Positive, Negative, or Mixed? Mining Blogs for Opinions

Xiuzhen Zhang, Zhixin Zhou, Mingfang Wu
School of CS & IT, RMIT University
GPO Box 2476v, Melbourne 3001, Australia
{*xiuzhen.zhang, zhixin.zhou, mingfang.wu*}*@rmit.edu.au*

Abstract *The rich non-factual information on the blogosphere presents interesting research questions. In this paper, we present a study on analysis of blog posts for their sentiment by using a generic sentiment lexicon. In particular, we applied Support Vector Machine to classify blog posts into three categories of opinions: positive, negative and mixed. We investigated the performance difference between global topic-independent and local topic-dependent opinion classification on a collection of blogs. Our experiment shows that topic-dependent classification performs significantly better than topic-independent classification, and this result indicates high interaction between sentiment words and topic.*

Keywords blog, sentiment analysis, opinion classification, opinion words, information retrieval

1 Introduction

With the wide availability of broadband network facilities, internet has become an indispensable channel for people to communicate. More and more people are publishing their own experience and opinions, as well as seeking other people's opinions. With the explosive amount of information generated daily, it is almost impossible for people to read through all the information even on a narrow topic. This demands for new techniques to help track sentiment trends and search for various opinions, a task that is very different from factual information task as in traditional information retrieval; and sentiment analysis is a key component of such techniques.

Sentiment analysis is the technology to evaluate a text and predicate the text's subjectivity (subjective versus objective) and/or sentiment (positive versus negative). A general approach is to find out those keywords from the text that are of evaluative feature or sentiment orientation to represent the text, and to use a classification method to predicate the text's probability of belonging into pre-defined categories; the classifier is usually trained on a set of labeled texts.

The above approach has shown success in some earlier work where sentiment analysis was used to reviews from a certain domain, such as a movie review or a product review [8], while less success was seen in those studies conducted within the TREC (Text REtrieval Conference) Blog Track on polarity search task [6]. On the other hand, existing studies have also shown that mixed sentiment is especially challenging [5] given its uncertainty in nature.

The TREC Blog Track's polarity task is to identify the polarity of the opinions in the retrieved documents (blogs) in respond to a search topic. A problem with the evaluation of this task is that sentiment analysis is mingled with topic search and rank task, as a result, it is hard to ascertain the effectiveness of a certain sentiment analysis method.

This paper presents a study on analysis of blog posts for their sentiments, or opinions. Specifically a blog post is analyzed and classified into three categories: positive sentiment, negative sentiment or mixed sentiment. We adopt a dictionary-based approach by using a generic sentiment lexicon developed by a linguistic study [12]. We propose to represent blog posts as bags of sentiment words and use the Support Vector Machine (SVM) [3] learning model to classify blog posts. Given that both the sentiment lexicon and the classification model are generic, our research question is: if a *global* classification of blog posts accross topic genres would achieve a similar performance as a *local* classification of blog posts of a certain topic genre.

The remainder of this paper is organised as follows. We review some related work in Section 2 and describe the sentiment lexicon used in this study in Section 3. We present our classification approach in Section 4, and experiment setup and evaluations in 5. We discuss the experiment result and conclude the paper in Section 6.

2 Related Work

Sentiment analysis was initially applied to a corpus of documents that are from the same genre, such as a corpus of movie reviews or a corpus of product review. The task of sentiment analysis is to specify if a document (or a review) expresses a positive or negative opinion. Naturally most studies adopted machine learning classification approaches [1, 8, 11]. Pang et al [8] applied and compared three machine learning methods, naive bayes, maximum entropy and support vector machines,

Proceedings of the 14th Australasian Document Computing Symposium, Sydney, Australia, 4 December 2009.

on a corpus of movie reviews with uniform class distribution. Their results showed that the support vector machine model generally performed the best.

While most sentiment analyses classify comments or documents into two categories: positive versus negative, Koppel and Schler [5] argued that there were other comments that might express a mixed or neutral sentiment. Their study showed that by incorporating neutral category can lead to significant improvement in overall classification accuracy, and this is achieved by properly combining pairwise classifiers.

