Feature Weighting Strategies in Sentiment Analysis

Olena Kummer and Jacques Savoy

Rue Emile-Argand 11, CH-2000 Neuchâtel
{olena.zubaryeva,jacques.savoy}@unine.ch
http://www2.unine.ch/iiun

Abstract. In this paper we propose an adaptation of the Kullback-Leibler divergence score for the task of sentiment and opinion classification on a sentence level. We propose to use the obtained score with the SVM model using different thresholds for pruning the feature set. We argue that the pruning of the feature set for the task of sentiment analysis (SA) may be detrimental to classifiers performance on short text. As an alternative approach, we consider a simple additive scheme that takes into account all of the features. Accuracy rates over 10 fold cross-validation indicate that the latter approach outperforms the SVM classification scheme.

Keywords: Sentiment Analysis, Opinion Detection, Kullback-Leibler divergence, Natural Language Processing, Machine Learning

1 Introduction

In this paper we consider sentiment and opinion classification on a sentence level. Sentiment analysis of user reviews, and short text in general could be of interest for many practical reasons. It represents a rich resource for marketing research, social analysts, and all interested in following opinions of the mass. Opinion mining can also be useful in a variety of other applications and platforms, such as recommendation systems, product ad placement strategies, question answering, and information summarization.

The suggested approach is based on a supervised learning scheme that uses feature selection techniques and weighting strategies to classify sentences into two categories (opinionated vs. factual or positive vs. negative). Our main objective is to propose a new weighting technique and classification scheme able to achieve comparable performance to popular state-of-the-art approaches, and to provide a decision that can be understood by the final user (instead of justifying the decision by considering the distance difference between selected examples).

The rest of the article is organized as follows. First, we present the review of the related literature in Section 2. Next, we present the adaptation of the Kullback-Leibler divergence score for opinion/sentiment classification in Section 3. Section 4 provides a description of the experimental setup and corpora used. Sections 5 and 6 present experiments and analysis of the proposed weighting

measure with the SVM model, and additive classification scheme respectively. Finally, we give conclusions in Section 7.

2 Related Literature

Often as a first step in machine learning algorithms, like SVM, naïve Bayes, k-Nearest Neighbors, one uses feature weighting and/or selection based on the computed weights. The selection of features allows decrease of the dimensionality of the feature space and thus the computational cost. It can also reduce the overfitting of the learning scheme to the training data. Several studies expose the feature selection question. Forman [1] reports an extensive evaluation of various schemes in text classification tasks. Dave *et al.* [2] give an evaluation of linguistic and statistical measures, as well as weighting schemes to improve feature selection.

Kennedy *et al.* [3] use General Inquirer [4] to classify reviews based on the number of positive and negative terms that they contain. General Inquirer assigns a label to each sense of the word out of the following set: *positive*, *negative*, *overstatement*, *understatement*, or *negation*. Negations reverse the term polarity while overstatement and understatements intensify or diminish the strength of the semantic orientation.

In the study carried out by Su *et al.* [5] on MPQA (Multi-Perspective Question Answering) and movie reviews corpora it is shown that publicly available sentiment lexicons can achieve the performance on par with the supervised techniques. They discuss opinion and subjectivity definitions across different lexicons and claim that it is possible to avoid any annotation and training corpora for sentiment classification. Overall, it has to be noted that opinion words identified with the use of the corpus-based approaches may not necessarily carry the opinion itself in all situations. For example, *He is looking for a good camera on the market*. Here, the word *good* does not indicate that the sentence is opinionated or expresses a positive sentiment.

Pang *et al.* [6] propose to first separate subjective sentence from the rest of the text. They assume that two consecutive sentences would have similar subjectivity label, as the author is inclined not to change sentence subjectivity too often. Thus, labeling all sentences as objective and subjective they reformulate the task of finding the minimum s-t cut in a graph. They carried out experiments on the movie reviews and movie plot summaries mined from the Internet Movie DataBase (IMDB), achieving an accuracy of around 85%.

A variation of the SVM method was adopted by Mullen *et al.* [7] who use WordNet syntactic relations together with topic relevance to calculate the subjectivity scores for words in text. They report an accuracy of 86% on the Pang *et al.* [8] movie review dataset. An improvement of one of the IR metrics is proposed in [9]. The so-called "Delta TFIDF" metric is used as a weighting scheme for features. This metric takes into account how the words are distributed in the positive vs. negative training corpora. As a classifier, they use SVM on the movie review corpus.

Paltoglou *et al.* [10] explore IR weighting measures on publicly available movie review datasets. They have good performance with BM25 and smoothing, showing that it is important to use term weighting functions that scale sublineary in relation to a number of times a term occurs in the document. They underline that the document frequency smoothing is a significant factor.

