Automatic Document Classification Temporally Robust

Thiago Salles[1] ***(Author)*, Marcos André Gonçalves**[1] ***(Advisor)*,**
Leonardo Rocha[2] ***(Co-Advisor)***

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

[2]Departamento de Ciência da Computação – Universidade Federal de São João Del-Rei
São João Del-Rei – MG – Brazil

{tsalles,mgoncalv}@dcc.ufmg.br, lcrocha@ufsj.edu.br

1. Motivation

Text classification is still one of the major information retrieval problems, and developing robust and accurate classification models continues to be in great need as a consequence of the increasing complexity and scale of current application scenarios, such as the Web. The task of Automatic Document Classification (ADC) aims at creating models that associate documents with semantically meaningful categories. These models are key components for supporting and enhancing a variety of other tasks such as automated topic tagging (that is, assigning labels to documents), building topic directories, identifying the writing style of a document, organizing digital libraries, improving the precision of Web searching, and even helping users to interact with search engines.

Similarly to other machine learning techniques, ADC usually follows a supervised learning strategy, where a training set of already classified documents is employed to learn a classifier. Such a classifier is then used to predict the classes of new unclassified documents. The majority of ADC algorithms consider that all training documents provide equally important information to learn an accurate classifier. However, this may not hold in practice due to several factors such as the document's creation time, the venue in which it was published, its authors, among other factors [Palotti et al. 2010].

In the master dissertation, we were particularly concerned with the impact that the document's creation time may have on ADC algorithms. Due to several factors, such as the dynamics of knowledge and even the dynamics of languages, the characteristics of a textual dataset may change over time. In such scenarios, the classification effectiveness may deteriorate over time, since the general assumption of static distribution may not hold. This becomes specially important nowadays, due to the availability of large datasets that span for long time periods. Therefore, the temporal dynamics of the data is an important aspect that must be taken into account in order to learn more accurate classifiers.

In fact, [Mourão et al. 2008] provided a characterization of these changes in terms of three main *temporal effects*: *(i)* the class distribution variation (CD), that accounts for the variations on the relative frequencies of the classes; *(ii)* the term distribution variation (TD), which refers to changes in the representativeness of the terms with respect to the classes as time goes by; and, *(iii)* the class similarity variation (CS), which considers how the similarity among classes, as a function of the terms that occur in their documents, changes over time. As reported in that work, two real textual datasets are indeed affected by the temporal effects, which negatively impacted the SVM classifier effectiveness.

Based on the above discussion, we took a step further and hypothesized that, besides the known negative impact of the temporal effects on classification effectiveness, *(i)* distinct textual datasets present differing dynamical behaviors; *(ii)* different ADC algorithms may be distinctively affected by the temporal evolution of data; and, finally, *(iii)* the temporal evolution of data may be explored to devise more effective classification models. Our claim was that, in order to come up with a classification strategy robust to the temporal effects, one should first better understand their extent in textual datasets, as well as their impact on existing traditional classifiers. In other words, the best strategy to handle temporal effects may depend on the specific characteristics of both the dataset and the ADC algorithm used, thus turning the learning of more accurate classifiers, that deal with these effects, an even more challenging task.

That said, the specific contributions of the master dissertation were two-fold. First, as summarized in Section 2, we provided a *quantification* of the impact of three main temporal effects in four widely used ADC algorithms. More specifically, we proposed a methodology to enable a deeper study of the three temporal effects, by means of a series of factorial experimental designs aimed at uncovering how each temporal effect affects each ADC algorithm and textual dataset. We instantiated that methodology considering three real textual datasets and four ADC algorithms, and provided a detailed study regarding the impact of the temporal effects on them. Second, as summarized in Section 3, we proposed strategies to *minimize* the impact of the temporal effects in ADC algorithms, in the light of our quantification study. Again, more specifically, we introduced a temporal weighting function to capture the varying behavior of textual datasets, and proposed two strategies to devise it. We also extended three well known ADC algorithms to incorporate such a function, devising what we called the temporally-aware algorithms for ADC. We performed an extensive experimental analysis in order to assess the benefits of considering the temporal dynamics of data.

During the development of the master dissertation, we published a paper in the world leading Information Retrieval conference, namely, the International ACM SIGIR Conference on Research & Development of Information Retrieval (Qualis A1) [Salles et al. 2010b]. We also published in the Simpósio Brasileiro de Bancos de Dados (Qualis B3) [Salles et al. 2009], in the IADIS WWW/Internet conference (Qualis B2) [Salles et al. 2011b] and a short paper in the ACM SIGKDD Workshop on Knowledge Discovery from Uncertain Data (a workshop on a conference Qualis B1) [Zadrozny et al. 2009]. Furthermore, we published two journal papers in the Journal of Information and Data Management (Qualis B3) [Salles et al. 2010a, Salles et al. 2011a] and submitted a paper to the Knowledge and Information Systems (Qualis B1). We also had some work related to the text classification topic (but not directly related to the temporal issues) already published. Specifically, we published a paper in the IEEE Congress on Evolutionary Computation (Qualis A1) [Palotti et al. 2011] and a paper in the IADIS Applied Computing conference (Qualis B3). We also published a paper in the Information Systems journal (Qualis A2) [Figueiredo et al. 2011] and another one in the Journal of Information and Data Management [Palotti et al. 2010].

