Finding Opinionated Blogs Using Statistical Classifiers and Lexical Features

Feifan Liu, Bin Li and **Yang Liu**
The University of Texas at Dallas
800 W. Campbell Road, Richardson, TX 75080
{ffliu,leroy,yangl}@hlt.utdallas.edu

Abstract

This paper systematically exploited various lexical features for opinion analysis on blog data using a statistical learning framework. Our experimental results using the TREC Blog track data show that all the features we explored effectively represent opinion expressions, and different classification strategies have a significant impact on opinion classification performance. We also present results when combining opinion analysis with the retrieval component for the task of retrieving relevant and opinionated blogs. Compared with the best results in the TREC evaluation, our system achieves reasonable performance, but does not rely on much human knowledge or deep level linguistic analysis.

Introduction

Opinion analysis[1] has drawn much attention in natural language processing community. There are many previous studies on sentiment analysis in some specific domains such as movie and other product reviews (Turney 2002; Dave, Lawrence, & Pennock 2003; Pang, Lee, & Vaithyanathan 2002), as well as cross-domain combination (Li & Zong 2008). Compared with other online resources, blogs are more flexible in their content and styles, which gives rise to new challenges in analyzing their opinion. Although opinion analysis on blogs has been greatly advanced by the annual Text REtrieval Conference (TREC) since 2006, performance is still far from perfect.

Many previous studies have been done on correct identification of the sentiment carrier using different levels of granularity, such as word, phrase, or sentence-level. Early research from (Hatzivassiloglou & McKeown 1997) suggested that adjectives are important indicators of sentiment orientation. (Benamara, Cesarano, & Reforgiato 2007) explored using adverb and adjective combinations to evaluate the polarity degree. (Gamon & Aue 2005) and (Turney 2002) built their sentiment vocabularies according to co-occurrences of candidate terms and hand-picked sentiment seed words. (Wilson, Wiebe, & Hoffmann 2005) presented a classifier-based system to identify phrase-level contextual sentiment polarity using a lexicon containing 8,000 single subjectivity words. (Riloff & Wiebe 2003) proposed a pattern learning technique based on pre-defined syntactic forms that proved to be useful for opinion detection.

In this paper we investigate opinion analysis in the context of the TREC Blog Track, and use a statistical classification approach for this problem. Our goal is to systematically explore various lexical features that are derived from both statistical analysis and heuristic knowledge. In addition, we examine the effect of different classification settings. Our experimental results show that the features we explored prove to be very helpful to improve both the classification accuracy and the MAP score in the TREC framework. Our system obtains comparable results with the best one in the TREC evaluation, but our system does not require a large opinion vocabulary or performing deep level linguistic analysis such as parsing.

TREC Blog Track and System Overview

This paper is focused on two tasks in the Blog track of TREC 2008: [2] opinion finding task and polarized opinion finding task. Both can be considered as a ranking task based on whether a blog is opinionated and whether it is relevant to the given query. For the polarized task, positive opinionated and negative opinionated blog posts should be ranked separately. The data used in this track is the Blog06 data (Ounis, Macdonald, & Soboroff 2008).

Our system of finding opinionated blogs contains four parts: preprocessing, topic retrieval, opinion analysis, and re-ranking based on their combination. We preprocessed the original permalink documents in the Blog collection by removing noisy html tags and non-English blogs. This paper focuses only on the opinion analysis module, for which we use a statistical classification approach. In 2008, for the 50 test topics, TREC provided five different baseline retrieval results (1000 blogs for each topic) for participants to use

[1] In literature, this is sometimes also called sentiment analysis.

[2] http://ir.dcs.gla.ac.uk/wiki/TREC-BLOG/

in the opinion finding and polarized opinion finding task. We use these TREC provided baselines as input in this paper. For each topic relevant blog, a posterior probability is assigned according to the classifier output, indicating how likely it is opinionated.

Since our ultimate goal is to rank those blogs higher that are more relevant and more opinionated, we believe it is reasonable to conduct opinion analysis using only the topic-relevant part of a blog rather than the entire blog text. Therefore, we first split a blog into sentences[3], and used Lemur toolkit[4] to retrieve the top five relevant sentences corresponding to the topic. For each retrieved sentence, we also extracted its preceding and following sentences. Thus for each blog, we used a maximum 15 sentences to perform classification.

