A Thorough Exploitation of Distance-Based Meta-Features for Automated Text Classification

Sergio Canuto[1], **Marcos André Gonçalves, Thierson Couto Rosa**

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

sergiodaniel@dcc.ufmg.br,mgoncalv@dcc.ufmg.br, thierson@inf.ufg.br

Abstract. *The definition of a set of informative features capable of representing and discriminating documents is paramount for the task of automatically classifying documents. In this doctoral dissertation, we present the most comprehensive study so far on the role of meta-features (high-level features built from lower-level ones) as an alternative for representing documents. We start by proposing new sets of (meta-)features that exploit distance measures in the original (bag-of-words) feature space to summarize potentially complex relationships between documents. We then (i) analyze the discriminative power of such meta-features with novel multi-objective feature selection strategies; (ii) provide new GPU implementations to reduce computational time; (iii) enrich distance relationships with labeled or context-specific information; (iv) adapt the proposed meta-features for tasks as hard as sentiment analysis. Our experimental results show that our meta-features can achieve remarkable classification results by distance exploitation, being the state-of-the-art in many situations and scenarios.*

1. Introduction

1.1. Problem Statement

Automated Text Classification (ATC) is one of the primary applications of supervised machine learning. Given a set of training documents, the ATC task corresponds to automatically learning how to classify new (unclassified) documents, using a combination of features of these documents that associates them with categories. ATC is an important component on topics of (i) Document Management and Classification, (ii) Information Retrieval Models and Techniques, (iii) and Text Database, of the SBBD Call for Papers.

Although the classification technique itself plays an important role in ATC, the features that represent documents are equally important to determine effectiveness. In particular, representing documents in a feature space is a pre-requisite for ATC methods, since they are designed to discover discriminative patterns on these features. A recent trend that has emerged in ATC, that works in the *data engineering* level instead of in the algorithmic level, is the introduction of meta-level features that can replace or work in conjunction with the original set of (bag-of-words-based) features [Canuto et al. 2014, Canuto et al. 2015, Yang and Gopal 2012, Pang et al. 2015]. These meta-features capture insightful new information about the unknown underlying data distribution that relates the observed patterns with categories.

Particularly, in this PhD dissertation, we focus on distance-based meta-features based on the hypothesis that distance measures are capable of summarizing potentially complex relationships between text documents. This summarized information can

then be exploited in different ways to generate robust and informative meta-features for text classification. For example, let us consider the meta-features driven from the projection of highly-dimensional documents into a low-dimensional space spanned by category centroids. In this scenario, the meta-features are the distance scores between the document and each category centroid. This specific group of meta-features, taken from category centroids, is capable of mitigating issues related to imbalanced class distributions and irrelevant or noisy term features [Pang et al. 2015]. Indeed recent work [Canuto et al. 2014, Yang and Gopal 2012, Pang et al. 2015] (including ours) has proposed to represent discriminative patterns for text classification with several different groups of meta-features. Each of these groups is engineered to extract complementary information from a particular ATC aspect, such as similarity scores, class distributions, entropy, and the within-class cohesion observed in the k nearest neighbors of a given test document. Enriching distance relationships with labeled data or context-specific information has also the potential to generate additional discriminative and complementary new groups. Their combination, as illustrated in Figure 1, produces a more compact and informative feature space that reportedly improves classification effectiveness.

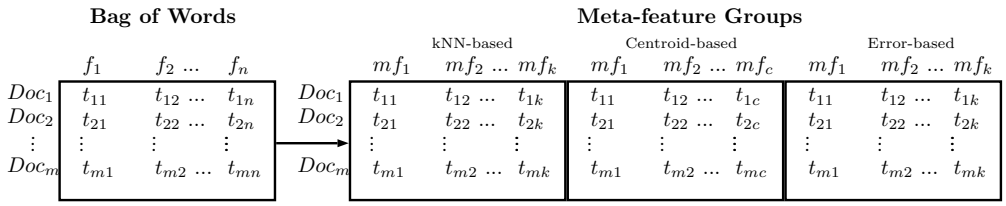

Figura 1. Transforming a dataset represented with bag-of-words into concatenation of meta-feature groups.

