Sentiment classification improvement using semantically enriched information

Ricardo B. Scheicher
Roberta A. Sinoara
ricardoxem@usp.br
rsinoara@usp.br
Institute of Mathematics and Computer Science
University of São Paulo
São Carlos - SP, Brazil

Jonas C. Felinto
jonas@itera.com.br
Itera Innovation and Technological Development
São Carlos - SP, Brazil

Solange O. Rezende
solange@icmc.usp.br
Institute of Mathematics and Computer Science
University of São Paulo
São Carlos - SP, Brazil

ABSTRACT

The emergence of new and challenging text mining applications is demanding the development of novel text processing and knowledge extraction techniques. One important challenge of text mining is the proper treatment of text meaning, which may be addressed by incorporating different types of information (e.g., syntactic or semantic) into the text representation model. Sentiment classification is one of the challenging text mining applications. It may be considered more complex than the traditional topic classification since, although sentiment words are important, they may not be enough to correctly classify the sentiment expressed in a document. In this work, we propose a novel and straightforward method to improve sentiment classification performance, with the use of semantically enriched information derived from domain expressions. We also propose a superior scheme for generating these expressions. We conducted an experimental evaluation applying different classification algorithms to three datasets composed by reviews of different products and services. The results indicate that the proposed method enables the improvement of classification accuracy when dealing with reviews of a narrow domain.

CCS CONCEPTS

• **Computing methodologies** → *Knowledge representation and reasoning*.

KEYWORDS

Sentiment analysis, Text classification, Text semantics.

1 INTRODUCTION

Nowadays we are witnessing an increasing interest on the development of different techniques for text processing and knowledge extraction. Several organizations have interest in taking advantages of the available data and a powerful data source that emerges in this context is user reviews. Reviews play an important role in helping people to acquire information about products and services. Besides, they also may be an important data source for companies that offer those reviewed products or services.

Sentiment analysis (or opinion mining) can be defined as a computational field of study that has the goal of organizing and classifying opinions of users (or customers) about products and services [13], which are normally expressed in unstructured texts. Thus, sentiment analysis deals with opinionated texts, which contains people's opinions, appraisals, emotions and sentiments related to entities [1]. One of the different tasks of sentiment analysis is sentiment classification. This task focuses on the categorization of texts according to sentiment orientations and can be conducted in different approaches [7] based on the characteristic of the sentiment classes, which can be binary, ternary, n-ary in the form of stars, and "thumbs up" or "thumbs down", etc. For instance, the binary polarity classification corresponds to the task of labeling an opinionated document as *positive* or *negative*. In ternary polarity classification, the *neutral* class is also considered. Beyond polarity, there are sentiment classification targeting classes of emotional states, such as *angry*, *sad*, and *happy*.

Classification of sentiment polarity or emotional state is more complex than traditional classification tasks, such as topic classification. In topic classification, the set of individual words that occur in the document, represented as a bag-of-words (BOW), is often sufficient to define the document class [5, 12]. For example, in a collection of sports' news, the frequent occurrence of some words such as "tire", "car", "grid", "circuit" are enough to define that the news document belongs to Formula 1 category. In the case of sentiment classification, although sentiment words are important, they may not be enough to correctly classify the sentiment expressed in a document. Thus, solving the problem of sentiment classification may require more than word frequencies and it is important to also incorporate information related to text meaning and knowledge about the text domain.

Recently, good results of sentiment classification has been obtained using word embeddings [5, 14, 15]. The use of such representations based on latent semantics achieves high performance classification, but it negatively affects the interpretability of the text representation features and therefore the explainability of certain classification models. Although classification performance is important, aspects of interpretability or explainability may be crucial for some text mining applications [2, 3]. Thus, alternative methods for the incorporation of text semantics might also be developed.

On the other hand, in certain application domains, users or domain experts have relevant knowledge about the content of the text

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

collection and about the existent classes. This knowledge can be expressed in the form of expressions of the domain, which appears as an alternative approach to obtain semantic information from texts of a specific domain and improve text representation. In this context, previous works have proposed a text representation model based on expressions of domain, named *generalized Bag of Expressions of Domain* (gBoED), which carries privileged information about a specific domain [6, 11]. The authors defined an expression of domain as a pair of terms composed by a domain term and a class identifier term. gBoED is vector-space matrix, whose features are expressions of the domain, and it is suitable for problems whose classes are already known and there are terms that are representative for each class. This is the case of the sentiment classification problem.

