Improving Blog Polarity Classification via Topic Analysis and Adaptive Methods

Feifan Liu
University of Wisconsin, Milwaukee
liuf@uwm.edu

Dong Wang, Bin Li, Yang Liu
The University of Texas at Dallas
dongwang,yangl@hlt.utdallas.edu

Abstract

In this paper we examine different linguistic features for sentimental polarity classification, and perform a comparative study on this task between blog and review data. We found that results on blog are much worse than reviews and investigated two methods to improve the performance on blogs. First we explored information retrieval based topic analysis to extract relevant sentences to the given topics for polarity classification. Second, we adopted an adaptive method where we train classifiers from review data and incorporate their hypothesis as features. Both methods yielded performance gain for polarity classification on blog data.

1 Introduction

Sentimental analysis is a task of text categorization that focuses on recognizing and classifying opinionated text towards a given subject. Different levels of sentimental analysis has been performed in prior work, from binary classes to more fine grained categories. Pang et al. (2002) defined this task as a binary classification task and applied it to movie reviews. More sentiment classes, such as document objectivity and subjectivity as well as different rating scales on the subjectivity, have also been taken into consideration (Pang and Lee, 2005; Boiy et al., 2007). In terms of granularity, this task has been investigated from building word level sentiment lexicon (Turney, 2002; Moilanen and Pulman, 2008) to detecting phrase-level (Wilson et al., 2005; Agarwal et al., 2009) and sentence-level (Riloff and Wiebe, 2003; Hu and Liu, 2004) sentiment orientation. However, most previous work has mainly focused on reviews (Pang et al., 2002; Hu and Liu, 2004), news resources (Wilson et al., 2005), and multi-domain adaptation (Blitzer et al., 2007; Mansour et al., 2008). Sentiment analysis on blogs (Chesley et al., 2005; Kim et al., 2009) is still at its early stage.

In this paper we investigate binary polarity classification (positive vs. negative). We evaluate the genre effect between blogs and review data and show the difference of feature effectiveness. We demonstrate improved polarity classification performance in blogs using two methods: (a) integrating topic relevance analysis to perform topic specific polarity classification; (b) adopting an adaptive method by incorporating multiple classifiers' hypotheses from different review domains as features. Our manual analysis also points out some challenges and directions for further study in blog domain.

2 Features for Polarity Classification

For the binary polarity classification task, we use a supervised learning framework to determine whether a document is positive or negative. We used a subjective lexicon, containing 2304 positive words and 4145 negative words respectively, based on (Wilson et al., 2005). The features we explored are listed below.

(i) Lexical features (LF)

We use the bag of words for the lexical features as they have been shown very useful in previous work.

(ii) Polarized lexical features (PL)

We tagged each sentiment word in our data set with its polarity tag based on the sentiment lexicon ("POS" for positive, and "NEG" for negative), along with its part-of-speech tag. For example, in the sentence "It is good, and I like it", "good" is tagged as "POS/ADJ", "like" is tagged as "POS/VRB". Then we encode the number of the polarized tags in a document as features.

(iii) Polarized bigram features (PB)

Contextual information around the polarized words can be useful for sentimental analysis. A word may flip the polarity of its neighboring sentiment words even though this word itself is not necessarily a negative word. For example, in "Given her sudden celebrity with those on the left..." (a sentence in a political blog), "sudden" preceding "celebrity" implies the author's negative attitude towards "her". We combine the sentiment word's polarized tag and its following and preceding word or its part-of-speech to comprise different bigram features to represent this kind of contextual information. For ex-

Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL, pages 309–312,
Los Angeles, California, June 2010.

ample, in "I recommend this.", "recommend" is a positive verb, denoted as "POS/VRB", and the bigram features including this tag and its previous word "I" are "I_POS/VRB" and "pron_POS/VRB".

(iv) Transition word features (T)

Transition words, such as "although", "even though", serve as function words that may change the literal opinion polarity in the current sentence. This information has not been widely explored for sentiment analysis. In this study, we compiled a transition word list containing 31 words. We use the co-occurring feature between a transition word and its nearby content words (noun, verb, adjective and adverb) or polarized tags of sentiment words within the same sentence, but not spanning over other transition words. For example, in "Although it is good", we use features like "although_is","although_good" and "although_POS/ADJ", where "POS/ADJ" is the PL feature for word "good".