Though discriminative by themselves, the combined use of meta-features may degrade effectiveness and efficiency of classification algorithms due to overfitting, and the computational time to compute distances [Canuto et al. 2014]. To use as few meta-feature groups as possible, it is necessary to consider the adequacy of the selected groups to a particular dataset, or to find a "common" reduced core of meta-features capable of providing effective results for different datasets. The solution to these problems can produce insights about the unknown behavior of different combinations of meta-feature groups, providing a starting point for new meta-features and their application in different contexts.

1.2. PhD Dissertation Hypotheses

As discussed, the potential of meta-features might be diminished by the aforementioned problems. This motivated a through investigation of meta-features to provide evidence for the main hypotheses of this dissertation, which include:

1. *The combined use of meta-features can improve ATC effectiveness through the exploitation of discriminative patterns obtained with statistics drawn from document distances.*
2. *Multi-objective optimization techniques can be used as effective and efficient strategies for meta-feature selection and the analysis of informative combinations of meta-features.*
3. *Enriching distance relationships with labeled information can improve meta-features.*
4. *Meta-features can improve the classification effectiveness on other applications through the combined exploitation of context-specific information and document distances.*
5. *Given the intrinsically paralyzable nature of the generation of distance-based meta-features, it is possible to considerably reduce the execution time with GPU architectures.*

1.3. Research Proposals and Contributions

To the best of our knowledge, there are no studies that provide a thorough analysis of the combined use of distance-based meta-feature groups. Accordingly, in our dissertation we proposed a methodology to search for the most informative combinations of meta-features focusing on the analysis of such combinations, as well as the efficiency and effectiveness of our combination strategies. To potentially improve the effectiveness of our combinations, we propose new meta-features that exploit discriminative patterns obtained from different statistics drawn from distances between documents. We aim at exploiting the neighborhood of documents and enriching their distance relationships with labeled information. Moreover, we present strategies to improve the classification effectiveness on other applications, such as sentiment analysis, through the combined exploitation of distances and context-specific information. Finally, we tackle the intrinsically parallelizable nature of the generation of distance-based meta-features with manycore GPU architectures.

In summary, in our dissertation we provide five novel contributions: (1) the proposition and thorough evaluation of new distance-based meta-features [Canuto et al. 2014], (2) the proposal of efficient and effective strategies to evaluate different combinations of meta-feature groups that provide the core information to classify documents [Canuto et al. 2018], (3) the proposal of effective strategies to enrich distance relationships with labeled data for meta-feature generation [Canuto et al. 2019], (4) the exploitation of meta-features designed to take into account the idiosyncrasies of sentiment analysis [Canuto et al. 2016] and (5) a GPU implementation of kNN for high dimensional and sparse data to reduce the computational time of meta-features [Canuto et al. 2015].

Dissertation Related Publications. We published on all main world leading Information Retrieval conferences, with articles on the ACM SIGIR *(h5-index:55, A1)*, ACM CIKM *(h5-index: 48, A1)*, ACM WSDM *(h5-index: 48, A1)* and ECIR *(h5-index: 26, A2)*, completing the *Information Retrieval Grand Slam*. Such papers summarize some of our main contributions [Canuto et al. 2014, Canuto et al. 2015, Canuto et al. 2016, Canuto et al. 2019] and applications of such contributions [Penha et al. 2019, Sousa et al. 2016, Viegas et al. 2019]. Our thorough analysis of meta-features [Canuto et al. 2018] was published on the *IEEE Transactions on Knowledge and Data Engineering (TKDE) (h5-index: 77, Imp. Fac.:3.85, A1)*, which is considered the top journal on Databases and Information Systems[1]. Further developments of our work include the exploitation of our multi-objetive framework in different tasks on *ACM Transactions on Information Systems (TOIS) (h5-index: 24, Imp. Fac.: 2.31, A1)* [Sousa et al. 2019] and a classification pipeline on *Information Processing & Managemnt (IP&M) (h5-index:39, Imp. Fac:3.89, A1)* [Cunha et al. 2020], the top journal in Information Retrieval. The combined *h5-index* of all aforementioned publications is 317.

