Comparison of Feature Selection Methods for Sentiment Analysis

Chris Nicholls and Fei Song

Dept. of Computing and Information Science
University of Guelph
Guelph, Ontario, Canada N1G 2W1
{cnicholl,fsong}@uoguelph.ca

Abstract. Sentiment analysis is a sub-field of Natural Language Processing and involves automatically classifying input text according to the sentiment expressed in it. Sentiment analysis is similar to topical text classification but has a significant contextual difference that needs to be handled. Based on this observation we propose a new feature selection method called Document Frequency Difference to automatically identify the words which are more useful for classifying sentiment. We further compare it to three other feature selection methods and show that it can help improve sentiment classification performance.

Keywords: Sentiment Analysis, Feature Selection, Text Classification, Natural Language Processing, Maximum Entropy Modeling.

1 Introduction

Sentiment analysis (SA) is concerned with automatically classifying pieces of text according to the opinions expressed in them – positive or negative. With the growth of user generated content on the web, the need for SA has increased. Popular applications of SA are summarization of online customer reviews or social media monitoring and analytics software that help organizations manage their reputation.

SA is often viewed as a special case of topical text classification [1, 2] and machine learning techniques can be used to classify documents. However, there are contextual differences that need to be modeled in order for any SA solution to work effectively. An important difference is that people tend not to repeat the same sentiment-carrying words in the same context. For example, when people write reviews for a digital camera they are less likely to write something like "The camera is good. The LCD screen is good and it takes good pictures" but more likely to write something like "The camera is great. The LCD screen is clear and it takes stunning pictures".

In this paper, we are focused mainly on the representation of documents for classification. We explore different methods for feature selection that have had success for topical text classification and propose a new method specifically for SA.

A. Farzindar and V. Keselj (Eds.): Canadian AI 2010, LNAI 6085, pp. 286–289, 2010.

2 Feature Selection

Feature Selection (FS) ranks all features based on a metric of how much they contribute to a class and removes all features below a specified threshold. The reduced set of features will not only improve the efficiency of the training and testing procedures, but also increase the classification performance since the least relevant features have been removed. In our study, we compare three existing FS methods: χ^2 [3], Optimal Orthogonal Centroid [4], and Count Difference [5]. We also propose a new metric based on our observations of polar (*positive* or *negative*) text.

Our proposed FS metric, called Document Frequency Difference (DFD), is calculated as follows:

$$score_t = \frac{|DF_+^t - DF_-^t|}{D} .$$

where DF_+^t is the number of documents in the *positive* class that term *t* occurs in, DF_-^t is the number of documents in the *negative* class that *t* occurs in and *D* is the number of documents in the training set.

An advantage of DFD is that the value it produces is normalized on a scale of 0 to 1. This means that scores will be proportional to other features, but furthermore, we feel that this makes the score suitable to be used as the feature weight itself. DFD will also not score rare terms highly. A drawback is that it requires an equal, or nearly equal, number of documents in each of the *positive* and *negative* classes. However, since we are only working with two classes, we feel that using the same number of documents in each class for training is necessary anyway.

3 Experiments and Results

3.1 Dataset and Evaluation

We use a publicly available dataset for our experiments. It is a set of movie reviews produced by Pang et al. [1, 2], which includes 1,000 positive and 1,000 negative review. This dataset can be seen as homogeneous since all of the reviews are from the movie domain despite genre differences.

To test our feature selection and weighting methods for SA, we implemented a classifier based on Maximum Entropy Modeling (MEM) as described in [6]. We prefer MEM to other machine learning methods because it has been shown to work well for topical text classification [5] and SA [1].

Since our dataset is relatively small, we employ cross-validation to make full use of them. The entire dataset is first partitioned into five folds. One fold is used as the validation set while the other four folds are joined and re-partitioned into ten-folds: nine for training and one for testing. So there are a total of 50 runs for each cross-validation process. The classification performance is computed on the validation set after every iteration of the IIS (Improved Iterative Scaling) algorithm for MEM. Training is terminated when all parameters have converged or a performance decrease is seen on the validation set.

Our baseline results are obtained by using all of the features seen in the training dataset. They are measured by P (precision), R (recall), and F (F-measure). For the positive class, P, R, and F are 0.791, 0.817, and 0.802, respectively, and for the negative class, P, R, and F are 0.811, 0.782, and 0.795, respectively. So the average-F for both classes is 0.799. Note that for both baseline and subsequent experiments, we filtered out all words that are not nouns, verbs, adjectives or adverbs along with the removal of a small list of 40 stop-words.