2. A Quantitative Analysis of Temporal Effects on ADC

In order to achieve a better understanding about the influence of the temporal dimension in classification effectiveness, a key aspect to be analyzed concerns the peculiar manifestations of each temporal effect in different datasets. For example, while some datasets may present large class distribution variations over time (a.k.a. CD effect), other datasets may, in contrast, present a more significant variability on term distribution (a.k.a. TD effect), or a more significant variability on inter-class similarities (a.k.a. CS effect). Moreover, different ADC algorithms may be distinctively affected by these effects due to their sensitivity or robustness to each specific effect. Hence, two important (and previously untackled) questions must be answered in order to better understand the impact of temporal effects: *(i) Which temporal effects are strongest in each dataset? (ii) What is the behavior of each ADC algorithm when faced with different levels of each temporal effect?*

In fact, it has already been established that these temporal effects do exist in some datasets and affect negatively one specific algorithm, namely the SVM classifier [Mourão et al. 2008]. We re-visited the characterization reported in that work, by including a third textual dataset belonging to the news articles domain in order to reinforce the existence of the temporal effects. Furthermore, we took a step further towards answering the posed questions, by proposing a factorial experimental design aimed at quantifying the impact of the temporal effects in four representative ADC algorithms (namely, Rocchio, KNN, Naïve Bayes and SVM), considering three textual datasets (ACM-DL, MEDLINE and AG-NEWS, detailed in the dissertation) with differing characteristics in their temporal evolution.

Briefly, given k factors, which can assume n levels (possible values), and a response variable, a full factorial $n^k r$ experimental design (with r replications) aims at quantifying the impact (effect) of each individual factor as well as of all inter-factor interactions (of all orders) on a given response variable, by means of a series of experiments carefully designed to cover all possible configurations of factor levels. We proposed a methodology to conduct such an experimental design aimed at quantifying the impact of the three temporal effects (CD, TD and CS). Details regarding the main steps of the factor isolation procedure and the experimental setup can be found in the dissertation.

Our characterization results showed that, contrary to the assumption of static data distribution on which all four explored algorithms are based, each reference dataset has a *specific* temporal behavior w.r.t. the temporal effects, exhibiting changes in the underlying data distribution with time. Such temporal variations do indeed limit the classification effectiveness. According to our results, the ACM-DL and AG-NEWS datasets are much more dynamic than the MEDLINE dataset, resulting in the four explored ADC algorithms being more impacted by the temporal aspects in the first two datasets.

In addition to such findings, our proposed methodology allowed us to answer the two posed questions. Regarding the extent of each effect w.r.t. the datasets, in the ACM-DL dataset, the impact of the observed temporal variations in the distribution of class sizes and in the pairwise class similarities are statistically equivalent to the impact of the observed variations in the term distribution on most classifiers (SVM being an exception). MEDLINE and AG-NEWS, on the other hand, are clearly more impacted by the first two temporal aspects. With 99% confidence, we obtained the following partial orderings regarding the extent of the temporal effects in each reference dataset: $CD_{MEDLINE} < CD_{ACM-DL} \sim CD_{AG-NEWS}$, $CS_{MEDLINE} < CS_{ACM-DL} \sim CS_{AG-NEWS}$ and $TD_{MEDLINE} < TD_{ACM-DL} < TD_{AG-NEWS}$. These findings reveal that the challenges imposed by the temporal effects are in fact data dependent and that developing strategies to handle them in ADC algorithms is a promising research direction.

Regarding the behavior of each ADC algorithm w.r.t. each temporal effect, we found that all four explored ADC algorithms suffer a negative impact of the temporal effects in terms of classification effectiveness, being the most significant impacts observed when these algorithms are applied to the most dynamic datasets (i.e., ACM-DL and AG-NEWS). Furthermore, the SVM classifier was shown to be more robust to the term distribution variation, while still being impacted by the other two effects. The other three algorithms, on the other hand, were shown to be very sensitive to *all* three effects. These relationships reinforce that, apart from being negatively impacted by all three temporal effects, the explored classifiers exhibit distinct behavior when faced with datasets with specific temporal dynamics, as revealed by the conducted factorial designs. Thus, the temporal dimension turns out to be an important aspect that has to be considered when learning accurate classifiers. Table 1, referring to the ACM-DL dataset, summarizes such findings. The complete set of results, along with an extensive study regarding the strengths and weaknesses of each classifier w.r.t. the temporal effects can be found in the dissertation.

