Dual Word and Document Seed Selection for Semi-supervised Sentiment Classification

Shengfeng Ju[†] Shoushan Li[†*] Yan Su[†] Guodong Zhou[†] Yu Hong[†] Xiaojun Li [‡]

[†]Natural Language Processing Lab
Soochow University, Suzhou, China
{shengfeng.ju,
shoushan.li, yansu.suda} @gmail.com,
{gdzhou, hongy}@suda.edu.cn

[‡] College of Computer and Information Engineering
Zhejiang Gongshang University, Hangzhou, China
lixj@mail.zjgsu.edu.cn

ABSTRACT

Semi-supervised sentiment classification aims to train a classifier with a small number of labeled data (called seed data) and a large amount of unlabeled data. a big advantage of this approach is its saving of annotation effort by using the unlabeled data which is usually freely available. In this paper, we propose an approach to further minimize the annotation effort of semi-supervised sentiment classification by actively selecting the seed data. Specifically, a novel selection strategy is proposed to simultaneously select *good* words and documents for manual annotation by considering both of their annotation costs and informativeness. Experimental results demonstrate the effectiveness of our approach.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing – *Linguistic processing*; I.2.7 [**Artificial Intelligence**]: Natural Language Processing – *Text analysis*; I.5.2 [**Pattern Recognition**]: Design Methodology – *Classifier design and evaluation*;

Keywords

Opinion Mining, Seed Selection, Semi-supervised, Sentiment Classification

1. INTRODUCTION

Research in sentiment analysis has been progressed tremendously in recent years (Hu and Liu, 2004; Wiebe et al., 2005; Pang and Lee, 2008). In this research area, sentiment classification is a fundamental task which aims to identify the sentimental categories (e.g., positive or negative) of a natural language text towards a given topic (Pang et al., 2002; Turney, 2002). This task has become the core component of many important applications in sentiment analysis (Cui et al., 2006; Lloret et al., 2009; Zhang and Ye, 2008; Li et al., 2011a).

In the literature, machine learning (ML) approaches have been proved to be promising and widely used in sentiment classification. Generally speaking, the ML approaches could be grouped into three main categories: (1) lexicon-based learning: which employs a lexicon containing a certain number of opinion words to conduct a classifier. The popularly used term-counting approach is a typical example of such approaches. (Turney and Littman, 2002). (2) Corpus-based learning: which employs annotated corpus containing many labeled samples to conduct a machine learning-based classifier (Pang et al., 2002). (3) Joint lexicon-corpus leaning: which employs both a lexicon and annotated corpus to perform sentiment classification (Melville et al., 2009). Generally, the learning approaches of the third group achieve better performances due to the full consideration of the classification knowledge in both types of resources.

However, whatever ML approach is employed, the good performance mainly relies on manually labeled data, which is sometimes rather time-consuming and expensive to get. To cope with this problem, a promising solution is to perform a semi-supervised learning activity in which only a small amount of labeled data (i.e., the seed data) to train a classifier, together with a large amount of unlabeled data.

One key issue to further minimize the annotation cost in semi-supervised learning is how to get the seed data. A simple way to achieve this is to randomly select some samples for annotation as the seed data. However, this simple selection strategy is problematic when the learning approaches of the third group are employed due to the following reasons. First, in joint lexicon-corpus learning, both words and documents are required to be annotated. The types of samples (words or documents) and the amount of them preferred to be selected respectively remains an untouched problem. Second, random selection is possibly to select many helpless samples for sentiment classification, e.g., the stop words like "a", "the" and "as". In contrast, the clever selection strategies should be applied to get some opinion words like "wonderful", "excellent" and "poor" which are thought to be more valuable for sentiment classification.