Let Opi, Pos, Neg be the probability of being opinionated, positive, and negative respectively, and Rel for being relevant from the retrieval module. We rerank all the blogs using a linear interpolation between the opinionated measurement and the relevance score from topic retrieval as follows:

$$Final = \lambda * Rel + (1 - \lambda) * Opi\ (Pos \text{ or } Neg) \quad (1)$$

where λ is a parameter used to adjust the balance between being relevant and opinionated.

Features Used for Blog Opinion Analysis

Lexical Features

These include n-gram of words and part-of-speech (POS) tags. In the following description, we use w_i and p_i to represent a word and its POS tag.

- Unigram features: This feature only carries information of an individual word, including a combination of word identify with its POS tag, for example, w_i, $w_i p_i$.
- Bigram and trigram of words and POS tags: These features are expected to capture the phrasal level feature and some syntactic patterns of opinion expressions. Examples of these features are: $w_{i-1}w_i$, $w_{i-1}w_i w_{i+1}$, $p_{i-1}w_i p_{i+1}$, $w_{i-1}p_i w_{i+1}$, $p_{i-1}p_i p_{i+1}$.

Sentiment Scores Based on Distributional Association Among Sentimental Expressions

The following steps describe how we extract these features.

(A) Generate sentiment terms

We started with a small set of sentiment seed terms that we think are context independent and are good indicators of opinions, e.g., *good, excellent, bad, terrible*. Then we automatically identify adjectives that have a high co-occurrence with these sentimental words based on a collection of reviews.[5] For a reliable estimation, we used a large co-occurring frequency threshold of ten in a context window of length three. Then a native English speaker manually examined the generated list of sentiment terms and kept 50 positive sentimental terms and 50 negative ones, for example, *delicious, glorious, problematic, stupid*.

(B) Calculate MI score for adjectives

We compute the MI scores between each of the sentimental terms we compiled above and any adjective in our blog training data. This is used as a measurement for the polarity strength (positive and negative) of an adjective. The positive score for an adjective w_i is obtained as follows:

$$MI^{+}_{w_i} = \frac{1}{N} \sum_{t \in S_+} C(w_i, t, win) \quad (2)$$

where S_+ is the set of positive sentiment terms with size of N; $C(w_i, t, win)$ is the frequency that w_i co-occurs with a sentiment term t within a contextual window size of win (five in our system). Similarly we calculate a word's negative score using the negative sentiment terms.

(C) Compute sentiment score features

Finally, we calculate the sentiment score for each sentence by simply adding the corresponding MI scores of all the adjectives in this sentence. Based on that, the following features are derived.

- Mean of sentence sentiment scores for positive and negative respectively.
- Mean of the difference between positive and negative scores among all the sentences.
- Mean of the ratio of positive and negative scores among all the sentences.

Polarized Features

We also explore the polarized features by combining the sentiment terms' polarization tags with their neighboring words and part-of-speech tags. We expect this to represent more opinion indicative patterns. For example, "good" becomes "POS" (positive), and polarized features include trigrams such as $w_{i-1}POSw_{i+1}$, $p_{i-1}POSp_{i+1}$.

Experiments

Classification Setting

In the TREC reference data, there are four opinion tags: "1" denotes non-opinionated, "2" negative opinionated, "3" mixed opinionated, and "4" positive opinionated. In the

[3]This was done using the "mxterminator" sentence boundary detection toolkit, developed by Adwait Ratnaparkhi.

[4]http://www.lemurproject.org/

[5]This corpus comprises of movie reviews from (Pang, Lee, & Vaithyanathan 2002), custom reviews from (Hu & Liu 2004), and some hotel reviews.

training data, the percentages of these four classes are 42.17%, 17.28%, 18.24%, and 22.31% for tag "1", "2", "3" and "4" respectively.

For the opinion finding task, we compare the following two classification paradigms (binary vs. 4-way).