In this work, we present a novel method to improve sentiment classification results based on semantically enriched information derived from expressions of the domain. The experimental evaluation, conducted with three datasets and four classification algorithms, indicates that our method is suitable for collections of reviews relating to entities of same nature. In summary, the main contributions of this work are: (i) proposal of a new weighting scheme for gBoED's domain expressions, which presented better performance than the original weighting scheme; and (ii) proposal of the gBoED-based classification improvement method, which applies domain expressions to improve the classification of documents with low predictive confidence (e.g., a classifier prediction probability), maintaining the interpretability and explainability of the solution. The proposals are presented in the next section. In Section 3, we present the experimental evaluation and, in Section 4, the concluding remarks.

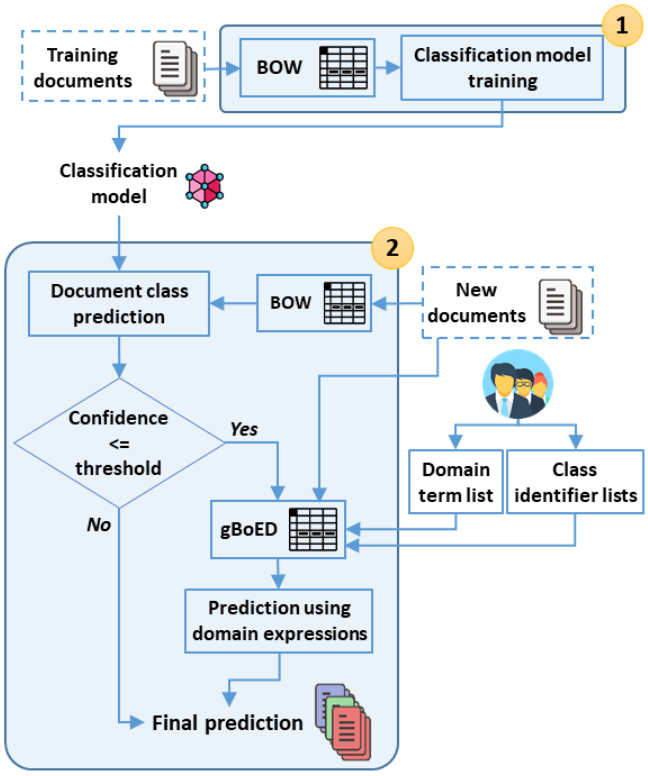

Figure 1: gBoED-based classification improvement method.

2 CLASSIFICATION METHOD

In this paper we propose a method for improving sentiment classification, illustrated in Figure 1. Our method consists of two steps: (1) classification model training, and (2) improvement of classification results using semantically enriched information, obtained from domain expressions. The proposed method applies this semantic information in the classification of instances that are most difficult to classify, i.e., instances that the traditional classifiers have low predictive confidence.

The step (1) corresponds to the training of a classification model based on a bag-of-words representation of the text collection. In this step any classification algorithm may be applied to obtain a classification model. BOW is used since this traditional model achieves good results on simple text classification scenarios.

In the step (2), the semantically enriched information of gBoED is applied to improve classification results for those documents whose prediction confidence is less than or equal to a defined global threshold. First, the new documents are prepared for classification, i.e., a BOW representation is built. Then it is presented to the classification model trained on step (1) to predict the documents' polarity. The model output is a predicted class and a prediction confidence value. If the confidence is higher than a defined threshold, the predicted class is considered as the final prediction to the document or, otherwise, the classification improvement is performed.

For the classification improvement, semantically enriched information are extracted from the documents whose prediction confidence is considered low (i.e., it is less than or equal to the defined threshold). This is performed building a gBoED representation, which is a vector-space model representation whose features are expressions of the domain. An expression of the domain is composed by the union of a domain term and a class identifier. Domain terms are important terms in that domain, whereas class identifies are terms that are important for a specific class. For instance, considering reviews of restaurants, "food" and "service" would be among domain terms, and "good" and "terrible" would be, respectively, among class identifiers for the classes positive and negative. In this case, "food_good" and "service_good" would be domain expressions related to the positive class, whereas "food_terrible" and "service_terrible" would be domain expressions of the negative class. The lists of possible domain terms and identifiers of each specific class are inputs to this step and must be previously built, usually by domain specialists.