3.2 Feature Selection Results

To evaluate the performance of the FS ranking methods, we iterate over different feature cut-off thresholds and observe the performance for each one. Fig. 1 plots the results for the movie dataset. The best performance is achieved using the DFD metric with an average F of 0.851 with 900 or 1000 features. The CD metric achieves an average F of 0.848 with each of 800, 1000 and 1200 features. We perform the one-tailed two-sample z-test on these results and find that the performance increase over the baseline (0.848 or 0.851 vs 0.799) is statistically significant with over 99% certainty.

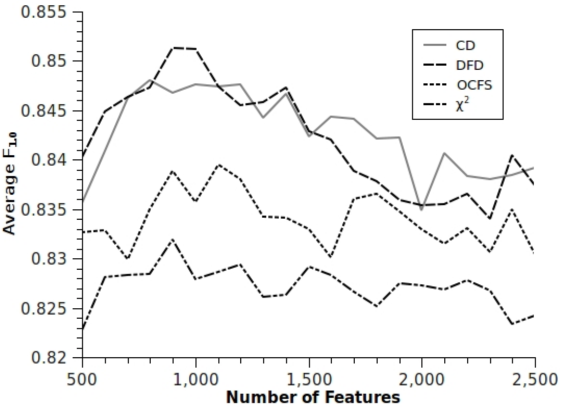

Fig. 1. Plot of feature selection results using the movie reviews dataset. The feature cut-offs in the range of 100-2,500 are plotted on the horizontal axis and the average F on the vertical axis.

3.3 Feature Score Weighting Results

In an attempt to weigh terms on their relevance, we further tried using the actual *score* computed by each FS method as the feature weight. The MEM feature function thus takes the form[7]:

$$f_{i,c'}(d,c) = \begin{cases} \frac{score_i}{\sum_{j\in d} score_j}, & c = c' \\ 0, & \text{otherwise} \end{cases} .$$

The average-F for DFD, CD, OCFS, and ÷^{2} are 0.869, 0.862, 0.809, and 0.857, respectively. All of them are better than the baseline of 0.799 and interestingly, using the DFD and CD scores in this fashion further increases classification performance beyond that achieved using them for feature selection. We did try to couple the feature score weights with feature selection, but the performance actually decreased slightly. We believe this is because we are already getting the most information that we can get out of the feature ranking metrics by using them as the scores.

4 Conclusions

We compared four FS methods and showed that they can help the classification performance of SA. The CD metric and our proposed DFD metric, both of which focus on the document frequency of a feature, have been shown to work well for SA. This matches our observation that for SA, frequently occurring terms in a particular review are not necessarily more helpful as is the case for topical text classification. This is further evidenced by the fact that when using the DFD score as the value of the feature function, performance was increased even higher. Although the movie dataset is about one domain, the automatic FS methods we have experimented with can be easily adapted to new domains.

References

1. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? Sentiment classification using machine learning techniques. In: Proceedings of the ACL 2002 Conference on Empirical Methods in Natural Language Processing, vol. 10, pp. 79–86 (2002)
2. Pang, B., Lee, L.: A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics (2004)
3. Yang, Y., Pedersen, J.O.: A comparative study on feature selection in text categorization. In: Machine Learning, pp. 412–420 (1997)
4. Yan, J., Liu, N., Zhang, B., Yan, S., Chen, Z., Cheng, Q., Fan, W., Ma, W.Y.: OCFS: Optimal orthogonal centroid feature selection for text categorization. In: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 122–129 (2005)
5. Cai, J., Song, F.: Maximum Entropy Modeling with Feature Selection for Text Categorization. In: Li, H., Liu, T., Ma, W.-Y., Sakai, T., Wong, K.-F., Zhou, G. (eds.) AIRS 2008. LNCS, vol. 4993, pp. 549–554. Springer, Heidelberg (2008)
6. Nigam, K., Lafferty, J., McCallum, A.: Using maximum entropy for text classification. In: IJCAI 1999 Workshop on Machine Learning for Information Filtering (1999)
7. Nicholls, C., Song, F.: Improving sentiment analysis with part-of-speech weighting. In: Proceedings of the International Conference on Machine Learning and Cybernetics (2009)