3 KL Score

In our experiments we adopted a feature selection measure described in [11] that is based on the Kullback-Leibler divergence (KL-divergence) measure. In this paper, the author seeks to find a measure that would lower the score of the features that have different distribution in the individual training documents of a given class from the distribution in the whole corpus. Thus, the scoring function would allow to select features that are representative of all documents in the class leading to more homogeneous classes. The scoring measure based on KL-divergence introduced in [11] yields an improvement over MI with naïve Bayes on Reuters dataset, frequently used as a text classification benchmark.

Schneider [11] shows how we can use the KL-divergence of a feature f_t over a set of training documents $S = d_1, ..., d_{|S|}$ and classes c_j, $j = 1, ..., |C|$ is given in the following way:

$$KL_t(f) = \tilde{K_t}(S) - \tilde{KL_t}(S) \tag{1}$$

where $\tilde{K_t}(S)$ is the average divergence of the distribution of f_t in the individual training documents from all training documents. The difference $KL_t(f)$ in the Equation 1 is bigger if the distribution of a feature f_t is similar in the documents of the same class and dissimilar in documents of different classes.

$\tilde{K_t}(S)$ is defined in the following way:

$$\tilde{K_t}(S) = -p(f_t) \log q(f_t) \tag{2}$$

where $p(f_t)$ is the probability of occurrence of feature f_t (in the training set). This probability could be estimated as the number of occurrences of f_t in all training documents, divided by the total number of features. Let N_{jt} be the number of documents in c_j that contain f_t, and $N_t = \sum_{j=1}^{|C|} N_{jt}/|S|$. Then $q(f_t|c_j) = \sum_{j=1}^{|C|} N_{jt}/|c_j|$ and $q(f_t) = N_t/|S|$. The second term from 1 is defined as follows:

$$\tilde{KL_t}(S) = -\sum_{j=1}^{|C|} p(c_j) p(f_t|c_j) \log q(f_t|c_j) \tag{3}$$

where $p(c_j)$ is the prior probability of category c_j, and $p(f_t|c_j)$ is the probability that the feature f_t appears in a document belonging to the category c_j. Using the maximum likelihood estimation with a Laplacean smoothing, Schneider [11] obtains:

$$p(f_t|c_j) = \frac{1 + \sum_{d_i \in c_j} n(f_t, d_i)}{|V| + \sum_{t=1}^{|V|} \sum_{d_i \in c_j} n(f_t, d_i)} \tag{4}$$

where $|V|$ is the training vocabulary size or the number of features indexed, $n(f_t, d_i)$ is the number of occurrences of f_t in d_i. It is important to note that the afore mentioned average diversion calculations are really approximations based on two assumptions: the number of occurrences of f_t is the same in all documents containing f_t, and all documents in the class c_j have the same length. These two assumptions may turn detrimental for long extract text classification as noted by the author himself [11], but turn out quite effective for a sentence classification setup where a phrase mostly consists of features that occur once, with usually low variations in sentence length. It is important to note that the computation of $p(f_t|c_j)$ should be done on a feature set with removed outliers, since they occur in all or almost all sentences in the corpora.

In sentence-based classification the pruning of the feature set can turn out quite detrimental to the classification accuracy. This is true if the size of the training set is not big enough in order to be sure that some important for classification features are not discarded. Thus, we propose to modify the KL-divergence measure for sentiment and opinion classification. In [11] it calculates the difference between the average divergence of the distribution of f_t in individual training documents from the global distribution, all this averaged over all training documents in all classes. For the sentiment/opinion classification task it is interesting to calculate the difference between the average divergence in one class from the distribution over all classes. Therefore, we can obtain the average divergence of the distribution of f_t for each of the classification categories ($j \in POS, NEG$):

$$\tilde{KL_t}^{j}(S) = N_{jt} \cdot \tilde{p}_d(f_t|c_j) log \frac{\tilde{p_d}(f_t|c_j)}{p(f_t|c_j)} \tag{5}$$

Substituting $\tilde{KL_t}^{POS}(S)$ and $\tilde{KL_t}^{NEG}(S)$ in Equation 1 for each category we obtain measures that evaluate how different is the distribution of feature f_t in one category from the whole training set.

$$KL_t^{POS}(f) = \tilde{K_t}^{POS}(S) - \tilde{KL_t}^{POS}(S) \tag{6}$$

$$KL_t^{NEG}(f) = \tilde{K_t}^{NEG}(S) - \tilde{KL_t}^{NEG}(S) \tag{7}$$

This way, we obtain two sums $\sum KL_t^{POS}(f)$ and $\sum KL_t^{NEG}(f)$ over the features present in the sentence. The final difference of the two sums (denoted further as KL score) can serve as a prediction score of to which category the sentence is most similar.

4 Experimental Setup and Dataset Description

We use the setup with unigram indexing, short stop word elimination (several prepositions and verb forms: *a, the, it, is, of*) and the use of the Porter stemmer [12]. All reported experiments use 10 fold cross-validation setup.

