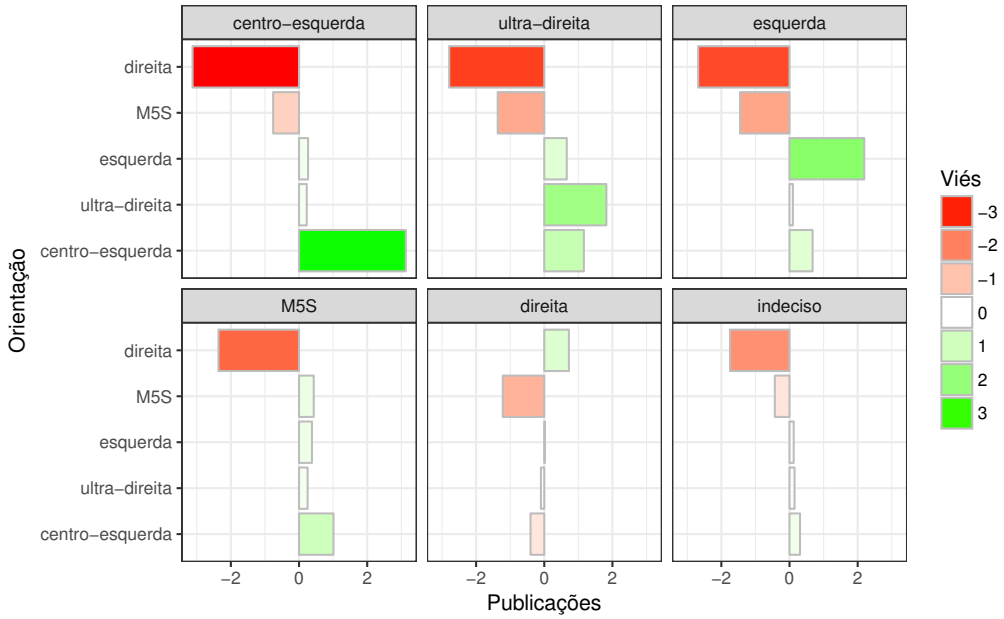


Figura 7. Viés por usuário e por orientação política

60% das pessoas não tinham conhecimento da existência dos filtros de personalização do Facebook. Em [Bakshy et al. 2015] foi identificado que as escolhas individuais são o maior fator de influência sobre o nível de exposições a posições políticas opostas no Facebook. O nosso trabalho, por outro lado, mostrou que o algoritmo reforça as preferências dos usuários. Esse reforço de preferências pode ajudar a criar as *filter bubbles* identificadas em [Pariser 2011].

Neste trabalho, mostramos como modelos de *caches* e filas podem ser usados no estudo de *timelines*. Existe uma vasta literatura sobre ambos [Martina et al. 2014, Dehghan et al. 2016, Harchol-Balter 2013]. Acreditamos que esta conexão aqui estabelecida permita estender-se resultados destes domínios para as *timelines*.

Existe uma vasta literatura focada em fazer engenharia reversa dos algoritmos por trás de *timelines* [Andreou et al. 2018]. Entretanto, a literatura de modelos analíticos nessa área é bem mais escassa. O trabalho de [Altman et al. 2013] foi o primeiro a modelar uma timeline com um fila Fifo. O nosso trabalho propôs uma solução do sistema capaz de obter além da visibilidade, a ocupação das fontes. Em particular, não é de nosso conhecimento nenhum trabalho anterior que tenha feito uso de medições reais de *timelines*, conectando tais medições com modelos analíticos.

8. Conclusões

O Facebook afeta milhões de usuários da Internet todos os dias, e qualquer decisão algorítmica sobre as *timelines* pode ter importantes impactos sociais e políticos. Neste trabalho, propusemos uma metodologia envolvendo medições e um modelo analítico para quantificar métricas de ocupação, visibilidade e viés em *timelines*. Indicamos que o modelo proposto tem poder preditivo, e que permite fazer análise contrafactual de dados. Acreditamos que este seja um importante passo no sentido de garantir maior transparência para os usuários ao torná-los mais informados sobre os processos de filtragem aos quais as publicações por eles visualizadas estão submetidas.

Agradecimentos Este projeto foi em parte conduzido pelo time associado do projeto THANES, com recursos do INRIA (França) e da FAPERJ (UFRJ/Brasil), tendo sido parcialmente financiado também pelo CNPq, CAPES e FAPESP.

Referências

Altman, E., Kumar, P., Venkatramanan, S., and Kumar, A. (2013). Competition over timeline in social networks. *ASONAM*, pages 1352–1357.

Andreou, A., Venkatadri, G., Goga, O., Gummadi, K. P., Loiseau, P., and Mislove, A. (2018). Investigating ad transparency mechanisms in social media. NDSS.

Arnaudo, D. (2017). Computational Propaganda in Brazil : Social Bots during Elections.

Bakshy, E., Messing, S., and Adamic, L. A. (2015). Exposure to ideologically diverse news and opinion on Facebook. *Science*, 348(6239):1130–1132.

Dehghan, M., Massoulie, L., Towsley, D., Menasché, D., and Tay, Y. C. (2016). A utility optimization approach to network cache design. In *Proceedings - IEEE INFOCOM*, volume 2016-July, pages 1–10.

Epstein, R. and Robertson, R. E. (2015). The search engine manipulation effect and its possible impact on the outcomes of elections. *Nat. Academy of Sciences of the United States of America*, 112(33):E4512–21.

Eslami, M., Rickman, A., Vaccaro, K., Aleyasen, A., Vuong, A., Karahalios, K., Hamilton, K., and Sandvig, C. (2015). I always assumed that I wasn't really that close to [her]. *Human Factors in Computing (HCI)*, pages 153–162.

Facebook (2018a). Graph api reference /user-id/home.

Facebook (2018b). Log de alterações da graph api. Retrieved April 6, 2018 from: `https://developers.facebook.com/docs/graph-api/changelog`.

Harchol-Balter, M. (2013). *Performance Modeling and Design of Computer Systems*. Cambridge University Press, Cambridge.

Martina, V., Garetto, M., and Leonardi, E. (2014). A unified approach to the performance analysis of caching systems. *IEEE INFOCOM*, pages 2040–2048.

O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown Publishing Group, New York, NY, USA.

Pariser, E. (2011). *The Filter Bubble: How the New Personalized Web Is Changing What We Read and How We Think*. Penguin Publishing Group.

TechCruch (2016). How facebook news feed works. Retrieved December 31, 2017 from: `https://techcrunch.com/2016/09/06/ultimate-guide-to-the-news-feed/`.

Tsaparas, P. (2017). Online social networks and media. `http://www.cs.uoi.gr/~tsap/teaching/cs-l14/references.html`.

Woolley, S. C. and Howard, P. N. (2017). Computational propaganda worldwide: executive summary.

Analisando a governabilidade presidencial a partir de padrões de homofilia na Câmara dos Deputados: Estudos de Casos no Brasil e nos EUA

Breno de Sousa Matos[1], **Carlos H. G. Ferreira**[1], **Jussara M. Almeida**[1]

[1]Departamento de Ciência da Computação - Universidade Federal de Minas Gerais

{brenomatos, chgferreira, jussara}@dcc.ufmg.br

Abstract. *In this work, we propose an approach to analyze the governability of the president based on roll call votes in the House of Representatives. In order to identify patterns of homophily between the government group and the remaining deputies, we model the roll call votes as a graph to observe how the homophily process impacts on the governability of the president in two countries, i.e., Brazil and United States.*

Resumo. *Neste trabalho, propomos uma abordagem para analisar a governabilidade do presidente com base nas votações realizadas na Câmara dos Deputados. Para isso, modelamos as votações em forma de um grafo, de modo a identificar padrões de homofilia entre a base do governo e os demais deputados em dois países, isto é, Brasil e Estados Unidos.*

1. Introdução

Nos últimos anos, vários países têm vivenciado episódios polêmicos na política. No Brasil é possível citar o *impeachment* da ex-presidente Dilma Rousseff, marcado por uma intensa disputa entre os poderes legislativo e executivo [Fearnside 2016]. Já nos Estados Unidos (EUA), mesmo tendo a maioria das cadeiras ocupadas por deputados do seu partido, ao assumir a presidência em 2017, o presidente Donald Trump apresentou em seus primeiros meses de governo uma profunda dificuldade em manter um bom relacionamento com a Câmara dos Deputados [Washington-Post 2017].

A governabilidade constitui uma importante característica que diz respeito à capacidade de um presidente da república de governar e aplicar as políticas públicas desejadas [Santos 1997]. Porém, em países como o Brasil, isso não é tão simples, devido aos problemas no sistema político, tais como, a alta fragmentação partidária, presente em pelo menos outros 57 países [Wehner 2010], e a permutação de deputados nos partidos durante o governo [Figueiredo 2007]. Contudo, é comum que os presidentes tentem manter o maior número possível de partidos e deputados na base governista para, consequentemente, aumentar a sua governabilidade [Figueiredo 2007]. Já nos Estados Unidos, a existência de uma Câmara dos Deputados dividida entre dois partidos com ideologias cada vez mais extremas, desafia o equilíbrio entre a base governista e oposição [Andris et al. 2015].

As relações entre a base aliada do governo e os demais deputados durante as votações podem ser analisadas por meio da homofilia. A homofilia, também conhecida como assortatividade, é a tendência dos nós presentes em uma rede se conectarem a outros indivíduos similares segundo algum critério [Newman 2003]. Assim, quanto mais a alta

a assortatividade em uma votação na Câmara dos Deputados, menor é a interação entre a base governista do presidente e os demais deputados, o que pode impor dificuldades ao governo do presidente em aprovar as votações de seu interesse. Neste contexto, nós propomos uma abordagem para analisar a governabilidade política do presidente baseada na homofilia em redes de votações na Câmara dos Deputados. Mais especificamente, estamos interessados em verificar como as relações entre os deputados da base aliada e os demais deputados estão associadas à estabilidade política do presidente em exercício. Para isso, avaliamos as votações do plenário da Câmara dos Deputados do Brasil, considerando as coalizões realizadas pelo governo nesse período. De maneira semelhante, nós aplicamos a nossa abordagem nos Estados Unidos.

2. Materiais e Métodos

2.1. Trabalhos Relacionados

A homofilia no contexto político tem sido amplamente estudada em redes sociais [Vergeer 2015]. No Twitter, usuários foram avaliados na eleição americana de 2016 em [Colleoni et al. 2014]. Já em [Halberstam and Knight 2016], são investigadas comunicações políticas em redes sociais caracterizadas por homofilia e tamanho de grupos ideológicos. Durante a votação do *impeachment* da ex-presidente brasileira Dilma Rousseff, um estudo revela que mensagens de um grupo raramente chegam ao grupo de ideias contrárias, caracterizando o processo de homofilia [de França et al. 2018]. Um método foi proposto em [Barber 2014] para estimar a ideologia dos usuários de mídias sociais. Em paralelo, um estudo realizado em [Vaz de Melo 2015] mostra que os partidos políticos brasileiros são altamente redundantes. Assim, o número de partidos que o país deveria ter é muito menor do que o existente. Diferentemente dos trabalhos analisados na literatura, este trabalho busca investigar o processo de homofilia no contexto político, propondo uma nova abordagem aplicada às votações na Câmara dos Deputados.

2.2. Bases de Dados

Os dados das votações nominais do plenário da Câmara dos Deputados do Brasil são disponibilizados através de uma interface de programação de aplicações (API)[1]. Foram coletadas as votações realizadas entre a 53ª e a 55ª legislatura, que compreende o período de 2007 a 2017. Já nos Estados Unidos, os dados das votações foram coletados pela API ProPublica[2]. O período considerado foi de 2009 a 2017, período compreendido entre o 111° e o 115° congresso. Nós também utilizamos a base de dados do Centro Brasileiro de Análise e Planejamento (CEBRAP), que fornece entre as várias informações, início, fim e partidos das coalizões realizadas pelo governo [CEBRAP 2017]. A Tabela 1 sumariza os dados das votações finais utilizadas no Brasil e Estados Unidos.

2.3. Modelagem das Votações em Grafo

No Brasil, um voto pode ser do tipo: *Sim, Não, Abstenção* ou *Obstrução*. De maneira similar, nos Estados Unidos um voto pode ser do tipo *Yes*, *No* e *Not Voting*. No Brasil,

[1] http://www2.camara.leg.br/transparencia/dados-abertos/dados-abertos-legislativo

[2] https://projects.propublica.org/api-docs/congress-api/

Tabela 1. Dados das Votações na Câmara dos Deputados

Brasil							
Legislatura	**Ano**	**Presidente**	**Partido**	**# de Votações**	**# de Votos**	**# de Partidos**	**# de Deputados**
53ª	2007	Lula	PT	209	76649	21	532
53ª	2008	Lula	PT	142	48667	20	542
53ª	2009	Lula	PT	148	50163	19	541
53ª	2010	Lula	PT	75	24700	19	544
54ª	2011	Dilma	PT	93	33809	23	552
54ª	2012	Dilma	PT	73	24883	23	554
54ª	2013	Dilma	PT	137	47180	23	542
54ª	2014	Dilma	PT	81	27686	24	556
55ª	2015	Dilma	PT	224	111897	25	550
55ª	2016	Dilma/Temer	PT/PMDB	163	78521	27	562
55*	2017	Temer	PMDB	170	79473	26	534

Estados Unidos							
Congresso	**Ano**	**Presidente**	**Partido**	**# de Votações**	**# de Votos**	**# de Partidos**	**# de Deputados**
111º	2009	Barack Obama	D	991	403549	3	442
111º	2010	Barack Obama	D	664	273088	3	446
112º	2011	Barack Obama	D	948	393321	2	439
112º	2012	Barack Obama	D	658	270137	2	440
113º	2013	Barack Obama	D	641	257367	2	438
113º	2014	Barack Obama	D	563	229596	2	438
114º	2015	Barack Obama	D	704	286938	2	437
114º	2016	Barack Obama	D	621	255086	2	438
115º	2017	Donald Trump	R	710	226712	2	441

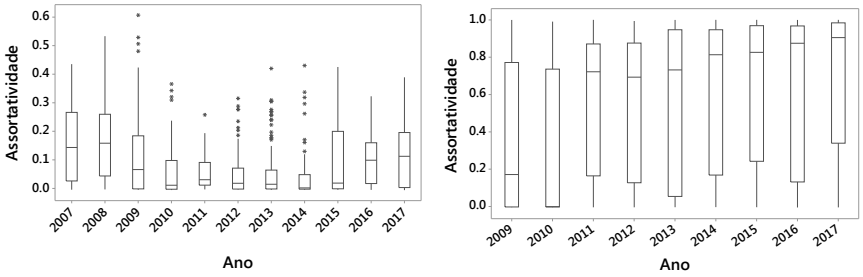
apenas os votos *Sim*, *Não* ou *Obstrução* foram considerados para construção dos grafos. Já nos Estados Unidos, somente os votos *Yes* e *No* foram utilizados. Então, modelamos as votações de cada país (Brasil e EUA) em grafos da seguinte forma: Para cada votação v analisada, foi criado um grafo $G^v(V, A)$ não direcionado e não ponderado onde $V = \{v_1, v_2, ...v_n\}$ é um conjunto de vértices que representam os deputados, e $A = \{a_1, a_2, a_3, ..., a_n\}$ um conjunto de arestas. Uma aresta (v_i, v_j) liga dois vértices (deputados) v_i e v_j ($i \neq j$), se, na votação sendo modelada, os dois deputados votaram igualmente. Por fim, cada deputado foi associado a um atributo que especifica o seu grupo, utilizado para o cálculo da assortatividade. No Brasil, os deputados foram divididos no grupo da coalizão do governo, composto por deputados de partidos que estavam na coalizão estabelecida no período da votação, e um segundo grupo composto pelos deputados cujo partido não pertencia à coalizão. Já nos Estados Unidos, foram mantidos os partidos originais.

2.4. Métricas Avaliadas

Após a modelagem dos grafos, foram calculadas as seguintes métricas: **Assortatividade**: mensura, através de um coeficiente entre -1 e 1, a preferência de nós de uma mesma rede se conectarem entre si com base em características similares [Newman 2003]; **Número de vitórias e derrotas para o partido do presidente**: para cada votação, verificou-se se o posicionamento da maioria do partido do presidente foi condizente com o posicionamento que venceu a votação; **Número de vitórias e derrotas para a coalizão**: para cada votação, foi verificada a lista de partidos que faziam parte da coalizão na data da votação, de acordo com a base de dados do CEBRAP, verificando se a coalizão venceu a votação; **Tamanho da maior componente**: para cada votação, verificou-se o número de deputados na maior componente do grafo.

3. Resultados

Iniciamos nossa análise pelo Brasil, onde a Tabela 2 mostra, por legislatura, ano e presidente, o número de derrotas e vitórias (como explicado na subseção 2.4) nas votações analisadas, em relação ao partido do presidente e a coalizão estabelecida. Por último, é apresentado o percentual médio do tamanho da maior componente para as vitórias e



(a) Brasil (b) Estados Unidos

Figura 1. Assortatividade calculada nas votações.

Tabela 2. Descrição das vitórias e derrotas observadas nas votações do Brasil

Legislatura	Ano	Presidente	Partido do Presidente		Coalizão do Presidente		% Média da Maior Componente	
			# de vitórias	# de derrotas	# de vitórias	# de derrotas	Vitórias	Derrotas
53ª	2007	Lula	198	11	202	7	80,77±1,25	55,86±2,63
53ª	2008	Lula	140	2	141	1	82,47±1,72	61,89±67,21
53ª	2009	Lula	137	11	142	6	82,56±1,92	65,8±6,8
53ª	2010	Lula	70	5	73	2	85,82±3	71,56±16,08
54ª	2011	Dilma	90	3	91	2	82,86±1,89	64,8±36,59
54ª	2012	Dilma	63	10	67	6	83,3±3,55	60,4±4,89
54ª	2013	Dilma	115	22	126	11	80,89±2,48	62,2±4,91
54ª	2014	Dilma	62	19	78	3	88,44±3,47	68,52±4,05
55ª	2015	Dilma	163	61	208	16	74,21±2,38	63,58±2,03
55ª	2016	Dilma/Temer	72	91	155	8	80,24±2,72	74,02±1,36
55ª	2017	Temer	92	78	165	5	81,29±3,02	67,77±1,61

derrotas considerando o posicionamento do partido do presidente em exercício, com um intervalo de confiança de 95%.

Na 53ª legislatura, a dispersão dos valores de assortatividade calculados é relativamente alta nos dois primeiros anos do período citado (Figura 1a) quando comparados aos dois últimos da mesma Legislatura. Com assortatividade maior, há menor interação e concordância entre a oposição e os deputados da coalizão do governo. Consequentemente, isso resulta em número de derrotas superior a todos os valores registrados até o ano avaliado. No último ano de governo, é possível observar uma menor dispersão das assortatividades calculadas, além das maiores médias de deputados na maior componente, em derrotas e vitórias, considerando todos os anos de governo do ex-presidente Lula, indicando uma maior estabilidade. Na 54ª Legislatura, é possível observar menor dispersão nos valores de assortatividade calculados durante os três primeiros anos (em relação à 53ª legislatura). Além disso, há um crescimento no número de derrotas proporcional nas votações, para o partido do presidente e para a coalizão. A presença de mais valores extremos e uma maior variação no percentual médio do tamanho da maior componente, para vitórias e para derrotas, indicam o início de um período de maior instabilidade, considerando todos os anos analisados até o momento.

Por último é analisada a 55ª Legislatura, período governado parcialmente por Dilma (até seu *impeachment* em 2016) e Michel Temer. Em 2015, há enfraquecimento do governo, com maior frequência de maiores valores das assortatividades calculadas, indicando aumento no processo de homofilia, quando comparados aos demais anos sob governo da ex-presidente Dilma. A alta assortatividade associada ao crescente número de derrotas, tanto do partido do presidente quanto da coalizão estabelecida, como mostra a Tabela 2, definem um período crítico para o governo. Também em 2016, o partido do pre-

Tabela 3. Descrição das vitórias e derrotas observadas nas votações dos EUA

Congresso	Ano	Presidente	Partido do Presidente		% Média da Maior Componente	
			# de vitórias	# de derrotas	# de vitórias	# de derrotas
111º	2009	Obama	907	22	75,38±1,16	58,37±2,36
111º	2010	Obama	622	9	77,46±0,53	57,36±4,89
112º	2011	Obama	275	633	75,98±1,79	56,71±0,32
112º	2012	Obama	210	415	77,48±2,14	56,43±0,41
113º	2013	Obama	205	389	80,86±2,13	54,67±0,43
113º	2014	Obama	179	352	78,29±2,27	53,61±0,37
114º	2015	Obama	209	453	79,32±2,11	56,29±0,33
114º	2016	Obama	176	412	82,66±2,29	55,26±0,32
115º	2017	Trump	155	370	80,17±2,81	53,91±0,27

sidente tem maior número de derrotas do que vitórias, um comportamento atípico em todo o período analisado. Em 2017, inteiramente sob o governo do presidente Michel Temer, o fato de que 75% das assortatividades observadas nas votações de 2017 serem menores ou iguais a 0,2 e o alto número de vitórias da coalizão, indicam uma alta governabilidade por parte do presidente. Apesar disso, o maior número de derrotas em relação ao número de vitórias, considerando o partido do presidente, indica uma instabilidade dentro do próprio partido, como mostra a Tabela 2.

Já nos Estados Unidos, o boxplot da Figura 1b apresenta a distribuição dos valores das assortatividades calculados nas votações. A Tabela 3 mostra, por congresso, ano e presidente, o número de derrotas e vitórias (Como explicado na subseção 2.4) nas votações analisadas (em relação ao partido do presidente). Por último, é apresentado o percentual médio do tamanho da maior componente para as vitórias e derrotas, considerando o posicionamento do partido do presidente em exercício, com um intervalo de confiança de 95%. Durante o primeiro governo do ex-presidente Barack Obama, 111º e 112º congresso, é possível observar na Figura 1b que aproximadamente 50% das votações têm assortatividade menor ou igual a 0,2 em 2009 e 0 em 2010, respectivamente. Dessa forma, é possível identificar um governo estável, em que o processo de homofilia é menos intenso, alinhado à maior razão de vitórias sob derrotas nos anos avaliados, apresentada na Tabela 3. No 112º, há crescimento na porcentagem de votações com assortatividade com valores em torno de 0,7, denotando governo mais instável em relação ao congresso anterior, também evidenciando crescente processo de homofilia. Além disso, o número de derrotas obtidas pelo partido do presidente passa a ser maior que o número de vitórias.

A instabilidade observada no fim do primeiro governo do ex-presidente Barack Obama estendeu-se para o seu segundo mandato (113º e 114º congresso). Neste período, 50% das votações possuem uma assortatividade maior que 0,8. Além disso, o número de derrotas é sempre superior ao número de vitórias. Isto descreve o processo de homofilia em função da baixa interação entre o partido do presidente e a oposição. Entretanto, maiores valores do percentual médio do tamanho da maior componente mostram que, em alguns anos, o ex-presidente Barack Obama teve mais vitórias com uma maior margem de votos e, por outro lado, derrotas com uma menor margem de votos. No 115º congresso, com o presidente Donald Trump, é possível observar frequência maior de votações com alta assortatividade, sem precedentes no período avaliado, em que 50% das votações têm assortatividade maior ou igual 0,89. Apesar da mudança de governo e partido ocorrida, a razão entre vitórias e derrotas do partido do presidente assemelha-se ao último ano de governo do ex-presidente Barack Obama. Isso também resulta no pior primeiro ano de governo para um presidente no período analisado, com uma Câmara dos Deputados cada vez mais heterogênea, indicando aumento no processo de homofilia.

4. Conclusão

Este trabalho propôs avaliar o comportamento dos deputados na Câmara dos Deputados de modo a inferir padrões sob governabilidade do presidente. No Brasil, foi possível ver como os valores da assortatividade podem ser relacionados à estabilidade política do governo, observando também o número de derrotas e vitórias. Já nos Estados Unidos, em que os partidos políticos são ideologicamente polarizados, foi possível observar que o extremismo ideológico tem aumentado nos últimos anos.

Referências

Andris, C., Lee, D., Hamilton, M. J., Martino, M., Gunning, C. E., and Selden, J. A. (2015). The rise of partisanship and super-cooperators in the u.s. house of representatives. *PLOS ONE*, 10(4):1–14.