3 Feature Effectiveness on Blogs and Reviews

The blog data we used is from the TREC Blog Track evaluation in 2006 and 2007. The annotation was conducted for the 100 topics used in the evaluation (blogs are relevant to a given topic and also opinionated). We use 6,896 positive and 5,300 negative blogs. For the review data, we combined multiple review data sets from (Pang et al., 2002; Blitzer et al., 2007) together. It contains reviews from movies and four product domains (kitchen, electronics, books, and dvd), each of which has 1000 negative and 1000 positive samples. For the data without sentence information (e.g., blog data, some review data), we generated sentences using the maximum entropy sentence boundary detection tool[1]. We used TnT tagger to obtain the part-of-speech tags for these data sets.

For classification, we use the maximum entropy classifier[2] with a Gaussian prior of 1 and 100 iterations in model training. For all the experiments below, we use a 10-fold cross validation setup and report the average classification accuracy. Table 1 shows classification results using various feature sets on blogs and review data. We keep the lexical feature (LF) as a base feature, and investigate the effectiveness of adding more different features. We used Wilcox signed test for statistical significance test. Symbols "†" and "§" in the table indicate the significant level of 0.05 and 0.1 respectively, compared to the baseline performance using LF feature setup.

For the review domain, most of the feature sets can significantly improve the classification performance over the baseline of using "LF" features. "PB" features yielded more significant improvement than "PL" or "T" feature categories. Combining "PL" and "T" features resulted in some slight further improvement, achieving the best accuracy of 83.94%. We notice that incorporating our proposed transition feature (T) always achieves some gain on different feature settings, suggesting that those transition features are useful for sentimental analysis on reviews.

Feature Set	Blogs	Reviews
LF	72.07	81.67
LF+PL	70.93	81.93
LF+PB	72.44	83.62†
LF+T	72.17	81.76
LF+PL+PB	70.81	83.61†
LF+PL+T	72.74	82.13§
LF+PB+T	72.29	83.73†
LF+PL+PB+T	71.85	83.94†

Table 1: Polarity classification results (accuracy in %) using different features for blogs and reviews.

From Table 1, we can see that overall the performance on blogs is worse than on the review data. We hypothesize this may be due to the large variation in terms of contents and styles in blogs. Regarding the feature effectiveness, we also observe some differences between blogs and reviews. Adding the polarized bigram feature and transition feature (PB and T) individually can yield some improvement; however, adding both of them did not result in any further improvement – performance degrades compared to LF+PB. Interestingly, although "PL" feature alone does not seem to help, by adding "PL" and "T" together, the performance achieved the best accuracy of 72.74%. We also found that adding all the features together hurts the performance, suggesting that different features interact with each other and some do not combine well (e.g., PB and T features). In addition, all the improvements here are not statistically significant.

Note that for the blog data, we randomly split them for the cross validation experiments regardless of the queries. In order to better understand whether the poor results on blog data is due to the effect of different queries, we performed another experiment where for each query, we randomly divided the corresponding blogs into training and test splits. Only 66 queries were kept for this experiments – we did not include those queries that have fewer than 10 relevant blogs. The results for the query balanced split on blogs are shown in Figure 1. We also include results for the five individual review data sets in order to see the topic effect. We present results using four representative feature sets chosen according to Table 1. For the review data, we notice some difference across different data sets, suggesting their inherent difference in terms of task difficulty. We observe slight performance increase for some feature sets using the query balanced setup for blog data, but overall it is still much worse than the review data. This shows that the query unbalanced training/test split does not explain the performance gap between blogs and reviews. This is consistent with (Zhang et al., 2007) that found that a query-independent classifier performs even better than query-dependent one. We expect that the

[1] http://stp.ling.uu.se/~gustav/java/classes/MXTERMINATOR.html

[2] http://homepages.inf.ed.ac.uk/s0450736/maxent_toolkit.html

query unbalanced setup is more realistic, therefore, in the following experiments, we continue with this setup.

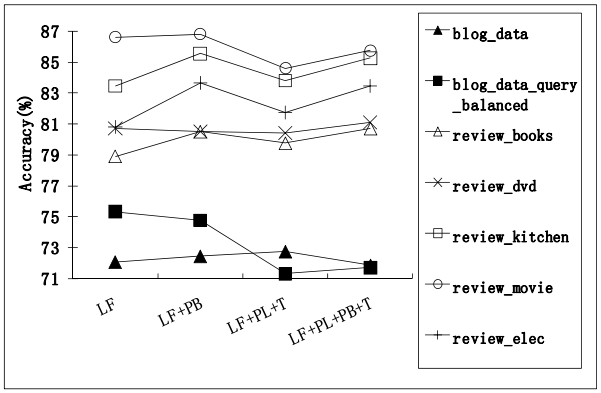

Figure 1: Polarity classification results on query balanced blog data and five individual review data sets.