2. Main Results

2.1. Meta-Features: Proposition, Evaluation and Selection

Unlike previous works that exploit raw distance scores between documents and class centroids [Pang et al. 2015] or raw distances (Euclidean and Cosine) among documents [Yang and Gopal 2012], our meta-features [Canuto et al. 2014] exploit the

[1] https://scholar.google.com/citations?view_op=top_venues&vq=eng_databasesinformationsystems

neighborhood of a document extracting sophisticated statistics from them. Specifically, we propose the following meta-features based on the cosine similarity: (i) normalized number of neighbors from each class; (ii) sample of similarity scores between a document and its neighbors; (iii) deviance from expected values from the neighborhood; (iv) similarity between neighbors and category centroids; (v) entropy observed in the neighborhood and (vi) fisher correlation between similarity scores. As shown in Table 1 there is no clear indication about the best combination of meta-features for each dataset. For instance, the combination of the proposed and existing (literature) meta-features (allMF) achieves the best results on 4UNI, but leads to classification overfitting on REUT.

	Canuto et al.	Yang and Gopal	Pang et al.	Bag of Words	allMF
4UNI	78.9(1.6)	75.6(1.2)	67.6(1.1)	70.7(0.8)	**80.3(1.2)**
REUT	71.5(0.9)	**77.9(1.2)**	71.8(0.8)	65.7(0.7)	71.9(1.3)

Tabela 1. Average MicroF$_1$ effectiveness of SVM with different document representations on two datasets.

To select and analyze the possible combinations of various meta-features for each dataset, we proposed a new adaptive methodology based on multi-objective optimization [Canuto et al. 2018]. We aim at evaluating various combinations of meta-feature groups from a perspective of many objective criteria. Based on the evolutionary multi-objective strategy SPEA2, which presented successful results in related machine learning tasks [Sousa et al. 2019], we proposed two strategies: (1) **SPEA2SVM** – provides the most relevant combinations of meta-feature groups (Pareto frontier) by considering the trade-off between two objectives: (i) maximizing the SVM classification effectiveness and (ii) minimizing the number of meta-feature groups. When considering both objectives, it is possible to guide the method to the most promising regions of the search space; and (2) **SPEA2fast** – efficient evaluation of the trade-off between the maximization of the effectiveness on the following three classification methods as objectives: (i) Naive Bayes, (ii) Nearest Centroid and (iii) Extreme Randomized Trees. These methods were chosen due to their speed and diversity.

According to Table 3, SPEA2SVM and SPEA2fast are capable of finding effective combinations of meta-feature groups without the expensive Brute-force evaluation of all combinations. The combinations found by our methods not only significantly improve over the traditional Bag of Words, but also keep or improve the results of the combination allMF using significantly less meta-feature groups. Figure 2 illustrates the benefits of analyzing effective combinations found by considering the minimization of the number of meta-feature groups with SPEA2SVM. The combination of only two groups that exploit raw centroid distances and cosine similarities among neighbors provides most of the significant information for classification. In this case, the inclusion of two of our statistics (information gain and relationship between centroids and neighbors) produces a combination of four meta-feature groups responsible for almost all of the effectiveness achieved by the best results in 4UNI. This simple analysis allow us to to identify the combined information necessary for effective classification on scenarios of imbalanced and noisy data such as the 4UNI dataset.

2.2. Distance-based Meta-features Enriched with Label Information

Traditional distances aim at summarizing discriminative evidence based on simple manipulations of term weights, which might thwart the importance of relevant discriminative terms. To tackle this problem, we proposed to focus on extending the underlying distance

	20NG	4UNI	REUT	ACM
SPEA2SVM	90.0↑	79.9↕	77.4↑	76.3↕
SPEA2fast	89.9↑	79.6↕	77.0↑	76.1↕
Brute-force	90.1↑	79.8↕	*	76.4↕
Bag of Words	87.6↓	70.7↓	65.7↓	72.1↓
allMF	89.0	80.3	71.9	76.2

Tabela 2. Average MicroF$_1$ effectiveness of SVM on document representations. ↑, ↓ and ↕ correspond to significances over allMF.