- Binary classification: All the instances with tags "2,3,4" can be grouped together as positive class, and tag "1" corresponds to negative class.
- 4-way classification: Obviously, we can simply train a 4-way classifier based on the four tags, and then we assign blogs labeled with "2,3,4" hypotheses as opinionated.

For the polarized opinion finding, we evaluate three classification strategies:

- One stage with 4-way classification (1S+4W): A 4-way classifier was trained to distinguish blogs as no-opinion, negative opinion, mixed opinion, and positive opinion. Then based on the classifier's hypothesis, we can generate a positive and negative ranked list respectively with the corresponding posterior probabilities.
- Two-stage with successive binary classification and 3-way classification (2S+B+3W): The first stage was simply a binary classification for opinion finding. Then in the second step, a 3-way classifier trained using blog instances with "2,3,4" tags determines the polarity tag for the opinionated blogs generated from the first stage. Blogs classified as "2" and "4" are selected for the final negative and positive lists.
- One stage with 3-way classification (1S+3W): A 3-way classifier trained to distinguish tags "2,3,4" is applied directly to all the retrieved relevant blogs, generating the positive and negative lists.

In addition, considering that some blogs classified as mixed polarity (tag 3) might also belong to the positive or negative ones, we select the ones labeled as "mixed" tag with high posterior probabilities for positive tags and add them to the end of the existing positive ranked list in the order of the posterior probability, until we reach 400 blogs (400 is an empirical number we have chosen). The same rule is also used for the negative ranked list.

We used the Maximum Entropy classifier[6] in our experiments where the Gaussian prior was 0.1, the number of iterations was 100, and the other parameters were the default ones. We used two evaluation metrics: conventional classification accuracy and mean average precision (MAP) in TREC.

Effects of Different Linguistic Features

In the following experiments, we used the topics in 2006 as our training data, and topics in 2007 and 2008 as the development and test data respectively. Table 1 shows the 5-fold cross validation classification accuracy on the training data using different features.

Features	binary	3-way	4-way
unigram	**60.60**	42.29	37.71
+ bigram and trigram	60.35	42.83	40.24
+ sentiment score	60.40	43.42	40.62
+ polarized features	57.11	**44.50**	**40.87**

Table 1: Effect of different features on opinion classification using 5-fold cross validation on the training data.

We can see from Table 1 that the features we added gradually improve the accuracy for both 3-way and 4-way classification. This is consistent with our expectation. However, a different pattern is observed for binary classification. Adding higher order n-gram features or sentiment score features does not help. In particular, there is a noticeable performance degradation after adding polarized features. We will later use the development set and the MAP metric to draw further conclusions on the effect of polarized features.

Comparison among Different Classification Settings in the TREC Framework

Next we compare different classification strategies in the TREC framework. Performance is measured using the MAP score on the development data.

- Opinion finding task

  Table 2 shows the results for opinion finding on the TREC 2007 topics using binary or 4-way classification strategies, as well as with or without polarized features (PF). We can see that the binary classification framework outperforms the 4-way classification. This may be because that binary setting can help the classifier better distinguish opinionated features from non-opinionated features. We also notice that adding the polarized features in binary classification yielded an improvement on the MAP score. This is mostly likely due to the class distribution in the data set, where the non-opinionated blogs are the majority class. The MAP score is a more appropriate measurement for this task with skewed data.

Classification setting	with PF	Map
4-way	No	0.384
binary	No	0.40
binary	Yes	0.45

Table 2: Opinion finding results on TREC Blog 2007 topics.

- Polarized task

  Polarity results on the development data are shown in Table 3. The classification strategy of one-stage with 3-way classification (1S+3W) obtained the best results. This one-stage approach can effectively avoid the error prop-

[6] Available at http://homepages.inf.ed.ac.uk/s0450736/maxent_toolkit.html

agation caused by using two stages. The 3-way classification could also alleviate the problem arisen from the imbalanced training data as in 4-way classification (the non-opinionated class is the majority one). This yielded a significant gain in negative polarity MAP, from 0.07 to 0.16. Again, adding polarized features is useful in the polarity task, especially for the positive class. The expansion using blogs with mixed tags based on their corresponding posterior probability yielded significant performance gain (see the last two rows). Since there is a large number of blogs with multiple polarity opinions, the classifier often predicts one instance as mixed class rather than positive or negative class. Therefore a postprocessing step to handle the mixed hypotheses helps improve performance.