With so many opinionated documents available on the Web, people are actively seeking other people's opinion toward a certain topic. In this case, we need to do more than sentiment analysis: we need first to retrieve a set of documents that are about the topic, then judge if a document indeed contains any opinion at all - so called *subjectivity analysis*, then analyse if a subjective opinion or sentiment is positive, negative, or mixed. Such an opinion polarity finding task was introduced in TREC 2007 conference [6]. A commonly adopted approach by participants is to use baseline search engines to search topic-relevant documents first, and then use polarity-finding heuristics to re-rank documents for polarity. Machine learning models have not been widely used to improve the polarity classification accuracy.

3 A Generic Sentiment Lexicon

Identification of sentiment words is fundamental to sentiment analysis and classification. There are two broad methods to identify sentiment words and build sentiment lexicon. One method is through manual construction in which annotators manually annotate a list of words or phrases [9] or find and annotate sentiment words from a given corpus [8, 12].

Another method is to build a lexicon from a small number of seed words with pre-determined sentimental polarity, and then populate the seed list through learning or other relationships. For example, Hatzivassiloglou and McKeown [2] expanded a seed list by adding those words that are linked to seed words through conjunction such as *and, or, but, either-or,* or *neither-or*; while Kim and Hovey made use of WordNet to populate seed words through synonym and antonym relationships [4].

In our study, we use the sentiment lexicon developed by Wiebe et al. [12]. This lexicon list has 8221 annotated words resulted from manual annotation of a 10,000-sentence corpus of news articles of various topics. The following is an example of such an annotation:

> type=strongsubj len=1 word1=admire pos1=verb stemmed1=y priorpolarity=positive

The property *prior polarity* indicates the attitude being expressed by the word *admire* and has three values: *positive*, *negative* and *neutral*. The neutral tag are those subjective expressions that do not have positive or negative polarity. The property *type* indicates the expression intensity and here it has binary values: *strong* or *weak*. As annotation was done within context of a sentence, the grammar function of a word is also annotated, for example, the word *admire* here is a verb. Thus a word may occur twice or more in the list depending on which grammar function a word acts in the original text for annotation, for example, the word "cooperation" is annotated as *adjective* and *none*. This list also includes words with multiple morphemes, for example, cooperate, cooperation, cooperative, and cooperatively.

4 Opinion Classification

This section presents our classification method.

4.1 Support Vector Machine

Support Vector Machine (SVM) has been widely used in text categorisation, and with reported success [3]. In an SVM model, objects are represented as vectors. In learning a model to classify two classes, the basic idea of SVM is to find a hyperplane, represented by a vector, that separates objects of one class from objects of other classes at a maximal margin. When using a linear kernel, SVM learns a linear threshold function. With polynomial and radial basis kernels, SVM can also be used to learn polynomial and radial basis classifiers.

SVM$_{multiclass}$ [1] is an implementation of the multiclass SVM, and is based on Structural SVMs [10]. Unlike regular SVMs, structural SVMs can predict complex objects like trees, sequences, or sets. SVM$_{struct}$ can be used for linear-time training of binary and multiclass SVMs under the linear kernel. Features extracted jointly from inputs and outputs are used to form an optimal separation plane.

4.2 Opinion Word Extraction

To apply a classification model effectively, a key issue is feature selection, i.e. what input will be given to a classification model. The feature selection is application dependent - how do we want to classify a set of documents, and what are prominent features from a set of documents that can separate them from each other. For the sentiment classification task, it is intuitive that we identify those opinion words from a set of documents as classification features.

In this study, we simply treated opinion words as tokens and do not apply natural language processing methods such as Part-Of-Speech tagging to analyse the grammatical function of those words. We applied Porter stem method to the list and group different forms of the same word, and this leaves us 4919 "words".

A closer look at the stemmed opinion words reveals some interesting facts. There are 103 words that are of contradictory polarities. After we removed these words, we had 4816 words with unique sentiment

[1] Avaiable at `http://svmlight.jochims.org/svm_multiclass.html`

polarity. However, there are also some words that have mixed levels of strength. In lieu of this, we created a new level of strength and named it "contextual strength"; there are a total of 194 in this category. The distribution of opinion words in term of polarity and strength is summarised in Table 1.