Barber, P. (2014). How social media reduces mass political polarization. *Evidence from Germany, Spain, and the US, New York University*.

CEBRAP (2017). Centro brasileiro de pesquisa e planejamento. `http://cebrap.org.br/`.

Colleoni, E., Rozza, A., and Arvidsson, A. (2014). Echo chamber or public sphere? predicting political orientation and measuring political homophily in twitter using big data. *Journal of Communication*, 64(2):317–332.

de França, F. O., Goya, D., and Penteado, C. C. (2018). Analysis of the twitter interactions during the impeachment of brazilian president. In *Proceedings of the 51st Hawaii International Conference on System Sciences*.

Fearnside, P. M. (2016). Brazilian politics threaten environmental policies. *Science*, 353(6301):746–748.

Figueiredo, A. C. (2007). Government coalitions in brazilian democracy. *Brazilian Political Science Review*, 1(2):182–216.

Halberstam, Y. and Knight, B. (2016). Homophily, group size, and the diffusion of political information in social networks: Evidence from twitter. *Journal of Public Economics*, 143:73–88.

Newman, M. E. (2003). Mixing patterns in networks. *Physical Review E*, 67(2):026126.

Santos, M. H. d. C. (1997). Governabilidade, Governança e Democracia: Criação de Capacidade Governativa e Relações Executivo-Legislativo no Brasil Pós-Constituinte. *Dados*, 40.

Vaz de Melo, P. O. S. (2015). How many political parties should brazil have? a data-driven method to assess and reduce fragmentation in multi-party political systems. *PLOS ONE*, 10(10):1–24.

Vergeer, M. (2015). Twitter and political campaigning. *Sociology Compass*, 9.

Washington-Post (2017). 6 months of president trump, in 7 issues. `https://www.washingtonpost.com/news/the-fix/wp/2017/07/20/6-months-of-president-trump-in-7-issues`.

Wehner, J. (2010). Institutional constraints on profligate politicians: The conditional effect of partisan fragmentation on budget deficits. *Comparative Political Studies*.

Análises de Dados de Sistemas Crowdsourcing: estudo de caso de avaliações de estabelecimentos realizadas no Yelp

Mateus P. Silveira[1], **Wender Z. Xavier**[1], **Humberto T. Marques-Neto**[1]

[1]Programa de Pós-Graduação em Informática
Departamento de Ciência da Computação
Pontifícia Universidade Católica de Minas Gerais (PUC-MG)
Belo Horizonte – MG – Brasil

{mateus.parreiras, wender.xavier}@sga.pucminas.br, humberto@pucminas.br

Abstract. *This work does an analysis of the Yelp database, a commercial appraisal platform, which is popular in Europe and the United States. A characterization and analysis of feelings were made to determine the behavior of the users of this platform to help not only the improvement of the services provided by the establishments but also to contribute with a better understanding of the dynamics of the use of the services in a city.*

Resumo. *Este trabalho faz uma análise da base de dados do Yelp, uma plataforma para avaliação de estabelecimentos comerciais, muito popular na Europa e EUA. Realizou-se uma caracterização e uma análise de sentimentos para determinar o comportamento dos usuários desta plataforma para auxiliar não somente o aprimoramento dos serviços prestados pelos estabelecimentos como também contribuir com um melhor entendimento da dinâmica de utilização dos serviços em uma cidade.*

1. Introdução

A compreensão do comportamento humano acerca da utilização dos serviços da cidade é um tema amplamente pesquisado para melhoria de serviços e no desenvolvimento de cidades inteligentes [Batty et al. 2012]. O desenvolvimento dessas aplicações podem ainda, a partir do comportamento humano, avaliar e indicar melhorias no cotidiano das pessoas de forma a evitar riscos de acidentes e hábitos prejudiciais à saúde [Gustafson et al. 2014].

Aplicações de *Crowdsourcing* possuem um papel muito importante na coleta e na distribuição de informações. Estas aplicações são responsáveis por distribuir uma ou mais tarefas para que uma comunidade de pessoas, podendo ser questionários, tirar fotos e fazer resenhas (e.g Wikipédia[1] - Enciclopédia escrita de maneira colaborativa, ReclameAqui[2] - Site de reclamações contra empresas sobre atendimento, compra, serviços, etc.).

Com objetivo de analisar padrões de comportamento humano e verificar características de bases de dados de aplicações *Crowdsourcing*, utilizamos neste trabalho dados disponibilizados pela plataforma *Yelp*. Esta plataforma conta com um site e um aplicativo de avaliação de estabelecimentos comerciais.Em 2014 a *Yelp* disponibilizou parte de sua base de dados para que a academia pudesse utilizar esses dados em pesquisas. Então,

[1]http://www.wikipedia.org
[2]http://www.reclameaqui.com.br

uma análise da base de dados foi feita com o intuito de melhor entender o conjunto de dados. Assim, foi possível encontrar certos padrões como o grande número de avaliações 5 estrelas, o crescimento das de 1 estrela nos últimos anos e que normalmente os usuários tendem a fazer somente uma avaliação ao invés de várias. Além disso, uma análise de sentimento foi empregada para determinar se existia diferenças em textos com quantidade de estrelas dadas, que demonstrou que normalmente as avaliações se mantêm neutras.

Este trabalho está organizado da seguinte maneira. Seção 2 apresenta alguns trabalhos relacionados à aplicações de *Crowdsoucing*. Seção 3 apresenta a base de dados Yelp, principais informações contidas e propostas de análise. Seção 4 apresenta resultados obtidos a partir da análise da base do Yelp. A conclusão e trabalhos futuros são apresentados na Seção 5.

2. Trabalhos Relacionados

O trabalho [Zhang 2015] enfatiza a importância de incorporar revisões textuais para recomendação através de análise de sentimento de nível de frase e investigar ainda mais o papel que os textos desempenham em diversas tarefas importantes de recomendação. Em [Yu et al. 2014], os autores prepõem combinar informações de relacionamento heterogêneas para cada usuário de forma diferente e fornecer resultados de recomendação personalizados de alta qualidade usando dados de feedback implícitos do usuário e modelos de recomendação personalizados.

[McClanahan and Gokhale 2016] propõem uma nova abordagem para entender as relações entre os clientes e as empresas e o tipo de informação que pode ser inferida a partir dessas relações. Um grafo é gerado com os nós sendo os negócios e o peso das arestas a quantidade de clientes em comum, e concluiu-se que os clientes preferem visitar empresas que são geograficamente próximos e/ou possuem produtos e serviços similares. Em [Bhowmick et al. 2017] os autores definem métricas de popularidade para rotular diversas empresas no conjunto de dados do *Yelp*, a principal foi a difusão da informação. Com isso, foi desenvolvido um modelo de recomendação que sugere as principais regiões para os empresários para iniciar negócios populares.

Os trabalhos apresentados tem como objetivo criar modelos de recomendação para o usuário, levando em consideração vários fatores. Nenhum dos trabalhos fazem uma comparação entre a análise de sentimento das avaliações e a quantidade de estrelas dadas comparando a base toda. Além disso, este trabalho faz uma caracterização da base de dados utilizada para que seja possível ter um melhor entendimento de maneia geral de como é o comportamento dos usuários.

3. Metodologia

Nesta Seção, as etapas da metodologia aplicadas no trabalho são apresentadas. Na primeira Subseção é apresentada a base de dados do *Yelp*. Em seguida é mostrado principais características e os processos de análise empregado nas avaliações.

3.1. Base de Dados *Yelp*

Em 2014 a empresa *Yelp* iniciou o desafio *Yelp Dataset Challenge*[3], disponibilizando uma base de dados contendo um subconjunto dos estabelecimentos comerciais, as avaliações e

[3]https://www.yelp.com/dataset/challenge

os usuários que utilizaram o aplicativo. O desafio já teve várias rodadas, na qual a base de dados era atualizada. A base de dados possui 5.200.000 avaliações, 174.000 negócios de 11 regiões metropolitanas de 4 países e 1.300.000 usuários, com dados de 2004 a 2017.

3.2. Caracterização da Base

Uma análise inicial na base de dados foi feita com intuito de compreender o seu conteúdo. Dentre as 20 cidades com maior número de estabelecimentos na base de dados, Las Vegas é a cidade que possui o maior conjunto de dados possuindo cerca de 25 mil estabelecimentos e 1,6 milhões de avaliações, seguida de Phoenix e Toronto. Para compreender o comportamento de usuário do *Yelp*, foi verificado primeiramente o número de avaliações por usuário. Destas 52,7% dos usuários do *Yelp* postam somente uma avaliação, 16,6% postaram duas avaliações, e 8,49% postaram três avaliações e somente 0,043% dos usuários fizeram mais de 250 avaliações, cerca de somente 500 no total. Assim, grande parte dos usuários tende utilizar a plataforma para conhecer informações e saber da experiência de outros usuários do que propriamente gerar conteúdo para a rede.

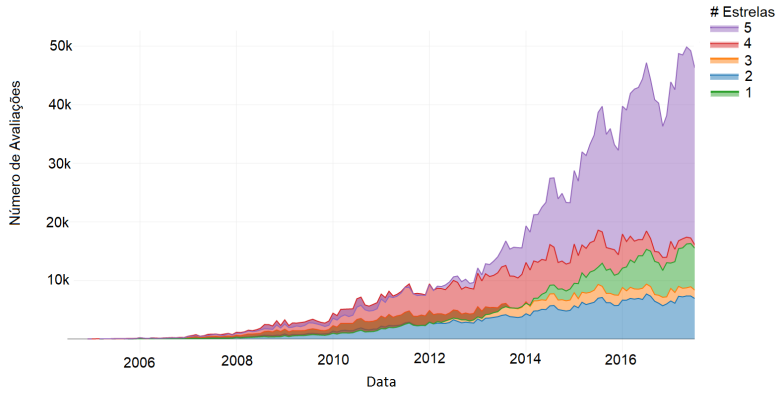


Figura 1. Número de avaliações por mês por estrela.

Os usuários podem dar notas de 1 a 5 estrelas para os estabelecimentos, sendo 1 experiencia ruim e 5 ótima. Na Figura 1 é possível perceber picos anuais entre os meses de Julho e Agosto, e o menor número de avaliações entre os meses de Novembro e Dezembro. Isso pode estar associado ao fato de os meses de Julho à Agosto serem períodos de férias escolares nos EUA, e os meses de Novembro e Dezembro as festas de fim-de-ano. Outro comportamento que pode ser visualizado é o número de estrelas atribuídas ao estabelecimento. A partir de 2014 existe uma tendência dos usuários darem mais avaliações de 1 estrela, enquanto notas de 2 a quatro tiveram pouca alteração. Com isso, pode-se perceber que os usuários estão tendendo a postar notas de 1 ou 5 estrelas para estabelecimentos que gostaram ou não gostaram.

Cada avaliação possui atributos como *usefull*, *cool* e *funny* que representam o número de usuários que acharam que a avaliação foi útil, interessante e engraçada respectivamente. Na Figura 2 é possível observar a correlação dentre estes atributos e o tamanho da avaliação (*text_length*). Uma correlação positiva muito forte existe entre os atributos *funny* e *usefull*, de 0,98, desta forma avaliações engraçadas costumam ser avaliações consideradas como úteis. Existe outras fortes correlações uma de 0,86 entre o tamanho do texto e o atributo *usefull*, e 0,76 entre *text_length* e *funny* demonstrando que o tamanho

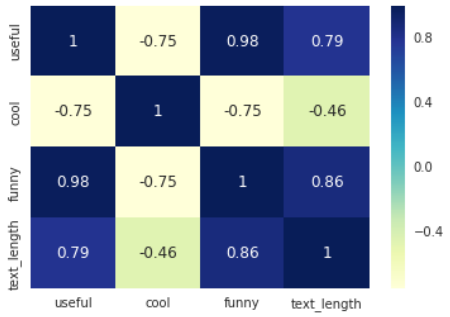


Figura 2. Correlação entre os atributos dos usuários.

do texto é importante para uma avaliação ser considerada como útil ou engraçada. Já a relação de *cool* e *usefull* é negativa, ou seja, se um aumenta o outro tende a diminuir.

A base de dados foi filtrada por selecionado as categorias de bar, restaurantes e comida de estabelecimentos nos EUA. Chegou-se à um conjunto de 49.914 estabelecimentos comerciais, 166.196 usuários e 569.291 avaliações. Então foi feita uma análise textual das avaliações para definir: quais são as palavras mais utilizadas por usuários dividida pela nota dada. Isto foi feito para verificar se avaliações com notas inferiores são mais negativas se comparadas com as avaliações com as notas mais altas. Para efetuar essas análises, foram utilizadas as bibliotecas do Python *NLTK*[4] *(Natural Language Toolkit)* e *TextBlob*[5]. Também foi feita a análise para definir se o sexo do usuário pode influenciar na maneira como ele posta comentários, utilizando a biblioteca *Gender-Guesser*[6]. Finalmente, para a criação do mapa de palavras foi utilizado o pacote Wordcloud[7] do R. Os resultados do processamento textual são apresentados na próxima Seção.

4. Resultados

Para efetuar a análise textual, cada avaliação passou por um processo de tratamento dos dados. Primeiramente foram retirados pontuação e em seguida todo o texto da avaliação foi transformado para letra minúscula. Em seguida, foram retiradas palavras de parada (i.e. *Stopwords*). Finalmente, o texto foi submetido à um processo de reduzir palavras flexionadas ao seu tronco (i.e. *stemming*). Então um mapa de palavras foi gerado para as avaliações considerando o número de estrelas (Figura 3) e três palavras se destacam em todos, *Food* (comida), *Place* (local), *Service* (serviço). Dito isto, percebeu-se que os três principais pontos que devem ser levados em conta para um estabelecimento de alimentação são a qualidade da comida, ambiente, e serviço prestado ao cliente, já que estes são as palavras mais escritas nas avaliações.

A Figura 4 mostra os resultados da aplicação dos algoritmos *TextBlob* e *NLTK* na base de dados. Ao analisar os resultados da aplicação do *TextBLob* (ver Figura 4(a)) percebe-se que em todas as categorias de estrelas o algoritmo encontrou valores variando entre -1 e 1, demonstrando que o número de estrelas de certa forma não está relacionado ao sentimento da avaliação. Entretanto, percebe-se que com o aumento do número de

[4]https://www.nltk.org/
[5]http://textblob.readthedocs.io/en/dev/
[6]https://pypi.python.org/pypi/gender-guesser/
[7]https://cran.r-project.org/package=wordcloud

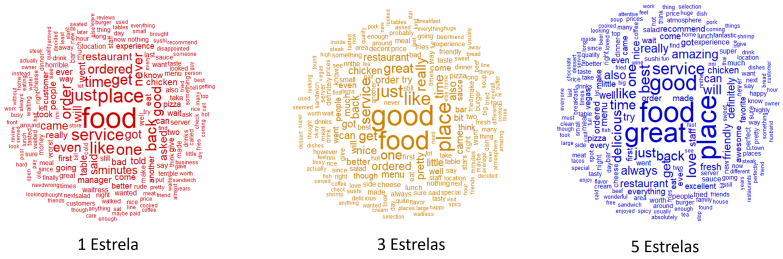
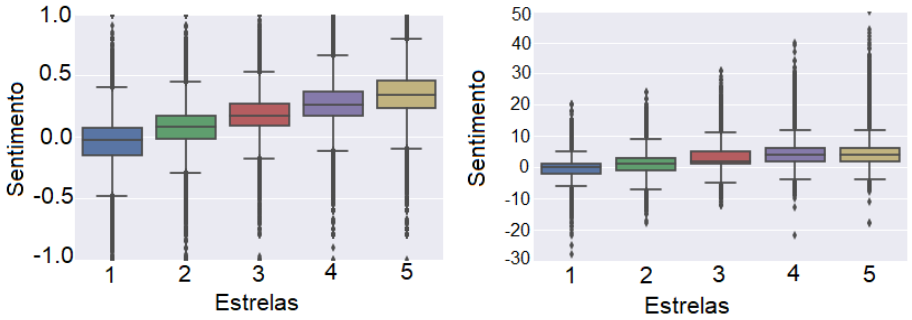


Figura 3. Mapa de Palavras das avaliações divididos pelo número de estrelas

estrelas atribuídas, cresce sutilmente o sentimento relacionado à avaliação. Grande parte das avaliações concentra-se em valores próximos à 0 para as avaliações com estrelas 1 e 2 e próximos à 0.5 para as avaliações de 5 estrelas. O mesmo comportamento pode ser observado nos resultados da análise de sentimento pelo *NLTK*, onde grande parte das avaliações se concentra próximo à valores de sentimentos neutros. Existem entretanto avaliações que chegaram à valores negativos como -20, e valores positivos próximos à 50 demonstrando que mesmo em avaliações boas dos estabelecimentos, os usuários podem estar utilizando de palavras consideradas negativas pelos algoritmos.



(a) Análise de sentimento utilizando o TextBlob.

(b) Análise de sentimento utilizando o NLTK.

Figura 4. Análise de sentimento das avaliações divididos pelo número de estrelas.

Uma última análise utilizando a biblioteca *Gender-Guesser* foi feita para verificar se o gênero das pessoas afeta na quantidade de estrelas das avaliações dadas. Esta biblioteca contém uma base de dados de nomes, e retorna uma estimativa do gênero da pessoa entre masculino, feminino e indefinido. Após a execução dos algoritmos, foi retornado 217.428 nomes masculinos, e 244.900 femininos. Considerando este valor, pôde-se perceber que ambos os gêneros utilizam a aplicação de forma parecida. Avaliando as estrelas atribuídas a cada avaliação este padrão se mantém, tendo o grupo feminino atribuído mais estrelas à uma quantidade maior de estabelecimentos comparado ao grupo masculino.

5. Conclusões e Trabalhos Futuros

Neste trabalho realizamos uma análise de base de dados de aplicações *Crowdsourcing* utilizando base de dados disponibilizada pelo *Yelp*. Identificamos os principais tópicos abordados na elaboração das avaliações dos clientes utilizando análise textual. Além

disso, Identificamos padrões de postagens ao longo do tempo que poderiam ser utilizados pelos estabelecimentos para realizar promoções para atrair maior número de clientes.

Além disso, foi apresentada uma análise inicial relacionado ao comportamento dos usuários. Verificamos que poucos usuários tendem a postar várias avaliações de estabelecimentos de diversas categorias, enquanto grande parte dos usuários tende postar somente uma avaliação. A cada avaliação pode ser definidos atributos em relação à humor, interesse e utilidade. Utilizando-se de bibliotecas do Python realizamos análise de sentimentos das avaliações esperando encontrar padrões entre a quantidade de estrelas e o sentimento do texto. Descobrimos que independente do número de estrelas atribuídas, as avaliações tendem a ter sentimento próximo à 0 (i.e. neutro).

Como trabalhos futuros existe a possibilidade de verificar séries temporais das avaliações e das estrelas atribuídas, verificando quando um restaurante pode prosperar ou não. Por fim, classificar os restaurantes por tipos mais específicos podem mostrar comportamentos diferentes de usuários para estabelecimentos diferentes. Novas modelagens de redes também representam um desafio na compreensão do comportamento de usuários do *Yelp* devido à complexidade e número de estabelecimentos e usuários na rede.

6. Agradecimentos

Este trabalho foi financiado por MASWeb (processo FAPEMIG/PRONEX APQ-01400-14), FAPEMIG (processo APQ-02924-16), PUC-Minas, CNPq, CAPES e STIC AmSud 18-STIC-07.

Referências

Batty, M., Axhausen, K. W., Giannotti, F., Pozdnoukhov, A., Bazzani, A., Wachowicz, M., Ouzounis, G., and Portugali, Y. (2012). Smart cities of the future. *The European Physical Journal Special Topics*, 214(1):481–518.

Bhowmick, A. K., Suman, S., and Mitra, B. (2017). Effect of information propagation on business popularity: A case study on yelp. In *2017 18th IEEE International Conference on Mobile Data Management (MDM)*, pages 11–20.

Gustafson, D. H., McTavish, F. M., Chih, M.-Y., Atwood, A. K., Johnson, R. A., Boyle, M. G., Levy, M. S., Driscoll, H., Chisholm, S. M., Dillenburg, L., et al. (2014). A smartphone application to support recovery from alcoholism: a randomized clinical trial. *JAMA psychiatry*, 71(5):566–572.

McClanahan, B. and Gokhale, S. S. (2016). Centrality and cluster analysis of yelp mutual customer business graph. In *2016 IEEE 40th Annual Computer Software and Applications Conference (COMPSAC)*, volume 1, pages 592–601.

Yu, X., Ren, X., Sun, Y., Gu, Q., Sturt, B., Khandelwal, U., Norick, B., and Han, J. (2014). Personalized entity recommendation: A heterogeneous information network approach. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining*, WSDM '14, pages 283–292, New York, NY, USA. ACM.

Zhang, Y. (2015). Incorporating phrase-level sentiment analysis on textual reviews for personalized recommendation. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, WSDM '15, pages 435–440, New York, NY, USA. ACM.

Detecção de traços de narcisismo em conversas com predadores sexuais

Leonardo Ferreira dos Santos[1], **Gustavo Paiva Guedes**[1]

[1]CEFET/RJ - Centro Federal de Educação Tecnológica Celso Suckow da Fonseca
Av. Maracana, 229 - Rio de Janeiro - RJ - Brasil.

leonardo.santos@eic.cefet-rj.br, gustavo.guedes@cefet-rj.br

Abstract. *Sexual predators exploit the people to whom they relate in order to achieve their different individual goals. The profile of the sexual predator presents several mental disorders, among them, the narcissism stands out. In this scenario, the present work aims to detect traces of narcissism in the communication of sexual predators. To achieve this goal, we used the LIWC to analyze the use of personal pronouns in texts of these predators. In addition, a set of data from PAN-2012 is used. Results found present significant correlations between some categories related to personal pronouns and predatory conversations, consistent with results presented in the literature.*

Resumo. *Predadores sexuais exploram as pessoas com quem se relacionam a fim de atingir seus diferentes objetivos individuais. O perfil do predador sexual apresenta variados transtornos mentais, dentre eles, o destaca-se o narcisismo. Nesse cenário, o presente trabalho tem o objetivo de detectar traços de narcisismo na comunicação dos predadores sexuais. Para atingir esse objetivo, utilizamos o LIWC para analisar o uso de pronomes pessoais em textos desses predadores. Além disso, é utilizado um conjunto com os dados disponibilizados pela PAN-2012. Resultados encontrados apresentam correlações significativas entre algumas categorias relacionadas aos pronomes pessoais e conversas predatórias, coerentes com resultados apresentados na literatura.*

1. Introdução

Segundo dados recentes, 9% dos menores de idade já sofreram algum tipo de abuso sexual na internet e, aproximadamente, 1 em cada 25 foi abordado por predadores sexuais para encontros fora do meio virtual ou contactados diretamente por meio de cartas, telefonemas ou *in loco* [Mitchell et al., 2014]. No entanto, esses números podem ser ainda maiores, visto que diversos casos de abusos sexuais em crianças não são reportados [Barbara et al., 2017]. Isso ocorre, primeiramente, porque crianças abusadas sexualmente tendem a ficar em silêncio, seja por culpa, vergonha ou ameaças [Bagley and King, 2003]. Além disso, o abuso sexual destrói a espontaneidade e a liberdade da criança, causando um terror solitário [Bagley and King, 2003].

Os predadores sexuais fazem uso de um padrão de comunicação enganosa que, inicialmente, busca criar uma falsa relação de confiança com o menor de idade para então, ter uma oportunidade de contato fora do meio virtual e executar o abuso sexual [Olson et al., 2007]. Comportamentos enganosos no meio virtual podem caracterizar a presença de transtornos mentais [Crossley, 2016]. Nesse cenário, algumas pesquisas na área de

saúde têm estudado o comportamento dos usuários, o que vem permitido uma série de avanços na análise e compreensão de distúrbios mentais [Ayers et al., 2014]. Por exemplo, textos provenientes de salas de bate-papo e redes sociais têm sido objeto de estudo nos últimos anos, permitindo a identificação de transtornos mentais [Preoţiuc-Pietro et al., 2015; Coppersmith et al., 2014; Rodrigues et al., 2017].