In this paper, we focus on the seed selection issue in semi-supervised sentiment classification where the joint lexicon-corpus learning approaches are employed. To address the above two problematic issues, we first investigate the annotation costs of annotating a word or a document; Then, we propose a uniform

[*] Corresponding author

CIKM'12, October 29-November 2, 2012, Maui, HI, USA.
Copyright 2012 ACM 978-1-4503-1156-4/12/10…$15.00.

measurement to define the informativeness of a word or a document; Finally, both the annotation cost and the informativeness measurement are taken into account to decide a selection strategy to select good words and documents for manual annotation.

The remainder of this paper is organized as follows. Section 2 overviews the related work on semi-supervised sentiment classification. Section 3 proposes our strategy for selecting samples as the seed data. Section 4 reports the experimental results. Finally, section 5 draws the conclusion.

2. RELATED WORK

Although supervised learning methods for sentiment classification have been extensively studied (Pang et al., 2002), the studies on semi-supervised sentiment classification are relatively new. Most related studies applied the learning approaches belonging to the second group, i.e., corpus-based learning, such as, Dasgupta and Ng (2009), Li et al. (2010) and Li et al. (2011b). As far as the third group of learning approaches, i.e., joint lexicon-corpus learning, is concerned, Sindhwani and Melville (2008) firstly propose a semi-supervised sentiment prediction algorithm that utilizes lexical prior knowledge in conjunction with unlabeled examples based on a document-word bipartite graph. More recently, Li et al. (2009) propose a non-negative tri-factorization approach to jointly learning classifier with a lexicon, labeled and unlabeled documents for semi-supervised sentiment classification. Unlike both studies mentioned above, our work focuses on the selection strategy of the seed data when a semi-supervised learning approach is employed, which has not been addressed in sentiment classification yet.

3. Selection Strategy for the Seed Data

3.1 Annotation Cost

Table 1: The numbers of annotated words and documents by each annotator who takes half hour for annotation.

Annotator	**A**	**B**	**C**	**D**
N_{Word}	986	1025	819	759
N_{Doc}	54	59	48	45
N_{Word} / N_{Doc}	18.3	17.3	17.0	16.9

A distinguishing feature of the third-group learning approaches is its requirement of annotating both words and documents. Instinctively the cost of acquiring a labeled word is different from a labeled document. To investigate their real annotation costs, four annotators, named **A**, **B**, **C**, and **D**, are asked to take half an hour to annotate some words and documents. Note that all the annotators are with background knowledge of sentiment analysis, which makes the annotation work reliable and fast. The documents for annotating are from the multi-domain sentiment classification corpus, collected by Blitzer et al. (2007)[1]and the words for annotating are extracted from the corpus. Table 1 shows the annotation results where N_{Word} and N_{Doc} denote the number of the annotated words and documents respectively.

From table 1, we can see that in a certain time, much more words can be annotated than documents. On average, the time of annotating a document equals that of annotating 16-18 words.

3.2 Informativeness Measurement

Informative samples are encouraged to be selected as seed data for a good performance of a semi-supervised learning approach. In sentiment classification, the informativeness of a word is influenced by two main factors. First, the POS of a word is an important prior knowledge for evaluating its informativeness. For example, adjectives are more likely to be an opinion word and thus thought to be more informative than other words like verbs and nouns. To make the POS information computational, we evaluate the informativeness value of a sample as follows. 200 words are firstly randomly selected from the corpus and the proportion of opinion words are calculated for each POS tag[2]. The proportion is served as the informativeness value of the word belonging to the same POS tag. Table 2 shows the detailed proportions.

Table 2: Proportion of the opinion words in each POS category

POS	JJ	RB	VB	NN	Others
Proportion	0.44	0.11	0.08	0.06	0.01

Second, the frequency of a word is another important prior knowledge for evaluating its informativeness. More frequently occurring words are believed to be more informative.

In summary, the informativeness value of a word w is defined as follows:

$$Infor(w) = V(POS(w)) \times \log\left(F(w)\right) \qquad (1)$$

Where $POS(w)$ is the POS tag of the word w and $V(x)$ is the informativeness value of the POS tag x. For example, if the word is a adjective, the POS value is set to V(JJ)=0.44 as shown in Table 2. $F(w)$ is the occurring frequency of the word w in the corpus.