Temporal Effect	Dataset		
	ACM-DL	MEDLINE	AG-NEWS
CD	SVM > NB ~ KNN ~ RO	RO > SVM > NB > KNN	RO ~ KNN > SVM ~ NB
CS	SVM > KNN ~ RO > NB	RO > SVM ~ NB > KNN	RO ~ KNN ~ NB > SVM
TD	SVM ~ KNN ~ RO ~ NB	SVM > RO ~ NB ~ KNN	RO > NB > KNN > SVM

Table 1. A Comparative Study on the Impact of the Temporal Effects on each ADC Algorithm—Rocchio (RO), SVM, Naïve Bayes (NB) and KNN.

3. Temporally-Aware Algorithms for Automatic Document Classification

The performed quantitative analysis brought some key aspects to be considered towards the development of ADC algorithms temporally robust. As reported in the master dissertation, all three datasets are composed by both stable and unstable documents (measured by what we called the Document Stability Level metric). Two common approaches to deal with data varying distributions are window-based and instance weighting strategies [Klinkenberg 2004]. The first consists of considering a temporal window where documents created at time points outside it are simply discarded. However, stable data may contribute positively to classification effectiveness and, if outside the temporal window, it would be unwittingly discarded. This motivates us to consider an instance weighting approach—a smoother way to tackle the varying data distribution issue. Common instance weighting strategies, on the other hand, apply a weight to documents according to their creation points in time. Usually, such strategies consider ad-hoc weighting functions, such as an exponential decayment function. However, according to the performed quantitative analysis, distinct datasets present peculiar varying behavior, and the choice of an ad-hoc weighting scheme, without considering the characteristics of the dataset, is not general enough. Thus, our weighting function should not only consider the temporal distance between training and test documents, but also the varying behavior of the datasets.

We first proposed a methodology to model a temporal weighting function (TWF) that captures changes in term-class relationships for a given period of time, based on a series of statistical tests [Salles et al. 2010b]. For the ACM-DL and MEDLINE datasets, we showed that the TWF follows a lognormal distribution, whose parameters may be easily determined using statistical methods. For the AG-NEWS dataset, on the other hand, we showed that the same adopted hypothesis testing procedures failed, implying that its associated TWF follows a distinct (yet unknown) distribution. Thus, in order to guarantee a wider applicability of our strategies, we proposed an automatic procedure to determine the TWF, based on meta-learning principles. Such a strategy was able to effectivelly determine the TWF for all three datasets. Having determined the dataset specific TWF, we then proposed three strategies to incorporate the TWF to classifiers, summarized in the following.

TWF in documents: This strategy weights each training document by the TWF according to its temporal distance to the test document d'.

TWF in Scores: Let $\mathbb{P}$ be the set of observed creation points in time. Each training document class c is associated with the corresponding creation point in time $p \in \mathbb{P}$, generating a new class defined as $\langle c, p \rangle$. A traditional classifier is then used to generate scores for each new class $\langle c, p \rangle$. Finally, the test document d' is classified by a traditional classifier applied to this new training set, ultimately generating scores for each $\langle c, p \rangle$. To decide to which class d' should be assigned to, the learned scores for each $\langle c, p \rangle$ are summed up, for all $p \in \mathbb{P}$, weighting them by the $TWF(\delta)$, where $\delta = p - p_r$ corresponds to the temporal distance between p and the creation time p_r of d'. d' is then assigned to the class with highest score.

Extended TWF in Scores: This strategy employs a series of classifiers to associate scores with each class considering only documents belonging to each point in time independently, but belonging to all classes. The scores obtained by each classifier are aggregated with the corresponding TWF weight, according to the temporal distance between the point in time associated to each classifier and the creation time of the test document.

The three strategies were implemented considering three traditional classifiers, namely Rocchio, KNN, and Naïve Bayes. Our experimental evaluation considering the three reference datasets showed that, with 99% confidence, the temporally-aware classifiers outperform their traditional counterparts (for both strategies to devise the TWF). Some of the results are reported in Table 2, and the complete set of results can be found in the dissertation. We refer the reader to the dissertation for a detailed discussion regarding the strenghts and weaknesses of each approach. Also, the best performing temporally-aware classifiers achieved better results than the state-of-the-art SVM classifier, in the ACM-DL and MEDLINE datasets, with a runtime an order of magnitude smaller (as can be found in Table 3, for ACM-DL).

