An Information Theoretic Approach to Text Sentiment Analysis

David Pereira Coutinho[1,3] and Mário A. T. Figueiredo[2,3]
[1]*Instituto Superior de Engenharia de Lisboa, Lisboa, Portugal*
[2]*Instituto Superior Técnico, Lisboa, Portugal*
[3]*Instituto de Telecomunicações, Lisboa, Portugal*

Keywords: Sentiment Analysis, Binary Categorization, Ziv-Merhav Method, Cross Parsing, Lempel-Ziv Algorithm.

Abstract: Most approaches to text sentiment analysis rely on human generated lexicon-based feature selection methods, supervised vector-based learning methods, and other solutions that seek to capture sentiment information. Most of these methods, in order to yield acceptable accuracy, require a complex preprocessing stage and careful feature engineering. This paper introduces a coding-theoretic-based sentiment analysis method that dispenses with any text preprocessing or explicit feature engineering, but still achieves state-of-the-art accuracy. By applying the *Ziv-Merhav method* to estimate the relative entropy (Kullback-Leibler divergence) and the cross parsing length from pairs of sequences of text symbols, we get information theoretic measures that make very few assumptions about the models which are assumed to have generated the sequences. Using these measures, we follow a dissimilarity space approach, on which we apply a standard support vector machine classifier. Experimental evaluation of the proposed approach on a text sentiment analysis problem (more specifically, movie reviews sentiment polarity classification) reveals that it outperforms the previous state-of-the-art, despite being much simpler than the competing methods.

1 INTRODUCTION

The task of automatically classifying a text, not in terms of topic, but according to the overall sentiment it expresses, is the objective of text *sentiment analysis* (SA). A particular instance of this task is that of determining whether a user review (*e.g.*, of a movie, or a book) is positive or negative, that is, determining the so-called *sentiment polarity*. To solve this binary categorization problem, different approaches have been proposed in the literature. Most of those approaches rely on human-generated lexicon-based feature selection methods, based on which it is possible to build supervised vector-based learning methods. The key drawback of those methods is that they demand a complex preprocessing stage and can only achieve acceptable accuracy with careful lexicon and feature design/engineering.

In this paper, we propose a new information-theoretic approach to text sentiment analysis, and illustrate it in the particular case of binary sentiment polarity categorization. The proposed method does not use any of the classical text preprocessing steps, such as stop-word removal or stemming. The proposed method follows earlier work (Pereira Coutinho and Figueiredo, 2005) in that it is based on the *Ziv-Merhav method* (ZMM) for the estimation of relative entropies (or Kullback-Leibler divergences) between pairs of sequences of text symbols, with these estimates serving as features, based on which a classifier (*e.g.*, a support vector machine – SVM) can be built.

The seminal work on the text sentiment analysis problem was published in 2002 by Pang and Lee (Pang et al., 2002), who focused on movie review sentiment polarity categorization. The method proposed by those authors is based on a human-generated lexicon, based on which bag-of-words (BoW) descriptions of the texts were obtained and used as feature vectors by an SVM classifier. Due to the success of Joachims (Joachims, 1998) in dealing with text classification problems by combining SVM classifiers with BoW-based vector space models, many researchers have followed similar approaches.

In this work, we aim at dispensing with the human-generated lexicon for building BoW features, or the need for any other feature design or engineering. For that purpose, we partially follow previous work (Pereira Coutinho and Figueiredo, 2005) in that we use the ZMM as a *model-free* feature extractor that doesn't require any human intervention. We adopt

Pereira Coutinho D. and A. T. Figueiredo M. (2013).
An Information Theoretic Approach to Text Sentiment Analysis.
In *Proceedings of the 2nd International Conference on Pattern Recognition Applications and Methods*, pages 577-580
DOI: 10.5220/0004269005770580

the dissimilarity space approach (see (Pekalska et al., 2001), (Pekalska and Duin, 2002), and references therein); in particular, we characterize each text by the vector of its ZMM-based dissimilarity values with respect to (all or a subset of the) other texts in the training set. Finally, a standard SVM is used as a classifier. We stress again that the crucial aspect of the proposed approach is that it dispenses with any preprocessing (such as stop-word removal and word stemming) or any human-based feature design. Still, as shown in the experiments reported below, our approach establishes a new state-of-the-art accuracy in a benchmark movie review sentiment polarity categorization dataset.