4 Improving Blog Polarity Classification

To improve the performance of polarity classification on blogs, we propose two methods: (a) extract only topic-relevant segments from blogs for sentiment analysis; (b) apply adaptive methods to leverage review data.

4.1 Using topic-relevant blog context

Generally a review is written towards one product or one kind of service, but a blog may cover several topics with possibly different opinions towards each topic. The blog data we used is annotated based on some specific topics in the TREC Blog Track evaluation. Take topic 870 in the data as an example, "Find opinions on alleged use of steroids by baseball player Barry". There is one blog that talks about 5 different baseball players in issues of using steroids. Since the reference opinion tag of a blog is determined by polarity towards the given query topic, it might be confusing for the classifier if we use the whole blog to derive features. Recently topic analysis has been used for polarity classification (Zhang et al., 2007; Titov and McDonald, 2008; Wiegand and Klakow, 2009). We take a different approach in this study.

In order to obtain a topic-relevant context, we retrieved the top 10 relevant sentences corresponding to the given topic using the Lemur toolkit[3]. Then we used these sentences and their immediate previous and following sentences for feature extraction in the same way as what we did on the whole blog. In addition to using all the words in the relevant context, we also investigated using only content words since those are more topic indicative than function words. We extracted content words (nouns, verbs, adjectives and adverbs) from each blog in their original order and apply the same feature extraction process as for using all the words.

[3]http://www.lemurproject.org/lemur/

Table 2 shows the blog polarity classification results using the whole blog vs. relevant context composed of all the words or only content words. For the significance test, the comparison was done for using relevant context with all the words vs. using the whole blog; and using content words only vs. using all the words in relevant context. Each comparison was with respect to the same feature setup. We observe improved polarity classification performance when using sentence retrieval based topic analysis to extract relevant context. Using all the words in the topic relevant context, all the improvements compared to using the original entire blog are statistically significant at the level of 0.01. We also notice that unlike on the entire blog document, the "PL" features contribute positively when combined with "LF". All the feature settings with "PL" perform very well. The best accuracy of 75.32% is achieved using feature combination of "LF+PL" or "LF+PL+T". This suggests that polarized lexical features suffered from the off-topic content when using the entire blog and are more useful within contexts of certain topics.

When using content words only, we observe consistent gain across all the feature sets. Three feature settings, "LF+PB","LF+T" and "LF+PL+PB+T", achieve statistically significant further improvement (compared to using all the words of relevant contexts). The best accuracy (75.6%) is achieved by using the "LF+PB" features.

Feature Set	Whole Blog	Relevant Context	
		All Words	Content Words
LF	72.07	74.92†	75.14
LF+PL	70.93	75.32†	75.34
LF+PB	72.44	75.03†	75.6†
LF+T	72.17	75.01†	75.35§
LF+PL+PB	70.81	75.27†	75.35
LF+PL+T	72.74	75.32†	75.41
LF+PB+T	72.29	75.17†	75.42
LF+PL+PB+T	71.85	75.21†	75.45§

Table 2: Blog polarity classification results (accuracy in %) using topic relevant context composed of all the words or only content words.

4.2 Adaptive methods using review data

Domain adaptation has been studied in some previous work (e.g., (Blitzer et al., 2007; Mansour et al., 2008)). In this paper, we evaluate two adaptive approaches in order to leverage review data to improve blog polarity classification. In the first approach, in each of the 10-fold cross-validation training, we pool the blog training data (90% of the entire blog data) together with all the review data from 5 different domains. In the second method, we augment features with hypotheses obtained from classifiers trained using other domain data. Specifically, we first trained 5 classifiers from 5 review domain data sets respectively, and encoded the hypotheses from different classifiers as features for blog training (together with the original features of the blog data). Results of these two approaches are shown in Table 3. We use the topic rele-

vant context with content words only in this experiment, and present results for different feature combinations (except the baseline "LF" setting). The significance test is conducted in comparison to the results using only blog data for training, for the same feature setting.

