Classificação de filmes: uma abordagem utilizando o LIWC

Rian Tavares[1], **Gustavo Paiva Guedes**[1]

[1]Departamento de Informática – Centro Federal de Educação Tecnológica
Celso Suckow da Fonseca (CEFET/RJ) – Rio de Janeiro - RJ – Brasil

rian.tavares@gmail.com,gustavo.guedes@cefet-rj.br

Abstract. *This article aims to present an approach to classify movies based on their subtitles and information extracted from social networks. The methodology developed uses LIWC program, which contains a dictionary of words that allows extracting linguistic, psychological and social characteristics of texts. Preliminary results were very satisfactory, indicating promising directions for this study.*

Resumo. *Esse artigo tem o objetivo de apresentar uma abordagem para classificação de filmes com base em suas legendas e informações extraídas de redes sociais. A metodologia desenvolvida utiliza o programa LIWC, que contém um dicionário de palavras que permite extrair características linguísticas, psicológicas e sociais de textos. Os resultados preliminares foram bastante satisfatórios, indicando direções promissoras para esse trabalho.*

1. Introdução

A Computação Afetiva (CA) é uma área de pesquisa que compreende a criação de sistemas capazes de reconhecer, interpretar e simular emoções [Picard 1997]. É uma área multidisciplinar que envolve conceitos provenientes da Ciência da Computação, Psicologia e Ciências Cognitivas. Na Psicologia, por exemplo, sabe-se que escrever sobre as emoções em experiências pessoais pode trazer melhorias na saúde mental e psicológica [Pennebaker and Seagal 1999]. Na área da computação, estudos estão auxiliando usuários de redes sociais a selecionar documentos com base em suas emoções [Bao et al. 2012].

No âmbito da Ciência da Computação, a CA engloba dois tópicos de pesquisa distintos: Análise de Sentimentos (AS) e Reconhecimento de Emoções (RE) [Poria et al. 2017]. Já se sabe pela neurociência que sentimentos e emoções representam um papel importante na forma que os indivíduos se comportam [Marg 1995]. Isso pode ser evidenciado por alguns trabalhos que desenvolvem pesquisas relacionadas a esses tópicos, por exemplo auxiliando aos usuários no entendimento de sentimentos, opiniões e emoções expressos em textos [Nascimento et al. 2012].

Outro cenário que abrange a detecção de emoções e sentimentos é proveniente da indústria cinematográfica. Conforme descrito em [Oliveira et al. 2011], os filmes são, por excelência, uma forma de arte que envolve atividade afetiva. Nesse aspecto, os filmes têm sido bastante explorados na área da psicologia para a indução de emoções [Ashby et al. 2002]. Isso também ocorre na área da computação, em que alguns trabalhos utilizam as legendas dos filmes para encontrar o gênero dos filmes [Wortman 2010].

Nesse panorama, o presente trabalho apresenta um estudo sobre a utilização de legendas de filmes para a classificar a qualidade de filmes nas classes *excelente* e *ruim*, conforme adotado em [Mullen and Collier 2004, Ye et al. 2006]. Além disso, esse estudo examina se informações provenientes do Facebook podem auxiliar no processo de classificação. Vale ressaltar que a maior motivação desse trabalho consiste em auxiliar usuários a encontrar filmes de interesse, dada a grande quantidade de filmes disponíveis.

Esse artigo é dividido em mais cinco seções. Na Seção 2 são discutidos alguns trabalhos relacionados. Na Seção 3 são descritas a metodologia desse estudo e a criação dos conjuntos de dados. Na Seção 4 são exibidos os resultados experimentais. Na Seção 5 são discutidas as conclusões e algumas perspectivas para trabalhos futuros. Por fim, na Seção 6, são feitos os agradecimentos.

2. Trabalhos Relacionados

O trabalho proposto em [Ye et al. 2006] se enquadra na área de mineração de opiniões em filmes, em que são utilizados os textos de avaliações realizadas por usuários. Esse trabalho desenvolve uma abordagem de orientação semântica para o chinês. Dessa maneira, o maior objetivo é realizar a classificação de filmes nas classes `excelente` e `ruim`. Vale ressaltar que a orientação semântica das palavras é calculada de acordo com a distância semântica entre as palavras dos textos das avaliações e um conjunto de dados padrão.