The outline of the paper is has follows. Section 2 discusses some previous work and results in text sentiment analysis. Section 3 introduces the fundamental tools used in our approach and provides details about our categorization method. Our experiments and analysis of the results are presented in section 4, and finally conclusions are presented in Section 5.

2 RELATED WORK

Starting with the seminal work of Joachims (Joachims, 1998), SVM classifiers have been one of the weapons of choice when dealing with topic-based text classification. These SVM classifiers typically work on vector spaces where each text is characterized by a bag of words (BoW) or bag of pairs of words (word bi-grams). It was thus not surprising that the initial attempts at addressing text sentiment analysis (which of course is just a special type of text categorization) were also based on SVM tools and BoW-type features (Pang et al., 2002). The early work of Pang and Lee, using this type of approach, provided a strong baseline accuracy of 82.9% in a task of movie reviews sentiment polarity (binary) classification.

Since then, the movie review dataset (also known as the sentiment polarity dataset) used in (Pang et al., 2002), (Pang and Lee, 2004) has become a benchmark for many sentiment classification studies. We now recall some of the best result to date on this dataset.

Whitelaw and collaborators (Whitelaw et al., 2005), reported an accuracy of 90.2%. Their method is based on so-called *appraisal groups*, which are defined as coherent groups of words around adjectives that together express a particular opinion, such as "very funny" or "not terribly surprising". After building an apraisal lexicon (manually verified) it uses a combination of different types of appraisal group features and BoW features for training an SVM classifier. The state-of-the-art accuracy was established by Matsumoto and collaborators (Matsumoto et al., 2005). They proposed a method where information about word order and syntactic relations between words in a sentence is used for training a classifier. Thus using the extracted word sub-sequences and dependency sub-trees as features for SVMs training they attained an accuracy of 93.7%.

More recently Yessenalina and colleagues (Yessenalina et al., 2010) proposed a supervised multi-level structured model based on SVMs, which learns to jointly predict the document label and the labels of a sentence subset that best explain the document sentiment. The authors treated the sentence-level labels as hidden variables so the proposed model does not required sentence-level annotation for training, avoiding this way the lowerlevel labellings cost. They formulate the training objective to directly optimize the document-level accuracy. This multi-level structured model achieved 93.22% document-level sentiment classification accuracy on the movie review dataset.

These results and references are summarized in Table 1. Further examples can be found in the survey paper (Vinodhini and Chandrasekaran, 2012), but none with better accuracy results for this dataset than those mentioned above, and all usually involving complex preprocessing stages and careful feature engineering.

Table 1: Baseline and best reported classification accuracies in the literature over the same collection of movie reviews.

Method	Accuracy [%]
(Pang et al., 2002)	82.9
(Pang and Lee, 2004)	87.2
(Whitelaw et al., 2005)	90.2
(Matsumoto et al., 2005)	93.7
(Yessenalina et al., 2010)	93.2
Proposed approach (linear SVM)	96.9
Proposed approach (7-NN)	95.6

3 PROPOSED APPROACH

3.1 The Ziv-Merhav Method

The Ziv-Merhav Method (ZMM) was introduced in 1993 (Ziv and Merhav, 1993) for measuring relative entropy between pairs of sequences of symbols. It is based on the incremental Lempel-Ziv (LZ) parsing algorithm (Ziv and Lempel, 1978) and on a variation thereof, known as cross parsing. Combining these two algorithms, the authors defined an estimate of

the relative entropy that can be used as a dissimilarity measure.

The LZ algorithm is a well-known tool for text compression (Salomon and Motta, 2010), which in recent years has also been used for text/sequence classification purposes; for example, in (Pereira Coutinho and Figueiredo, 2005), LZ-based dissimilarity measures were used to achieve state-of-the-art performance in a specific text classification task (authorship attribution).