In this work we propose the use of two ways of building gBoED representation. Scheicher et al. [11] build gBoED considering the expression's frequency, which is counted when both a domain term and a class identifier occur in the same sentence. In the present work we also propose a novel method to build gBoED. The weight of each feature (domain expression) is given by the multiplicative inverse of the number of words between the corresponding domain term and the class identifier in a sentence. The more words exist between a domain term and a class identifier, the smaller is the weight assigned to the domain expression. When a expression occurs multiple times in the same document, its weight is given by the sum of the weight of its occurrences.

As previously stated, each feature of gBoED is related to a specific class. A global weight for each class is calculated summing the weights of the domain expressions of that class that occurs in the document. Thus, the prediction using domain expressions is based on the global weights of the classes. The predicted class is set as the class that has the higher global weight. If there is a tie between the classes, the original prediction is maintained.

Based on the two versions of gBoED construction, that differ in the way the feature weights are computed, we propose two versions of the gBoED-based classification improvement method. The first version, "*gBoED_Freq*", uses the frequency, i.e., the total number of occurrences of expressions of the domain in a specific document. The second version, "*gBoED_Dist*", uses the weights based on the distance between domain terms and class identifiers.

3 EXPERIMENTAL EVALUATION

In this section we present the experimental setup and discuss the results. Details of the experimental evaluation, including the datasets, the tested parameters and the results, are available at http://sites.labic.icmc.usp.br/ricardoxem/doceng2019.

3.1 Experimental setup

The experimental evaluation was conducted using three datasets.

HuLiu2004: composed by reviews of five different products (two different digital cameras, a cell phone, an MP3 player and a DVD player) [4]. The original dataset has annotations assigning polarities (positive, neutral or negative) to each entity aspect on the reviews. Thus, we considered the most frequent polarity among the evaluated product aspects as the label of the whole review document. In this scenario, *HuLiu2004* dataset has 186 positive and 110 negative reviews.

SemEval2014: composed of reviews of restaurants and laptops, it was created for the SemEval-2014 Aspect Based Sentiment Analysis task 4 [9]. As the original dataset provides annotation of aspect polarities, we determine the global review polarity the same way as *HuLiu2004* dataset, resulting in 1,836 positive and 1,073 negative reviews.

SemEval2015: composed of reviews of restaurants, laptops and hotels, it was created for the SemEval-2015 Aspect Based Sentiment Analysis task 12 [10]. The global review polarity was determined in the same way as the previous two datasets, resulting in 555 positive and 246 negative reviews.

In the first step of the proposed method, we evaluated the use of four traditional classification algorithms, corresponding to different machine learning paradigms: K-Nearest Neighbors (KNN), Multinomial Naive Bayes (MNB), Decision Tree (C4.5) and Support Vector Machine (SVM), all of them implemented in python SciKit Learn library [8]. For each algorithm, we applied several parameter variations. For the second step, the class identifier lists were prepared based on the positive and negative words used by [4] and the domain term list of each dataset were composed by entity aspects available in the respective dataset. We evaluated eleven values for confidence threshold (between 50% and 100%). For each configuration, we only applied the threshold values that were higher than the lower confidence of the respective trained model.

As baseline, we considered the classification performance of the model obtained with the BOW. The classification accuracies were obtained using the 10-fold cross-validation procedure.

3.2 Results and Discussion

The defined experimental evaluation setup resulted in 1,804 classification performance results: 609 of *SemEval2014*, 607 of *HuLiu2004*, and 588 of *SemEval2015*. The differences in number of results for each dataset is a consequence of the predictive confidences obtained by the classification models. We can note that *SemEval2015* presented, in general, a lower number of classification models with predictions with low confidences, thus we had a lower number of executions for this dataset (i.e., low thresholds were not executed for some configurations since every prediction was higher than such threshold). We can also note that *SemEval2015* achieved higher classification accuracies among the different algorithms when compared to the other datasets.