O perfil do predador sexual apresenta variados transtornos mentais, dentre eles, o narcisismo [Crouch et al., 2015]. Algumas vezes definido, simplesmente, como alta autoestima [Gray, 2011], clinicamente o narcisismo é um transtorno de personalidade caracterizado por: acreditar ser "especial" e único; baixa empatia em relacionamentos e exploração das relações interpessoais com o objetivo de atender as próprias necessidades [Association et al., 2014].

O padrão de comunicação do agressor sexual e o caráter explorativo das relações interpessoais presentes no narcisismo definem a principal motivação para o presente trabalho, que tem como o principal objetivo a detecção traços de narcisismo em predadores sexuais. Para isso, é importante notar que a literatura destaca a existência de relação entre o narcisismo e utilização de pronomes pessoais [McGregor, 2010; Baryshevtsev and McGlone, 2018].

Até o momento dessa pesquisa não foram encontrados trabalhos com o objetivo de detectar traços de narcisismo na comunicação de predadores sexuais no meio virtual. Para realizar essa detecção, é empregado o dicionário linguístico LIWC que, com base na contagem de palavras presentes em textos, permite a extração de categorias psicolinguísticas [Pennebaker et al., 2001]. Os experimentos são realizados com os dados disponibilizados pela PAN-2012[1] para a competição de identificação de predadores sexuais.

As demais seções estão dispostas da seguinte maneira: na seção 2, são levantados trabalhos semelhantes e relacionados ao tema; na seção 3 é apresentado a metodologia para análise do conjunto de dados; na seção 4 são discutidos os resultados encontrados; por último, a seção 5 apresenta as conclusões, limitações e discussão sobre trabalhos futuros.

2. Trabalhos relacionados

O trabalho desenvolvido por Baryshevtsev and McGlone [2018] explorou a presença do uso de pronomes na comunicação de um predador sexual. Foram consideradas 561 transcrições de conversas ocorridas em meio virtual realizadas entre predadores sexuais e pseudo-vítimas (e.g., agentes federais agindo como menores de idade), disponibilizadas pelo site Perverted Justice[2], em que foi possível constatar um uso maior de pronomes de segunda pessoa, comparado aos de primeira pessoa. Essa discrepância é justificada em [Olson et al., 2007], que indica que o objetivo de colocar a vítima como o foco da conversação é estabelecer uma relação de confiança, para que, em seguida, seja iniciado o abuso sexual, físico ou virtual.

No estudo apresentado em Drouin et al. [2017], os autores analisaram 590 conversas entre predadores sexuais e pseudo-vítimas, extraídas do site Perverted Justice[2], com o objetivo de encontrar correlações com as seguintes categorias do LIWC: sexualidade,

[1] https://pan.webis.de/clef12/pan12-web/author-identification.html
[2] http://www.perverted-justice.com/

influência e quantidade de palavras por sentença. Resultados apontaram que, perante suas pseudo-vítimas, predadores sexuais tendem a usar 91% mais palavras de cunho sexual, 66% mais palavras por sentença e 82% exerceram postura mais influente nas conversas quando comparado às pseudo-vítimas. Também foi destacada a utilidade do LIWC para o uso forense, permitindo uma análise objetiva de características psicológicas de predadores sexuais.

Em O'Reilly III et al. [2014], o narcisismo é identificado como uma característica existente nos líderes. Nesse estudo, o LIWC foi utilizado para analisar transcrições de teleconferências sobre lucros e cartas destinadas a acionistas. Foi possível identificar significativa correlação com o uso de pronomes pessoais, destacando-se pronomes de primeira pessoa.

Muitos trabalhos realizados buscam um melhor entendimento do perfil do agressor sexual em meio virtual, não somente psicológico mas também de atuação. No entanto, não foram encontrados estudos que busquem detectar traços de narcisismo em predadores sexuais em meio virtual. Sendo assim, o presente trabalho se enquadra nesse cenário, buscando contribuir com a detecção de traços de narcisismo em predadores sexuais com o uso do LIWC.

3. Detecção de traços de narcisismo

Esta seção descreve a metodologia utilizada para detectar a existência de traços de narcisismo em usuários predadores sexuais em comparação com usuários não-predadores. Para isso, são utilizados o conjunto de dados PAN-2012 e o dicionário do LIWC. Inicialmente, na subseção 3.1, são descritos o PAN-2012 e o LIWC. Em seguida, na subseção 3.2, é descrita a metodologia utilizada nesse estudo.

3.1. Conjuntos de dados

PAN-2012: O conjunto de dados PAN-2012 consiste de conversas provenientes de alguns sites, dentre eles, o `perverted-justice.com` e `irclog.org`. Há um total de 66.914 conversas envolvendo 97.671 usuários únicos. Desses usuários, 142 foram rotulados como predadores. Existem 2.015 conversas em que os predadores estavam envolvidos. Nenhuma conversa possui mais de um predador e 68% das conversas continham dois usuários. O predador mais ativo se envolveu em 182 conversas.

LIWC: O *Linguistic Inquiry and Word Count* (LIWC) é um programa capaz de contabilizar palavras em diversas categorias psicolinguísticas [Pennebaker et al., 2001]. A versão do dicionário do LIWC utilizada nesse estudo possui mais de 6.400 palavras categorizadas em mais de 70 categorias [Pennebaker et al., 2015]. É interessante destacar que uma palavra pode estar associada a diversas categorias, por exemplo, a palavra *amazing* é categorizada como uma palavra afetiva e com emoção positiva.

3.2. Metodologia

A metodologia utilizada nesse estudo consiste em detectar a existência de traços de narcisismo – por meio da análise dos pronomes pessoais – em textos escritos por predadores sexuais. Isso é feito com a análise das diferenças existentes entre textos de predadores e texto de não-predadores. Os dados são provenientes do conjunto de dados PAN-2012 e as categorias utilizadas são oriundas do LIWC. A análise realizada é inspirada no estudo

proposto por [McGregor, 2010]. Segundo esse estudo, (*i*) não existe correlação entre os narcisistas e a utilização de pronomes na primeira pessoa do plural (e.g., nós, nosso); (*ii*) não existe correlação entre o narcisismo e o uso de pronomes de segunda pessoa (e.g., você); (*iii*) não existe correlação entre o narcisismo e o uso de pronomes na terceira pessoa (e.g., ele).

Nesse cenário, o objetivo é utilizar o LIWC para avaliar as categorias de pronomes para evidenciar a diferença de uso pronominal entre predadores e não-predadores. Isso é efetuado com base na equação que define o tamanho de efeito *d* de Cohen [Rosnow and Rosenthal, 1996], representada na Equação 1. As categorias do LIWC são representadas por $(0 \leq i < 5)$, em que i representa uma das cinco categorias de pronomes pessoais presentes no LIWC: *i*, *we*, *you*, *shehe* e *they*, representando pronomes na 1^a pessoa do singular, 1^a pessoa do plural, 2^a pessoa, 3^a pessoa do singular e 3^a pessoa do plural, respectivamente.

A média simples $\bar{X}_P^i$ é dada pela i-ésima componente do vetor de frequência dos textos dos predadores. Analogamente, $\bar{X}_{NP}^i$ representa a média simples da i-ésima componente do vetor de frequência dos textos dos não-predadores. Observando o denominador, SD_P^i e SD_{NP}^i representam os desvios padrão da i-ésima componente do vetor de frequência dos textos dos predadores e não-predadores, respectivamente. Valores positivos de d_i indicam que mais palavras na categoria *i* foram utilizadas pelos predadores e valores negativos, pelos não-predadores. Os valores de referência de d são considerados a partir de [Cohen, 1988], em que $d = 0.20$ indica um pequeno efeito, $d = 0.50$ um médio efeito e $d = 0.80$ um grande efeito.

$$d_i = \frac{\bar{X}_P^i - \bar{X}_{NP}^i}{\sqrt{((SD_P^i)^2 + (SD_{NP}^i)^2)/2}} \tag{1}$$

4. Resultados

Nesse estudo, são exibidos os resultados encontrados para a diferença entre a utilização de pronomes pessoais por predadores sexuais e não-predadores. Para isso, é exibido o tamanho de efeito *d* de Cohen para cada classe de pronomes do LIWC. Além disso, para ambos os grupos, são apresentados os valores para a média e o desvio padrão. A Tabela 1 apresenta os resultados encontrados.

Tabela 1. Tamanho do efeito *d* de Cohen em cada uma das classe de pronomes do LIWC. Os resultados são apresentados na forma de média (M), desvio padrão (D) e (*d*) para o tamanho de efeito.

		Predadores Sexuais		Não-predadores		
Dimensão	Exemplo	M	D	M	D	(d)
i	*I* , *mine*	6.65	6.38	3.17	5.00	**0.61**
we	*we*, *our*	0.31	1.21	0.24	1.41	0.06
you	*you*, *your*	10.74	10.90	4.20	6.49	**0.73**
shehe	*her*, *him*	0.22	0.76	0.17	1.35	0.05
they	*they*, *their*	0.15	0.67	0.13	0.72	0.03

Os valores em em negrito destacam valores significantes de tamanho de efeito. Pode-se evidenciar um tamanho de efeito médio no uso de pronomes na primeira pessoa (i.e., 0.61) e no uso de pronomes na segunda pessoa (i.e., 0.73). Para os demais pronomes não foram encontrados valores de efeito significantes.

5. Discussão

O estudo aqui apresentado objetivou detectar traços de narcisismo em indivíduos identificados como predadores sexuais. Essa tarefa foi fundamentada na premissa de que esses predadores tendem a ser narcisistas. Para efetuar esse estudo, foi utilizado um conjunto de dados de predadores sexuais, denominado PAN-2012.

Os resultados encontrados corroboram com duas conclusões apresentadas em [McGregor, 2010]: (*i*) não existe correlação entre os narcisistas e a utilização de pronomes na primeira pessoa do plural (e.g., nós, nosso); (*iii*) não existe correlação entre o narcisismo e o uso de pronomes na terceira pessoa. Isso pôde ser evidenciado pelos tamanhos de efeito apresentados para a categoria de pronomes pessoais na primeira pessoa e na terceira pessoa.

Porém, nosso estudo encontrou um tamanho médio de efeito para os pronomes da terceira pessoa, não apresentando o mesmo resultado encontrado em [McGregor, 2010] que destaca que (*ii*) não existe correlação entre o narcisismo e o uso de pronomes na segunda pessoa. Entretanto, conforme evidenciado em estudo recente, indivíduos que interagem de forma *online* com menores de idade, com objetivos de encontrá-los pessoalmente, tendem a utilizar mais pronomes na segunda pessoa do singular, indicando que isso pode ser uma maneira de transmitir interesse e atenção à criança [Baryshevtsev and McGlone, 2018]. O estudo ainda destaca que focar mais na criança pode transmitir cuidado e preocupação (e.g., "você é melhor que isso"), resultando em uma maior receptividade por parte da criança.

Embora esse estudo ofereça resultados preliminares, é interessante destacar que todos os resultados encontrados foram coerentes com resultados apresentados na literatura. Para trabalhos futuros, é considerada a inclusão de outras categorias do LIWC para análise no conjunto de dados usado no presente artigo. Um bom exemplo seria a utilização de palavras de cunho sexual e xingamentos.

Referências

Association, A. P. et al. (2014). *DSM-5: Manual diagnóstico e estatístico de transtornos mentais*. Artmed Editora.

Ayers, J. W., Althouse, B. M., and Dredze, M. (2014). Could behavioral medicine lead the web data revolution? *Jama*, 311(14):1399–1400.

Bagley, C. and King, K. (2003). *Child sexual abuse: The search for healing*. Routledge.

Barbara, G., Collini, F., Cattaneo, C., Facchin, F., Vercellini, P., Chiappa, L., and Kustermann, A. (2017). Sexual violence against adolescent girls: labeling it to avoid normalization. *Journal of Women's Health*, 26(11):1146–1149.

Baryshevtsev, M. V. and McGlone, M. S. (2018). Pronoun usage in online sexual predation. *Cyberpsychology, Behavior, and Social Networking*, 21(2):117–122.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences 2nd edn.

Coppersmith, G., Dredze, M., and Harman, C. (2014). Quantifying mental health signals in twitter. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 51–60.

Crossley, L. N. (2016). *The influence of the Dark Triad and communication medium on deceptive outcomes*. PhD thesis, University of British Columbia.

Crouch, J. L., Hiraoka, R., Rutledge, E., Zengel, B., Skowronski, J. J., and Milner, J. S. (2015). Is narcissism associated with child physical abuse risk? *Journal of Family Violence*, 30(3):373–380.

Drouin, M., Boyd, R. L., Hancock, J. T., and James, A. (2017). Linguistic analysis of chat transcripts from child predator undercover sex stings. *The Journal of Forensic Psychiatry & Psychology*, 28(4):437–457.

Gray, P. (2011). The decline of play and the rise of psychopathology in children and adolescents. *American Journal of Play*, 3(4):443–463.

McGregor, S. (2010). The analysis of personality through language: Narcissism predicts use of shame-related words in narratives.

Mitchell, K., Jones, L., Finkelhor, D., and Wolak, J. (2014). Trends in unwanted sexual solicitations: Findings from the youth internet safety studies. *Youth Internet Safety Survey Bulletin*.

Olson, L. N., Daggs, J. L., Ellevold, B. L., and Rogers, T. K. (2007). Entrapping the innocent: Toward a theory of child sexual predators' luring communication. *Communication Theory*, 17(3):231–251.

O'Reilly III, C. A., Doerr, B., Caldwell, D. F., and Chatman, J. A. (2014). Narcissistic ceos and executive compensation. *The Leadership Quarterly*, 25(2):218–231.

Pennebaker, J. W., Boyd, R. L., Jordan, K., and Blackburn, K. (2015). The development and psychometric properties of liwc2015. Technical report.

Pennebaker, J. W., Francis, M. E., and Booth, R. J. (2001). Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001.

Preoţiuc-Pietro, D., Eichstaedt, J., Park, G., Sap, M., Smith, L., Tobolsky, V., Schwartz, H. A., and Ungar, L. (2015). The role of personality, age, and gender in tweeting about mental illness. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 21–30.

Rodrigues, R. G. R., Pereira, W. W. P., Bezerra, E. B., and Guedes, G. P. G. (2017). Inferência de idade utilizando o liwc: identificando potenciais predadores sexuais. In *Proceedings of the Brazilian Workshop on Social Network Analysis and Mining*. SBC.

Rosnow, R. L. and Rosenthal, R. (1996). Computing contrasts, effect sizes, and counternulls on other people's published data: General procedures for research consumers. *Psychological Methods*, 1(4):331.

Emoções em português do Brasil: um conjunto de dados e resultados de base

Gabriel Nascimento[1], **Fellipe Duarte**[2], **Gustavo Paiva Guedes**[1]

[1]CEFET/RJ - Centro Federal de Educação Tecnológica Celso Suckow da Fonseca
Av. Maracanã, 229 - Rio de Janeiro - RJ - Brasil.

[2]UFRRJ - Universidade Federal Rural do Rio de Janeiro
Av. Gov. Roberto Silveira, s/n - Moquetá - Nova Iguaçu - Brasil

gabriel.nascimento@eic.cefet-rj.br, duartefellipe@ufrrj.br,
gustavo.guedes@cefet-rj.br

Abstract. *This paper presents a new dataset for sentiment analysis in Brazilian Portuguese. The texts were extracted from a Brazilian social network named Meu Querido Diário. In this social network, users often share feelings and emotions associated with everyday life. The main difference of this data set is that, in this social network, the user himself can inform the emotion associated with his entry. Preliminary experiments were performed with some classification models, creating the first baseline results. The model that obtained the best result was the SVM with linear kernel using bigrams.*

Resumo. *Este artigo apresenta um novo conjunto de dados para análise de sentimentos em português do Brasil. Os textos foram extraídos de uma rede social brasileira denominada Meu Querido Diário. Nessa rede social, os usuários frequentemente compartilham sentimentos e emoções associados ao dia-a-dia. O principal diferencial deste conjunto de dados é que, nessa rede social, o próprio usuário pode informar a emoção associada à sua entrada. Foram realizados experimentos preliminares com alguns modelos de classificação, criando os primeiros resultados de base. O modelo que obteve melhor resultado foi o SVM com kernel linear utilizando bigramas.*

1. Introdução

Há um vasto número de indivíduos que compartilham suas opiniões/sentimentos sobre diversos assuntos (e.g., produtos, pessoas, notícias) em redes sociais, *blogs*, *microblogs* e sites de *reviews* [Rosenthal et al., 2015]. Essa grande quantidade de dados tem despertado, cada vez mais, o interesse de pesquisadores em Análise de Sentimentos (AS), um importante campo da área de Processamento de Linguagem Natural (PLN) [Pang et al., 2008].

O objetivo da AS é analisar opiniões, atitudes, emoções, sentimentos e avaliações expressas por usuários em textos [Liu, 2012]. Logo, muitas empresas investem em pesquisas nessa área, seja para analisar informações políticas ou para investigar a percepção acerca de um produto [Cambria et al., 2013]. Dentre as tarefas existentes na AS, destaca-se a classificação de polaridade, que consiste em classificar um texto como *um* entre *dois* sentimentos opostos [Cambria et al., 2013].

A detecção de sentimentos em textos permite que os pesquisadores adquiram informações valiosas em grande escala [Rosenthal et al., 2015]. Isso reduz o tempo e aumenta a quantidade de informação que os pesquisadores podem absorver, visto que a detecção ocorre de maneira automática. No entanto, é importante destacar que a maior parte dos estudos em classificação de polaridade se concentra em textos na língua inglesa [Wiegand et al., 2010]. Dessa maneira, ainda existe uma carência de estudos concentrados no Português do Brasil (PB) [Guedes et al., 2016].

Nesse panorama, a principal contribuição desse trabalho consiste em criar um conjunto de dados de emoções em PB. Para isso, foram extraídas entradas de uma rede social brasileira denominada Meu Querido Diário (MQD)[1]. Nessa rede, os usuários podem compartilhar sentimentos/emoções sobre seus dias. Além disso, podem associar a emoção que estão sentindo no momento em que estão escrevendo suas entradas. Não foram encontrados conjuntos de dados semelhantes na literatura para o PB. Após a criação do conjunto de dados, foi produzido um resultado de base (*baseline*) nesse conjunto de dados, aplicando os algoritmos-padrão utilizados na área de AS, o que pode ser utilizado em trabalhos futuros nesse tópico.

O restante deste artigo está organizado como segue. A seção 2 discute alguns trabalhos relacionados ao tema proposto. A seção 3 descreve o conjunto de dados criado. A seção 4 descreve a metodologia de avaliação e a seção 5 apresenta a conclusão.

2. Trabalhos Relacionados

Existem na literatura alguns trabalhos que apresentam novos conjuntos de dados de redes sociais para realização de análise de sentimentos. Podemos destacar o trabalho realizado por Saif et al. [2013], que apresentou um novo conjunto de dados baseado no Twitter. Este conjunto de dados foi nomeado *STS-Gold* e nele, todos os *tweets* foram rotulados com sentimentos: positivo, negativo, neutro, misto, outro. Os *tweets* que foram rotulados como "misto" são aqueles que possuem mais de um tipo de sentimento envolvido enquanto os rotulados com "outro" são difíceis de classificar.

Brum and Nunes [2017] construíram um conjunto de dados em PB, anotado a partir do Twitter, com base nas hashtags referentes a programas de televisão brasileiros. Cada *tweet* foi rotulado manualmente com as classes de sentimento positivo, negativo ou neutro. Contudo, apesar do conjunto de dados apresentar os *ids* dos Tweets, os dados não foram disponibilizados em sua integridade por conta da política de privacidade do Twitter, porém foram disponibilizados os *ids* dos Tweets. Além disso, os conjuntos de dados anotados apresentam uma limitação em relação a quantidade de texto disponível por Tweet visto que um Tweet pode ter até 280 caracteres[2].

De Pelle et al. de Pelle and Moreira [2017] criaram um conjunto de dados em PB, anotado para tarefas de detecção de discurso de ódio com base em comentários de usuários da internet. A coleta dos dados foi efetuada coletando comentários do site de notícias *g1.globo.com* onde, foi realizado um processo de julgamento utilizando um sistema[3], com três juízes, para rotular a classe dos comentários. Cada comentário foi classificado como ofensivo ou não, e caso classificado como ofensivo, poderia possuir de uma a seis

[1] http://www.meuqueridodiario.com.br

[2] eram 140 antes de novembro de 2017

[3] http://inf.ufrgs.br/~rppelle/hatedetector/

classes: racismo, sexismo, homofobia, xenofobia, intolerância religiosa ou xingamento. Os experimentos realizados consistiram em classificar automaticamente se um comentário contém discurso de ódio ou não.

Embora existam alguns conjuntos de dados textuais rotulados em PB (e.g., postagens no Twitter [Nascimento et al., 2012], notícias [de Pelle and Moreira, 2017]), o conjunto de dados aqui proposto se diferencia por não limitar a quantidade de caracteres permitidos (no caso do Twitter, são até 280 caracteres) e pelo próprio usuário ser o rotulador do seu texto, expressando, de forma voluntária, seu sentimento ao escrever.

3. Conjunto de dados para Análise de Sentimentos em PT-BR

Conforme mencionado nas seções anteriores, a maioria dos estudos no âmbito da AS estão concentrados em textos em inglês, o que motivou o presente trabalho a propor um conjunto de dados em PB. Para isto, foram extraídos dados textuais das entradas dos usuários da rede social brasileira MQD. Nesta rede social, os usuários compartilham seus sentimentos e emoções relacionados aos seus dias. O diferencial deste conjunto de dados é destacado pelo fato de o próprio usuário, que escreveu a entrada, ter realizado a classificação da emoção. Isso ocorre porque os usuários podem associar às suas entradas *uma* entre seis emoções (i.e., felicidade, tristeza, raiva, medo, nojo e surpresa).

No momento da extração dos dados[4], o MQD possuía 69.452 usuários cadastrados e 191.042 entradas. Do total de usuários, 32.546 possuem 1 ou mais entradas das quais foram selecionadas as entradas dos usuários que decidiram associar uma emoção ao realizar a escrita. Desta maneira, foram selecionadas 79.523 entradas e o número de entradas associada a cada emoção é distribuído da seguinte maneira: felicidade, 32.672; tristeza, 27.642; raiva, 6.323; medo, 6.112; surpresa, 5.262; nojo, 1.512;

O conjunto de dados apresentado foi denominado `MQDEmotion2018`[5]. Ao todo foram contabilizas 18.601.010 palavras e um vocabulário de 265.741 termos (palavras distintas) no conjunto de dados. Vale destacar que o `MQDEmotion2018` é um conjunto de dados para AS em PB, extraído diretamente de uma atividade cotidiana e, portanto, um dos maiores desafios desse conjunto de dados é a quantidade de palavras escritas com grafia incorreta, pois gera um grande número de palavras com apenas uma ocorrência.

4. Metodologia de avaliação

Os modelos de pré-processamento e seleção de atributo utilizados nos experimentos são apresentados nesta seção que é dividida em duas subseções: a Subseção 4.1 descreve a metodologia empregada para pré-processar o conjunto de dados utilizado nos experimentos, ou seja, o `MQDEmotion2018`. Em seguida, a Subseção 4.2 apresenta a metodologia adotada para a execução dos algoritmos de classificação.

4.1. Pré-processamento

O pré-processamento dos dados foi dividido em três fases. A primeira fase consistiu em selecionar um subconjunto dos dados do `MQDEmotion2018`. Isso ocorreu devido à tarefa proposta neste artigo, ou seja, a classificação de polaridade. Desta maneira, para os experimentos deste trabalho, foram utilizadas apenas entradas que representam as emoções *felicidade* e *tristeza*, o que corresponde a 60.314 entradas do

[4] A extração foi realizada em 10 de janeiro de 2018.

[5] Disponível em https://github.com/LaCAfe/MQDEmotion2018

`MQDEmotion2018`. Do total de entradas, 32.672 são rotuladas com a emoção *felicidade* e 27.642 com a emoção *tristeza*. Com isto, o número de palavras totalizou $13.926.356$ e o total de termos únicos, 220.540.