Setups	With PF	Expansion	Pos (Map)	Neg (Map)
1S+4W	No	No	0.26	0.07
2S+B+3W	No	No	0.16	0.14
2S+B+3W	Yes	No	0.20	0.13
1S+3W	Yes	No	0.26	0.16
1S+3W	Yes	Yes	0.35	0.27

Table 3: Results of polarity task on TREC Blog 2007 topics.

Performance on 2008 Test Data

We tested the opinion analysis system on the TREC 2008 data. The classifiers were trained using the reference annotation from the 2006 and 2007 data. For the opinion finding task, the best MAP result our system achieved is 0.3844 (using baseline 4 provided by TREC), comparable with the best result 0.4155 (using the same baseline input) in the TREC evaluation (Ounis, Macdonald, & Soboroff 2008). This is reasonable because the best system used deep level features such as parsing information, yet ours is only based on some easily extracted lexical features. We also found that different baselines yielded quite different performance, suggesting that the quality of topic retrieval has a great impact on overall opinion retrieval system. In addition, we examined the performance curve while the weight λ changes when interpolating the opinion score and the relevance score. We found that for three baselines (3, 4 and 5) a bigger λ is preferred, indicating a more dominant role of the topic retrieval component. This is not true for the other two baselines, which we believe is because of the different quality of the retrieval system as well as the appropriateness of the relevance scores.

For the polarity task in TREC 2008, our best result (0.135/0.096) was obtained using baseline 4, which is again slightly worse than the best TREC evaluation result (0.161/0.148). We also observed that the performance for polarized opinion finding is much worse than only opinion finding. It suggests that the polarity task is more challenging, and for some blogs with mixed opinions, it is difficult to determine whether the overall opinionated orientation is positive or negative, even for human subjects.

Conclusion and Future Work

In the context of TREC Blog track, we have examined various lexical features for opinion finding and polarized opinion finding tasks. In addition, we compared different classification settings. For opinion finding (whether a blog is opinionated or not), we found that adding more features does not improve classification performance based on accuracy metric; however, using all the features that we investigated proved to be useful according to the MAP scores, for both opinion finding and polarized task. Our experiments also show that different classification settings significantly impacted the system performance for the two tasks. The best system result in TREC 2008 is slightly better than ours; however, our approach is much more simple and does not need much human knowledge to create a large opinion vocabulary or perform deep linguistic analysis such as parsing.

One of our future work is to investigate the characteristics of blogs and incorporate more effective features to better identify opinions in blogs. We also plan to find a better approach to determine the polarity of a blog, especially for those containing mixed opinions.

References

Benamara, F.; Cesarano, C.; and Reforgiato, D. 2007. Sentiment analysis: Adjectives and adverbs are better than adjectives alone. In *Proceedings of ICWSM*.

Dave, K.; Lawrence, S.; and Pennock, D. M. 2003. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of WWW*.

Gamon, M., and Aue, A. 2005. Automatic identification of sentiment vocabulary: Exploiting low association with known sentiment terms. In *Proceedings of ACL Workshop on Feature Engineering for Machine Learning in NLP*.

Hatzivassiloglou, V., and McKeown, K. R. 1997. Predicting the semantic orientation of adjectives. In *Proceedings of ACL*.

Hu, M., and Liu, B. 2004. Mining and summarizing customer reviews. In *Proceedings of ACM SIGKDD*.

Li, S., and Zong, C. 2008. Multi-domain sentiment classification. In *Proceedings of ACL*.

Ounis, I.; Macdonald, C.; and Soboroff, I. 2008. Overview of the trec-2008 blog track. In *Proceedings of TREC*.

Pang, B.; Lee, L.; and Vaithyanathan, S. 2002. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of EMNLP*.

Riloff, E., and Wiebe, J. 2003. Learning extraction patterns for subjective expressions. In *Proceedings of EMNLP*.

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

Wilson, T.; Wiebe, J.; and Hoffmann, P. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of HLT-EMNLP*.