[Mullen and Collier 2004] utilizam uma abordagem semelhante à supracitada. O objetivo principal também consiste em atribuir valores semânticos aos textos de avaliações de filmes. De forma análoga, os autores buscam classificar filmes nas classes `excelente` e `ruim`. No entanto, os autores desenvolvem uma abordagem baseada em Máquina de Vetores de Suporte (MVS).

O estudo desenvolvido por [Wortman 2010] utiliza legendas de filmes para a classificação de gênero (e.g. drama, suspense, terror). A abordagem proposta pelo autor também utiliza o LIWC para produzir vetores representando as legendas dos filmes. Em seguida, é calculada a similaridade entre esses vetores e modelos de gêneros de filmes propostos pelo autor. Os resultados alcançados nesse trabalho são bastante satisfatórios.

Diversos trabalhos existentes na literatura estudam a classificação de filmes. Entretanto, não foram encontrados estudos que utilizem o LIWC para realizar a classificação da qualidade dos filmes. Dessa maneira, o presente trabalho utiliza conceitos de alguns trabalhos existentes na literatura para propor uma metodologia que envolve a utilização do LIWC, as legendas dos filmes e informações provenientes do Facebook para classificar os filmes nas classes `excelente` e `ruim`.

3. Metodologia

A metodologia proposta para esse trabalho é descrita em 4 fases: extração dos dados, seleção dos dados, integração/transformação dos dados e seleção de atributos.

3.1. Extração dos dados

A extração dos dados desse estudo foi efetuada a partir de dois conjuntos de dados: o primeiro, referente à legendas de filmes e o segundo, relacionado à informações coletadas de um banco de dados de filmes. O conjunto de dados de legendas de filmes pode ser encontrado em [Wortman 2010] e compreende legendas de $1,184$ filmes na língua inglesa.

O conjunto de dados com informações coletadas dos filmes provém do IMDB. O Internet Movie Database (IMDB: www.imdb.com) disponibiliza informação sobre filmes, como diretor, atores, orçamento, ano, etc. Assim, o conjunto de dados utilizado, denominado IMDB5000[1], possui informações de filmes provenientes do IMDB e do Facebook. É composto de dados de 5.043 filmes e possui 28 atributos relacionados a cada filme.

O LIWC [Pennebaker and Seagal 1999] foi desenvolvido com o intuito de fornecer um método eficiente para estudar fatores emocionais, cognitivos, estruturais, entre outros, presentes em amostras de falas verbais e escritas de indivíduos. Seu núcleo é constituído por um dicionário de palavras que fornece informações sobre os fatores supracitados. Nesse trabalho utilizamos a versão em inglês do dicionário LIWC, que possui, aproximadamente, 6.540 palavras, em cada palavra é associada a uma ou mais categorias dentre as 73 categorias presentes no dicionário.

3.2. Seleção dos dados

Inicialmente, foram selecionados todos os nomes de filmes existentes tanto no conjunto de dados de legendas quanto no conjunto de dados IMDB5000, o que totalizou 672 filmes. Em seguida, selecionamos um subconjunto de 4 atributos do conjunto de dados IMDB5000. Esses atributos fazem referência à nota do filme no IMDB (*imdb_score*), à quantidade de *likes* recebidos pelo diretor do filme no Facebook (*director_facebook_likes*), à quantidade de *likes* recebidos pelo filme no Facebook (*movie_facebook_likes*) e à quantidade de *likes* que o elenco do filme recebeu no Facebook (*cast_total_facebook_likes*). Em seguida, foram efetuadas algumas transformações nos 672 filmes e no atributo *imdb_score*, descritas na Subseção 3.3.

3.3. Integração/transformação dos dados

Um usuário que deseja avaliar um filme no IMDB pode inserir uma nota entre 0 e 10. Dessa maneira, a nota final de um filme é composta pela média das notas que todos os usuários atribuíram. Com isso, as notas finais são representadas por valores contínuos. Esses valores são representados pelo atributo *imdb_score* no conjunto de dados IMDB5000. Nesse estudo, esses valores foram generalizados e definidos em 4 classes distintas, conforme em [Asad et al. 2012]. A Tabela 1 apresenta as classes utilizadas para cada faixa de valor.

Tabela 1. Faixas correspondente a cada classe utilizada para classificar filmes.

Classe	Faixa
Excelente	7.5 - 10
Bom	5.0 - 7.4
Ruim	2.5 - 4.9
Péssimo	1.0 - 2.4

Esse estudo é centralizado na classificação de filmes das classes excelente e ruim, conforme mencionado na Seção 1. Com isso, serão considerados apenas os filmes dessas classes, o que totaliza 172 filmes: 125 rotulados com a classe excelente e 47 com a classe ruim.