In our study we use three publicly available datasets that we chose based on their popularity as benchmark datasets in SA research. The first one is Sentence Polarity dataset v1.0[1] [13]. It contains 5331 positive and 5331 negative snippets of movie reviews, each review is one sentence long. The Subjectivity dataset contains 5000 subjective and 5000 objective sentences [6].

As a third dataset, that contains newspaper articles, we use the MPQA dataset[2] [14]. The problem with the MPQA dataset is that the annotation unit is at the phrase level, which could be a word, part of a sentence, a clause, a sentence itself, or a long phrase. In order to obtain a dataset with a sentence as a classification unit, we used the approach proposed by Wilson *et al.* [14]. They define the sentence-level opinion classification in terms of the phrase-level annotations. A sentence is considered opinionated if:

1. It contains a "GATE_direct-subjective" annotation with the attribute intensity not in ['low', 'neutral'] and not with the attribute 'insubstantial';
2. The sentence contains a "GATE_expressive-subjectivity" annotation with attribute intensity not in ['low'].

Here is the information on corpus statistics as reported in [14]: there are 15,991 subjective expressions from 425 documents, containing 8,984 sentences. After parsing, we obtained 6,123 opinionated and 4,989 factual sentences.

5 KL Score and SVM

We were interested in evaluating the features selected by our method with the use of the SVM classifier. As pointed out in [15], SVM is able to learn a model independent of the dimension of the space with few irrelevant features present. The experiments on text categorization task show that even the features, that are ranked low according to their IG, are still relevant and contain the information needed for successful classification. Another particularity of the text classification tasks in the context of the SVM method is the sparsity of the input vector, especially when the input instance is a sentence, and not a document.

Joachims [15] observed that the text classification problems are usually linearly separable. Thus, a lot of the research dealing with text classification uses linear kernels [1]. In our experiments we used SVM^{light} implementation with the linear kernel with the soft-margin constant $cost = 2.0$ [16]. We chose *cost* value based on the experimental results. Generally, the low *cost* value (by default 0.01) indicates a bigger error tolerance during training. With the growth of the *cost* value the SVM model assigns larger penalty for margin errors.

We also experimented with other types of kernels, namely with the radial basis function kernel. From our experiments, learning of the SVM model with this kernel takes substantially longer time and gives approximately the same level of the performance as the linear kernel.

[1] http://www.cs.cornell.edu/people/pabo/movie-review-data
[2] http://www.cs.pitt.edu/mpqa/

Dataset	% of features		
	60%	*80%*	*100%*
Polarity	65.93	65.93	65.88
Subjectivity	84.72	84.72	84.69
MPQA	68.42	68.42	68.46

Table 1. Accuracy of SVM^{light} with the linear kernel ($\gamma = 2.0$) and different percentage of features.

We prune the ranked features by the score, accounting for at least 60% of the feature set. This is due to the fact that further pruning of features leads to drastic degradation in accuracy. Further elimination of features from the training model leads to the situation when some testing sentences are represented with one or two features only. The pruning of the feature set up to 60% and 80% of top ranked features did not ameliorate the accuracy of the KL score and the SVM model. In the next section, we discuss possible reasons for degradation in accuracy when pruning the feature set for the task of SA classification on a sentence level, and propose a simple additive classification scheme.

6 Additive Classification Model Based on KL Score

In text classification, after calculating the scores between every feature and every category the next steps are to sort the features by score, choose the best k features and use them later to train the classifier. For the task of sentiment classification on a sentence-based level the pruning of the feature set may lead to the elimination of infrequent features (several occurrences) and may cause the loss of important information needed for classification of the new instances. Here are some differences in the aspects of use of the feature selection measures in text classification and opinion/sentiment analysis contexts. First, the aim in topic text classification is to look for the set of topic-specific features that describe the classification category. In sentiment classification, though, the markers of the opinion could be carried by both topic-specific and context words that may also have small differences in distributions across categories due to the short text length. If we look at the opinion review domain, the topic-specific features would be *movie*, *film*, *flick* and context words would be (*long*, *short*, *horror*, *satisfy*, *give up*).

Second, the usual text classification methods are designed for documents consisting of at least several hundreds of words, assuming that the features that could aid in classification repeat across the text several times. The format of a sentence does not let us make the same assumption. The opinion or sentiment polarity can be expressed with the help of one word/feature. There is substantial evidence from several studies that the presence/absence of a feature is a better indicator than the tf scores [17].

Thus, for effective classification, the model should identify features that are strong indicators of opinion/sentiment, take into account the relations between

the features in each category, and be able to adjust scores of the features that were not frequent enough in order to expand the set of features that are strong indicators of the sentiment.