An implementation of the cross parsing algorithm was proposed in (Pereira Coutinho and Figueiredo, 2005), based on a modified LZ77 (Ziv and Lempel, 1977) algorithm, where the dictionary is static and only the lookahead buffer slides over the input sequence, as shown in Figure 1 (for more details, see (Pereira Coutinho and Figueiredo, 2005)). This very same implementation, using a 2 Mbyte dictionary and a 256 byte look ahead buffer, was used in the experiments reported below.

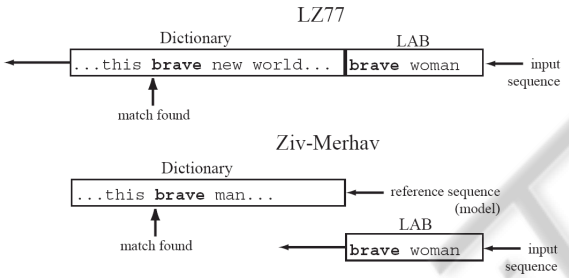

Figure 1: The original LZ77 sliding window and the modified implementation for cross parsing.

Whereas in (Pereira Coutinho and Figueiredo, 2005), the ZMM was applied to compute text dissimilarities, which were then used by a K-nearest-neighbors (K-NN) classifier, here we propose to use the ZMM to build a dissimilarity space representation of the texts, following the framework proposed in (Pekalska et al., 2001), (Pekalska and Duin, 2002), and reviewed in the next subsection.

3.2 Dissimilarity-based Classification

Let us consider a given training set of objects (movie reviews, in the example considered in this paper) $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_n\}$, where each object belongs to some set $\mathcal{X}$ (*e.g.*, the set of finite length strings of some finite alphabet) and some dissimilarity measure between pairs of objects, $D : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. In the dissimilarity-based approach, each object (either in the training set or a new object to be classified after training) is represented by the vector of its dissimilarities with respect to the elements of $\mathbf{X}$ (or a subset thereof). That is, the training set in the so-called dissimilarity space becomes

$$\mathcal{D} = \{\mathbf{d}_1, ..., \mathbf{d}_n\},$$

where

$$\mathbf{d}_i = \begin{bmatrix} D(\mathbf{x}_i, \mathbf{x}_1) \\ \vdots \\ D(\mathbf{x}_i, \mathbf{x}_n) \end{bmatrix} \in \mathbb{R}^n.$$

An important aspect of dissimilarity-based approaches is that very few conditions are put of the dissimilarity measure; namely, it doesn't have to be a metric, it doesn't even need to be symmetric (Pekalska et al., 2001), (Pekalska and Duin, 2002). Dissimilarity representations can also be based on a subset of the training set, in which case the dissimilarity space has dimension equal to the cardinal of that subset; some work has been devoted to methods for selecting this subset (Duin and Paclk, 2006).

In this paper, we propose to build the dissimilarity-based representation by using the relative entropy estimate between pairs of sequences of symbols, as measured by the Ziv-Merhav method described in the previous subsection.

Once in possession of a dissimilarity-based representation of a training set, any standard classification method that works on vector spaces can be used. In this paper, we report preliminary results by using (linear) SVM and K-NN classifiers. As shown in the experiment results below, this simple approach already achieves results that outperform the previous state-of-the-art, although it is conceptually much simpler and requires much less human intervention. In future work, even better results may be obtained by exploring other possibilities, such as other kernels, tuning of the SVM C parameter, and better strategies to select a subset of objects with respect to which the dissimilarity representations are obtained.

4 EXPERIMENTAL SETUP

In the experimental evaluation of the proposed approach, we use the polarity dataset[1] v2.0, introduced by (Pang and Lee, 2004); this (human classified) dataset includes 1,000 positive and 1,000 negative movie reviews. The dataset is split into a training set with 900 examples per class and then into 10 cross-validation (CV) folds. We report CV accuracy estimates, following the same protocol of (Pang and Lee, 2004), where in each run, 1800 examples are used to train and 200 examples to test. We stress, that we don't use any text preprocessing.

We use K-NN and SVM classifiers (with linear kernel), implemented by the PRTools Matlab toolbox

[1] www.cs.cornell.edu/people/pabo/movie-review-data

for pattern recognition [2] (version 4). The value of the C parameter in the SVM was set to one.