Algorithm	Rocchio		KNN		Naïve Bayes	
Metric	macF$_1$(%)	microF$_1$(%)	macroF$_1$(%)	microF$_1$(%)	macroF$_1$(%)	microF$_1$(%)
Baseline	57.39	68.24	58.48	71.84	57.27	73.24
TWF *in documents*	60.21 (+4.91) ▲	70.70 (+3.60) ▲	60.08 (+2.74) ▲	73.88 (+2.84) ▲	61.38 (+7.18) ▲	74.60 (+1.86) ●
TWF *in scores*	60.47 (+5.47) ▲	72.90 (+6.83) ▲	61.88 (+5.81) ▲	74.53 (+3.74) ▲	45.16 (-26.82) ▼	64.55 (-13.46) ▼
TWF *in scores ext.*	59.96 (+4.48) ▲	71.99 (+5.49) ▲	59.80 (+2.26) ▲	73.95 (+2.94) ▲	56.28 (-1.76) ●	72.73 (-0.70) ●

Table 2. Results Obtained with the *Estimated* TWF—ACM-DL.

4. Summary

The main focus of the master dissertation was Automatic Document Classification (ADC) in face of data varying distributions, a challenging issue frequently observed in real world

Algorithm	Metric		
	macF$_1$(%)	micF$_1$.(%)	Runtime (s)
SVM	59.91	73.88	144.10±5.30
KNN with TWF *in scores*	61.88 (+3.29) ▲	74.53 (+0.88) ●	10.10±0.31
Naïve Bayes with TWF *in documents*	61.38 (+2.45) ▲	74.60 (+0.97) ●	9.10±0.32

Table 3. Best Performing Temporally-Aware Classifiers *versus* SVM—ACM-DL.

textual datasets. We tackled this issue by first characterizing it by means of a quantitative analysis on the impact of three main temporal effects in distinct datasets and ADC algorithms. This enabled us to gain a deeper understanding regarding these effects (both in terms of the extent of their manifestations in textual datasets and in terms of the strenghts and weaknesses of traditional classifiers w.r.t. these effects). Such an analysis was of fundamental importance to figure out better strategies to overcome these effects. We showed that the impact of the temporal effects is dataset and classifier specific, and this aspect must be considered when developing effective classifiers to handle these effects.

In the light of our quantitative analysis, we were able to develop three strategies to overcome the negative impact of the temporal effects, which we called the temporally-aware classifiers. We have shown that these classifiers are not only effective (with statistically significant gains over traditional classifiers, such as the state of the art SVM classifier) but also very efficient (with runtime one order of magnitude smaller then that of the SVM classifier). This highlights the quality of our proposed solutions to the problem.

References

Figueiredo, F., Rocha, L., Couto, T., Salles, T., Gonçalves, M., and Meira Jr., W. (2011). Word co-occurrence features for text classification. *Inf. Sys.*, 36(5):843–858.

Klinkenberg, R. (2004). Learning drifting concepts: Example selection vs. example weighting. *Intell. Data Anal.*, 8(3):281–300.

Mourão, F., Rocha, L., Araújo, R., Couto, T., Gonçalves, M., and Meira Jr., W. (2008). Understanding temporal aspects in document classification. In *WSDM*, pages 159–170.

Palotti, J., Salles, T., Pappa, G., Gonçalves, M., and Meira Jr., W. (2011). Assessing documents' credibility with genetic programming. In *IEEE CEC*, pages 200–207.

Palotti, J. M., Salles, T., Pappa, G., Arcanjo, F., Gonçalves, M., and Meira Jr., W. (2010). Estimating the credibility of examples in automatic document classification. *JIDM*, 1(3):439–454.

Salles, T., Cardoso, T., Oliveira, V., Rocha, L., and Gonçalves, M. (2011a). Tackling temporal effects in automatic document classification through cascaded temporal smoothing. *JIDM*, 2(3):417–432.

Salles, T., Rocha, L., Mourão, F., Pappa, G., Cunha, L., Gonçalves, M., and Meira Jr., W. (2010a). Automatic document classification temporally robust. *JIDM*, 1(2):199–212.

Salles, T., Rocha, L., Pappa, G., Mourão, F., Gonçalves, M., and Meira Jr., W. (2009). Classificação automática de documentos robusta temporalmente. In *SBBD*, pages 106–119.

Salles, T., Rocha, L., Pappa, G., Mourão, F., Gonçalves, M., and Meira Jr., W. (2010b). Temporally-aware algorithms for document classification. In *SIGIR*, pages 307–314.

Salles, T., Sandin, I., Oliveira, L., Rocha, L., and Gonçalves, M. (2011b). Exploiting document stability level for efficient temporally-aware classification. In *IADIS WWW/Internet*, pages 309–316.

Zadrozny, B., Pappa, G., Meira Jr., W., Gonçalves, M., Rocha, L., and Salles, T. (2009). Exploiting contexts to deal with uncertainty in classification. In *KDD Workshop on Knowledge Discovery from Uncertain Data*, pages 19–22.