Combining Learn-based and Lexicon-based Techniques for Sentiment Detection without Using Labeled Examples

Songbo Tan[1], Yuefen Wang[2] and Xueqi Cheng[1]
[1]Information Security Center, Institute of Computing Technology, Chinese Academy of Sciences, China
[2]Information Center, Chinese Academy of Geological Sciences, China
tansongbo@software.ict.ac.cn, tansongbo@gmail.com

Abstract

In this work, we propose a novel scheme for sentiment classification (without labeled examples) which combines the strengths of both "learn-based" and "lexicon-based" approaches as follows: we first use a lexicon-based technique to label a portion of informative examples from given task (or domain); then learn a new supervised classifier based on these labeled ones; finally apply this classifier to the task. The experimental results indicate that proposed scheme could dramatically outperform "learn-based" and "lexicon-based" techniques.

Categories and Subject Descriptors

I.2.7 [**Artificial Intelligence**]: Natural Language Processing

General Terms

Algorithms; Performance; Experimentation

Keywords

Sentiment Classification; Opinion Mining; Information Retrieval

1. Introduction

Up to now, many researches have been conducted on document sentiment classification. These researches have fallen into two categories [1]. The first ("supervised techniques" or "machine learning techniques" or "learn-based method") [1][2] attempts to train a sentiment classifier based on occurrence frequencies of the various words in the documents. Pang's researches [1] indicate that standard machine learning methods perform very well, even definitively outperform human-produced baselines. However, supervised sentiment classifier typically requires a large amount of labeled training examples. As a result, when confronted with a task (or domain) without any labeled examples, supervised learning doesn't work at all.

The other approach ("unsupervised techniques" or "semantic orientation" or "lexicon-based method") [3-5] is to classify words into two classes, i.e., "positive" or "negative", and then count an overall positive/negative score for the text. If a document contains more positive than negative terms it is deemed as positive, and if the number of negative terms exceeds the number of positive terms it is assigned as negative. It is clear that this method is unsupervised learning because it doesn't require any labeled examples. Accordingly, it seems to be well suited for sentiment analysis task without providing any labeled training data. In most cases, however, its prediction accuracy is definitively limited because it heavily depends on lingual experts.

We see that both supervised and unsupervised techniques have their strength and weakness. In this work, we propose a novel scheme for sentiment classification which combines the strengths of both approaches as follows: we first use unsupervised technique to label a portion of informative examples from given task (or domain); then learn a new supervised classifier based on these labeled ones; finally apply this classifier to this task. Our scheme is very flexible, which only needs one supervised and unsupervised technique, and there is no change to the two techniques in any way. As a result, any machine learning and semantic orientation techniques can work together effectively under this general framework.

SIGIR'08, July 20-24, 2008, Singapore.
ACM 978-1-60558-164-4/08/07.

2. Methodology

2.1 Combination of Supervised and Unsupervised Techniques

Detailed algorithm for proposed scheme is presented in Figure 1. The basic idea is to use one unsupervised technique to label some informative unlabelled examples in new domain and train a new supervised classifier over these selected examples. In this scheme, the parameter "*Ratio*" indicates what percentage of new-domain data shall be picked out as informative examples.

1 Load semantic lexicon, new-domain unlabelled data (NU), and parameter "Ratio";
2 Acquire NU*Ratio most informative examples in new domain using semantic orientation;
3 Learn a new classifier using these selected examples;
4 Use the new classifier to classify examples on new domain.

Figure 1: The outline of proposed scheme

It is believed to be possible that proposed scheme could work effective and robust if the following conditions hold:

1. New-domain unlabeled data is abundant enough for selection of informative ones.

2. Unsupervised technique is effective to pick out informative ones from new-domain unlabeled data.

3. Supervised learning techniques can train a high-precision classifier over these selected examples, even when they consists of some outliers.

2.2 Detecting Informative Examples

In order to effectively detect informative examples using lexicon-based method, we make a simple assumption: for one example, the larger the negative term number T^N, the more likely it is drawn from negative class; the larger the positive term number T^P, the more likely it is taken from positive class.

However, this presumption doesn't hold when the length difference among different reviews is very large, because it is often the case that the larger the length of one review, the larger the T^N or T^P. To tackle this problem, we normalize (or divide) T^N and T^P so that the adverse effect of length difference can be counteracted to a high degree. This is the basic idea of relative term number. Formally, we define Negative Relative Term

Number (T^{RN}) and Positive Relative Term Number (T^{RP}) as following,

$$T^{RP} = \frac{T^P}{(T^N + T^P)/2},\ T^{RN} = \frac{T^N}{(T^N + T^P)/2}.$$

Up to this point, we can make a refined **assumption** that, for one example, the larger the T^{RN}, the more likely it is drawn from negative class; the larger the T^{RP}, the more likely it is taken from positive class. According to this assumption, we propose Relative Term Number Ranking method (*RTNR*): we first rank T^{RN} of all examples, and assign top $n/2$ largest examples as negative; then rank T^{RP}, and label top $n/2$ largest ones as positive. (n is a pre-defined number indicating how many examples in new domain shall be picked out as informative ones)

A crucial problem underling this method is to acquire a semantic lexicon. The commonly used method is to make use of existent sentiment lexicon or dictionary, such as General Inquirer (GI) [4] or Chinese Network Sentiment Dictionary (CNSD)[1] or NTU Sentiment Dictionary (NTUSD) [5]; another more complicated method is to use semantic orientation to construct a new semantic lexicon, such as Turney's method [3]. In the following, we elaborate NTUSD.