As far as a document d is concerned, the informativeness value is defined as follows:

$$Infor(d) = \frac{\sum_{w \in d} Infor(w)}{\log\left(L(d)\right)} \qquad (2)$$

Where $\sum_{w \in d} Infor(w)$ means the summation of the informativeness values of all containing words. L(d) is the length of the document, i.e., the total number of the containing words. The document containing more words possibly takes more helpless words and thus is thought to be less informative.

3.3 Informativeness Measurement

Our strategy for selecting the samples as the seed data considers both the annotation cost and the informativeness values. Principally, the samples with less annotation cost and higher informativeness value are encouraged to be labeled. The samples

[1] http://www.seas.upenn.edu/~mdredze/datasets/sentiment/

[2] Oliver Mason's QTag is used for POS tagging: http://www.english.bham.ac.uk/staff/oliver/software/tagger/index.htm

are ranked according to their scores which are calculated as follows:

$$Score(s) = \frac{Infor(s)}{Cost(s)} \quad (3)$$

Where *Infor*(*s*) is the informativeness value of the sample s and *Cost*(*s*) is annotation cost of the sample *s*. In this study, if the sample is a word, the informativenss value is calculated by formula (1) and the annotation cost is set to 1/16. If the sample is a document, the informativenss value is calculated by formula (2) and the annotation cost is set to 1.

4. Experimentation

We systematically evaluate our seed selection approach for semi-supervised sentiment classification on the multi-domain dataset as mentioned in Section 3.1.

4.1 Experimental Setting

- Dataset: This dataset contains product reviews from four different domains: Book, DVD, Electronic and Kitchen appliances, each of them contains 1000 positive and 1000 negative labeled reviews. In the experiments, 200 documents in each category are served as testing data and the remaining ones are served as initial labeled data (i.e., the seed data) and unlabeled data.
- Features: Word unigram features are used. Each review text is treated as a bag-of-words and transformed into binary vectors encoding the presence or absence of one word feature.
- Classification algorithm: The semi-supervised classification algorithm based on document-word bipartite graph is adopted (Sindhwani and Melville, 2008). Note that we change the original algorithm a bit. In the testing phrase, the linear classification model is changed into maximum entropy (ME) model due to the better performance of ME. The ME model is implemented with the help of the public tool, Mallet Toolkits[3].

Table 2: Statistics of the annotated words with polarity labels.

Domain	Book	DVD	Electronic	Kitchen
Positive	582	380	356	269
Negative	482	425	372	267
Neutral	6744	6785	3621	3351

To perform the lexicon-corpus learning, the polarities of the words are necessarily known before. Thus, we manually annotate the words that occurs more than twice in the corpus. The detailed statistics of the annotated words with polarity labels (i.e., positive, negative, and neutral) are shown in Table 3. From this table, we can see that most words are neutral.