Em seguida, na segunda fase do pré-processamento, foram removidas as palavras com *uma* ocorrência, dado que elas podem piorar a tarefa de classificação Zhu and Chen [2005]. Desta maneira, o número de palavras totalizou $13.796.872$ e o número de termos únicos correspondeu a 91.056. Por fim, na terceira fase do pré-processamento, os textos dos documentos foram, inicialmente, convertidos para letras minúsculas. Em seguida, os documentos foram representados como vetores TF-IDF (*Term frequency - inverse document frequency*), uma vez que esta é a representação vetorial mais bem sucedida para na tarefa de categorização de textos Salton et al. [1975]. Estes são os dados utilizados no decorrer do presente artigo. Para melhor entendimento das demais seções, o conjunto de dados resultante das três fases de pré-processamento é denominado `MQDEmotion2018ft`.

4.2. Classificadores

Os experimentos utilizam o classificador Multinomial Naïve Bayes (MNB), por ser considerado o *baseline* por alguns trabalhos [Arias et al., 2013; Hassan and Mahmood, 2017]. Também utiliza o *Linear Support Vetor Machines* (LSVC), por ter apresentado bons resultados em tarefas de AS [Arias et al., 2013; Hassan and Mahmood, 2017]. Desta maneira, esses classificadores são empregados para avaliar o `MQDEmotion2018ft`.

O classificador *Naïve Bayes* supõe que todos os atributos extraídos das amostras fornecidas são independentes, dada uma hipótese em um contexto de classificação [McCallum et al., 1998]. Os autores ainda destacam que, embora essa presunção não seja verdade para a maioria dos problemas reais, o classificador *Naïve Bayes* tende a ter um bom desempenho. O *Multinomial Naïve Bayes* (MNB) é uma implementação do classificador *Naïve Bayes* e assume que os dados seguem uma distribuição multinomial, utilizando como informação o número de vezes que o termo ocorre em cada documento [McCallum et al., 1998].

O *Support Vector Machine* (SVM) tem como objetivo encontrar um hiperplano ótimo entre os dados em um espaço vetorial, dividindo-os em classes [Cortes and Vapnik, 1995]. Este algoritmo tem apresentado grande sucesso na tarefa de classificação em texto por ser efetivo em espaços de alta dimensionalidade e em tarefas que a dimensionalidade dos dados é maior que o número de amostras disponíveis [Forman, 2007]. Logo, utilizando os documentos de uma classe como pontos em um espaço n-dimensional, um classificador LSVC separa linearmente o espaço em que os pontos das classes devem pertencer.

4.3. Análise Experimental

Para realizar os experimentos, o trabalho foi organizado como segue. Os classificadores foram treinados e avaliados a partir da estratégia de validação cruzada de 10 partições (*10-fold cross-validation*).

Conforme procedido em Hassan and Mahmood [2017], além dos classificadores MNB e SVM linear (LSVC), foram utilizadas duas representações de N-gramas: unigrama (1-grama) e bigramas (2-gramas). Essas representações são encontradas na Tabela 1. A tarefa de classificação do MNB com 1-grama e 2-gramas foi denominada MNB^1

e MNB^2, respectivamente. Analogamente, a tarefa de classificação do LSVC com 1-grama e 2-gramas foi denominada $LSVC^1$ e $LSVC^2$. Os classificadores MNB foram avaliados com diferentes valores de *smoothing*(α), com $\alpha = 0$, $\alpha = 0.5$ e $\alpha = 1$.

Atributos	MNB	LSVC
1-grama	MNB^1	$LSVC^1$
2-grama	MNB^2	$LSVC^2$

Tabela 1. Rótulos de identificação da combinação entre a representação dos atributos e dos classificadores.

A Tabela 2 exibe a média dos resultados de acurácia, *F1 score* e seus respectivos desvios padrões obtidos para os modelos treinados. O modelo $LSVC^2$ obteve os melhores resultados, tanto na acurácia, quanto no *F1 score*. Ademais, o uso de 2-gramas provou ser melhor na tarefa, aumentando o *F1 score* e a acurácia de todos os modelos.

Classificador	Acurácia	F1
$MNB^1(\alpha = 0)$	$0,72 \pm 0,02$	$0,69 \pm 0,02$
$MNB^1(\alpha = 0,5)$	$0,78 \pm 0,02$	$0,77 \pm 0,02$
$MNB^1(\alpha = 1)$	$0,78 \pm 0,02$	$0,77 \pm 0,02$
$LSVC^1$	$0,81 \pm 0,01$	$0,79 \pm 0,01$
$MNB^2(\alpha = 0)$	$0,74 \pm 0,02$	$0,70 \pm 0,02$
$MNB^2(\alpha = 0,5)$	$0,79 \pm 0,01$	$0,78 \pm 0,02$
$MNB^2(\alpha = 1)$	$0,79 \pm 0,02$	$0,78 \pm 0,02$
$LSVC^2$	**0,82** $\pm 0,01$	**0,80** $\pm 0,01$

Tabela 2. Acurácia e *F1 score* obtidos com os modelos MNB e LSVC na tarefa de AS.

5. Conclusão

Este artigo apresenta como principal colaboração um novo conjunto de dados para análise de sentimentos. Esse conjunto de dados foi retirado de uma rede social brasileira denominada Meu Querido Diário. O principal diferencial desse conjunto de dados é que o próprio usuário fornece, de forma espontânea, a emoção associada ao seu texto no momento em que o estava escrevendo.

Os *baseline* escolhidos foram os algoritmos *Multinomial Naive Bayes* e *Linear Support Vector Classifier* para realizar as tarefas de classificação. De forma geral, os classificadores que utilizaram 2-gramas apresentaram melhorias na acurácia e no *F1 score* de todos os classificadores avaliados onde o classificador que obteve melhor desempenho foi o *Linear Support Vector Classifier* com 2-gramas assim como apresentado em Hassan and Mahmood [2017] para a tarefa de AS em inglês.

Em trabalhos futuros serão estudados outros modelos para realizar a classificação neste conjunto de dados. Com o objetivo de melhorar o desempenho das tarefas de análise de sentimentos, também serão analisadas melhores técnicas de pré-processamento e outras formas de representação de atributos, como por exemplo, *word embeddings*.

Referências

Arias, M., Arratia, A., and Xuriguera, R. (2013). Forecasting with twitter data. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(1):8.

Brum, H. B. and Nunes, M. d. G. V. (2017). Building a sentiment corpus of tweets in brazilian portuguese. *arXiv preprint arXiv:1712.08917*.

Cambria, E., Schuller, B., Xia, Y., and Havasi, C. (2013). New avenues in opinion mining and sentiment analysis. *IEEE Intelligent Systems*, 28(2):15–21.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297.

de Pelle, R. P. and Moreira, V. P. M. (2017). Offensive comments in the brazilian web: a dataset and baseline results. In *Congresso da Sociedade Brasileira de Computação-CSBC*.

Forman, G. (2007). Feature selection for text classification. *Computational methods of feature selection*, 1944355797.

Guedes, G. P., Bezerra, E., Ferrari, L., and Duarte, F. (2016). Gender differences in the use of portuguese in social networks: Evidence from liwc. In *Proceedings of the 22nd Brazilian Symposium on Multimedia and the Web*, pages 339–342. ACM.

Hassan, A. and Mahmood, A. (2017). Deep learning for sentence classification. In *Systems, Applications and Technology Conference (LISAT), 2017 IEEE Long Island*, pages 1–5. IEEE.

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167.

McCallum, A., Nigam, K., et al. (1998). A comparison of event models for naive bayes text classification. In *AAAI-98 workshop on learning for text categorization*, volume 752, pages 41–48. Citeseer.

Nascimento, P., Aguas, R., Lima, D., Kong, X., Osiek, B., Xexéo, G., and Souza, J. (2012). Análise de sentimento de tweets com foco em notícias. In *Brazilian Workshop on Social Network Analysis and Mining*.

Pang, B., Lee, L., et al. (2008). Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2):1–135.

Rosenthal, S., Nakov, P., Kiritchenko, S., Mohammad, S., Ritter, A., and Stoyanov, V. (2015). Semeval-2015 task 10: Sentiment analysis in twitter. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pages 451–463.

Saif, H., Fernandez, M., He, Y., and Alani, H. (2013). Evaluation datasets for twitter sentiment analysis: a survey and a new dataset, the sts-gold.

Salton, G., Wong, A., and Yang, C.-S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620.

Wiegand, M., Balahur, A., Roth, B., Klakow, D., and Montoyo, A. (2010). A survey on the role of negation in sentiment analysis. In *Proceedings of the workshop on negation and speculation in natural language processing*, pages 60–68. Association for Computational Linguistics.

Zhu, J. and Chen, W. (2005). Some studies on chinese domain knowledge dictionary and its application to text classification. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*.

Identificação de *fake news*: uma abordagem utilizando métodos de busca e *chatbots*

Yara de Lima Araujo[1], Anderson Cordeiro Chares[2], Jonice de Oliveira Sampaio[3]

[1,2,3]Instituto de Informática
Universidade Federal do Rio de Janeiro (UFRJ)

araujo.yara93@gmail.com, andersoncordeironf@gmail.com,
jonice@dcc.ufrj.br

Abstract. *Given the growth of social media, dissemination of information occurs in a fast and scalable way. This dynamic makes evaluation of informations' veracity a task that consumes time and breaks the social interaction flow. Therefore, rumors or fake news also spread faster on internet. In this paper, we propose a chatbot from Facebook Messenger that retrieves messages' keywords for possible rumors in a dataset, which contains data from Brazilian websites that gather news and classify them into true or fake. To do so, it utilizes some information retrieval techniques.*

Resumo. *Diante do grande crescimento das mídias sociais, a disseminação de informações ocorre de maneira mais rápida e escalável. Esse dinamismo transforma a avaliação da veracidade de uma informação em tarefa que demanda tempo e quebra do fluxo contínuo da interação social. Isto faz com que rumores também se espalhem com maior velocidade na rede. Nesse trabalho, propõe-se a utilização de um chatbot para o Facebook Messenger que, através de técnicas de busca e recuperação da informação, pesquisa em um dataset por palavras-chave de possíveis rumores que recebe de usuários, respondendo com links que possam auxiliar na validação da informação. Esse dataset é composto por conteúdo de sites brasileiros que reúnem notícias classificadas como falsas ou verdadeiras.*

1. Introdução

Nos últimos anos, o exponencial crescimento das mídias sociais intensificou a interação humana na internet permitindo que diversas informações fossem disseminadas rapidamente. Apesar da notória contribuição social trazida pelas tecnologias e ferramentas sociais, a dificuldade em lidar com o excesso de informação contribuiu para o também crescente surgimento de conteúdo não verificado (rumores) e muitas vezes falso (boatos).

A velocidade na comunicação impede que uma avaliação possa ser feita no conteúdo que está sendo trafegado e, além disso, os embates sociais resultantes de divergências sócio-políticas impulsionam a prática da produção de conteúdo duvidoso que sirva de alicerce para críticas ou fomente discussões na rede. Identificar esses boatos se apresenta como um desafio; estabelecer a confiabilidade de informações online é um desafio assustador mas crítico [Conroy, Rubin, & Chen 2015]. De acordo com Vosoughi & Roy [2017], um rumor pode ser definido como uma afirmação não-

verificada que começa em uma fonte (ou mais) e começa a se espalhar ao longo do tempo para diversos nós na rede.

Rumores podem ser classificados em alguns tipos. De acordo com Rubin, Chen & Conroy [2015], três tipos são identificados. Primeiramente, as notícias que provém de jornais sensacionalistas, que fabricam notícias sobre escândalos, pessoas famosas, crimes, com o objetivo de obter audiência; muitas notícias são fabricadas, falsificadas ou exageradas, o que leva a muitos rumores. O segundo tipo são definidos como *hoaxes* (embustes/enganos), rumores criados nas mídias sociais com a finalidade de enganar as pessoas como notícias inverídicas. Estes rumores podem ser validados por engano por meios de notícia tradicionais. Diferente de uma simples brincadeira, *hoaxes* podem causar danos reais a alguém.

Este trabalho está organizado da seguinte forma: a seção 2 resume os trabalhos relacionados, a seção 3 demonstra a aplicação proposta e todas as etapas de seu desenvolvimento, a seção 4 apresenta os resultados preliminares e a seção 5 apresenta a conclusão.

2. Trabalhos Relacionados

Atualmente, diversas técnicas têm emergido no intuito de detectar rumores na internet. O trabalho de Castillo *et al* [2011] utiliza técnicas de aprendizado supervisionado de máquina para classificar um tópico no Twitter como notícia ou conversa pessoal e dentro da classe notícia classificá-la como crédula ou não crédula. A classificação é feita por humanos e um algoritmo de aprendizado extrai os padrões que tweets dessas classes possuem. O trabalho conclui que tópicos de notícias costumam possuir links e uma árvore de propagação maior.

Já Buntain [2015] avalia a crescente de palavras-chave em épocas de crise, relacionando a localização de um determinado evento com o surgimento de mensagens a seu respeito e, utilizando aprendizagem de máquina, analisou a técnica proposta em mídias sociais durante eventos esportivos e adaptou-os a situações de risco como a ocorrência de terremotos. Seus experimentos demonstraram que, dados estes cenários, identificar palavras importantes e verificar suas origens pode auxiliar na avaliação da credibilidade de informação.

No trabalho de Shou *et al*. [2016], os autores desenvolveram uma plataforma chamada *Hoaxy* para coletar notícias de diversas mídias sociais, assim como sites de notícias através de *crawlers* e APIs. Após a coleta, a plataforma somente rastreia atualizações através de *RSS.* Todas as notícias são armazenadas em um banco de dados. Como trabalhos futuros, os autores indicam o desenvolvimento de uma interface web interativa para análise das notícias.

Apesar da variedade de técnicas, ferramentas e plataformas utilizadas para a detecção de rumores em outros países, no Brasil ainda há uma escassez de fontes de dados públicos que possam servir de consulta na tarefa de verificação de informação. Além disso, o controle automático de conteúdo quanto à sua veracidade permanece como uma tarefa difícil às máquinas por requerer níveis de abstração e criatividade inerentes a seres humanos. Sabendo disso, soluções que envolvam o desenvolvimento colaborativo de conhecimento como crowdsourcing e o acesso à fontes de dados

abertas, podem servir de alicerce para o desenvolvimento de soluções frente ao desafio da identificação de rumores. Por apresentar esse perfil, acredita-se que a aplicação desenvolvida neste trabalho possa contribuir efetivamente para o problema de identificação de rumores.

3. Aplicação proposta

Nesta seção são apresentados os componentes da aplicação proposta e descritos os processos da criação do dataset de rumores, da API de utilização do dataset e do chatbot desenvolvido para validar a ferramenta de busca.

A Figura 1 apresenta a estrutura da aplicação como um todo. As interações ocorrem com o usuário através do *FakeChatBot*, são recebidas pela API, transformadas em buscas diretas ao dataset de notícias (busca.py); após isso resultados são retornados ao usuário. É possível identificar os módulos que compõem todo o processo, como os *crawlers* e a camada na qual o modelo vetorial é desenvolvido, detalhados na seção 3.4.

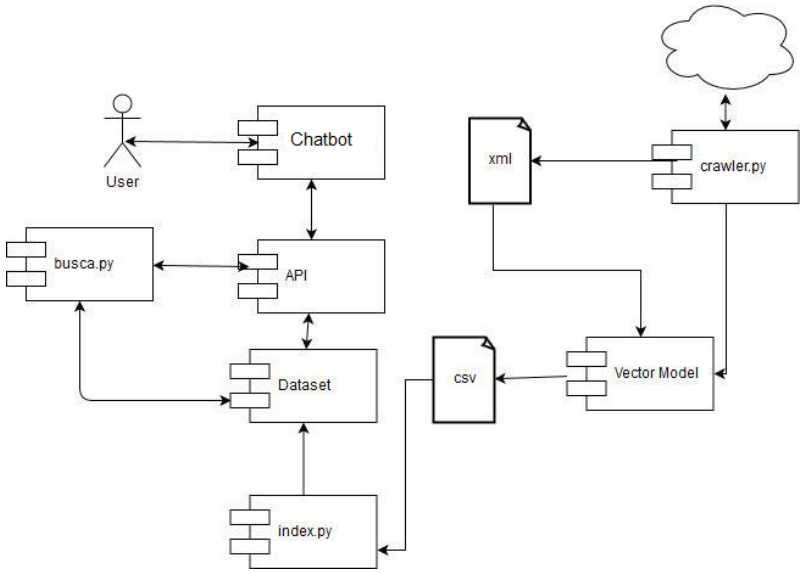


Figura 1. Estrutura da aplicação

3.1 Dataset de Rumores

Com o objetivo de reunir em um *dataset* rumores já conhecidos e assim disponibilizar tal fonte de dados para consultas, foram criados *crawlers* que percorrem sites brasileiros conhecidos por trabalhar com verificação de fatos, fazendo o download do conteúdo e armazenando no *dataset* de maneira estruturada para facilitar a compreensão do conteúdo da notícia e do resultado da avaliação realizada pelo site (falso ou verdadeiro). Foram utilizados os sites "boatos.org"[1],"e-farsas"[2] e "é ou não é"[3]. Para realizar esse processo, *crawlers* desenvolvidos em Python extraem os dados dos sites utilizando a biblioteca *Beautiful Soup*, e os escrevem em um arquivo .xml.

O conteúdo dos arquivos .xml é então submetido a técnicas de normalização de texto para que possam ser removidos termos irrelevantes ao mecanismo de busca, como artigos e pronomes -*stop words*. Notícias duplicadas também foram removidas. Após essa etapa, outro *script* é acionado para a criação de um arquivo .csv estruturado com campos que seguem o padrão do .xml como "*summary*" com uma descrição da notícia, e

"*check*", campo booleano que indica se a notícia foi identificada como verdadeira ou falsa. Após a organização e definição da estrutura, os conteúdo dos arquivos é armazenado no dataset *online* e disponibilizado ao método de busca que está descrito na seção 3.3.

3.2 API

Para facilitar a utilização do dataset, e incentivar o desenvolvimento de aplicações que pudessem utilizá-lo como fonte de dados, foi desenvolvida uma API (*Application Programming Interface*), cujas funções permitem a interação com os registros do dataset.

Através da API é possível listar todos os registros armazenados pelos *crawlers*, utilizar filtros para identificar o que foi classificado como boato ou verdadeiro nos sites especializados e realizar consultas utilizando de um processo bem definido de Busca e Recuperação de Informação. Disponível para acesso em www.fakepedia.org, a API possui uma breve documentação sobre suas funcionalidades e a descrição das chamadas para interação com o *dataset*.

3.3. Métodos de busca e recuperação da informação

Uma das funcionalidades da API a ser utilizada é a busca por um termo ou expressão utilizando o Modelo Vetorial de Busca e Recuperação de Informação. No Modelo Vetorial, um documento é representado por um conjunto de termos indexados e associados a um valor normalizado que indica o seu grau de relevância para o documento. Optou-se pelo Modelo Vetorial por ser bastante difundido e se mostrar uma solução eficaz para o problema de recuperação de texto, como em Senin & Malinchik [2013].

Neste processo, o grau de similaridade (relevância) é calculado através de pesos atribuídos aos termos da consulta e do documento em si. Para o cálculo destes pesos, implementou-se um script que cria uma lista invertida a partir dos arquivos .xml gerados pelos *crawlers*, contendo cada termo e o conjunto de notícias correspondentes. As listas auxiliam no cálculo do TFxIDF dos termos, sendo TF ("*term frequency*") o número de vezes que um termo aparece em um documento e IDF ("*inverse document frequency*") a freqüência com que um termo ocorre em todo conjunto de documentos. Maiores detalhes sobre TDF-IDF podem ser encontrados em Ramos [2013].

3.4. FakeChatBot

O uso de *chatbots* permite um alcance instantâneo a uma grande quantidade de pessoas [Schlicht 2016]. Dada a popularização e alcance desse tipo de abordagem, o desenvolvimento de um *chatbot* como interface entre o usuário e a API mostrou-se como uma excelente alternativa para a validação do uso da ferramenta de identificação de rumores.

Optou-se por utilizar o Facebook Messenger como plataforma para o desenvolvimento do chatbot devido a sua popularidade, pois é utilizado por mais de um bilhão de pessoas [Constine 2016].

O chatbot, chamado de *FakeCheckBot*, foi desenvolvido em Python, utilizando algumas bibliotecas como *Flask (desenvolvimento web), requests*, e gunicorn – espécie

de servidor Python para web. Em seguida, foi criada uma página no *Facebook* de mesmo nome para vincular o chatbot a um perfil da plataforma.

4. Resultados

Após a definição das ferramentas e plataformas a serem utilizadas, bem como a criação da estrutura necessária para receber requisições através do *Facebook Messenger*, testes foram realizados a fim de verificar a qualidade do retorno obtido através das consultas realizadas no *dataset*.

Apesar de carecer de experimentos para validação, resultados preliminares demonstram que o *FakeChatBot* é capaz de identificar trechos de boatos reconhecidos e retornar links com conteúdo já verificado acerca do assunto. A Figura 2 representa o momento em que uma pessoa envia uma pergunta sobre um boato ao *chatbot*, a resposta enviada pelo chatbot após todo o processo de busca e recuperação de informação e métodos aqui descritos.

Notou-se que algumas variações e textos mais longos demandam tempo maior de processamento e, por vezes, não retornam o conteúdo exato por conta das limitações do plano gratuito do servidor utilizado. Também foi notado que a busca de termos muito curtos, contendo uma ou duas palavras, retorna resultados pouco precisos.

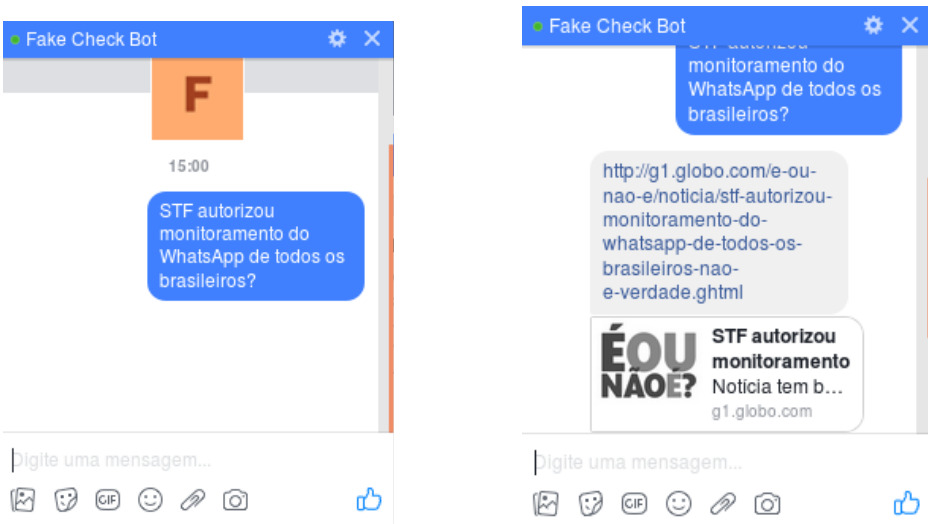


Figura 2. Mensagem enviada pelo usuário e respondida pelo chatbot

5. Conclusão

Diante do crescente uso da internet e das mídias sociais, um grande volume de dados é criado e espalhado rapidamente pela rede. Por consequência, muitos rumores também são disseminados. Entretanto, percebe-se que há dificuldade de avaliação da veracidade de informações por parte de usuários, que não desejam parar suas interações para verificar, por exemplo, se uma notícia é verdadeira.

Neste sentido, muitas abordagens surgiram no intuito de identificar rumores na rede. Neste trabalho, foi proposta uma solução que permite a busca de rumores da internet a partir de um chatbot. Para tal, foram recuperadas informações de sites brasileiros especializados em verificação da informação, através de *crawlers* e reunidos em um único dataset. Para facilitar o acesso, uma API foi desenvolvida e disponibilizada online.

Utilizando Python, foi desenvolvido um chatbot para o *Facebook Messenger* que recebe um texto do usuário e verifica, através da similaridade dos termos, se existe algum boato correspondente no dataset.

Por fim, através desta abordagem foi possível prover uma solução funcional que permite que usuários busquem por rumores de fontes verificadas e descubram se determinada notícia é um rumor ou não. Além disto, propõe-se um *dataset* unificado de bases em português, o que dificilmente está disponível de acordo com um levantamento realizado.

Como trabalhos futuros, pretende-se melhorar o processo de busca através da adoção de técnicas de paráfrase capazes de torná-lo mais inteligente. A partir disto, será possível inferir variações nos termos buscados, que podem ser um mesmo rumor escrito de maneira distinta.