[1] https://www.kaggle.com/deepmatrix/imdb-5000-movie-dataset

Em seguida, os textos das legendas dos filmes e os 3 atributos relacionados ao Facebook (ver Seção 3.2) foram utilizados para para produzir 172 vetores (um para cada filme) de dimensão $n = 76$, representando as 73 categorias do LIWC e os 3 atributos provenientes do Facebook. Para produzir um vetor $\vec{v}$ representando um filme f, foram consideradas as palavras existentes nas legendas desse filme. Cada palavra p existente no filme f foi pesquisada no dicionário do LIWC e suas categorias foram identificadas. Para cada categoria identificada x_i em p, sua posição correspondente no vetor $\vec{v}$ foi incrementada em 1. Assim foram formadas as 73 posições iniciais do vetor $\vec{v}$. As 3 últimas posições foram compostas pelos valores existentes nos seguintes atributos provenientes do `IMDB5000`: *movie_facebook_likes*, *director_facebook_likes* e *cast_total_facebook_likes*. A Figura 1 ilustra o processo utilizado para representar cada filme.

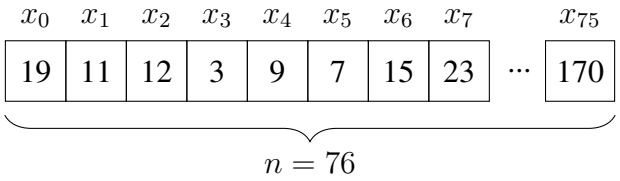

Figura 1. Vetor representando um filme utilizando as categorias do LIWC e 3 atributos provenientes do `IMDB5000`.

Podemos notar que $x_0 = 19$, assinalando que ocorreram 19 palavras no filme f que se enquadraram nessa categoria. Para o caso de x_5, é possível perceber que 7 palavras pertencem a essa categoria. Com relação às 3 últimas posições, os valores representam o número de curtidas. Como exemplo, a posição x_{75} corresponde ao número total de curtidas obtidas pelos atores no Facebook (*cast_total_facebook_likes*).

Em seguida, cada filme foi associado a sua respectiva classe (i.e. excelente ou ruim). O conjunto de dados resultante foi denominado `FILM-172-76`. Esse conjunto é composto por 172 instâncias (125 da classe *excelente* e 47 da classe *ruim*) e 76 atributos.

3.4. Seleção de atributos

O processo descrito na Seção 3.3 foi responsável pela produção do conjunto de dados `FILM-172-76`. Dado que diversos atributos podem atrapalhar a tarefa de classificação [Jain and Zongker 1997], foram selecionados 20% dos atributos com maior importância seguindo o Princípio de Pareto [Koch 1999], também denominado princípio 80-20. Existem diversas maneiras de calcular a importância dos atributos de um conjunto de dados. Nesse estudo, utilizamos o Ganho de Informação (Information Gain) [Han et al. 2011] para selecionar 15 (20%) dos atributos com maior importância. Esse conjunto de dados foi denominado `FILM-172-15`.

4. Resultados experimentais

Os resultados experimentais desse trabalho foram produzidos com a utilização do conjunto de dados `FILM-172-76` e `FILM-172-15`, descritos na Seção 3. Para realizar os experimentos foi utilizada a plataforma WEKA [Hall et al. 2009]. Essa plataforma possui uma série de algoritmos de aprendizagem de máquina bem conhecidos na literatura. Nesse trabalho, foram utilizados 4 algoritmos que apresentam bons resultados na área de

mineração de texto, são eles: Random Forest (RF), Naive Bayes (NB), NB Multinomial e Sequential Minimal Optimization (SMO). O ZeroR foi utilizado como *baseline*.

A Tabela 2 ilustra os resultados obtidos para os dois conjuntos supracitados. Com exceção do *baseline*, os algoritmos obtiveram melhores resultados no conjunto de dados `FILM-172-15`. Os valores em negrito indicam o algoritmo que obteve o melhor F1 para o conjunto de dados em questão. O algoritmo RF se comportou melhor no conjunto de dados `FILM-172-76` e o SMO apresentou melhores resultados para o conjunto de dados `FILM-172-15`.