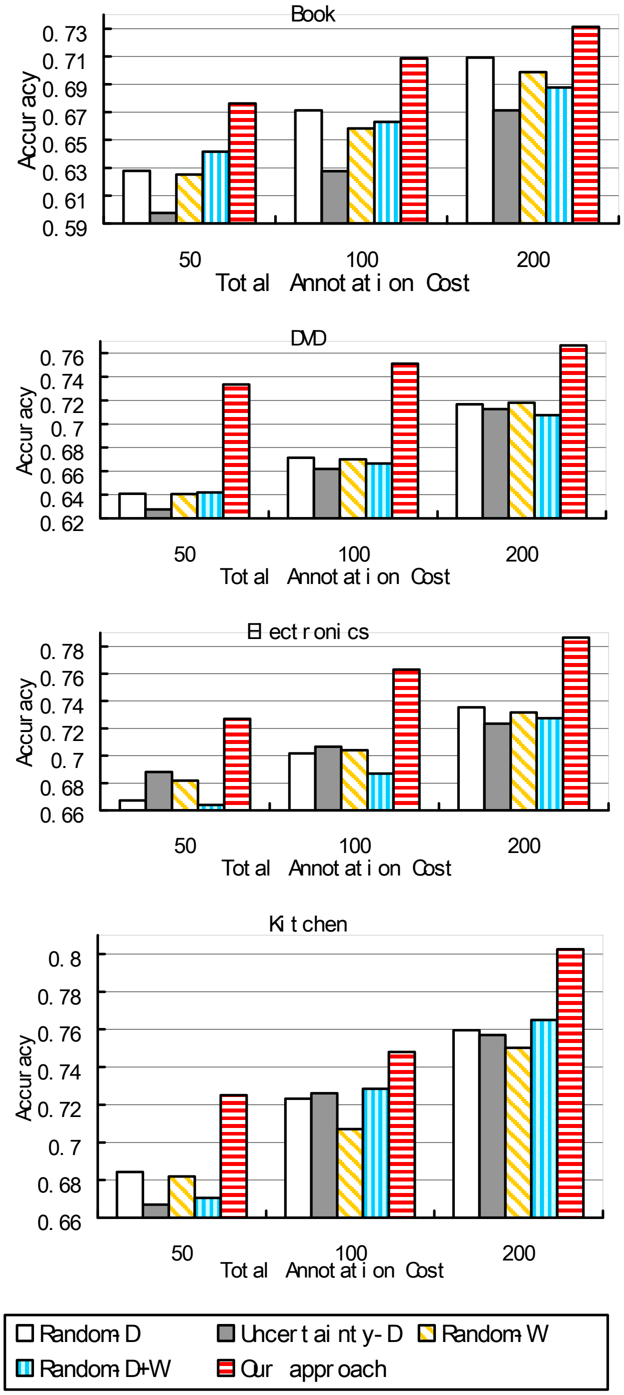

Figure 1: Performances of different selection strategies with 50, 100, and 200 annotation costs

4.2 Experimental Results

- **Random-D:** randomly selecting only documents for annotation.
- **Uncertainty-D:** employing the active learning procedure as described in Tong and Koller (2002). Specifically, we begin with training a ME classifier on one labeled sample from

[3] http://mallet.cs.umass.edu/

each class, iteratively labeling the most uncertain unlabeled sample in each class.

- **Random-W:** randomly selecting only words for annotation.
- **Random-D+W:** randomly selecting both documents and words for annotation.

Figure 1 illustrates the performances of different selection strategies where 50, 100, and 200 annotation costs are used. Here, 50 annotation costs means the cost of annotating 50 documents or 800 (50*16) words. From this figure, we can see that **Uncertainty-D** performs no better than **Random-D**, which indicates that uncertainty is useless for the seed selection for semi-supervised sentiment classification, although it is shown to be effective for many active learning task. Our approach significantly outperforms other approaches (p-value<0.01). Especially, when only a few annotation effort (annotation cost equals 50) is used, our approach performs remarkably better. Taking a look into the seed samples, we find that they contains both words (most are adjectives) and documents. For example, in DVD domain, the seed samples contains 544 words (including 119 positive words and 91 negative words) and 16 documents. This result demonstrates the importance of labeling both opinion words and informative documents when a limited annotation cost is given.

5. CONCLUSION

In this paper, we investigate the annotation costs of labeling a word and a document and propose a measurement to uniformly define the informativeness of a word and a document. The annotation cost and the informativeness measurement are both considered into the selection strategy for selecting *good* seed data for semi-supervised sentiment classification of document-level. Experimental results show that our selection strategy could significantly improve the performance of semi-supervised sentiment classification.

Acknowledgments

The research work described in this paper has been partially supported by three NSFC grants, No.61003155, No.60873150 and No.90920004, one National High-tech Research and Development Program of China No.2012AA011102, Open Projects Program of National Laboratory of Pattern Recognition, and the NSF grant of Zhejiang Province No.Z1110551. We also thank the two anonymous reviewers for their helpful comments.