Table 1 presents the best accuracies obtained by the classification model based on BOW and by the two versions of the proposed method (gBoED_Freq and gBoED_Dist). The accuracies are presented for each dataset and for each algorithm among all tested parameters. Values greater than the baseline BOW are highlighted in bold, cells in gray correspond to the best accuracy of each line, and underlined values indicate the higher value between gBoED_Freq and gBoED_Dist. The header line of each dataset corresponds to the best results for the respective dataset.

Analyzing the best accuracies achieved for each dataset and algorithm configurations (Table 1), we can note that the best results of the proposed method was achieved for the dataset *HuLiu2004*. For this dataset, the proposed method improved the highest classification accuracy of BOW in all tested algorithms. The highest improvements obtained by gBoED_Dist were the cases of the algorithm C4.5, in which the best accuracy was improved from 0.68552 (entropy) and 0.68908 (gini) to 0.79425 (entropy and gini). In the case of gBoED_Freq, the highest improvement were achieved for the SVM-rbf, whose best BOW accuracy was 0.79069 and the application of gBoED_Freq improved to 0.87586 (the best accuracy obtained for the dataset).

Considering only the best accuracies for the datasets *SemEval2014* and *SemEval2015*, there was only one case that our method improved the best value obtained by BOW: SVM-poly for *SemEval2014*. When considering each tested configuration individually, there were some cases that gBoED_Freq or gBoED_Dist improved the BOW accuracy. All of these cases were configurations of SVM-poly or SVM-rbf, and in most of them the BOW accuracy was lower than 0.7. It is worth noting that these two datasets have an important difference when compared to *HuLiu2004*. While *HuLiu2004* contains reviews of electronic products, *SemEval2014* and *SemEval2015* contains reviews of both products and services (laptops, restaurants and hotels). This fact may impact the performance of the proposed method since the diversity of entity types of the two SemEval datasets increases the occurrences of inconsistencies among the class identifiers. For example, the term "long" may be positive for battery life and negative for a waiting time in a restaurant queue.

Table 1: Best accuracies for each dataset and algorithm.

Algorithm	BOW	gBoED_Freq	gBoED_Dist
HuLiu2004	0.83092	**0.87586**	**0.84291**
C4.5-entropy	0.68552	**0.76011**	**0.79425**
C4.5-gini	0.68908	**0.76011**	**0.79425**
KNN-cosine	0.75391	**0.76724**	**0.79425**
KNN-euclidean	0.75391	**0.76724**	**0.79425**
MNB	0.79356	0.76011	**0.79425**
SVM-linear	0.82103	**0.83844**	**0.83844**
SVM-poly	0.83092	**0.87586**	**0.86513**
SVM-rbf	0.79069	**0.87586**	**0.86513**
SemEval2014	0.80440	0.79615	0.79649
C4.5-entropy	0.64076	0.58060	0.60363
C4.5-gini	0.65452	0.45686	0.47473
KNN-cosine	0.75594	0.73703	0.74563
KNN-euclidean	0.74905	0.72569	0.73325
MNB	0.80440	0.59400	0.60742
SVM-linear	0.78652	0.78549	0.78549
SVM-poly	0.78962	0.78996	0.78996
SVM-rbf	0.79649	0.79615	0.79649
SemEval2015	0.87142	0.85898	0.85898
C4.5-entropy	0.73412	0.62946	0.63444
C4.5-gini	0.75539	0.73176	0.73302
KNN-cosine	0.80276	0.77026	0.77276
KNN-euclidean	0.80151	0.77275	0.77525
MNB	0.87142	0.65918	0.65918
SVM-linear	0.84773	0.84648	0.84648
SVM-poly	0.85773	0.85773	0.85773
SVM-rbf	0.85898	0.85898	0.85898

Comparing the two versions of the proposal, gBoED_Dist achieved higher accuracy than gBoED_Freq in 1,629 cases. When the threshold was set to 65% or higher, the accuracy obtained by gBoED_Dist was higher than the accuracy of gBoED_Freq in more than 80% of the tested configurations. As the threshold selects the instances to be reclassified, higher thresholds lead to higher number of selected instances and, therefore, increase the coverage of the proposed method. Thus, the results indicates that the weighting scheme based on the distance between terms have a positive impact on the effectiveness of gBoED.