Tabela 2. Resultados da classificação dos filmes - Média F1

	ZeroR	RF	NB	NBM	SMO
`FILM-172-76`	0.446	**0.502**	0.454	0.405	0.464
`FILM-172-15`	0.446	0.580	0.506	0.470	**0.600**

É interessante destacar que todos os algoritmos se comportaram apresentando melhores resultados que o ZeroR, utilizado como *baseline*. Isso indica que a metodologia proposta nesse estudo e a utilização do LIWC são capazes de auxiliar na distinção de filmes excelentes e ruins. Os melhores resultados alcançados com o conjunto de dados com redução de atributos (`FILM-172-15`) são bastante fundamentados na literatura, visto que muitas vezes a redução de dimensionalidade pode não apenas melhorar os resultados, mas também reduzir o tempo de processamento [Jain and Zongker 1997].

5. Conclusão

Esse estudo teve como objetivo a utilização do dicionário de palavras do LIWC para classificar a qualidade de filmes a partir de suas legendas. Para isso, foram produzidos dois conjuntos de dados, denominados `FILM-172-76` e `FILM-172-15`. Em seguida, foram aplicados alguns algoritmos conhecidos na literatura para realizar os experimentos: RF, NB, NBM e SMO. Os resultados foram satisfatórios, indicando que o caminho seguido ainda pode ser explorado em trabalhos futuros.

Como trabalhos futuros, surgiram algumas ideias que terão sua viabilidade estudadas. Dentre elas, a possibilidade de buscar outras redes sociais (além do Facebook) capazes de fornecer dados que possam auxiliar na classificação. Isso pode envolver, também, textos de avaliações de filmes realizadas por usuários. Também é considerada a utilização das classes `bom` e `péssimo`. Como trabalho futuro também é vislumbrado o desenvolvimento de um sistema capaz de avaliar filmes antes mesmo do lançamento, visando assim a redução o tempo em que o filme é lançado até a coleta de um conjunto de avaliações e críticas. Para tanto, já estão sendo estudadas maneiras de viabilizar uma maior quantidade de legendas de filmes.

6. Agradecimentos

Os autores agradecem ao CEFET/RJ pela bolsa de iniciação cientifica concedida à Rian Tavares.

Referências

Asad, K. I., Ahmed, T., and Rahman, M. S. (2012). Movie popularity classification based on inherent movie attributes using c4. 5, part and correlation coefficient. In *Informatics, Electronics & Vision (ICIEV), 2012 International Conference on*, pages 747–752. IEEE.

Ashby, F. G., Valentin, V. V., et al. (2002). The effects of positive affect and arousal and working memory and executive attention: Neurobiology and computational models.

Bao, S., Xu, S., Zhang, L., Yan, R., Su, Z., Han, D., and Yu, Y. (2012). Mining social emotions from affective text. *Knowledge and Data Engineering, IEEE Transactions on*, 24(9):1658–1670.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18.

Han, J., Pei, J., and Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.

Jain, A. and Zongker, D. (1997). Feature selection: Evaluation, application, and small sample performance. *IEEE transactions on pattern analysis and machine intelligence*, 19(2):153–158.

Koch, R. (1999). *The 80/20 Principle: The Secret of Achieving More with Less*. A Currency book. Doubleday.

Marg, E. (1995). Descartes'error: Emotion, reason, and the human brain. *Optometry & Vision Science*, 72(11):847–848.

Mullen, T. and Collier, N. (2004). Sentiment analysis using support vector machines with diverse information sources. In *EMNLP*, volume 4, pages 412–418.

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*.

Oliveira, E., Martins, P., and Chambel, T. (2011). Ifelt: Accessing movies through our emotions. In *Proceddings of the 9th International Interactive Conference on Interactive Television*, EuroITV '11, pages 105–114, New York, NY, USA. ACM.

Pennebaker, J. W. and Seagal, J. D. (1999). Forming a story: The health benefits of narrative. *Journal of clinical psychology*, 55(10):1243–1254.

Picard, R. W. (1997). *Affective Computing*. MIT Press, Cambridge, MA, USA.

Poria, S., Cambria, E., Bajpai, R., and Hussain, A. (2017). A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, 37:98–125.

Wortman, J. (2010). *Film classification using subtitles and automatically generated language factors*. Technion-Israel Institute of Technology, Faculty of Industrial and Management Engineering.

Ye, Q., Shi, W., and Li, Y. (2006). Sentiment classification for movie reviews in chinese by improved semantic oriented approach. In *System Sciences, 2006. HICSS'06. Proceedings of the 39th Annual Hawaii International Conference on*, volume 3, pages 53b–53b. IEEE.