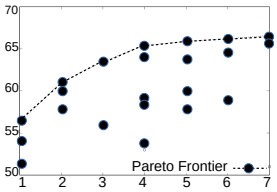

Figura 2. MacroF$_1$ Effectiveness vs number of meta-feature groups on combinations (dots) found by the SPEA2SVM search on the 4UNI dataset.

relationships between a document and its neighbors with supervised strategies that evaluate the relevance of similarity evidence [Canuto et al. 2019]. Particularly, we propose meta-features capable of correlating a set of similarity evidence of a pair of documents with the likelihood of these documents belonging to the same class. We use these likelihoods, predicted with SVM hyperplane distances, to build new meta-features, which are ultimately enriched with labels by the prediction model. We also estimate the level of prediction error introduced by these newly proposed meta-features with the error-rate among the prediction neighbors, providing meta-features to identify hard-to-classify documents.

Table 3 shows the obtained MicroF$_1$ effectiveness for the proposed meta-features enriched with label information, the best strategy so far to combine our previous meta-features (SPEA2SVM) and the recent deep learning CNN method for text (BOW-CNN) [Canuto et al. 2019] that exploits automatically generated and label-enriched meta-features represented in its convolutional layers. Our proposed meta-features consistently achieved the best results in **all** evaluated datasets, a remarkable result.

	20NG	4UNI	REUT	ACM	MEDLINE
Label-enriched *(Proposed)*	**91.6(0.5)**	**83.0(0.6)**	**79.7(1.0)**	**77.9(0.3)**	**87.8 (0.4)**
SPEA2SVM	90.0(0.7)	79.9(1.4)	77.4(1.5)	76.3(0.7)	84.4(0.5)
BOW-CNN	89.3(0.7)	81.3(0.8)	70.9(0.6)	74.9(0.4)	82.5(0.4)

Tabela 3. Average MicroF1 with different meta-feature strategies on five datasets.

2.3. Meta-features for Sentiment Analysis

We studied the application of meta-features for the specific context of sentiment analysis [Canuto et al. 2016], which poses new challenges due to the shortage of information in small messages, the potentially limited number of training samples and noisy texts. To tackle the new challenges, we made use of BM25 as similarity score, since it is an useful measure to rank documents with short messages as queries. We also exploited the neighborhood of a test example in a dataset containing 1.6 million tweets automatically labeled by its users with emoticons. The last additional evidence we exploited was taken from lexicon-based methods to infer the message's polarity towards a sentiment. We compared our proposal with the traditional set of Bag of Words and our best label-enriched meta-features in Table 4. The set of proposed meta-features for sentiment analysis achieved the highest effectiveness in most datasets. This is a strong evidence towards the robustness of the proposed meta-features for the specific task of sentiment analysis, and the potential benefits of adapting meta-features with context-specific information.

	Twitter				Reviews					Comments			
	narr	sand	semev	vader	digg	amazon	rw	movie	yelp	debate	bbc	youtube	myspace
Proposed	**88.8**	**86.5**	**85.8**	**97.2**	**82.1**	**78.0**	**79.8**	**78.6**	93.4	**80.0**	**88.6**	**86.1**	**88.4**
Label-Enrich	85.0	**84.2**	81.9	88.2	78.6	74.2	76.6	77.5	**94.2**	**78.5**	**86.9**	82.2	86.1
Bag of Word	81.4	**84.5**	80.2	84.0	77.6	74.1	76.0	76.9	93.3	77.4	**87.4**	81.2	86.2

Tabela 4. Average MicroF1 for document representations on sentiment analysis datsets.