5 RESULTS

We compare the accuracy of the proposed approach with respect to the methods described in Section 2 in the movie review sentiment polarity classification problem using the dataset described in the previous section. The results shown in Table 1 reveal that the proposed approach with the SVM outperforms the previous state-of-the-art, by achieving an average accuracy of 96.9%. Regarding the K-NN classifier, the best accuracy (95.55%) was obtained with $K = 7$, although the result for $K = 1$ (95.45%) is almost identical and also outperforms the previous state-of-the-art.

Finally, we also explored the random prototype selection method proposed by Duin et al. (Duin and Paclk, 2006); the results are shown in Figure 2. Notice that using only 10% of the prototypes for training (180) we still get an accuracy of 95.0%, better than the previous state-of-the-art.

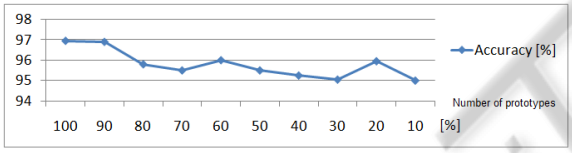

Figure 2: Average accuracy in the dissimilarity space as a function of the number of randomly prototypes.

6 CONCLUSIONS

In this paper, we have presented a new approach for text sentiment analysis, based on an information-theoretic dissimilarity measure, which is used to build dissimilarity representations on which SVM and K-NN classifiers are applied. The aim of our proposal was mainly to show that this type of approach allows achieving state-of-the-art results in hard text classification problems, while involving virtually no human intervention and no text preprocessing. We have illustrated the approach on a benchmark dataset, where the task is to perform movie review sentiment polarity categorization. Our methods outperform previous state-of-the-art results, although it is drastically simpler and requires much less human intervention.

[2] www.prtools.org/index.html

REFERENCES

Duin, R. P. W. and Paclk, P. (2006). Prototype selection for dissimilarity-based classifiers. *Pattern Recognition*, 39:189–208.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features.

Matsumoto, S., Takamura, H., and Okumura, M. (2005). Sentiment classification using word sub-sequences and dependency sub-trees. In *Proceedings of the 9th Pacific-Asia conference on Advances in Knowledge Discovery and Data Mining*, PAKDD'05, pages 301–311, Berlin, Heidelberg. Springer-Verlag.

Pang, B. and Lee, L. (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *In Proceedings of the ACL*, pages 271–278.

Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up? sentiment classification using machine learning techniques. In *IN PROCEEDINGS OF EMNLP*, pages 79–86.

Pekalska, E. and Duin, R. P. W. (2002). Dissimilarity representations allow for building good classifiers. *Pattern Recognition Letters*, 23(8):943–956.

Pekalska, E., Paclk, P., and Duin, R. P. W. (2001). A generalized kernel approach to dissimilarity-based classification. *Journal of Machine Learning Research*, 2:175–211.

Pereira Coutinho, D. and Figueiredo, M. (2005). Information theoretic text classification using the Ziv-Merhav method. *2nd Iberian Conference on Pattern Recognition and Image Analysis – IbPRIA'2005*.

Salomon, D. and Motta, G. (2010). *Handbook of Data Compression (5. ed.)*. Springer.

Vinodhini, G. and Chandrasekaran, R. (2012). Sentiment analysis and opinion mining: A survey. *International Journal of Advanced Research in Computer Science and Software Engineering*.

Whitelaw, C., Garg, N., and Argamon, S. (2005). Using appraisal taxonomies for sentiment analysis. In *Proceedings of CIKM-05, the ACM SIGIR Conference on Information and Knowledge Management*, Bremen, DE.

Yessenalina, A., Yue, Y., and Cardie, C. (2010). Multi-level structured models for document-level sentiment classification. In *In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP*.

Ziv, J. and Lempel, A. (1977). A universal algorithm for sequential data compression. *IEEE Transactions on Information Theory*, 23(3):337–343.

Ziv, J. and Lempel, A. (1978). Compression of individual sequences via variable-rate coding. *IEEE Transactions on Information Theory*, 24(5):530–536.

Ziv, J. and Merhav, N. (1993). A measure of relative entropy between individual sequences with application to universal classification. *IEEE Transactions on Information Theory*, 39:1270–1279.