Referências

Buntain, C. (2015) “Discovering Credible Events in Near Real Time from Social Media Streams”, *WWW 2015 Companion*, May 18–22, 2015, Florence, Italy

Conroy, N., Victoria L. and Chen, Y. (2015) “Automatic Deception Detection: Methods for Finding Fake News”, *ASIST 2015*, November 610, 2015, St. Louis, MO, USA.

Ferneda, E.Introdução aos Modelos Computacionais de Recuperação de Informação. [S.l.]: Editora Ciência Moderna, 2012

Ramos, J. (2003). Using tf-idf to determine word relevance in document queries. *Proceedings of the first instructional conference on machine learning* (Vol. 242, pp. 133-142).

Rubin, V. L., Chen, Y., & Conroy, N. J. (2015). Deception Detection for News: Three Types of Fake News. *Proceedings of the Association for Information Science and Technology*, *52*(1), 1–4.

Senin, P., & Malinchik, S. (2013). Sax-vsm: Interpretable time series classification using sax and vector space model. *IEEE 13th International Conference on Data Mining* (pp. 1175-1180). IEEE.

Shao, C., Ciampaglia, G. L., Flammini, A., & Menczer, F. (2016). Hoaxy: A Platform for Tracking Online Misinformation. https://doi.org/10.1145/2872518.2890098

Vosoughi, S., & Roy, D. (2017). Rumor Gauge : Predicting the Veracity of Rumors on Twitter. *ACM Transactions on Knowledge Discovery from Data (TKDD). In Press.* *11*(4), 1–38. https://doi.org/10.1145/3070644

Identificando Sinais de Comportamento Depressivo em Redes Sociais

Rodolpho da Silva Nascimento[1], **Pedro Parreira**[1],
Gabriel Nascimento dos Santos[1], **Gustavo Paiva Guedes**[1]

[1]CEFET/RJ - Centro Federal de Educação Tecnológica Celso Suckow da Fonseca
Av. Maracanã, 229 - Rio de Janeiro - RJ - Brasil.

{rodolpho.nascimento,pedro.cruz}@eic.cefet-rj.br,
gustavo.guedes@cefet-rj.br

Abstract. *There are over 300 million individuals suffering from depression in the world. Among the many consequences of this disorder, the most catastrophic is suicide. In this aspect, several individuals manifest in social networks the depressive signs before reaching the end of the suicide. In this scenario, the present study aims to construct a data set of words related to depression for the Portuguese language. Preliminary experiments indicate that the dataset created helps to classify emotions in texts of a Brazilian social network.*

Resumo. *Existem mais de 300 milhões de indivíduos que sofrem de depressão no mundo. Dentre as diversas consequências provenientes desse transtorno, a mais catastrófica é o suicídio. Nesse aspecto, diversos indivíduos manifestam em redes sociais os sinais depressivos antes de chegar ao extremo do suicídio. Nesse cenário, o presente estudo tem objetiva construir um conjunto de dados de palavras relacionadas à depressão para a língua portuguesa. Experimentos preliminares indicam que o conjunto de dados criado auxilia na classificação de emoções em textos de uma rede social brasileira.*

1. Introdução

A Organização Mundial de Saúde (OMS) destaca que a cada quatro pessoas no mundo, uma é afetada por algum tipo de problema mental em determinada fase de sua vida [Organization 2001]. Ainda segundo o estudo da OMS, a depressão é a principal causa de incapacidade global e ocupa o quarto lugar de destaque entre as dez principais doenças. Com isso, a OMS evidencia a relevância de um lançamento de campanhas por todo o mundo, objetivando prevenir sintomas da depressão ou consequências a ela associadas, como suicídio, epilepsia e esquizofrenia [WHO 2015].

Estima-se que mais de 300 milhões de indivíduos sofrem de depressão no mundo, tendo havido um aumento de 18% entre os anos de 2005 e 2015 [WHO 2017]. Dentre outras consequências negativas provenientes da depressão, a mais catastrófica é o suicídio, que é a segunda maior causa de morte entre os adolescentes [WHO 2018]. Em alguns casos de suicídio, as vítimas apresentaram, previamente, sinais depressivos em conteúdos postados nas redes sociais. Dentre eles, podemos destacar o caso de uma estudade de Rio Branco no Acre, que momentos antes de seu suicídio postou o seguinte texto *Já viram*

alguém morrer ao vivo?[1]. Também houve o caso de um jovem de Itaparica na Bahia, que antes de cometer o suicídio, publicou uma carta de despedida no Facebook[2].

Nesse cenário é interessante destacar que metade dos brasileiros utilizam a internet (i.e., 48%) e desses, 92% utilizam alguma rede social [SECOM 2015]. Dessa maneira, diversos esforços vêm sendo empreendidos para evidenciar sintomas depressivos em redes sociais [Goh and Huang 2009]. Frente a isso, é importante considerar a necessidade de intervenções nos casos em que esses sintomas sejam detectados [Moreno et al. 2011].

Dado o panorama apresentando nos parágrafos precedentes, o presente trabalho tem dois objetivos principais: (*i*) o primeiro corresponde à criação de um conjunto de dados de palavras relacionadas à depressão para o português do Brasil, visto que não foram encontrados semelhantes na literatura; (*ii*) o segundo objetivo refere-se à realização de experimentos iniciais com base no conjunto de dados criado, objetivando melhorar a detecção de textos com emoção negativa. Esse objetivo se fundamenta no estudo proposto em [Shen et al. 2013], que considera a "identificação de manifestação de emoções negativas" como primeiro passo para a detecção de potenciais candidatos à depressão.

O restante desse artigo está organizado da seguinte forma: a Seção 2 descreve trabalhos relacionados ao contexto de detecção de depressão em redes sociais; a Seção 3 discute os conjuntos de dados utilizados no presente estudo; a Seção 4 detalha a criação do conjunto de dados de depressão; a Seção 5 apresenta a metodologia utilizada para avaliar o conjunto de dados criado; a Seção 6 descreve os resultados alcançados no presente estudo e, por fim, a Seção 7 conclui e apresenta alguns cenários de trabalhos futuros.

2. Trabalhos relacionados

[Nguyen et al. 2014] apresentaram um trabalho no qual investigaram comunidades de interesse em depressão e estudaram seus fatores diferenciais para outras comunidades de outros temas. Três aspectos foram examinados: afeto, processos psicolinguísticos e tópicos dentro de cada conteúdo postado. Os autores utilizaram o conjunto de dados LiveJournal[3] para efetuar analise de sentimentos. Os resultados alcançados indicaram alta prevalência de palavras com características depressivas (e.g., *ódio, suicídio, dor*) nas comunidades de interesse em depressão em comparação com comunidades de outros temas.

[Nambisan et al. 2015] coletaram dados do Twitter para identificar *posts* relacionados à depressão. Este diagnóstico constituiu na formação de um conjunto de palavras ou frases, tais como "eu estou sofrendo E depressão", "eu E medicamentos depressão", "eu E terapia E depressão", dentre outras. Esse conjunto de palavras serviu como filtro de seleção dos *tweets* e assim, foi criada uma amostra com os termos desses *posts*. Seguidamente, elaboraram uma segunda amostra em que foram coletados *posts* aleatoriamente. Os resultados deste estudo indicaram que os usuários deprimidos do Twitter exibem as mesmas características comportamentais quando estão conectados (postam entradas) e quando não estão, especialmente em tópicos como sono, dor e pensamentos suicidas. O

[1]https://g1.globo.com/ac/acre/noticia/estudante-transmite-suicidio-ao-vivo-em-rede-social-ja-viram-alguem-morrer.ghtml

[2]http://www.itaberabanoticias.com.br/plantao/jovem-evangelico-comete-suicidio-apos-deixar-carta-de-despedida-no-facebook

[3]https://www.livejournal.com/

estudo também indicou que o Twitter é utilizado por indivíduos com depressão para expressarem pensamentos e refletirem sobre os sintomas e problemas que os incomodam.

Os trabalhos supracitados envolveram a elaboração de léxicos, que foram utilizados para análise de palavras que continham características depressivas, no entretanto, todos foram criados para a língua inglesa. Este estudo objetiva desenvolver um léxico com características semelhantes porém, com base em palavras do português do Brasil.

3. Conjuntos de dados existentes

ANEW-Br: O ANEW-br [Haag Kristensen et al. 2011] é um conjunto de dados composto por 1.046 palavras em português do Brasil. Cada uma dessas palavras é associada a valores de valência que correspondem o intervalo entre 1.16 a 8.80. Valores superiores a 5 indicam valência positiva. Analogamente, valores inferiores a 5 indicam valência negativa. Do total de palavras, 466 possuem valência negativa, que formam o conjunto de palavras relevante para o presente estudo.

LIWC: O Linguistic Inquiry and Word Count (LIWC) é um sistema proposto por James Pennebaker [Rude et al. 2004]. O LIWC classifica palavras em uma ou mais categorias psicolinguísticas com auxílio de um dicionário interno. A versão brasileira do dicionário do LIWC [Balage Filho et al. 2013], aqui denominada LIWC2007-pt-br, contém 127.149 palavras e 64 categorias psicolinguísticas. É importante destacar que, nesse estudo são utilizadas apenas as categorias *negemo*, *posemo* e as palavras associadas a essas categorias.

MQD: O Meu Querido Diário (MQD) é uma rede social brasileira criada em 2009 e tem o objetivo de funcionar como um diário online, em que os usuários podem descrever sobre o dia-a-dia. É interessante destacar que no momento em que escreve suas entradas, os usuários podem classificá-las com 1 entre as 6 emoções básicas propostas por Paul Ekman [Ekman 1992], essas são, *felicidade*, *tristeza*, *raiva*, *medo*, *nojo*, *surpresa*. O MQD possui atualmente 69.452 entradas. Dentre essas entradas, 32.244 estão associadas com a emoção *felicidade* e 26.921 estão associadas com a emoção *tristeza*. As entradas com a emoção *felicidade* e *tristeza* formam o conjunto de dados `MQD-FT`, que compreende 59.165 entradas.

4. Depress-pt-br

O conjunto de dados de depressão proposto neste artigo, denominado `Depress-pt-br`, foi construído em 3 fases, conforme ilustra a Figura 1. A primeira fase consistiu na aquisição de textos com palavras de cunho depressivo. Essa aquisição foi efetuada com base na coleta de dados provenientes do site Yahoo Respostas[4], conforme adotado em [De Choudhury et al. 2013] para a língua inglesa.

Devido a limitações técnicas do Yahoo Respostas, a coleta de dados restringiu-se a 1.000 resultados por *string* de busca. Desta maneira, as *strings* de busca utilizadas na coleta foram `depressão`, `depressão+suicídio` e `depressão+socorro`, respectivamente. Essa tarefa resultou em um total de 3.000 documentos únicos. Para cada resultado, foram extraídos o (*i*) título resumido da pergunta; (*ii*) o detalhe da pergunta (nem sempre informado) e (*iiii*) a melhor resposta para a pergunta, sendo essa indicada

[4]http://br.answer.yahoo.com

pelos próprios usuários do site. Essa coleta resultou em um conjunto de documentos de texto contendo termos relacionados à depressão.

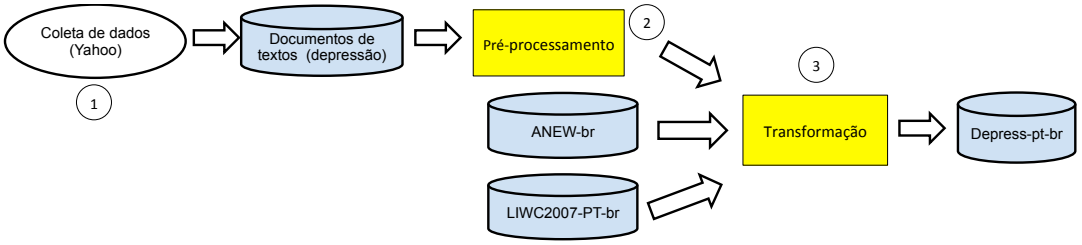


Figura 1. Fases para a construção do conjunto de dados `Depress-pt-br`.

A segunda fase consistiu no pré-processamento de todos os documentos de textos com termos relacionados à depressão. Foram efetuadas as tarefas de *tokenização* e *case folding* (conversão dos *tokens* para letra minúscula). É importante ressaltar que as pontuações foram removidas.

Por fim, a terceira fase consistiu em unificar o resultado da saída da fase 2 (pré-processamento) com os conjuntos de dados ANEW-br e LIWC2007-pt-br. Inicialmente, foram selecionadas todas as palavras do ANEW-br que possuem valência negativa, resultando em 466 palavras. Esse procedimento foi adotado com base na premissa de que indivíduos com depressão tendem a utilizar mais palavras com valência negativa [Rude et al. 2004].

Das 466 palavras com valência negativa do ANEW-br, foram sub-selecionadas as que estavam contidas no resultado do pré-processamento da fase 2, o que resultou em 335 palavras. Para finalizar a terceira fase, foram selecionadas todas as palavras do LIWC2007-pt-br para inclusão no `Depress-pt-br`, além das categorias *posemo* e *negemo*. Após essa inclusão, o `Depress-pt-br` continha 127.149 palavras e 2 categorias. Em seguida, foi incluída a categoria *depress* para representar as palavras relacionadas à depressão. Por fim, a categoria *depress* do `Depress-pt-br` foi associada a cada uma das 335 palavras sub-selecionadas do ANEW-br [5].

5. Metodologia

A metodologia proposta neste estudo consiste em utilizar o conjunto de dados descrito na Seção 4 (i.e., `Depress-pt-br`) para auxiliar na detecção de entradas positivas e negativas em um conjunto de dados proveniente do MQD. Esse conjunto de dados, aqui denominado de `MQD-FT`, consiste em 32.244 entradas com a emoção *felicidade* e 26.921 entradas com a emoção *tristeza*.

Inicialmente, os textos de cada entrada e tiveram todos os caracteres convertidos para letra minúscula. Em seguida, os textos foram transformados em vetores de contagem de palavras. Cada dimensão do vetor correspondente diz respeito a uma das 3 dimensões presentes no `Depress-pt-br` (i.e., *posemo*, *negemo* e *depress*). A Tabela 1 ilustra um exemplo contendo duas entradas. Pode-se notar, por exemplo, que a entrada representada por e_1 possui 43, 52 e 14 palavras associadas à categoria de emoções positivas, emoções negativas e depressão, respectivamente.

[5] A lista com as 335 palavras se encontra em https://github.com/LaCAfe/Depress-pt-br.

Entrada (e)	*posemo*	*negemo*	*depress*
e_1	43	52	14
e_2	50	11	1

Tabela 1. Representação das entradas em vetores de contagem de palavras.

Em seguida, esses vetores foram submetidos ao algoritmo J48 para que fosse realizada a tarefa de classificação em duas classes no MQD-FT: *felicidade* e *tristeza*. Foi adotado o procedimento de validação cruzada com 10 partições aleatórias (*10-fold cross-validation*). Os resultados foram analisados em termos de F1 e número de entradas classificadas corretamente. O algoritmo J48 se encontra na plataforma Weka[6].

6. Resultados

Os resultados deste estudo são exibidos com a avaliação dos atributos do conjunto de dados `Depress-pt-br` na tarefa de classificação de emoções no conjunto de dados `MQD-FT`. A Tabela 2 apresenta o resultado encontrado. Os atributos foram avaliados separadamente e combinados.

Atributo(s)	F1	Class. corret.
negemo	$0,676$	39.988
posemo	$0,552$	32.638
depress	$0,611$	36.153
negemo + posemo	0.679	40.242
negemo + depress	**0,681**	40.287
posemo + depress	$0,611$	36.379
negemo+ posemo + depress	**0,681**	**40.444**

Tabela 2. Resultados da classificação obtidos com a aplicação do algoritmo J48.

Na análise dos atributos separadamente, o atributo *negemo* apresenta o melhor F1 ao ser realizada a tarefa de classificação, ou seja, $0,676$. O segundo melhor atributo foi o *depress* que obteve F1 igual a $0,611$. A combinação de dois atributos indica que os atributos *negemo* e *depress* alcançam o melhor F1 $(0,681)$, ou seja, o mesmo valor alcançado pela combinação dos três atributos em questão. No entanto, a combinação dos 3 atributos possui maior número de acertos (40.444).

7. Conclusão

O presente estudo objetivou a criação de um conjunto de dados de palavras com conteúdo depressivo para o português do Brasil, denominado `Depress-pt-br`. Em seguida, utilizou esse conjunto para efetuar a classificação de emoções em textos das classes *felicidade* e *tristeza* com base em entradas de uma rede social brasileira. Os resultados indicam que o `Depress-pt-br` foi capaz de melhorar os resultados na tarefa de classificação, quando comparada a utilização dos atributos *negemo* e *posemo* do LIWC.

No conhecimento dos autores, não há trabalhos anteriores que criaram um conjunto de dados de palavras com conteúdo depressivo em português do Brasil. Dessa maneira, esse conjunto pode servir de *baseline* para novos estudos que objetivem a detecção

[6]https://www.cs.waikato.ac.nz/ml/weka/

de depressão em nessa língua. Como trabalhos futuros, serão adicionadas novas palavras ao conjunto de dados e será avaliada a atribuição de pesos às palavras.

Referências

Balage Filho, P. P., Pardo, T. A. S., and Aluísio, S. M. (2013). An evaluation of the brazilian portuguese liwc dictionary for sentiment analysis. In *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology*.

De Choudhury, M., Gamon, M., Counts, S., and Horvitz, E. (2013). Predicting depression via social media. *ICWSM*, 13:1–10.

Ekman, P. (1992). An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200.

Goh, T.-T. and Huang, Y.-P. (2009). Monitoring youth depression risk in web 2.0. *VINE*, 39(3):192–202.

Haag Kristensen, C., Azevedo Gomes, C. F., Reuwsaat Justo, A., and Vieira, K. (2011). Normas brasileiras para o affective norms for english words. *Trends in Psychiatry and Psychotherapy*, 33(3).

Moreno, M. A., Jelenchick, L. A., Egan, K. G., Cox, E., Young, H., Gannon, K. E., and Becker, T. (2011). Feeling bad on facebook: Depression disclosures by college students on a social networking site. *Depression and anxiety*, 28(6):447–455.

Nambisan, P., Luo, Z., Kapoor, A., Patrick, T. B., and Cisler, R. A. (2015). Social media, big data, and public health informatics: Ruminating behavior of depression revealed through twitter. *2015 48th Hawaii International Conference on System Sciences*.

Nguyen, T., Phung, D., Dao, B., Venkatesh, S., and Berk, M. (2014). Affective and content analysis of online depression communities. *IEEE Transactions on Affective Computing*, 5(3):217–226.

Organization, W. H. (2001). *The World Health Report 2001: Mental health: new understanding, new hope*. World Health Organization.

Rude, S., Gortner, E.-M., and Pennebaker, J. (2004). Language use of depressed and depression-vulnerable college students. *Cognition & Emotion*, 18(8):1121–1133.

SECOM (2015). Pesquisa brasileira de mídia 2015. url: http://www.secom.gov.br/atuacao/pesquisa/lista-de-pesquisas-quantitativas-e-qualitativas-de-contratos-atuais/pesquisa-brasileira-de-midia-pbm-2015.pdf/view.

Shen, Y.-C., Kuo, T.-T., Yeh, I.-N., Chen, T.-T., and Lin, S.-D. (2013). Exploiting temporal information in a two-stage classification framework for content-based depression detection. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 276–288. Springer.

WHO (2015). Comprehensive mental health action plan 2013–2020. url: http://www.who.int/mental_health/action_plan_2013/en/.

WHO (2017). "depression: let's talk"says who, as depression tops list of causes of ill health. url: http://www.who.int/mediacentre/news/releases/2017/world-health-day/en/.

WHO (2018). Depression. url: http://www.who.int/mediacentre/factsheets/fs369/en/.

Importância das Colaborações Interdisciplinares nas Redes de Coautoria Científica

Geraldo J. Pessoa Junior[1], **Thiago M. R. Dias**[2],
Thiago H. P. Silva[1], **Alberto H. F. Laender**[1]

[1]Universidade Federal de Minas Gerais 31270-901 Belo Horizonte, MG, Brasil

[2]Centro Federal de Educação Tecnológica 35503-822 Divinópolis, MG, Brasil

{geraldo.pessoa, thps, laender}@dcc.ufmg.br, thiago@div.cefetmg.br

Abstract. *How important are interdisciplinary collaborations in scientific coauthorship networks? This is a relevant question that has drawn the attention of scholars, since bridging academic relationships contributes to make scientific networks stronger. However, traditional studies have focused on characterizing specific groups rather than on studying a complete and robust scientific community. In this paper, we characterize such relationships by considering the eight major areas of the entire Brazilian scientific coauthorship network. Our results show that the Brazilian coauthorship network grew and became remarkably interdisciplinary.*

Resumo. *Quão importantes são as colaborações interdisciplinares nas redes de coautoria científica? Esta é uma questão relevante que tem chamado a atenção da comunidade acadêmica, já que o estreitamento das relações entre pesquisadores contribui para tornar as redes científicas mais fortes. Entretanto, trabalhos tradicionais têm se concentrado na caracterização de grupos específicos e não no estudo de uma comunidade científica completa e robusta. Neste artigo, caracterizamos essas relações considerando as oito grandes áreas de toda a rede de coautoria científica brasileira. Nossos resultados mostram que a rede de coautoria brasileira cresceu e tornou-se especialmente interdisciplinar.*

1. Introdução

Quão importantes são as colaborações interdisciplinares nas redes de coautoria científica? Esta é uma questão importante e ampla que tem chamado a atenção da comunidade acadêmica, uma vez que o estreitamento das relações entre pesquisadores contribui para tornar as redes de colaboração científica mais fortes. Esforços recentes têm focado nas colaborações externas (por exemplo, cooperação entre grupos de pesquisa, migrações de pesquisadores e influência de áreas de pesquisa) como sendo fatores relevantes na evolução das comunidades científicas [Kato and Ando 2013, Mena-Chalco et al. 2014, Mooney et al. 2013]. No entanto, tais estudos concentram-se principalmente na caracterização de áreas ou grupos de pesquisa específicos (e.g., [Kato and Ando 2013]) em vez de contemplar a análise de uma comunidade completa.

Assim, ao invés de analisarmos cenários particulares de grupos fechados (por exemplo, um determinado departamento acadêmico ou uma área de pesquisa específica), neste artigo avançamos um passo adiante construindo as redes de coautoria global e interdisciplinar de todo o Brasil, usando como fonte de dados a Plataforma Lattes[1]. Mantida

[1]Lattes Platform: http://lattes.cnpq.br

pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), a Plataforma Lattes é uma iniciativa de renome internacional [Lane 2010] que fornece um repositório de currículos vitae e grupos de pesquisa integrados em um único sistema. Uma vez que todos os pesquisadores que atuam no Brasil são obrigados a manter seus currículos atualizados nessa plataforma, ela fornece um grande volume de informação sobre as atividades de pesquisa e a produção científica desses indivíduos.

Neste artigo analisamos a importância das colaborações interdisciplinares usando como referência o nível superior do esquema de classificação das áreas do conhecimento proposto pelo CNPq[2], que considera as seguintes oito grandes áreas: Ciências Agrárias, Ciências Biológicas, Ciências da Saúde, Ciências Exatas e da Terra, Ciências Humanas, Ciências Aplicadas e Sociais, Engenharias, e Linguística, Letras e Artes. Note que este esforço envolve a rede completa de coautoria acadêmica brasileira, compreendendo todos os indivíduos com doutorado (264.731 em junho de 2017) ora atuando em atividades de pesquisa no país. Como veremos, 37,4% de todas as coautorias nessa rede são interdisciplinares, ou seja, envolvem pesquisadores de diferentes grandes áreas, o que enfatiza a importância das colaborações interdisciplinares no cenário atual de ciência e tecnologia.