We find that the first approach does not yield any gain, even though the added data is about the same size as the blog data. It indicates that due to the large difference between the two genres, simply combining blogs and reviews in training is not effective. However, we can see that using augmented features in training significantly improved the performance across different feature sets. The best result is achieved using "LF+T" features, 76.84% compared with the best accuracy of 75.6% when using the blog data only ("LF+PB" features).

Feature Set	Only Blog	Pool Data	Augment Features
LF+PL	75.34	75.05	76.12†
LF+PB	75.6	74.35	76.28†
LF+T	75.35	74.47	76.84†
LF+PL+PB	75.35	74.94	76.7†
LF+PL+T	75.41	74.85	76.32†
LF+PB+T	75.42	74.46	76.3†
LF+PL+PB+T	75.45	74.96	76.53†

Table 3: Results (accuracy in %) of blog polarity classification using two methods leveraging review data.

4.3 Error analysis

Notice that after achieving some improvements the performance on blogs is still much worse than on review data. Thus we performed a manual error analysis for a better understanding of the difficulties of sentiment analysis on blog data, and identified the following challenges.

(a) Idiomatic expressions. Compared to reviews, bloggers seem to use more idioms. For example, "Of course he has me over the barrel..." expresses negative opinion, however, there are no superficially indicative features.

(b) Ironic writing style. Some bloggers prefer ironic style especially when speaking against something or somebody, whereas opinions are often expressed using plain writing style in reviews. Simply using the surface word level features is not able to model these properly.

(c) Background knowledge. In some political blogs, the polarized expressions are implicit. Correctly recognizing them requires background knowledge and deeper language analysis techniques.

5 Conclusions and Future Work

In this paper, we have evaluated various features and the domain effect on sentimental polarity classification. Our experiments on blog and review data demonstrated different feature effectiveness and the overall poorer performance on blogs than reviews. We found that the polarized features and the transition word features we introduced are useful for polarity classification. We also show that by extracting topic-relevant context and considering only content words, the system can achieve significantly better performance on blogs. Furthermore, an adaptive method using augmented features can effectively leverage data from other domains, and yield improvement compared to using in-domain training or training on combined data from different domains. For our future work, we plan to investigate other adaption methods, and try to address some of the problems identified in our error analysis.

6 Acknowledgment

The authors thank the three anonymous reviewers for their suggestions.

References

Apoorv Agarwal, Fadi Biadsy, and Kathleen McKeown. 2009. Contextual phrase-level polarity analysis using lexical affect scoring and syntactic n-grams. In *Proc. of EACL*.

John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proc. of ACL*.

Erik Boiy, Pieter Hens, Koen Deschacht, and Marie-Francine Moens. 2007. Automatic sentiment analysis in on-line text. In *Proc. of ELPUB*.

Paula Chesley, Bruce Vincent, Li Xu, and Rohini K. Srihari. 2005. Using verbs and adjectives to automatically classify blog sentiment. In *Proc. of AAAI*.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proc. of ACM SIGKDD*.

Jungi Kim, Jin-Ji Li, and Jong-Hyeok Lee. 2009. Discovering the discriminative views: Measuring term weights for sentiment analysis. In *Proc. of ACL-IJCNLP*.

Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. 2008. Domain adaptation with multiple sources. In *Proc. of NIPS*.

Karo Moilanen and Stephen Pulman. 2008. The good, the bad, and the unknown: Morphosyllabic sentiment tagging of unseen words. In *Proc. of ACL*.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proc. of ACL*.

Bo Pang, Lilian Lee, and Shrivakumar Vaithyanathan. 2002. Thumbs up? sentiment classification using machine learning techniques. In *Proc. of EMNLP*.

Ellen Riloff and Janyce Wiebe. 2003. Learning extraction patterns for subjective expressions. In *Proc. of EMNLP*.

Ivan Titov and Ryan McDonald. 2008. Modeling online reviews with multi-grain topic models. In *Proc. of WWW*.

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

Michael Wiegand and Dietrich Klakow. 2009. Topic-Related polarity classification of blog sentences. In *Proc. of the 14th Portuguese Conference on Artificial Intelligence: Progress in Artificial Intelligence*, pages 658–669.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proc. of HLT-EMNLP*.

Wei Zhang, Clement Yu, and Weiyi Meng. 2007. Opinion retrieval from blogs. In *Proc. of CIKM*, pages 831–840.