As a classification model we use a simple additive score of the features in the sentence computed for each category. Our aim is to determine the behavior of the KL score for the task of sentence sentiment and opinion classification in terms of its goodness and priority in feature weighting based on feature distribution across classification categories.

Dataset	**Prec.**	**Recall**	**F1**	**Acc.**
Polarity	67.26%	72.01%	69.55%	68.48%
Subjectivity	91.17%	90.64%	90.90%	90.93%
MPQA	75.53%	61.39%	67.69%	65.07%

Table 2. Precision, recall, F1-measure, and accuracy of all metrics over the three corpora: Movie Review, Subjectivity and MPQA datasets.

From the results presented in the Table 2, we can see that a simple classification scheme based on computing the sum of the feature scores according to the classification category outperforms the SVM model on the sentence datasets. As we deal with a small number of features, it is advantageous to use all of them when taking a classification decision. Comparing with the results in Table 1, we have achieved an improvement in accuracy for the Polarity and Subjectivity datasets. Nevertheless, the SVM model gives better results for the MPQA corpus. This may be due to the stylistics and opinion annotation and expression differences in movie and newspaper domains. The former is usually much more expressive, containing more sentiment-related words, than the latter.

7 Conclusions

In this article we suggest a new adaptation of the Kullback-Leibler divergence score as a weighting measure for sentiment and opinion classification. The proposed score, named KL score, we use for feature weighting with the SVM model. The experiments showed that the pruning of the feature set does not improve the SVM performance. Taking into account the differences in topical and sentiment classification of short text, we proposed a simple classification scheme based on calculation of sum of the features present in the sentence according to each classification category. Surprisingly, this scheme yields better results than SVM.

Based on the three well-known test-collections in the domain (Sentence Polarity, Subjectivity and MPQA datasets), we suggested a new way of computing feature weights, that could be later used with SVM or other supervised classification schemes that use feature weight computation. The proposed score and classification model were successfully applied in two different contexts (sentiment and opinion) and two domains (movie review and news articles).

References

1. Forman, G.: An extensive empirical study of feature selection metrics for text classification. The Journal of Machine Learning Research, Special Issue on Variable and Feature Selection, vol. 3, pp. 1289–1305. (2003)
2. Dave, K., Lawrence, S., Pennock, D. M.: Mining the peanut gallery: opinion extraction and semantic classification of product reviews. In Proceedings of the WWW Conference, pp. 519–528. (2003)
3. Kennedy, A., Inkpen, D.: Sentiment classification of movie reviews using contextual valence shifters. In Journal of Computational Intelligence, vol. 22(2), pp. 110–125. (2006)
4. Stone, P.J.: The General Inquirer: a computer approach to content analysis. The MIT Press. (1966)
5. Su, F., Markert, K.: From words to senses: a case study of subjectivity recognition. In Proceedings of the 22nd International Conference on Computational Linguistics, pp. 825–832. (2008)
6. Pang, B., Lee, L.: A sentimental education: sentiment analysis using subjectivity summarization based on Minimum Cuts. In Proceedings of the 42nd Annual Meeting of the ACL. (2004)
7. Mullen, T., Collier, N.: Sentiment analysis using Support Vector Machines with diverse information sources. In Proceedings of the EMNLP Conference, pp. 412–418. (2004)
8. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up?: sentiment classifcation using machine learning techniques. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 79–86. (2002)
9. Martineau, J., Finin, T.: Delta TFIDF: an improved feature space for sentiment analysis. In Proceedings of the AAAI Conference on Weblogs and Social Media. (2009)
10. Paltoglou, G., Thelwall, M.: A study of information retrieval weighting schemes for sentiment analysis. In Proceedings of the 48th Annual Meeting of the ACL, pp. 1386–1395. (2010)
11. Schneider, K.M.: A new feature selection score for multinomial naïve Bayes text classification based on KL-divergence. Proceedings of the 42nd Annual Meeting of the ACL. (2004)
12. Porter, M. F.: Readings in information retrieval. Morgan Kaufmann Publishers Inc. (1997)
13. Pang, B., Lee, L.: Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales. In Proceedings of the 43rd Annual Meeting of ACL, pp. 115–124. (2005)
14. Wilson, T., Wiebe, J., Hoffmann, P.: Recognizing contextual polarity in phrase-level sentiment analysis. In Proceedings of the HLT and EMNLP, pp. 354–362. (2005)
15. Joachims, T.: Text categorization with Support Vector Machines: learning with many relevant features. In Proceedings of the European Conference on Machine Learning, pp. 137–142. (1998)
16. Joachims, T.: Making large-scale (SVM) learning practical. Advances in Kernel Methods - Support Vector Learning, pp. 169–184. MIT Press, Cambridge, MA. (1999)
17. Pang, B., Lee, L.: Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, vol. 2(1-2). (2008)