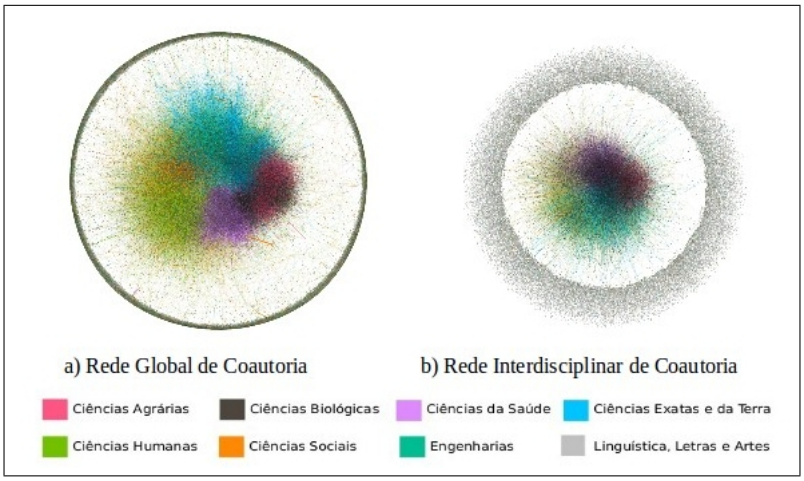


Figura 1. Redes Global e Interdisciplinar de Coautoria no Brasil

Para ilustrar nossa alegação, a Figura 1 revela o volume e a complexidade de tais colaborações nas diferentes comunidades científicas. O grafo à esquerda mostra a rede de coautoria global brasileira incluindo as colaborações dentro de cada grande área, enquanto que o grafo à direita mostra apenas as colaborações interdisciplinares. Como podemos observar, a rede global é mais densa com fortes conexões entre os nós do componente gigante, enquanto que a rede interdisciplinar possui um conjunto maior de componentes isolados, tendo em vista que ao serem removidas as conexões não interdisciplinares ela tornou-se menos densa, aumentando a quantidade de nós isolados que podem ser observados nas suas extremidades. Embora as grandes áreas estejam bem definidas quando se observa mais de perto a rede global, nota-se claramente um componente conectado bem distinto na rede interdisciplinar. Particularmente, existe uma forte sobreposição das grandes áreas, representada por uma mistura maior de cores na proximidade de cada uma

[2]Este esquema de classificação é organizado nos seguintes quatro níveis (veja http://bit.ly/1JM2j1k): *grande área* (e.g., Ciências Exatas e da Terra), *área* (por exemplo, Ciência da Computação), *subárea* (e.g., Teoria da Computação) e *especialidade* (e.g., Linguagens Formais e Autômatos).

delas, sendo que cada cor corresponde a uma grande área.

Conforme mencionado anteriormente, vários estudos tendem a se concentrar apenas no aspecto global, dando a mesma importância para as arestas que conectam partes específicas de uma rede. Em contraste, neste artigo enfatizamos a importância de minerar mais profundamente padrões interdisciplinares para explicar a força dessas ligações internas, isto é, possibilitar uma melhor compreensão da evolução da ciência [Mooney et al. 2013]. Neste contexto, este artigo tem como objetivo apresentar uma análise preliminar da importância das colaborações interdisciplinares na formação de redes de coautoria, tendo como escopo a comunidade científica brasileira.

O restante deste artigo está organizado da seguinte forma. A Seção 2 descreve como foi construído o conjunto de dados usado neste trabalho, a Seção 3 apresenta os resultados análise realizada e a Seção 4 apresenta algumas considerações finais.

2. Metodologia

Para gerar os dados necessários para a nossa análise, utilizamos um conjunto de currículos vitae (versões XML) de 264.731 pesquisadores brasileiros com título de doutor, coletado da Plataforma Lattes em junho de 2017. Posteriormente, os respectivos documentos XML foram transformados em arquivos CSV contendo apenas os campos necessários para identificar as colaborações científicas entre os pesquisadores, reduzindo assim a quantidade de dados a serem tratados. Além disso, também foram desconsiderados todos os currículos cujos pesquisadores não haviam indicado a grande área mais próxima de suas atividades de pesquisa, reduzindo assim o número total de currículos para 245.583.

Para construção e caracterização das redes necessárias ao nosso estudo, aplicamos a estratégia proposta por Dias & Moita [2015] para identificar as colaborações por comparação dos títulos das publicações encontradas nos currículos. Para fins das análises realizadas, foram consideradas apenas as coautorias identificadas a partir de artigos de periódicos, uma vez que este tipo de publicação é o mais comum na maioria das áreas de pesquisa. A distribuição final do número total de pesquisadores por grande área no conjunto de dados considerado é: Ciências Agrárias 24.957, Ciências Biológicas 33.756, Ciências da Saúde 42.392, Ciências Exatas e da Terra 36.752, Ciências Humanas 41.969, Ciências Sociais e Aplicadas 27.137, Engenharias 23.344, Linguística, Letras e Artes 15.276. Como pode ser visto, o número total de pesquisadores em cada grande área é bastante desigual e basicamente reflete a popularidade de suas respectivas áreas de atuação.

3. Análise das Redes

Nesta seção, analisamos as propriedades das redes brasileiras de coautoria global e interdisciplinar construídas a partir de dados da Plataforma Lattes. Além disso, descrevemos como essas redes estão interligadas, mostrando a proporção de colaborações em cada grande área. Por rede global, entende-se aquela cujos vértices correspondem aos 245.583 pesquisadores considerados e as arestas indicam as suas coautorias. Já a rede interdisciplinar possui o mesmo conjunto de vértices, sendo descartadas as arestas que correspondem a coautorias dentro da própria grande área dos pesquisadores. A Tabela 1 mostra os valores de algumas métricas usuais referentes às redes global e interdisciplinar. Primeiro, como se pode observar, 37,4% (758.392 de 2.027.511) das arestas correspondem a colaborações interdisciplinares. Este número é muito expressivo, já que se refere à rede de

uma comunidade científica completa. Além disso, o fato de que uma grande quantidade de arestas do componente gigante[3] da rede global ser mantida na rede interdisciplinar é bastante relevante, uma vez que 95,5% dessas arestas pertencem ao componente gigante. Isso reforça o quão próximas e importantes são tais colaborações no contexto acadêmico.

Tabela 1. Comparação entre as Redes Gobal e Interdisciplinar

Métrica	Global	Interdisciplinar
Número de Vértices	245.583	245.583
Número de Arestas	2.027.511	758.392
Grau Médio	9,9	5,7
Tamanho do Componente Gigante	194.021	145.255
Percentual de Vértices no Componente Gigante	73,3%	54,9%
Arestas no Componente Gigante	2.024.152	754.635
Densidade da Rede	5,79E-5	2,16E-5
Diâmetro da Rede	15	16
Caminho Mínimo Médio	5,4	6,7
Total de Componentes Isolados	65.362	113.509

Como esperado, o número de arestas e o grau médio dos nós diminuíram na rede interdisciplinar, enquanto que o número total de componentes isolados aumentou. Embora os diâmetros e os tamanhos do caminho mínimo médio das duas redes sejam próximos, nota-se que tais valores são ligeiramente maiores na rede interdisciplinar que possui menos arestas, tornando-a mais dispersa. Apesar disso, os tamanhos do caminho mínimo médio de ambas as redes são compatíveis com a propriedade do *mundo pequeno* [Milgram 1967]. Quanto à densidade das redes (a relação entre o número de arestas existentes e o número de arestas possíveis), a rede interdisciplinar tende a ser menos densa, uma vez que somente as colaborações interdisciplinares são consideradas.

Na literatura já existem alguns estudos acerca da topologia das redes de cada uma das grandes áreas definidas pelo CNPq. Em particular, Mena-Chalco et al. [2014] observaram padrões diferentes em todas elas, comprovando a hipótese por eles levantada que toda a rede brasileira de coautoria científica apresenta uma interdisciplinaridade natural. Por outro lado, Vanz & Stumpf [2012] também mostraram que as principais áreas de pesquisa no Brasil apresentam diferentes padrões de coautoria. Tais evidências mostram a necessidade de uma análise mais profunda dessas colaborações interdisciplinares, particularmente para entender a motivação dos pesquisadores envolvidos [Iglič et al. 2017].

Em particular, investigamos a proporção de colaborações interdisciplinares entre as oito grandes áreas. A Figura 2 apresenta o percentual de colaborações interdisciplinares para cada grande área. Por exemplo, Ciências Agrárias (primeiro gráfico) apresenta um percentual de colaborações com as demais grandes áreas de 58,38% (Ciências Biológicas), 11,62% (Ciências da Saúde) e assim por diante. Observe que as colaborações interdisciplinares mais relevantes envolvem as grandes áreas de Ciências Biológicas (barras pretas), Ciências da Saúde (barras cor-de-rosa), Ciências Exatas e da Terra (barras azuis) e Ciências Agrárias (barras vermelhas), enquanto que as colaborações envolvendo as grandes áreas de Linguística, Letras e Artes (barras cinzas) e Ciências Aplicadas e Sociais (barras laranjas) apresentam percentuais menos expressivos.

O maior percentual de colaborações interdisciplinares ocorre entre pesquisadores das Ciências da Saúde e das Ciências Biológicas (60,53%), seguido pelos percentuais entre os pesquisadores das Ciências Agrárias e Ciências Biológicas (58,38%), Linguística,

[3]Subgrafo máximo que inclui um caminho que conecta cada par de nós.

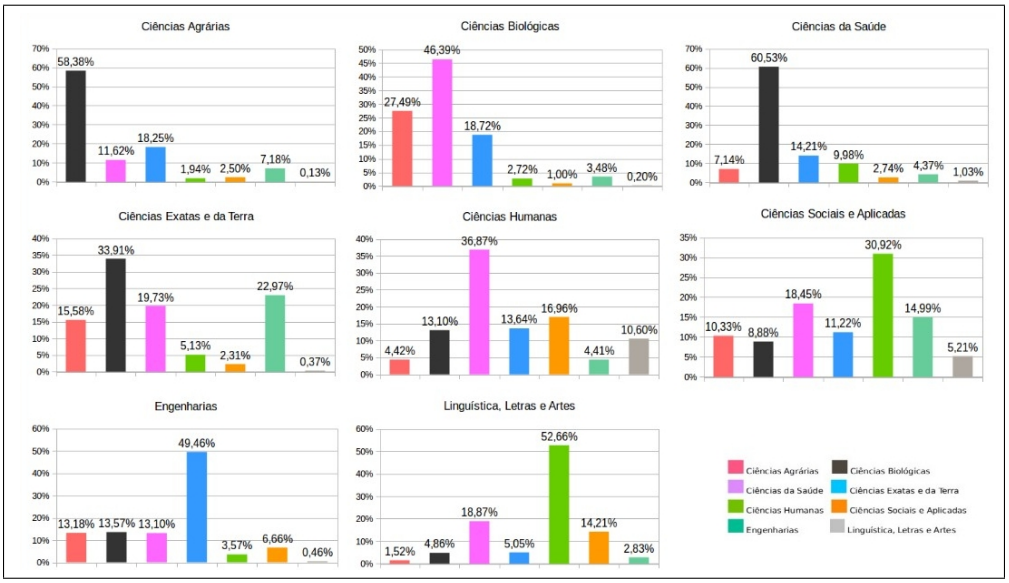


Figura 2. Percentual de Colaborações Interdisciplinares por Grande Área

Letras e Artes e Ciências Humanas (52,66%), Engenharias e Ciências Exatas e da Terra (49,46%), Ciências Biológicas e Ciências Agrárias (46,39%), e Ciências Humanas e Ciências Biológicas (36,87%). Nota-se que tais valores são bastante consistentes com a realidade dos pesquisadores envolvidos nessas colaborações, uma vez que essas grandes áreas possuem uma base teórica bastante próxima (por exemplo, há uma forte interação entre as grandes áreas de Engenharias e Ciências Exatas e da Terra). Vale ainda ressaltar que os pesquisadores da grande área de Ciências Biológicas são os que mais se envolvem em colaborações interdisciplinares. De fato, a grande área de Ciências Biológicas não só provê uma forte base teórica para a grande área de Ciências da Saúde, mas também serve como fundamentação para novas disciplinas como a Bioinformática.

Em relação à diversidade das colaborações, as Ciências Sociais e Aplicadas e as Ciências Humanas podem ser destacadas como as mais democráticas entre todas as grandes áreas, já que seus pesquisadores tendem a colaborar mais amplamente com colegas de outras áreas. Em contraste, os pesquisadores das Ciências Agrárias, Ciências da Saúde e Linguística, Letras e Artes tendem a concentrar as suas colaborações com colegas de uma grande área específica, no caso Ciências Biológicas (as duas primeiras) e Ciências Humanas, como mostra a Figura 2.

4. Considerações Finais

Neste artigo, procuramos salientar a importância das colaborações interdisciplinares nas redes de coautoria científica. Para isso analisamos as oito grandes áreas do conhecimento da rede de coautoria científica brasileira. De modo geral, nossos resultados mostram que que essa rede tem crescido, tornando-se mais interdisciplinar. Particularmente, mostramos que 37,4% de todas as colaborações são interdisciplinares. Além disso, nossos resultados mostram o quão integradas são as grandes áreas, enfatizando a força das colaborações entre elas. Considerando o potencial interdisciplinar das grandes áreas, Ciências Biológicas, a terceira mais populosa, destaca-se como a mais democrática. Porém, cabe ressaltar que o tamanho das grandes áreas não é fator decisivo para promover a interdisciplinaridade.

Como trabalhos futuros, pretendemos analisar como ocorre a formação daqueles docentes que transferem conhecimento para outras áreas através de novas colaborações [Mooney et al. 2013, Silva et al. 2016]. Uma outra perspectiva consiste em explicitar a força das colaborações [Brandão et al. 2016] para, por exemplo, destacar os principais atores (pesquisadores ou grupos de pesquisadores) que tornam uma rede mais integrada, como também para identificar aquelas colaborações que em função de sua força dão origem a novas comunidades acadêmicas [Leão et al. 2017]. Além disso, evidências obtidas nos levam a especular que há um compartilhamento do conhecimento interdisciplinar implícito entre as regiões do país. Assim, para melhor explicitar essas colaborações, seria importante identificar aquelas que demonstram a relevância das ligações externas para compartilhamento do conhecimento [Furtado et al. 2015].

Agradecimentos

Trabalho apoiado pelo projeto MASWeb (Proc. FAPEMIG/PRONEX APQ-01400-14) e por auxílios individuais de pesquisa concedidos pelo CNPq e CAPES. O primeiro autor agradece à UFV pela licença concedida para aperfeiçoamento profissional.

Referências

Brandão, M. A., Diniz, M. A., and Moro, M. M. (2016). Using Topological Properties to Measure the Strength of Co-authorship Ties. In *Proc. of the V Brazilian Workshop on Soc. Netw. Anal. and Mining*, pages 199–210, Porto Alegre, Brazil.

Dias, T. M. R. and Moita, G. F. (2015). A method for the identification of collaboration in large scientific databases. *Em questão*, 21(2):140–161.

Furtado, C. A., Davis, Jr., C. A., Gonçalves, M. A., and de Almeida, J. M. (2015). A Spatiotemporal Analysis of Brazilian Science from the Perspective of Researchers' Career Trajectories. *PLOS ONE*, 10(10):1–28.

Iglič, H., Doreian, P., Kronegger, L., and Ferligoj, A. (2017). With whom do researchers collaborate and why? *Scientometrics*, 112(1):153–174.

Kato, M. and Ando, A. (2013). The relationship between research performance and international collaboration in Chemistry. *Scientometrics*, 97(3):535–553.

Lane, J. (2010). Let's make science metrics more scientific. *Nature*, 464(7288):488–489.

Leão, J. C., Brandão, M. A., de Melo, P. O. V., and Laender, A. H. F. (2017). Classificação de Relações Sociais para Melhorar a Detecção de Comunidades. In *Proc. of the VI Brazilian Workshop on Soc. Netw. Anal. and Mining*, pages 647–657, São Paulo, Brazil.

Mena-Chalco, J. P., Digiampietri, L. A., Lopes, F. M., and Junior, R. M. C. (2014). Brazilian bibliometric coauthorship networks. *JASIST*, 65(7):1424–1445.

Milgram, S. (1967). The small-world problem. *Psychology Today*, 1(1).

Mooney, H. A., Duraiappah, A., and Larigauderie, A. (2013). Evolution of natural and social science interactions in global change research programs. *PNAS*, 110(1):3665–3672.

Silva, T. H. P., Laender, A. H. F., Davis Jr., C. A., da Silva, A. P. C., and Moro, M. M. (2016). The Impact of Academic Mobility on the Quality of Graduate Programs. *D-Lib Magazine*, 22(9/10).

Sentiment Analysis on Brazilian News Broadcast Data

Alexandre Martins da Cunha[1], Isabela Santos[1], Daniel Pedrosa[1], Francis F. Steen[2], Mark Turner[3], Maira Avelar[4], Lilian Ferrari[5], Gustavo Paiva Guedes[1]

[1]CEFET/RJ - Centro Federal de Educação Tecnológica Celso Suckow da Fonseca
Av. Maracanã, 229 - Rio de Janeiro - RJ - Brasil.

[2]University of California, Los Angeles - Los Angeles, CA 90095, USA

[3]Case Western Reserve University - 10900 Euclid Ave, Cleveland, OH 44106, USA

[4]UESB - Estrada Itapetinga/Itambé - s/n, Itapetinga - BA, 45700-000

[5]UFRJ - Universidade Federal do Rio de Janeiro
Av. Horácio de Macedo, 2151 - CEP 21941-917 - Rio de Janeiro - RJ - Brasil

alexandre.cunha@eic.cefet-rj.br, bela.rsantos@gmail.com,
daniel.souzapedroza@gmail.com,steen@comm.ucla.edu,
mark.turner@case.edu, mairavelar@uesb.edu.br,
lilianferrari@uol.com.br, gustavo.guedes@cefet-rj.br

Abstract. *This work aims to compare the occurrence of negative emotion words in Brazilian broadcast news JN and JR, and also analyzes Twitter posts related to them. We use the Brazilian Portuguese version of LIWC dictionary, which is a Sentiment Analysis software. The results indicate that both JN and JR tend to use negative emotion words, but in JR this tendency is greater. Nevertheless, Twitter posts direct more criticisms towards JN than JR.*

1. Introduction

The world is constantly changing and the relationship between public and press is also evolving over time. Information and communication technology (ICT) allowed internet users to access, transmit and manipulate information. In fact, information propagation has been growing in the free network environment in ways that would be unimaginable several years ago [Ma et al., 2013]. For example, online social networks (OSNs) provide a new way of spreading information which is far beyond "word-of-mouth" [Ma et al., 2008]. Recent studies, however, indicate that television still remains the main communication medium [Gamonar, 2015]. In Brazil, 95% percent of the people watch television regularly and 74% watch it every day [Vizeu, 2016], and telejournalism is the most important communication medium in the country [Vizeu, 2016].

Given that it has been attested that criminal-and violence-related news can increase the audience of a news channel [Junqueira et al., 2013], the preference for news with negative focus in Brazilian telejournalism comes as no surprise [Vaz and Medeiros, 2014]. In order to evaluate the exhibition of negative news in Brazilian telejournalism, this work analyzes Closed Captions (CC) from Brazilian news broadcasts. Since Sentiment Analysis (SA) has become a hot topic in recent years [Pang et al., 2008], we analyzed the CC with a SA software named Linguistic Inquiry and Word Count (LIWC) [Pennebaker et al.,

2001]. Finally, we applied *Cohen's effect size* to analyze psycolinguistic differences of word occurrences in the CC.

The data set analyzed in this work consists of CC from two of the most-watched news program in Brazilian television (i.e., Jornal Nacional (JN) and Jornal da Record (JR)) extracted from 05/22/2017 and 12/06/2017. The first one is exhibited by rede Globo® and the latter is exhibited by Rede Record®. Closed Captions were captured by the Brazilian Red Hen Capture Station[1], with the facilities of the Distributed Little Red Hen Lab.

Experimental results indicate that there are many negative news reports in both JN and JR. However, JR has more negative words (indicating anger, sadness, death) than JN. After obtaining these findings, we analyzed Twitter posts related to JN and JR, based on the assumption that there would be several criticisms of the negative content of Brazilian broadcast news [Ribeiro, 2015]. To investigate the extent of negative feelings about JN and JR, we checked Twitter posts for a range of negative keywords referring to JN and JR. For example, the negative word "desgraça" (disgrace) appeared in some collected twitter posts, as in "JN só mostra desgraça" and "JR só mostra desgraça", which roughly mean that news broadcasted in these channels report on bad things. Preliminary results indicate that there are more criticisms of JN than JR.

The remainder of this paper is organized as follows. Section 2 presents related work on Brazilian broadcast news analysis. Section 3 outlines main LIWC characteristics. Section 4 discusses data set acquisition. Section 5 explains the methodology used to analyze the data. Section 6 summarizes our experiments. We conclude in Section 7.

2. Related Word

Given the nature of our work, we focus on publications directly related to news broadcasts. In Junqueira et al. [2013], the authors analyze the representation of urban violence in telejournalism. They analyze content and discourse in urban violence news, using two newsletters: *Bom dia Goiás*, from tv Anhanguera, a subsidiary of Rede Globo® and *Direto da redação*, from Record Goiás, a subsidiary of Record®. They conclude that both newspapers violate the principle of human dignity; however, the newsletter *Direto da redação* produces more serious violations of citizenship. The authors conclude that the main goal of *Direto da redação* is to transmit rapid, superficial and decontextualized information.

Vaz and Medeiros [2014] focus on the news production process, which involves events selection. The aim is to investigate the negative aspect of facts converted into news. Their work performs an analysis in news extracted from *UOL* web site, *Folha de S.Paulo* newspaper and the *Jornal Nacional* broadcast news. The work concludes that the abusive use of negative aspects of facts should be questioned.

In Almeida [2017], the authors criticize the way telejournalism generally reports and discloses criminal trials. They emphasize that in some news broadcasts, journalists dramatize human pain in scenes of dead people, looking for a guilty party against whom society can turn. They conclude by pointing out that in many cases, telejournalists extrapolate their obligation to inform, and often exhibit only violence and death.

[1]http://www.redhenlab.org/

Our study also analyzes negative aspects in news. However, we focus on a direct comparison of CCs from two of the most watched news program in Brazilian television, exhibited by Rede Globo® and Record®. Our work also differs from those mentioned above by using a reputable sentiment analysis software. We also analyze twitter messages in order to evaluate audience's feelings about these two news program.

3. LIWC

LIWC (Linguistic Inquiry and Word Count) is a software that can obtain narrative, structural, emotional and intellectual elements from written texts [Pennebaker et al., 2001]. In the early days, LIWC was used to improve mental health treatment through analysis of patient narratives about negative experiences [Pennebaker et al., 2003]. In the course of time, new applications have been proposed, such as the transcription of daily narratives [Pennebaker et al., 2003].

LIWC encompasses a large collection of entries, in which each entry is associated with one or more categories. These categories are related to linguistic processes (e.g., pronoun, verb, article) and psychological processes (e.g., anxiety, negative emotion, swear words) [Pennebaker et al., 2003]. The Brazilian Portuguese version of LIWC is based on the 2007 LIWC English dictionary and has 127,149 entries divided into 64 categories [Balage Filho et al., 2013].

4. Data Set

In order to investigate the exhibition of negative news in Brazilian telejournalism, this work analyzes text from news broadcasts. To achieve this goal, we created a data set named `NewsBroadcast-PT-2017`, which comprises collected Closed Captions from JN and JR in the period between 05/22/2017 and 10/14/2017. The data were collected with the facilities of the Distributed Little RedHen Lab.

The data set comprehends CCs from 170 news programs. It contains 85 CCs from JN programs with duration between 44 and 60 minutes. On the other hand, it contains also 85 CCs from JR programs with duration between 45 and 55 minutes. It is important to mention that each CC from JN has, for the same date, the correspondent JR CC.

5. Methodology

This section describes the methodology adopted to explore the sentiment analysis in CCs. As described in Section 1, there is a preference for news with negative focus in Brazilian telejournalism, since news about crime and violence can increase the audience. Thus, we used the `NewsBroadcast-PT-2017` data set to analyze the use of the following LIWC categories: negation words (*negate*), negative emotion words (*negemo*) and positive emotion words (*posemo*). It is important to note that *negemo* is divided into three subcategories: anxiety words (*anx*), anger words (*anger*) and sadness words (*sad*).

In this scenario, we first created a frequency vector using the above mentioned categories for each CC, as illustrated in Figure 1. As an example, in this vector, the frequency of $negate$ words is 53.