	Positive	Negative	Neutral	Total
Strong	954	2061	107	3192
Contextual	81	98	14	194
Weak	544	783	163	1490
Total	1579	2942	284	4816

Table 1: Distribution of opinion words

4.3 Opinion Word Vectors

In information retrieval, each document is represented by all word tokens from a collection. However, for the purpose of opinion classification, we represent a document as a vector of opinion word tokens and ignore those words that do not express any sentiment. As in retrieval models, we weight each feature (an opinion word) of the document vector. The $tf \times idf$ weight of an opinion word f in a document d is:

$$w_{fd} = tf_{fd} \times log\frac{|D|}{|D_f|}$$

where tf_{fd} is the frequency of f in d. $|D|/|D_f|$ is inverse document frequency of f — $|D|$ is the number of documents in the collection, and $|D_f|$ is the number of documents containing f. We expect that this model is general enough to be applied to opinion classification.

5 Evaluation

5.1 Topic-independent versus Topic-dependent Classification

Opinion classification is usually applied to a set of documents that are of same genre or about a similar topic such as movie reviews and product reviews. With a huge number of opinionated documents on the Web and the nature of inexact match of a Web search engine, it is unlikely that we can always get a set of documents from the same genre to be classified. As a sentiment lexicon is independent of semantic topic of a document, we then investigate if there exists any difference between classification of documents that are about mixed topics and documents about a topic; we call these two types of document classification topic-independent (or global) classification and topic-dependent (or local) classification respectively.

5.2 Experiment Set-up

The TREC Blog track 2006 collection Blog06 [7] is a sample of the blogosphere crawled from 6 December 2005 to 21 February 2006. The collection is 148GB in total, and comprises three components: XML feeds of 38.6GB, which are the blogs, Permalink documents of 88.8GB, which are the blog posts with associated comments, and HTML homepages of 28.8GB, which are the entries to blogs. The permalink documents are the unit for the opinion finding task and polarity tasks.

Category	TREC-2006	TREC-2007
Negative	3,707 (32.15%)	1,844 (26.34%)
Mixed	3,664 (31.78%)	2,196 (31.37%)
Positive	4,159 (36.07%)	2,960 (42.29%)
Total	11,530	7,000

Table 2: Distribution of document categories in TREC-2006 and TREC-2007

The content of a blog post is defined as the content of the blog post itself and the contents of all comments to the post. A blog post is considered having subjective content if "it contains an explicit expression of opinion or sentiment about the target, showing a personal attitude of the writer" [7]. Fifty topics were selected by NIST from a collection of queries of a commercial search engine for the opinion retrieval task. For a topic, permalink documents are tagged with NIST relevance judgement, with the following categories (or scales) [7]: not judged(-1), not relevant(0), relevant(1), negative(2), mixed(3) and positive(4).

The Blog06 collection was used for both TREC-2006 and TREC-2007 Blog Track. Fifty (different) topics were used for each conference. For each topic, we selected documents with NIST assessor relevance judgement scale of 2 (negative), 3 (mixed - both positive and negative) and 4 (positive) for our study. Table 2 shows the distribution of documents in different categories in TREC-2006 and TREC-2007 respectively.

Zettair search engine [2] was used to index documents with the sentiment lexicon. Each document was converted into a vector of opinion words with the weighting scheme as described in Section 4.3.

5.3 Topic-independent Opinion Classification

To train the topic-independent opinion classification model, we pooled and indexed all documents from 50 topics in TREC-2006. SVM model was then trained on the converted opinion-word vectors with judgement scale $>=2$. Ten-fold cross validation experiment was conducted on all 10,737 documents of 50 topics. It showed an overall accuracy of 52.90$\pm$3%, that is 52.9% of documents correctly classified, with a standard deviation of 3%.

5.4 Topic-dependent Opinion Classification

To examine the interactions between topics and opinion classification accuracy, topics of TREC-2006 that contain at least 10 documents from each opinion category

[2] http:www.seg.rmit.edu.au/zettair/

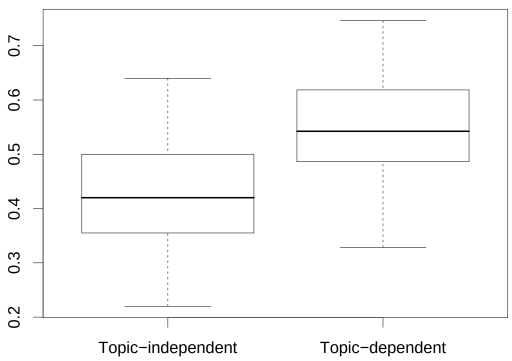

Figure 1: Classification accuracy: Topic-independent vs. topic-independent

(recall that there are 3 categories positive, negative or mixed) were extracted, this resulted in 36 topics and 9,771 documents in total.