2.4. Proposed Parallel Implementation

Since the proposed meta-features rely on the computation of distances between documents, unless this procedure is efficiently implemented, their use may have limited applicability. We advance both, the literature on CPU and GPU implementations for meta-features using the kNN method with strategies specially designed for highly dimensional and sparse data. The solution efficiently implements an inverted index in the GPU, by using a parallel counting operations, taking advantage of Zipf's law for saving space [Canuto et al. 2015]. At query time, this inverted index is used to quickly find the documents sharing terms with the query document. Our GPU implementation achieved speedups up to 140 and 16 times on previous CPU and GPU implementations,respectively.

3. Conclusions

This is the first work that thoroughly investigated the impact of different distance-based meta-feature groups for ATC. In this dissertation, we not only proposed a comprehensive set of new meta-features, but also performed a thorough analysis of different vector spaces drawn from them. We provided empirical evidence about the potential benefits of combining groups of meta-features that contain complementary discriminative information. We also investigated potential issues due to the use of more complex and highly dimensional meta-feature spaces. Furthermore, we enriched meta-features with labeled information, adapted meta-features for the sentiment analysis and parallelized the neighborhood search for documents, which is crucial for practical application. The quality of our research is evidenced by the number and excellence of our publications in top conferences and journals as well as in research opportunities being currently pursued in theses and dissertations in development based on research questions we left open or open up by our work.

Referências

Canuto, S., Gonçalves, M. A., and Benevenuto, F. (2016). Exploiting new sentiment-based meta-level features for effective sentiment analysis. In *WSDM*, pages 53–62. ACM.

Canuto, S., Marcos, G., Santos, W., Rosa, T., and Wellington, M. (2015). Efficient and scalable metafeature-based document classification using massively parallel computing. In *SIGIR*, pages 333–342.

Canuto, S., Salles, T., Gonçalves, M. A., Rocha, L., Ramos, G., Gonçalves, L., Rosa, T., and Martins, W. (2014). On efficient meta-level features for effective text classification. In *CIKM*, pages 1709–1718.

Canuto, S., Salles, T., Rosa, T. C., and Gonçalves, M. A. (2019). Similarity-based synthetic document representations for meta-feature generation in text classification. In *SIGIR*, pages 355–364. ACM.

Canuto, S., Sousa, D. X., Goncalves, M. A., and Rosa, T. C. (2018). A thorough evaluation of distance-based meta-features for automated text classification. *IEEE TKDE*, 30:2242–2256.

Cunha, W., Canuto, S., Rosa, T., Gonçalves, M. A., and Rocha, L. (2020). Extended pre-processing pipeline for text classification: On the role of meta-feature representations, sparsification and selective sampling. *Information Processing & Management*, 57(4):32.

Pang, G., Jin, H., and Jiang, S. (2015). Cenknn: a scalable and effective text classifier. *Data Mining and Knowledge Discovery*, 29(3):593–625.

Penha, G., Campos, R. R., Canuto, S. D., Gonçalves, M. A., and Santos, R. L. T. (2019). Document performance prediction for automatic text classification. In *ECIR*, volume 11438, pages 132–139.

Sousa, D., Canuto, S., Gonçalves, M. A., Rosa, T., and Martins, W. (2019). Risk-sensitive learning to rank with evolutionary multi-objective feature selection. *ACM Trans. Inf. Syst.*, 37(2):24:1–24:34.

Sousa, D., Canuto, S., Rosa, T., Martins, W., and Gonçalves, M. A. (2016). Incorporating risk-sensitiveness into feature selection for learning to rank. In *CIKM*, pages 257–266, New York, NY, USA. ACM.

Viegas, F., Canuto, S., Gomes, C., Luiz, W., Rosa, T., Ribas, S., Rocha, L., and Gonçalves, M. A. (2019). Cluwords: Exploiting semantic word clustering representation. In *WSDM*, pages 753–761.

Yang, Y. and Gopal, S. (2012). Multilabel classification with meta-level features in a learning-to-rank framework. *JMLR*, 88:47–68.