In our experiment, we only used NTUSD lexicon.

2.3 Learning a New Classifier

After picking out informative example from new domain, we treat sentiment classification simply as a special case of topic-based categorization (with the two "topics" being *positive sentiment* and *negative sentiment*), and employ typical classifiers, such as Centroid Classifier [6], to learn a new classifier that is suitable for new domain.

To implement Centroid Classifier, we use the standard bag-of-words framework. In this framework, each document d is considered to be a vector in the term-space. For term weight we employ normalized *TFIDF*.

3. Experiment Results

3.1 Experimental design

To validate the effectiveness and robustness of proposed scheme, we collected four domain-specific datasets: Movie Reviews (Mov), Computer Reviews (Comp), Education Reviews (Edu) and House Reviews (Hou). The details are listed in Table 1.

Table 1: The comparison of four datasets

	Negative	Positive	Average length	Vocabulary
Mov	562	430	251	14,622
Comp	390	544	120	4,725
Edu	1,012	254	600	19,150
Hou	868	296	300	12,674

To conduct our experiments, we use 50% of "Movie Reviews" as old-domain labeled training set, and halve each of the other datasets into unlabeled set and testing set. For supervised techniques, we train Centroid Classifier using only old-domain labeled data, i.e., "Movie Reviews". With respect to unsupervised techniques, we only use Ku's sentiment dictionary "NTUSD" [5]. For the sake of simplicity, we call this method "Lexicon based method" in the rest of this paper. Note that this method doesn't require any labeled or unlabeled data.

3.2 Does proposed scheme work?

Table 2 shows the results of experiments comparing proposed scheme with supervised techniques and unsupervised technique. For proposed scheme, we use Centroid Classifier as supervised technique and the *Ratio* is set to 0.4. As mentioned before, the parameter "*Ratio*" indicates what percentage of new-domain data shall be picked out as informative examples.

Table 2 shows that Centroid Classifier doesn't work at all on domains without labeled training data. This observation means that, in order to train a high-precision classifier, it is necessary to acquire a large amount of labeled data for one domain.

Although having not trained over old-domain labeled data and new-domain unlabeled data, Lexicon based method performs much better than supervised techniques. This result indicates that Lexicon based method doesn't rely on training data and is able to provide robust performance for domains without any labeled training data.

As expected, proposed scheme does indeed provide much better performance than supervised and unsupervised techniques. For example, the averaged accuracy of proposed scheme outperforms Centroid Classifier by 14 percent, and beats Lexicon based method by 8 percent. The result is very encouraging and of enormous value in sentiment-analysis applications that require high-precision classification but hardly have any labeled training data.

Table 2: Accuracy of different methods

	Centroid	Lexicon Based	Proposed Scheme
Hou	0.7405	0.7714	**0.8230**
Edu	0.8562	0.7883	**0.9210**
Comp	0.6509	0.8501	**0.9186**
Average	0.7492	0.8033	**0.8875**

4. Conclusion Remarks

The results described in this paper lead us to believe that the combination of supervised and unsupervised techniques is indeed useful for sentiment classification without labeled training data. We have shown that proposed scheme outperforms than supervised and unsupervised techniques.

Although the combination of supervised and unsupervised techniques indeed improves the classification accuracy using unlabeled data, there is a lot of room for improvement. For example, *RTNR* is not the best strategy for picking out informative examples; the size and quality of sentiment lexicon is severely dependent on some bias resources such as *WordNet* or Internet.

5. ACKNOWLEDGMENTS

This work was mainly supported by special fund of Chinese Academy of Sciences, "Research on Opinion Mining of Web Text", under grant number 0704021000 and one another project, i.e., 2004CB318109.

6. REFERENCES

[1] B. Pang, L. Lee, et al. Thumbs up? Sentiment classification using machine learning techniques. EMNLP, 2002.

[2] A. Aue and M. Gamon. Customizing Sentiment Classifiers to New Domains: a Case Study. RANLP. 2005.

[3] P. D. Turney. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. ACL, 2002.

[4] P Philip J. Stone, Dexter C. Dunphy, et al. The General Inquirer: A Computer Approach to Content Analysis. The MIT Press, 1966.

[5] L. Ku, Y. Liang, et al. Opinion Extraction, Summarization and Tracking in News and Blog Corpora. AAAI-2006 Spring Symposium on Computational Approaches to Analyzing Weblogs.

[6] E. Han and G. Karypis. Centroid-Based Document Classification Analysis & Experimental Result. PKDD 2000.

[1] http://134.208.10.186/WBB/EMOTION_KEYWORD/Atx_emtwordP.htm