To evaluate the accuracy of topic-dependent opinion classification, we individually indexed documents from the same topic, and applied ten-fold cross validation experiment to each topic collection accordingly. On average, the topic-dependent model achieved an accuracy of 63±13%, significantly higher than that achieved by the topic-independent model.

5.5 Blind Test of the Classification Model

The topic-independent classification model trained on documents with TREC-2006 judgments were blind tested on documents with TREC-2007 judgements. 27 topics that contain at least 10 documents in each category were used in our study. The model showed an accuracy of 42% on the whole collection. The drop in performance compared to that of 10-fold croass validation (52.9±3%) may be attributed to the change of topics between the two collections, which in turn suggests that there is strong correlation between topics and opinion words.

On the other hand, in the 10-fold cross validation experiment on the TREC-2007 collection, the topic-dependent model achieved an average accuracy of 55%. We extracted individual topic's accuracy for the topic-independent model, and used a paired Wilcoxon test to compare the difference in classification accuracy between the topic-independent model and the topic-dependent model. The improvement in classification accuracy of the topic-dependent model over that of the topic-independent model is statistically significant ($p < 0.001$). Figure 1 shows the summary of two models. As we can see that the topic-dependent model achieved higher accuracy than the topic-independent model.

6 Conclusion

In this paper we have described our research on opinion classification of blogs. We have investigated the difference of global classification of documents from mixed topics and local classification of documents from the same topic. Our experiment on the TREC Blog collections has shown that the local classification is significantly more accurate than the global classification. This might be because that documents from the same topic tended to have a similar set of sentiment words. Our future research will concentrate on developing topic-specific opinion classification models, especially it is anticipated that the annotation of opinion words tensity can be used to further improve such models.

Acknowledgements The authors would like to thank Steven Garcia for his various help with using Zettair.

References

[1] K. Dave, S. Lawrence and D. M. Pennock. Mining the peanut gallery: opinion extraction and semantic classification of product review. In *Proceedings of the 12th international conference on World Wide Web*, pages 519–528, 2003.

[2] V. Hatzivassiloglou and K.R.McKeown. Predicting the semantic orientation of adjectives. In *Proceedings of 8th Conference on European chapter of the association for computational linguistica*, pages 174–181, 1997.

[3] T. Joachims. Text categorisation with support vector machines: Learning with many relevant features. In *Proceedings of ECML-98*, 1998.

[4] S. M. Kim and E. Hovy. Determining the sentiment of opinions. In *Proc. of the Coling Conference*, 2004.

[5] M. Koppel and J. Schler. The importance of neutral examples for learning sentiment. *Computational Intelligence*, Volume 22, Number 2, 2006.

[6] C. MacDonald, I. Ounis and I. Soboroff. Overview of TREC-2007 blog track. In *Proceedings of TREC-2007*, Gaithersburg, USA, 2008.

[7] I. Ounis, M. de Rijke, C. MacDonald and I. Soboroff. Overview of trec-2006 blog track. In *Proceedings of TREC-2006*, Gaithersburg, USA, 2007.

[8] B. Pang, L. Lee and S. Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of EMNLP*, pages 79–86, 2002.

[9] P. Subasic and A. Huettner. Affect analysis of text using fuzzy semantic typing. *IEEE Transactions on Fuzzy systems*, Volume 9, pages 483–496, 2001.

[10] I. Tsochantaridis, T. Hofmann, T. Joachims and Y. Altun. Support vector learning for interdependent and structured output spaces. In *Proc.: of ICML-2004*, 2004.

[11] P. Turney. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the ACL*, 2002.

[12] J. Wiebe, T. Wilson and C. Cardie. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation (formerly Computers and the Humanities)*, Volume 1, Number 2, 2005.