4 CONCLUSION

In this work we proposed and evaluated a gBoED-based classification improvement method, which applies domain expressions to improve the classification of documents with low predictive confidence. The experimental evaluation indicates that our method is suitable when the reviews relates to entities of the same nature. Besides, we also proposed of a new weighting scheme for gBoED's domain expressions based on the distance between terms, which presented better performance than the original frequency-based weighting scheme. As future works, we intend to: (i) expand the experimental evaluation considering state-of-art algorithms and the use of other datasets, making a deeper analysis of the impact of entities diversity on domain expressions; (ii) improve the generation of domain expressions, considering syntactic and semantic role information in automatic terms extraction; and (iii) define a method to determine the best threshold for each dataset.

ACKNOWLEDGMENTS

This study was financed by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) – Finance Code 001 and by grant 2016/17078-0, São Paulo Research Foundation (FAPESP).

REFERENCES

[1] Sedef Cal and Sebnem Yllmaz Balaman. 2019. Improved decisions for marketing, supply and purchasing: Mining big data through an integration of sentiment analysis and intuitionistic fuzzy multi criteria assessment. *Computers & Industrial Engineering* 129 (2019), 315–332.

[2] Filip Karlo Dosilovic, Mario Brcic, and Nikica Hlupic. 2018. Explainable artificial intelligence: A survey. In *41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. 0210–0215.

[3] Bryce Goodman and Seth Flaxman. 2017. European Union Regulations on Algorithmic Decision Making and a "Right to Explanation". *AI Magazine* 38, 3 (2017), 50–57.

[4] Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 168–177.

[5] Hyunjun Ju and Hwanjo Yu. 2018. Sentiment Classification with Convolutional Neural Network using Multiple Word Representations. In *12th International Conference on Ubiquitous Information Management and Communication - IMCOM '18*. ACM Press, 1–7.

[6] Caio A. Nunes Marques, Ivone P. Matsuno, Roberta A. Sinoara, Solange O. Rezende, and Henrique Rozenfeld. 2015. An exploratory study to evaluate the practical application of PSS methods and tools based on text mining. In *20th International Conference on Engineering Design*. 7–311–7–320.

[7] Bo Pang and Lillian Lee. 2008. Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval* 2, 1–2 (2008), 1–135.

[8] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent D. Clermont, Jake Vanderplas, Alexandre Passos, and David Cournapeau. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[9] Maria Pontiki, Dimitrios Galanis, Ion Androutsopoulos, Suresh Manandhar, and Harris Papageorgiou. 2014. SemEval-2014 Task 4: Aspect Based Sentiment Analysis. In *8th International Workshop on Semantic Evaluation*. 27–35.

[10] Maria Pontiki, Dimitrios Galanis, Harris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. 2015. Semeval-2015 task 12: Aspect based sentiment analysis. In *9th International Workshop on Semantic Evaluation*. 486–495.

[11] Ricardo B. Scheicher, Roberta A. Sinoara, Newton J. Koga, and Solange O. Rezende. 2016. Uso de expressões do domínio na classificação automática de documentos. In *XIII Encontro Nacional de Inteligência Artificial e Computacional*. 625 – 636.

[12] Roberta A. Sinoara, Jose Camacho-Collados, Rafael G. Rossi, Roberto Navigli, and Solange O. Rezende. 2019. Knowledge-enhanced document embeddings for text classification. *Knowledge-Based Systems* 163 (2019), 955–971.

[13] Chenhao Tan, Lillian Lee, Jie Tang, Long Jiang, Ming Zhou, and Ping Li. 2011. User-level sentiment analysis incorporating social networks. *17th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (2011), 1397–1405.

[14] Duyu Tang, Furu Wei, Nan Yang, Ming Zhou, Ting Liu, and Bing Qin. 2014. Learning sentiment-specific word embedding for twitter sentiment classification. In *52nd Annual Meeting of the Association for Computational Linguistics*, Vol. 1. 1555–1565.

[15] Shufeng Xiong. 2016. Improving Twitter Sentiment Classification via Multi-Level Sentiment-Enriched Word Embeddings. *Computing Research Repository (CoRR)*.