Then, in order to investigate psycholinguistic differences between news from JN and JR, we conducted an effect size analysis using Cohen's d statistic [Rosnow and Rosenthal, 1996], as shown in Eq. 1. In this equation, i represents the LIWC category of

$negate$	$posemo$	$negemo$	anx	$anger$	sad
53	20	14	4	7	3

Figura 1. Frequency vector representing a CC with 6 dimensions.

examination ($0 \leq i < 6$), $\bar{X}^i_{JN}$ and $\bar{X}^i_{JR}$ simple averages of the ith component of the frequency vectors of JN and JR respectively. Likewise, SD^i_{JN} e SD^i_{JR} are the standard deviations of words for each category i.

$$d_i = \frac{\bar{X}^i_{JN} - \bar{X}^i_{JR}}{\sqrt{((SD^i_{JN})^2 + (SD^i_{JR})^2)/2}} \tag{1}$$

The intuition for the interpretation of this equation is described as follows. Positive values of d_i indicate that JN used more words in category i than those of JR. Negative values denote greater use by JR than JN. Cohen proposed the interpretation of $d = 0.20$ as small effects, $d = 0.50$ as medium effects and $d = 0.80$ as large effects [Cohen, 1988].

6. Results

In order to analyze the differences between CCs from JN and JR, Table 1 presents the results of the calculation of Cohen's effect size (Eq. 1) for each of the following LIWC categories: *negate*, *negemo*, *posemo*, *negemo*, *anx*, *anger* and *sad*. For these groups, mean values ($\bar{X}$) and standard deviation (SD) were also presented. Positive values of d indicate that CCs from JN presents more words than JR in the correspondent category. Analogously, negative values indicate that CCs from JR contains more words related to the correspondent category.

Tabela 1. Positive values indicate the JN used the category more than the JR. The size of the effect is represented by d, according to Eq. 1) $\bar{X}$ and SD represent the mean and standard deviation, respectively.

		JN		JR		
Category	Example	$\bar{X}$	SD	$\bar{X}$	SD	(d)
negate	not, never	0,86	0,30	0,62	0,27	0,82
posemo	love, best	3,36	0,39	3,07	0,35	0,80
negemo	afraid, cry	1,86	0,29	1,91	0,33	-0,18
anger	hate, raping	0,72	0,17	0,77	0,19	-0,27
sad	crying, sad	0,56	0,11	0,60	0,15	-0,27
death	kill, war	0,19	0,09	0,28	0,12	-0,85

It can be noted that JN has a tendency to use more negation words (*negative*) in news. On the other hand, it also shows that JN CCs present more words related to positive emotion (*posemo*). In contrast, JR presents more negative emotion words (*negemo*), anger words (*anger*, sadness words (*sad*) and death words (*death*). It is important to note that death-related words present a large effect, indicating that JR exhibits much more death-related words than JN.

We have selected some negative keywords to analyze negative feelings about JN and JR in Twitter posts. We combined these keywords with "Jornal Nacional", "Jornal da Record", and "só" (only). We also used "*" to achieve better results. Tweets were collected between January 2010 and February 2018. Table 2 shows results for queries "Jornal Nacional só * desgraça" and "Jornal da Record só * desgraça", which means roughly that these channels report only news with content related to "disgrace". Results show 156 twitter messages in the JN search and 46 in the JR search.

Table 3 shows the results combining the combination of four negative keywords with "Jornal Nacional" and "Jornal da Record". The keywords are: "notícia ruim" (bad news), "tragédia" (tragedy), "morte" (death) and "violência" (violency). Results shows 155 twitter messages in the JN search and 28 in the JR search.

Tabela 2. Twitter search query including keyword "desgraça".

News program	Twitter query	Twitter messages
JN	"Jornal Nacional só * desgraça"	**156**
JR	"Jornal da Record só * desgraça"	46

Tabela 3. Twitter search query including keywords "notícia", "ruim", "tragédia", "morte" and "violência".

News program	Twitter query	Twitter messages
JN	"jornal nacional só *" AND (notícia ruim OR tragédia OR morte OR violência)	**155**
JR	"jornal da record só * " AND (notícia ruim OR tragédia OR morte OR violência)	28

Preliminary results indicate that there is more criticism about JN than JR. In all, twitter users posted 311 messages criticizing the JN and 74 criticizing the JR. The number of messages criticizing JN is more than 400 percent greater than messages criticizing JR. These results can be explained by research which indicates that JN has an audience four times larger than JR [Ferreira, 2007].

7. Conclusion and Future Work

In this work, we used the LIWC dictionary to analyze words from news broadcasts. We analyzed Closed Captions from the two most watched Brazilian news broadcasts (i.e., JN and JR). Results indicate that JN CCs contain more negative words and positive emotions. In contrast, JR CCs contains more words related to negative emotions, anger, sadness and death.

We also analyzed Twitter posts criticizing JN and JR. Results reveal that there is much more criticism of JN than of JR. This can be explained by research which indicates that JN has a much larger audience. Thus, although news about crime and violence can increase the audience of a news channel, we confirmed that there are several criticisms of the negative content of Brazilian broadcast news.

Our future work will take advantage of the possibility of including other affective categories. Moreover, we intend to develop multimodal studies, correlating affective characteristics of words with facial analysis, in order to verify the emotion related to the facial expression in the process. We also expect to extend the analysis to other broadcast news channels.

Referências

Luanny Galvão Almeida. O descompasso entre a realidade midiática e a realidade processual e suas implicações para o julgamento criminal justo. *Revista Transgressões*, 5 (2):82–103, 2017.

Pedro Paulo Balage Filho, Thiago Pardo, and Sandra Aluísio. An evaluation of the Brazilian Portuguese LIWC dictionary for sentiment analysis. In Sandra Maria Aluísio and Valéria Delisandra Feltrim, editors, *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology (STIL)*, pages 215–219, Fortaleza-CE, Brazil, 21–23 October 2013. Sociedade Brasileira de Computação.

Jacob Cohen. Statistical power analysis for the behavioral sciences 2nd edn, 1988.

Fernanda Vasques Ferreira. As representações dos indivíduos anônimos no telejornalismo brasileiro: um estudo comparativo entre o jornal nacional e o jornal da record. 2007.

Flavia Daniele Oliveira Gamonar. Planejamento e prototipagem de uma rede social de gastronomia convergente com programas de tv e mídias sociais. 2015.

Juliana Junqueira et al. Telejornalismo e violência urbana: cidadania nas notícias sobre criminalidade: realidade possível? 2013.

Hao Ma, Haixuan Yang, Michael R Lyu, and Irwin King. Mining social networks using heat diffusion processes for marketing candidates selection. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 233–242. ACM, 2008.

Li Ma, Zhong Tian Jia, Hai Yan Sun, and Chuan Yu. An improved model of the internet public opinion spreading on mass emergencies. In *Applied Mechanics and Materials*, volume 433, pages 1760–1764. Trans Tech Publ, 2013.

Bo Pang, Lillian Lee, et al. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2):1–135, 2008.

James W Pennebaker, Martha E Francis, and Roger J Booth. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001, 2001.

James W. Pennebaker, Matthias R. Mehl, and Kate G. Niederhoffer. Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology*, 54(1): 547–577, 2003.

Jéssica Conceição Ribeiro. Sentidos sobre o jornalismo no twitter: uma análise do discurso dos integrantes sobre o jornalismo contemporâneo. 2015.

Ralph L Rosnow and Robert Rosenthal. Computing contrasts, effect sizes, and counternulls on other people's published data: General procedures for research consumers. *Psychological Methods*, 1(4):331, 1996.

Élida Mattos Vaz and Theresa Medeiros. Jornalismo e jornalistas na berlinda: Uma análise da abordagem negativa da imprensa e sua relação com a crise contemporânea da imprensa. In *XXXVII Congresso Brasileiro de Ciências da Comunicação*. Intercom, 2014.

Alfredo Pereira Vizeu. 65 anos de televisão: o conhecimento do telejornalismo e a função pedagógica/65 years of the television: the knowledge of the telejournalism and the pedagogical function. *Revista FAMECOS*, 23(3):1–17, 2016.

Uso de mineração de textos para a identificação de postagens com informações de localização

Silas F. Moreira[1], **Maruscia Baklizky**[1], **Luciano A. Digiampietri**[1]

[1]Escola de Artes, Ciências e Humanidades – Universidade de São Paulo (USP)

silas.moreira@usp.br, maruscia.baklizky@usp.br, digiampietri@usp.br

Abstract. *Information from social networks is very useful for different tasks. It is possible, for example, to quickly identify evidence of an epidemic, voting trends for an election, or user satisfaction with services or products. Several text mining related tasks that use information from social networks can take advantage of geographic information about the user's location. However, the majority of posts in social networks do not have an explicit reference to their geolocation. The goal of this paper is to compare the efficiency of text mining techniques to identify whether or not a post contains information about a location.*

Resumo. *As informações contidas nas redes sociais são cada vez mais úteis para diferentes tarefas. É possível, por exemplo, identificar de maneira bastante rápida indícios de uma epidemia, tendências de voto para uma eleição, ou a satisfação de usuários em relação a serviços ou produtos. Diversas tarefas de mineração de textos em redes sociais conseguem tirar vantagem de informações geográficas sobre a localização do usuário. Porém, a grande maioria das postagens em redes sociais não possui uma referência explícita à sua geolocalização. O objetivo deste trabalho é comparar a eficácia de algumas técnicas de mineração de textos para identificar se uma postagem contém ou não informações sobre uma localidade.*

1. Introdução

As redes sociais conquistaram uma alta popularidade atualmente, tornando-se um grande repositório de informações sobre seus usuários, por meio de seus perfis e de suas publicações. Por conta disso, a análise dos conteúdos provenientes dessas redes tem recebido cada vez mais atenção e tem possibilitado uma série de estudos envolvendo o comportamento de seus usuários.

Informações sobre a geolocalização de usuários são importantes para diversas tarefas, como recomendação de locais de interesse, de rotas e de produtos. Para algumas aplicações, como o traçado de rotas em tempo real, essa informação é fundamental. Por outro lado, outras aplicações, como sistemas de recomendação de produtos, podem ser mais efetivas com o uso da informação de localização, porém, podem funcionar adequadamente sem ela.

Com o uso de informações de redes sociais é possível, por exemplo, detectar de maneira antecipada locais com alta incidência de uma doença com base nas informações contidas nas trocas de mensagens. Porém, na maioria dessas redes, quando a informação de geolocalização não é requerida, observa-se que a frequência do uso de marcadores

precisos de localização (*geo-tags*) é bastante baixa. Assim, é interessante que sejam desenvolvidos sistemas computacionais que possam, com base nas informações postadas em redes sociais, inferir a geolocalização dessas postagens.

Nesse contexto, a mineração de texto apresenta métodos e ferramentas que podem auxiliar no processo de identificação de informações de localização [Berry and Castellanos 2004, Feldman and Sanger 2007, Manning et al. 2008, Aggarwal and Zhai 2012]. O presente artigo está focado na identificação se uma postagem possui ou não informação de localidade e nele foram comparadas diferentes estratégias, considerando-se as medidas de acurácia e precisão.

As três hipóteses que nortearam o desenvolvimento desta pesquisa são: (i) técnicas simples (baseadas apenas em frequência de palavras) podem alcançar resultados satisfatórios na indicação de que uma postagem possui ou não referência a uma localidade; (ii) algoritmos de classificação podem atingir resultados superiores aos das técnicas simples; (iii) a combinação dessas duas estratégias pode aumentar a acurácia dos resultados.

2. Materiais e Métodos

Neste trabalho foram utilizadas postagens públicas da rede social *Facebook*. Tais postagens não eram explicitamente geolocalizadas e foram manualmente classificadas com a indicação de contendo ou não alguma informação de localização. Foram consideradas informações de localização a citação explícita do nome de um determinado lugar, seja comercial (por exemplo, lojas e hotéis) ou político (por exemplo, cidades e países). Erros de grafia foram desconsiderados para a classificação, de forma que mesmo com a apresentação de erros de escrita nos nomes das localizações, a postagem ainda seria classificada como contendo informação de localização.

A tarefa completa de geolocalizar compreende gerar como resultado uma referência a uma localização no globo terrestre. Porém, o escopo do presente artigo restringe-se a identificar se uma determinada postagem contém ou não uma informação de localização, não tratando da geolocalização propriamente dita e dos problemas de ambiguidade relacionados.

Para testes e validação, foram utilizadas 187 postagens públicas de páginas do *Facebook*, sendo que 97 foram selecionadas da página "Loucos por Viagens", que apresenta informações e críticas sobre as diversas viagens de seus autores. As postagens públicas foram obtidas com o auxílio da API disponibilizada pela própria rede social e classificadas manualmente, indicando se contêm ou não alguma informação de localização.

A partir da classificação, utilizou-se a medida estatística TF-IDF [Jones 1972], que mede a frequência de cada palavra em uma postagem, em relação à sua frequência em todo o *corpus* de mensagens com ou sem informações de localidade. Com essas informações, tornou-se possível ordenar as mensagens de acordo com a soma das medidas de TF-IDF de cada palavra que compõe a mensagem. Neste trabalho, o cálculo da pontuação TF-IDF de uma postagem em relação a um *corpus* foi realizado de acordo com a equação 1, sendo t cada um dos termos pertencentes a postagem P, TF_t a frequência relativa do termo t na postagem P, e IDF_t o inverso da frequência relativa do termo t no *corpus* utilizado.

$$\sum_{t \in P} log\left(1 + TF_t * IDF_t\right) \tag{1}$$

A pontuação TF-IDF pode ser utilizada de diferentes formas. Duas pontuações para cada postagem foram calculadas e estas foram analisadas de três formas diferentes. Ambas as pontuações utilizaram a equação 1, porém, para uma o *corpus* utilizado foi formado apenas pelas postagens que contêm indicação de localidade, já para a outra, o *corpus* foi composto apenas por postagens que não continham indicação de localidade.

Com base nas duas pontuações atribuídas para cada postagem, três estratégias de classificação foram utilizadas. A primeira, denominada de *Max*, atribuiu para a respectiva postagem a classe que atingiu maior pontuação TF-IDF, ou seja, se o valor calculado para o *corpus* de mensagens que indicam localidade for maior, então a postagem será marcada como "indica localidade"; caso contrário, será classificada como "não indica localidade". A segunda estratégia, denominada de *Limiar*, utiliza um limiar considerando apenas a pontuação em relação ao *corpus* de mensagens que indicam localidade. Essa nota pode ser entendida como o pertencimento ou a adequação de uma postagem em relação às mensagens que indicam localidade. Para os experimentos, o limiar utilizado foi aquele que maximizou a acurácia. Por fim, analisou-se a relação entre a primeira pontuação obtida pela postagem e a soma das pontuações recebidas pela postagem. Esta estratégia foi denominada de *Relação* e a partir do valor da relação foi estabelecido um limiar para a classificação: são classificadas como "indicam localidade" as postagens cuja pontuação recebida no *corpus* que indica localidade dividida pela soma das duas pontuações foi maior do que um dado limiar.

Antes do computo das pontuações, foram realizadas combinações de diferentes estratégias de pré-processamento de forma a identificar as variações nos resultados pelo uso ou não destas estratégias. As estratégias utilizadas foram: remoção de *stop-words*, isto é, palavras que, a princípio, não agregam significado aos textos; e a radicalização (ou *stemming*). Adicionalmente, um filtro alternativo de palavras foi utilizado independente de um dicionário de *stop-words* ou de um algoritmo de radicalização: a exclusão de palavras pequenas (deixando apenas palavras com, pelo menos, um certo número de letras). O número mínimo de letras utilizado variou de 1 a 10.

A ideia de se medir a importância de palavras utilizando TF-IDF foi estendida para a importância de n-gramas. Assim, além do cálculo das pontuações para termos individuais (unigramas), também foram realizados os cálculos para bigramas e trigramas.

Uma abordagem alternativa para a representação das postagens, bastante utilizada na mineração de textos, é a de *bag-of-words*, na qual uma postagem é representada como um conjunto não ordenado de palavras. Com base nessa representação, diversas medidas de comparação podem ser calculadas entre diferentes postagens (por exemplo, a distância cosseno) ou outras estratégias, como o uso de classificadores, podem ser utilizadas. Esta representação foi utilizada como entrada para classificadores que tinham como objetivo identificar se uma postagem continha ou não uma informação de localidade. Assim, a classificação realizada foi binária. Para esta representação, duas estratégias de pré-processamento foram testadas: o uso ou não de radicalização e a seleção ou não das palavras (ou atributos) mais relevantes. Para esta abordagem optou-se pela utilização de um seletor de atributos, com o objetivo de identificar a importância relativa das palavras em relação às classes e selecionar o subconjunto "mais informativo" destes atributos. Foi utilizado um seletor baseado na correlação entre atributos, chamado *Correlation-based feature Subset Selection (CfsSubsetEval)* [Hall 2000].

Três classificadores foram utilizados: dois baseados no Teorema de Bayes (Rede Bayesiana - *Bayes Net* e *Naïve Bayes*) [John and Langley 1995] e um meta classificador que utiliza análise de componentes principais para a projeção dos atributos em um espaço no qual a variância está maximizada nas primeiras dimensões (*Rotation Forest*) [Rodriguez et al. 2006]. Todos os testes utilizaram validação cruzada com dez subconjuntos (*10-fold-cross-validation*). As implementações utilizadas do seletor de atributos e dos classificadores foram aquelas disponíveis no arcabouço Weka[1].

3. Resultados e Discussão

A combinação de diferentes estratégias de pré-processamento, com o uso de três abordagens diferentes para a classificação das postagens com base nas pontuações TF-IDF e o uso de unigramas, bigramas e trigramas produziu 360 resultados diferentes. A tabela 1 apresenta os resultados de acurácia para cada uma das abordagens e/ou combinações de estratégias de pré-processamento.

Tabela 1. Acurácia do uso de TF-IDF

		com radicalização									sem radicalização								
	Letras	Unigrama			Bigrama			Trigrama			Unigrama			Bigrama			Trigrama		
		Max	Limiar	Relação	Max	Limiar	Relação	Max	Limiar	Relação	Max	Limiar	Relação	Max	Limiar	Relação	Max	Limiar	Relação
sem remoção de stopwords	1	74.87%	83.96%	84.49%	81.82%	87.17%	83.96%	73.80%	74.87%	74.87%	72.73%	82.89%	84.49%	79.68%	86.63%	82.35%	75.40%	76.47%	76.47%
	2	75.40%	85.56%	84.49%	79.68%	82.35%	81.28%	70.05%	71.12%	70.59%	73.26%	85.03%	83.96%	81.28%	83.96%	81.82%	71.12%	72.19%	71.66%
	3	71.12%	84.49%	83.96%	79.68%	83.42%	82.35%	70.05%	70.05%	70.05%	74.33%	84.49%	80.75%	71.66%	73.80%	72.73%	64.71%	64.71%	64.71%
	4	76.47%	86.63%	85.56%	74.33%	75.94%	75.40%	66.84%	67.38%	67.38%	78.61%	86.63%	83.42%	69.52%	69.52%	69.52%	64.71%	64.71%	64.71%
	5	78.07%	85.56%	83.96%	73.26%	73.80%	73.80%	65.78%	65.78%	65.78%	85.03%	**87.70%**	85.03%	65.78%	65.78%	65.78%	64.71%	64.71%	64.71%
	6	83.42%	86.10%	82.35%	68.98%	69.52%	69.52%	65.78%	65.78%	65.78%	81.28%	84.49%	83.96%	66.84%	66.84%	66.84%	64.71%	64.71%	64.71%
	7	78.61%	82.89%	80.75%	66.84%	66.84%	66.84%	64.71%	64.71%	64.71%	77.01%	79.14%	78.07%	64.71%	64.71%	64.71%	64.71%	64.71%	64.71%
	8	79.68%	80.75%	80.75%	64.71%	64.71%	64.71%	64.71%	64.71%	64.71%	74.33%	75.40%	74.87%	64.71%	64.71%	64.71%	64.71%	64.71%	64.71%
	9	76.47%	77.01%	77.01%	64.71%	64.71%	64.71%	64.71%	64.71%	64.71%	70.59%	70.59%	70.59%	64.71%	64.71%	64.71%	64.71%	64.71%	64.71%
	10	72.19%	72.19%	72.19%	64.71%	64.71%	64.71%	64.71%	64.71%	64.71%	67.91%	67.91%	67.91%	64.71%	64.71%	64.71%	64.71%	64.71%	64.71%
com remoção de stopwords	1	80.75%	86.63%	84.49%	75.94%	78.07%	78.07%	67.91%	67.91%	67.91%	79.14%	85.03%	84.49%	80.21%	83.42%	83.42%	68.98%	68.98%	68.98%
	2	80.75%	87.17%	84.49%	76.47%	78.07%	78.07%	67.91%	67.91%	67.91%	78.61%	85.56%	84.49%	79.68%	81.82%	81.82%	70.05%	70.05%	70.05%
	3	80.75%	86.10%	83.42%	78.07%	79.68%	79.68%	67.91%	67.91%	67.91%	78.07%	85.03%	82.35%	71.66%	73.80%	72.73%	64.71%	64.71%	64.71%
	4	78.61%	87.17%	85.03%	72.19%	73.80%	73.26%	66.84%	67.38%	67.38%	78.07%	86.63%	83.42%	69.52%	69.52%	69.52%	64.71%	64.71%	64.71%
	5	81.82%	86.10%	83.96%	72.19%	72.73%	72.73%	65.78%	65.78%	65.78%	85.03%	**87.70%**	85.03%	65.78%	65.78%	65.78%	64.71%	64.71%	64.71%
	6	79.14%	85.03%	82.89%	68.98%	69.52%	69.52%	65.78%	65.78%	65.78%	81.28%	84.49%	83.96%	66.84%	66.84%	66.84%	64.71%	64.71%	64.71%
	7	77.01%	80.75%	79.14%	66.84%	66.84%	66.84%	64.71%	64.71%	64.71%	75.94%	79.14%	78.07%	64.71%	64.71%	64.71%	64.71%	64.71%	64.71%
	8	77.54%	78.61%	78.61%	64.71%	64.71%	64.71%	64.71%	64.71%	64.71%	74.33%	75.40%	74.87%	64.71%	64.71%	64.71%	64.71%	64.71%	64.71%
	9	74.33%	74.87%	74.87%	64.71%	64.71%	64.71%	64.71%	64.71%	64.71%	71.12%	71.12%	71.12%	64.71%	64.71%	64.71%	64.71%	64.71%	64.71%
	10	72.19%	72.19%	72.19%	64.71%	64.71%	64.71%	64.71%	64.71%	64.71%	68.45%	68.45%	68.45%	64.71%	64.71%	64.71%	64.71%	64.71%	64.71%

Duas células da tabela contêm o valor mais alto de acurácia observado (87,7%). Os dois resultados ocorreram com o uso de unigramas e utilizando como estratégia de classificação um limiar sobre o valor obtido pela pontuação TF-IDF em relação às postagens classificadas como "indicam localidade" nos quais o texto foi pré-processado por um processo de radicalização e excluindo-se palavras com menos de quatro letras. A remoção ou não das *stop-words* não alterou o valor deste resultado.

Em termos de características individuais, inicialmente destaca-se o uso de unigramas. Essa estratégia se sobressaiu ao uso de bigramas e trigramas, o que pode ter ocorrido devido ao *corpus* não ser muito grande, o que pode causar esparsidade entre as postagens no uso de bigramas e, principalmente, de trigramas. Em termos das três estratégias de classificação aplicadas (*Max*, *Limiar* e *Relação*), o uso do *Limiar* foi a estratégia que, na média, apresentou os melhores resultados de acurácia, seguido pela estratégia *Relação*. Uma vantagem do uso de *Limiar* é a não necessidade de formação de um *corpus* negativo (de postagens que não indicam localidade), precisando apenas de um *corpus* positivo e do estabelecimento de um limiar que pode ser adaptado para aumentar a precisão do sistema ou aumentar a revocação, de acordo com as necessidades do usuário.

[1]Weka 3: https://www.cs.waikato.ac.nz/ml/weka/ , acessado em 22/03/2018

O uso da radicalização, na média, apresentou resultados bastante parecidos com o não uso de radicalização (apesar de ter repercutido no melhor resultado de acurácia geral). A remoção de *stop-words*, na média, também aumentou a acurácia do sistema. Já os filtros de uso de palavras com apenas um número mínimo de letras teve seus melhores resultados para unigramas com tamanho entre quatro e seis letras.