6. REFERENCES

[1] Blitzer, J., Dredze, M., and Pereira, F. 2007. Biographies, Bollywood, B oom-boxes and Blenders: Domain Adaptation for Sentiment Classification. In *Proceedings of ACL-07*, 440-447.

[2] Cui, H., Mittal, V., and Datar, M. 2006. Comparative Experiments on Sentiment Classification for Online Product Reviews. In *Proceedings of AAAI-06*, 1265-1270.

[3] Dasgupta, S. and Ng, V. 2009. Mine the Easy, Classify the Hard: A Semi-Supervised Approach to Automatic Sentiment Classification. In *Proceedings of ACL-09*, 701–709.

[4] Hu M. and B. Liu. 2004. Mining and Summarizing Customer Reviews. In *Proceedings of SIGKDD-04*, pp.168–177.

[5] Li, S., Huang, C. and Zong, C. 2011a. Multi-domain Sentiment Classification with Classifier Combination. *Journal of Computer Science and Technology (JCST)*, 26(1): 25-33.

[6] Li, S., Huang C., Zhou, G., and Lee, S. 2010. Employing Personal/Impersonal Views in Supervised and Semi-supervised Sentiment Classification. In *Proceedings of ACL-10*, 414-423.

[7] Li, S., Wang, Z., Zhou, G., and Lee, S. 2011b. Semi-supervised Learning for Imbalanced Sentiment Classification. In *Proceeding of IJCAI-11*, 826-1831.

[8] Li, T., Zhang, Y., and Sindhwani, V. 2009. A Non-negative Matrix Tri-factorization Approach to Sentiment Classification with Lexical Prior Knowledge. In *Proceeding of ACL-IJCNLP-09*, 244–252.

[9] Lloret, E., Balahur, A., Palomar, M., and Montoyo, A. Towards Building a Competitive Opinion Summarization System. In *Proceedings of NAACL-09 Student Research Workshop and Doctoral Consortium*, 72-77.

[10] Melville, P., Gryc, W., and Lawrence, R. 2009. Sentiment Analysis of Blogs by Combining Lexical Knowledge with Text Classification. In *Proceedings of KDD-09*, 1275-1284.

[11] Pang, B. and Lee, L. 2008. Opinion Mining and Sentiment Analysis: Foundations and Trends. *Information Retrieval*, vol.2(12), pp.1-135.

[12] Pang, B., Lee, L., and Vaithyanathan, S. 2002.Thumbs up? Sentiment Classification using Machine Learning Techniques. In *Proceedings of EMNLP-02*, 79-86.

[13] Sindhwani, V. and Melville, P. 2008. Document-Word Co-Regularization for Semi-supervised Sentiment Analysis. In *Proceedings of ICDM-08*, 1025-1030.

[14] Tong, S. and Koller, D. 2002. Support Vector Machine Active Learning with Applications to Text Classification. *Journal of Machine Learning Research,* 2:45–66.

[15] Turney, P. 2002. Thumbs up or Thumbs down? Semantic Orientation Applied to Unsupervised Classification of reviews. In *Proceedings of ACL-02*, 417-424.

[16] Turney, P. and Littman, M. 2002. Unsupervised Learning of Semantic Orientation from a Hundred-billion-word Corpus. *Technical Report ERB-1094, National Research Council,* Institute for Information Technology, 2002.

[17] Wiebe, J., Wilson, T., and Cardie, C. 2005. Annotating Expressions of Opinions and Emotions in Language. *Language Resources and Evaluation*, pp.165-210.

[18] Zhang, M. and Ye, X. 2008. A Generation Model to Unify Topic Relevance and Lexicon-based Sentiment for Opinion Retrieval. In *Proceedings of SIGIR-0s8*, 411-418.