A tabela 2 apresenta os resultados da classificação utilizando como base a representação *bag-of-words*. Destaca-se que a radicalização apresentou melhores resultados em todos os testes (em comparação ao não uso de radicalização) exceto em uma execução específica do *Rotation Forest*. Já o uso da seleção de atributos apresentou resultados melhores do que o uso de todas as palavras em todos os casos testados.

Tabela 2. Acurácia da classificação utilizando a representação *bag-of-words*.

	todas as palavras		palavras selecionadas	
Classificador	sem uso de radicalização	com uso de radicalização	sem uso de radicalização	com uso de radicalização
Bayes Net	82,89%	84,49%	87,70%	88,24%
Naive Bayes	80,21%	81,82%	83,96%	88,24%
Rotation Forest	81,82%	77,01%	79,14%	82,35%

O uso do classificador baseado em Rede Bayesiana (*Bayes Net*) apresentou resultados melhores ou iguais aos demais classificadores testados. Destaca-se ainda que os melhores resultados (88,24% de acurácia geral) atingidos pelos classificadores *Bayes Net* e *Naïve Bayes* com o uso de radicalização e seleção de atributos são levemente melhores do que os melhores resultados atingidos com o uso de TF-IDF (87,7%).

Por fim, avaliou-se a combinação do uso da pontuação TF-IDF e a representação *bag-of-words*. Optou-se por realizar testes apenas combinando os melhores resultados obtidos em cada uma das estratégias: no caso da pontuação TF-IDF, foi aplicada a radicalização, removidas as *stop-words*, e consideradas apenas palavras com cinco ou mais letras; já para a representação *bag-of-words*, considerou-se apenas o uso da radicalização. A tabela 3 apresenta os seis resultados obtidos pelos classificadores.

Tabela 3. Acurácia da classificação combinando a pontuação TF-IDF com a representação *bag-of-words*.

Classificador	todas as palavras	palavras selecionadas
Bayes Net	85,03%	89,84%
Naive Bayes	83,96%	86,10%
Rotation Forest	88,24%	89,30%

Observa-se que, na maioria dos casos, a combinação das características apresentou resultados melhores do que os resultados equivalentes das medidas individuais. Destaca-se que o melhor resultado entre todos os experimentos foi o obtido pelo classificador *Bayes Net* combinando a pontuação TF-IDF com a representação *bag-of-words*, utilizando seleção de atributos. A acurácia geral atingida foi de 89,84%.

4. Conclusões e Trabalhos Futuros

Neste trabalho foram avaliadas diferentes estratégias de mineração de textos para identificar se uma postagem em rede social online possui ou não informação de localidade,

usando como estudo de caso postagens veiculadas no *Facebook*. Duas principais abordagens foram avaliadas: o uso de uma pontuação dada às postagens com base em valores TF-IDF e o uso de classificadores para avaliar as postagens representadas como *bag-of-words*. Adicionalmente, a combinação destas duas abordagens também foi testada.

As três hipóteses de pesquisa foram confirmadas pelos resultados obtidos: (i) é possível obter resultados satisfatórios para o problema tratado considerando-se apenas o uso de estratégias simples baseadas na importância relativa das palavras; (ii) o uso de classificadores aumentou a acurácia da solução; e (iii) a combinação das duas estratégias anteriores proporcionou resultados ainda mais acurados.

Destaca-se que o uso de um limiar sobre a pontuação TF-IDF a partir do qual as postagens são classificadas como "indicam localidade" além de ter atingido uma acurácia satisfatória ainda permite uma variação de valores por parte do usuário a fim de maximizar a precisão na identificação dos elementos positivos ou a revocação dos elementos positivos (de acordo com as necessidades da aplicação).

Como trabalhos futuros, pretende-se aplicar as mesmas abordagens testadas neste artigo a um conjunto de dados oriundo de outras redes sociais online de forma a verificar a variabilidade dos resultados em diferentes contextos.

Agradecimentos

Este trabalho foi parcialmente financiado pelo Programa de Educação Tutorial (PET) do Ministério da Educação, pela CAPES e pelo CNPq.

Referências

Aggarwal, C. C. and Zhai, C. (2012). *Mining text data*. Springer Science & Business Media.

Berry, M. W. and Castellanos, M. (2004). Survey of text mining. *Computing Reviews*, 45(9):548.

Feldman, R. and Sanger, J. (2007). *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge university press.

Hall, M. A. (2000). Correlation-based feature selection for discrete and numeric class machine learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, pages 359–366, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

John, G. H. and Langley, P. (1995). Estimating continuous distributions in bayesian classifiers. In *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence*, pages 338–345.

Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21.

Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA.

Rodriguez, J. J., Kuncheva, L. I., and Alonso, C. J. (2006). Rotation forest: A new classifier ensemble method. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(10):1619–1630.

Vamos falar sobre deficiência? Uma análise dos *Tweets* sobre este tema no Brasil

Fábio Manoel França Lobato[1], Marcelo da Silva[1], Krislen Coelho[2], Simone da Costa Silva[2], Fernando Pontes[2]

[1]Instituto de Engenharia e Geociências - Universidade Federal do Oeste Pará
Santarém, Pará, Brasil

[2]Núcleo de Teoria e Pesquisa do Comportamento - Universidade Federal do Pará
Belém, Pará, Brasil

fabio.lobato@ufopa.edu.br

{marcelo.t.pain, krisllenmayra2010, symon.ufpa, farp1304}@gmail.com

Resumo. *As deficiências estão mais relacionadas a um contexto social do que com condições médicas. No entanto, a falta de atenção ao assunto e de suporte social impactam negativamente na vida das pessoas com deficiência e seus familiares. Em um estudo exploratório percebeu-se uma tendência à depreciação dos deficientes em detrimento da construção de uma rede de apoio. À luz de tais fatos, este trabalho analisou postagens relacionados às deficiências mentais, físicas e intelectuais, a fim de identificar os principais temas discutidos e as circunstâncias de utilização.*

Abstract. *Disabilities are more related to a social context than to medical conditions. However, the lack of attention to the subject and social support negatively impacts the lives of people with disabilities and their relatives as well. In an exploratory study, it was perceived a trend towards the depreciation regarding this theme, instead of the construction of a support network. In light of these facts, this paper analyzed postings related to mental, physical and intellectual disabilities to identify the main topics discussed and the circumstances of use.*

1. Introdução

Estima-se que mais de 10% da população mundial sofre de algum tipo de deficiência física, mental ou intelectual, sendo que 80% dessas pessoas vivem em países em desenvolvimento [Setareh Forouzan et al. 2013]. De acordo com uma pesquisa conduzida pelo *Pew Research Center*, nestes mesmos países, observa-se que a maior parte dos usuários adultos de internet são engajados nas Redes Sociais Online como Facebook e Twitter [Poushter 2016]. Neste contexto, diversos estudos mostram que pessoas com deficiência e seus familiares estão cada vez mais presentes nas mídias digitais, compartilhando experiências e procurando aconselhamento de outras pessoas em condições semelhantes [Naslund et al. 2016].

As conexões geradas por meio da interação entre os pares online podem impactar positivamente no bem-estar dessas pessoas, aflorando o sentimento de pertencimento a grupos sociais coesos e afins, possibilitando também a troca de conhecimentos e estratégias para lidar com desafios do dia a dia [Naslund et al. 2016]. Apesar dos benefícios

dessas plataformas, fenômenos relacionados à segregação e depreciação envolvendo o tema deficiência são observados com frequência. Diversos trabalhos dedicaram-se ao estudo do discurso de ódio nesta temática. [Mckay et al. 2015] analisaram expressões categorizadas como símbolos depreciativos e que podem ter implicações para pessoas e suas experiências de vida, por exemplo, como expressões comuns na linguagem moderna, podemos citar: ”retardado”, que evoluiu de um diagnóstico médico à insulto; e também expressões ligadas ao suicídio, usada para estigmatizar e banalizar pessoas que experimentaram ideações ou experiências suicidas.

Não obstante, uma análise exploratória mostrou que este fenômeno é inconteste em diversos termos ligados à deficiência. Neste sentido percebeu-se uma lacuna na literatura quanto à análise do discurso sobre deficiências nas redes sociais, sendo este o foco do presente trabalho. Dessa forma, considerou-se salutar distinguir os padrões de postagens contendo discurso de ódio dos que são de caráter informativo, sobretudo identificando conteúdo (expressões) e discurso (construção da postagem) que atinja o público-alvo (pessoas com deficiência). Esta tarefa pode ser vista como a realização de detecção de comunidades e segmentação de mercado [Lobato et al. 2017]. Uma tarefa que antecede a segmentação é a extração e seleção de atributos [Silva et al. 2017], sendo que uma etapa opcional é a anotação de dados, o que permitiria a classificação direta das instâncias anotadas. Neste âmbito, [Magalhães et al. 2017] demonstra a identificação de comentários ofensivos em um portal de notícias usando aprendizado supervisionado, para tal, a etapa anterior foi a de anotação do conjunto de dados de acordo com categorias pré-definidas. Ainda, [Cirqueira and Vinícius 2017] propõe uma ferramenta baseada em gamificação para anotação de dados a serem utilizados em análise de sentimentos. E em [Cirqueira et al. 2017] os autores avaliam o desempenho de métodos de análise de sentimento para português brasileiro.

O problema de pesquisa surgiu da necessidade de se estabelecer uma interlocução por intermédio de mídias digitais entre um consórcio de pesquisa e grupos de suporte parental relacionados à pessoas com deficiência. O consórcio é formado por pesquisadores de diferentes áreas, contemplando enfermagem, terapia ocupacional, psicologia, sistemas de informação, ciência da computação e comunicação. Já os grupos de suporte são diversos e abrangem membros/familiares de deficientes, dos mais variados espectros. No entanto, alguns problemas foram observados no planejamento de conteúdo direcionados a um público tão diverso e na construção de estratégias de marketing efetivas para recrutamento de deficientes e familiares para participação em projetos de pesquisa.

Visando contornar esta lacuna, o presente estudo analisou postagens relacionados às deficiências mentais, físicas e de aprendizado, a fim de identificar os principais temas discutidos e as circunstâncias de utilização. Os resultados obtidos dão subsídio para a construção de estratégias de comunicação mais efetivas para se estabelecer uma interlocução com os deficientes e seus familiares.

O restante deste artigo encontra-se organizado como segue. A metodologia e os resultados obtidos são descritos nas Seções 2 e 3, respetivamente. Os *insights* obtidos, as dificuldades encontradas e sugestões de trabalhos futuros são abordadas na Seção 4.

2. Metodologia

Inicialmente um estudo exploratório visando identificar padrões de comunicação na temática deficiência para melhorar a interlocução consórcio-grupo foi conduzido. Os achados auxiliaram na construção do desenho da pesquisa, o qual pode ser resumido em dois passos, delimitação do escopo e análise dos dados.

A delimitação do escopo abarca a definição e/ou validação dos termos relacionados às deficiências mais trabalhadas pelo consórcio de pesquisa. Os termos foram definidos em reunião entre um analista de rede social, um profissional da comunicação e um psicólogo. A validação dos termos deu-se por meio de consultas usando a interface de busca do Twitter. Os seguintes termos foram adotados Deficiência, Deficiência Mental, Cego(s), Cegueira, Surdo(s), Surdez, Autismo, Autista(s), Deficiência Auditiva, Deficiência Física, Deficiência Intelectual, Amputação, Síndrome de Down, Deficiência Visual, Pessoa com Deficiência, Paralisia Cerebral, Lesão Medular, Espinha Bífida, Mielomeningocele e Baixa Visão. Com os termos validados deu-se início ao segundo passo da pesquisa, que consistiu na aquisição, análise e anotação dos dados obtidos por meio dos termos de pesquisa. A aquisição de dados nesta fase foi feita utilizando a *Search API*.

Os dados obtidos foram analisados visando identificar os tópicos mais comentados e também classificando-os de acordo com o seu discurso. Para a identificação dos tópicos mais frequentes utilizou-se a modelagem de tópicos, a qual se mostrou potencialmente similar à análise baseada em teoria fundamentada [Baumer et al. 2017]. O método de modelagem de tópico avalaiado foi o *Latent Semantic Allocation* (LSA). A classificação do discurso deu-se de forma manual, com três avaliadores para cada *Tweets* selecionado. Considerava-se de uma classe quando havia o consenso de pelo menos dois avaliadores, estratégia similar à adotada em [Magalhães et al. 2017]. Para tal, cinco categorias foram utilizadas para a classificação do discurso, a citar: “Informativo”, “Ofensivo e Pejorativo”, “Indignação e Denúncia”, “Relato de Experiências”, e “Outros”.

As classes foram propostas, discutidas e validadas consensualmente pelos autores. Na primeira classe, a “Informativo”, enquadraram-se os *Tweets* que possuíam como característica precípua a divulgação de informações sobre o tema. Os *Tweets* que utilizavam um dos termos sob escrutínio para denegrir, praguejar ou com conotação pejorativa foram incluídos na categoria “Ofensivo e Pejorativo”. A classe “Indignação e Denúncia” foi incorporada para definir os *Tweets* cujo conteúdo envolvia a denúncia de crimes relacionados ao tema deficiência ou reportando indignação frente à um evento/caso específico. A classe “Relato de Experiências” é auto-explicativa. Por fim, a categoria “Outras” representa *Tweets* que não se enquadram em nenhuma das classes propostas.

3. Resultados e Discussões

O intervalo de coleta foi compreendido de 6 à 13 de março, este período foi escolhido pois não incluía nenhuma data representativa qualquer Deficiência (*e.g.* Dia mundial do deficiente físico). Foram obtidos aproximadamente 200.000 *Tweets* sobre o tema. O número de amostras suficientes para o intervalo de confiança de 95% e taxa de erro de 5% para o total de dados coletados é de 377 postagens. Com o acréscimo dos 35% utilizados para aumento de confiabilidade, foram anotadas 509 amostras. O processo de anotação resultou em um conjunto de dados composto por 459 dos 509 *Tweets* totais, sendo que 5 haviam sido duplicados e em 36 não houve concordância entre os avaliadores.

A Figura 1 apresenta a distribuição das classes considerando os 459 *Tweets* com anotação consolidados.

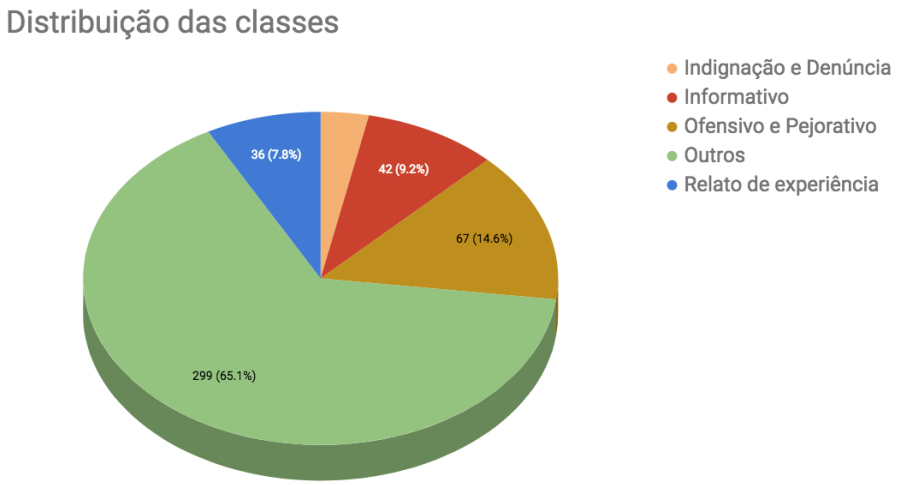


Figura 1. Distribuição das categorias para os *Tweets* classificados.

A maior parte dos *Tweets* foi classificada como "Outros" (299 *Tweets*), seguida de Ofensivo e Pejorativo (67), Informativo (42), Relato de experiência (36) e Indignação e Denúncia (15). Este fenômeno ocorreu justamente pela alta presença de ruído ligado à determinados termos utilizados como expressão popular, tais como: "Em terra de **cego**, quem tem olho é Rei", "O pior **cego** é aquele que não quer ver", "O amor é **cego**", e "Mais perdido do que **cego** em meio de tiroteio". Neste contexto, considerou-se que tais expressões não possuem caráter Ofensivo/Pejorativo aos deficientes visuais.

A análise da distribuição das classes, permite-nos concluir que, pelo menos no que tange ao Twitter, os usuários pouco postam denunciando ou demonstrando indignação. Se os exemplos ruidosos do *dataset* fossem excluídos (instâncias anotadas como "Outros"), menos de 10% dos *Tweets* representavam denúncia e indignação. Em contraste, cerca de 40% foram considerados Ofensivo e Pejorativo, e os 50% restantes quase que igualmente divididos entre Relato de Experiência e Informativo. Considerando a maior representatividade e coesão dos *Tweets* classificados como Outros, Ofensivo e Pejorativo, Informativo e Relato de experiência, aplicou-se neles a modelagem de tópicos para identificar os temas mais recorrentes. A Tabela 1 apresenta os resultados da modelagem de tópicos por categoria.

Por meio da análise da Tabela 1 é possível confirmar os achados na fase de anotação manual, já que a classe **"Outros"** é prevalente. Já sobre as impressões dos avaliadores acerca da classe **"Ofensivo e Pejorativo"**, por serem menos frequentes, os avaliadores não foram assertivos sobre os temas mais frequentes, justificando a utilização da modelagem de tópicos. Os *Tweets* anotados como **"Relatos de Experiência"** também contém ruídos, como a exposição de situações onde a pessoa se descreve em uma situação de perca de visão, necessitando de uma consulta em um oftalmologista (Tópico 2). Um achado interessante diz respeito ao Youtube como canal de comunicação bastante utilizado no nicho em estudo (Tópico 3), sendo que diversos canais são mantidos por deficientes e seus familiares com o intuito de compartilharem experiências, seguindo o estilo

Tabela 1. Tópicos mais frequentes por categoria.

Categoria	Tópico	Termos		
Outros	1	cego	tiroteio	amor
	2	surdo	ficar	mudo
	3	tiroteio	perdida	perdido
	4	amor	tiroteio	cupido
Ofensivo	1	cego	filho	melhor
	2	síndrome	down	nome
	3	autista	surdo	acho
	4	surdo	gritando	bocejar
Relato	1	síndrome	down	pessoas
	2	cego	ficando	oftalmo
	3	youtube	vídeo	gostei
	4	deficiência	aluno	auditiva
Informativo	1	deficiência	vagas	oferece
	2	down	síndrome	dia
	3	autismo	crianças	brasil
	4	pesquisa	sintomas	maconha

semelhante ao de um diário. No que tange aos *Tweets* **"Informativos"**, os temas mais discutidos sobre deficiência foram i) inserção no mercado de trabalho (Tópico 1); ii) divulgação de datas importantes (Tópicos 2 e 3); iii) e pesquisas relacionadas ao tema (Tópico 4). Tais achados permitem a construção de um *guideline* para se melhorar a eficácia da interlocução entre o consórcio de pesquisa e os deficientes e seus familiares.

4. Considerações Finais

É notório o crescimento de iniciativas que visam melhorar o bem-estar e qualidade de vida de pessoas com deficiência, neste ensejo, o suporte social por meio de plataformas computacionais tem ganhado destaque. Apesar dos esforços para uso consciencioso das redes sociais, é comum a presença de discurso de ódio ou de cunho pejorativo relacionado à deficiências, impactando negativamente na auto-estima de deficientes e seus familiares. Visando aprofundar as investigações nesta temática, este trabalho apresentou uma análise de *Tweets* sobre o tema em questão, identificando os principais tópicos discutidos e as circunstâncias de utilização.

Acerca das análises, a classificação do discurso deu-se de forma manual considerando-se cinco categorias, a citar: "Informativo", "Ofensivo e Pejorativo", "Indignação e Denúncia", "Relato de Experiências", e "Outros". Para a identificação dos tópicos mais frequentes utilizou-se o algoritmo *Latent Semantic Allocation*. Por meio da análise dos resultados alcançados, destacam-se as seguintes contribuições científicas tangíveis: i) Obtenção de um *dataset* anotado considerando termos relacionados à deficiências; ii) A identificação de tópicos mais frequentes, considerando as categorias propostas; iii) A disponibilização de informações para abalizar a construção de *guidelines* para comunicação com o deficientes e familiares por intermédio de mídias sociais.

As principais dificuldades encontradas estão relacionadas à alta presença de ruído nos dados analisados. No que tange às ameaças a validade do estudo, a não utilização de uma medida estatística para avaliar a concordância entre os avaliadores e também a quantidade de dados anotados são as mais notórias. A consideração de tais fatos impeliu-

nos a estabelecer como trabalhos futuros a anotação de um *dataset* sem incluir os termos de busca ruidosos, e a adoção do coeficiente kappa de Cohen para analisar a proporção de concordância observada. Também vislumbrou-se o estudo da estrutura do *Tweet* quanto à utilização de *hashtags* e *links* a fim de saber quais as características das postagens com maior engajamento

Agradecimentos

Este trabalho foi parcialmente financiado pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES). Agradecemos também aos voluntários, por seus esforços na anotação dos *datasets*, e aos revisores pelos comentários enriquecedores.

Referências

Baumer, E. P. S., Mimno, D., Guha, S., Quan, E., and Gay, G. K. (2017). Comparing grounded theory and topic modeling: Extreme divergence or unlikely convergence? *Journal of the Association for Information Science and Technology*, 68(6):1397–1410.

Cirqueira, D., Jacob, A., Lobato, F., de Santana, A. L., and Pinheiro, M. (2017). Performance Evaluation of Sentiment Analysis Methods for Brazilian Portuguese. In Abramowicz, W., Alt, R., and Franczyk, B., editors, *Business Information Systems Workshops: BIS 2016 International Workshops, Leipzig, Germany, July 6-8, 2016, Revised Papers*, pages 245–251. Springer International Publishing, Cham.

Cirqueira, D. and Vinícius, L. (2017). Opinion Label : A Gamified Crowdsourcing System for Sentiment Analysis Annotation. In *XVI Workshop de Ferramentas e Aplicações*.

Lobato, F., Pinheiro, M., Jacob, A., Reinhold, O., and Santana, Á. (2017). Social CRM: Biggest Challenges to Make it Work in the Real World. In Abramowicz, W., Alt, R., and Franczyk, B., editors, *Business Information Systems Workshops: BIS 2016 International Workshops, Leipzig, Germany, July 6-8, 2016, Revised Papers*, volume 263, pages 221–232. Springer International Publishing, Cham.

Magalhães, G. G. M. S., Lima, F., Santos, E. F., Junior, P., and Rosa, L. (2017). Seleção de Técnicas de Mineração de Dados para Segmentação de Mercado. In *6th Brazilian Workshop on Social Network Analysis and Mining*.

Mckay, K., Wark, S., and Mapedzahama, V. (2015). Sticks and stones : How words and language impact upon social inclusion. *Journal of Social Inclusion*, 6:146–162.

Naslund, J. A., Aschbrenner, K. A., Marsch, L. A., and Bartels, S. J. (2016). The future of mental health care: peer-to-peer support and social media. *Epidemiology and Psychiatric Sciences*, 25(2):113–122.

Poushter, J. (2016). Social networking very popular among adult internet users in emerging and developing nations. Technical report, Pew Research Center.

Setareh Forouzan, A., Mahmoodi, A., Jorjoran Shushtari, Z., Salimi, Y., Sajjadi, H., and Mahmoodi, Z. (2013). Perceived Social Support Among People With Physical Disability. *Iranian Red Crescent Medical Journal*, 15(8):663–667.

Silva, W., Santana, Á., Lobato, F., and Pinheiro, M. (2017). A Methodology for Community Detection in Twitter. In *Proceedings of the International Conference on Web Intelligence*, pages 1006–1009.

Patrocinador Diamante

GOVERNO
DO RIO GRANDE DO NORTE

Patrocinadores Bronze

ESIG
dynavideo
CABOTELECOM
Sistema FIERN
NATAL
PARNAMIRIM
Cuidando de você

Apoio Financeiro

CAPES
MINISTÉRIO DA EDUCAÇÃO
BRASIL
CNPq
Conselho Nacional de Desenvolvimento Científico e Tecnológico
nic.br
cgi.br