A combined sentiment classification system for SIGHAN-8

Qiuchi Li[1], Qiyu Zhi[2] and Miao Li[1]

[1]Multimedia Signal and Intelligent Information Processing Lab,
Department of Electronic Engineering, Tsinghua University, Beijing, P.R.China
[2]School of Information and Communication Engineering,
Beijing University of Posts and Telecommunications, Beijing, P.R.China
liqc@126.com, zhiqiyubupt@gmail.com, miao-li10@mails.tsinghua.edu.cn

Abstract

This paper describes our system (MSIIP_THU) used for Topic-Based Chinese Message Polarity Classification Task in SIGHAN-8. In our system, a lexicon-based classifier and a statistical machine learning-based classifier are built up, followed by a linear combination of these two models. The overall performance of the proposed framework ranks in the middle of all terms participating in the task.

1 Introduction

Sentiment analysis is becoming an alluring task in natural language processing(NLP). Since an increasing amount of data is available on the World Wide Web, sentiment analysis is playing an important role in lots of real-world applications. In particular, sentiment analysis on microblogs is especially essential as microblog becomes one of the most fashionable ways for people to communicate with each other, express opinions and acquire newest information. However, with limited length, sentiment analysis on microblog remains a challenging task.

In this paper, we focus on sentiment classification for Chinese microblog, i.e. Weibo. The research on Weibo sentiment starts later and produces higher challenges due to the complexity in Chinese language. On one hand, different from alphabetic languages such as English, word segmentation is needed and is more difficult for Chinese sentences. On the other hand, polysemy in Chinese is abundant.

As for existing works in the area of Weibo sentiment analysis, some methods are proposed on the lexicon basis. (Taboada et al., 2011) proposed Semantic Orientation CALculator (SO-CAL) using dictionaries of words annotated with their semantic orientation, (Baccianella et al., 2010) presented SENTIWORDNET 3.0, a lexical resource explicitly devised for supporting sentiment classification and opinion mining applications. Other researchers focus on machine learning approaches. (Mullen and Collier, 2004) introduced an approach to sentiment analysis which used support vector machines (SVMs) to bring together diverse sources of potentially pertinent information. The same framework is adopted in (Mohammd et al., 2013), where systematic experiments on a great variety of features were conducted, leading to the best-performed results in SEMEVAL-2013 Twitter Sentiment Classification competition. In this task, we combine these two typical methods to build our system.

The rest of this paper is organized as follows. Section 2 describes the topic-based sentiment classification task and its dataset. Section 3 introduces our preprocess procedure for Weibo. Section 4 and Section 5 respectively shows the lexicon-based model and the statistical model used in this task. Section 6 describes the combination method and the experimental results. Finally we conclude this paper in Section 7.

2 Task Description

The paper is targeted on the Topic-Based Chinese Message Polarity Classification. Given a message from Chinese Weibo Platform and a topic, one needs to classify whether the message is of positive, negative, or neutral sentiment towards the given topic. Each participant is required to submit two results based on the restricted resource and unrestricted resource respectively. The restricted resource includes restricted lexicon and corpus, which have been released together with the test data.

The given training corpus has around 5,000 Chinese Weibos from 5 different topics. After duplicate removal we obtain 4619 Weibos. The 3-class annotation of all Weibos are given in an-

other file. Moreover, we collected 43789 Weibos from NLPCC 2012,2013 and 2014 evaluation. These Weibos have no topic labels, but are annotated with 3-class labels. We use the collection as extra resource for the unrestricted resource task. The test data involves 19489 Weibos from 20 topics. These topics are different from the ones in the training corpus. The task is to annotate each Weibo in the test data.

The key measures for evaluation are overall accuracies and F-parameters for positive label and negative label. The mathematical formulations for these measures are omitted here because they are the most commonly used ones in sentiment analysis evaluations.

3 Preprocess Procedure

Although having a 140-character limitation, most Weibo has some unexpected characters, which poses an obstacle for us to extract features from the corpus and segment the sentence. Hence, preprocessing the Weibo data is a necessary step in sentiment analysis.

With regard to the corpus of this task, we first eliminate all the rare characters, then we extract all the punctuation, URLs and Weibo functional symbols such as "@" and "#". Finally we use NLPIR (Zhang et al., 2003) to segment the Weibo sentence.

4 Lexicon-based Approaches

Here we present our Lexicon-Based sentiment analysis approach. Sentiment lexicon is a simple, direct and efficient method to analyze sentiment by statistical method. In this section, the lexicon is firstly introduced, and then the algorithms for restricted and unrestricted lexicon are presented.

4.1 Basic Sentiment Lexicon

There are lots of lexicons that can be used for our task, such as Hownet Sentiment Dictionary (Dong, 2000), National Taiwan University Sentiment Dictionary (NTUSD) (Ku and Chen, 2007) and Chinese Emotion Word Ontology (CEWO) (Yan et al., 2008). We choose NTUSD as our base sentiment lexicon. The composition of this sentiment lexicon is shown in Table 1.

Table 1: NTUSD Sentiment Lexicon

Polarity	Number
1	2810
-1	8276

4.2 Weibo Emoticon Lexicon

Emoticon is proved to be important for Weibo sentiment classification task. Since sarcasm is common in Weibo expressions, a sentiment word may express the opposite emotion in sarcasm case, while the emoticons often reflect the real sentiment of the writer. We build a Weibo emoticon lexicon for unrestricted resources task. We extract all of the emoticons in training corpus, and then we incorporate common emoticons from Weibo platform. We manually label every emoticon in our lexicon with $10, -10, 1, -1, 0$. ± 10 represents the sentiment intensity for an emoticon strong enough to affect the sentiment of the whole sentence, while ± 1 refers to an emoticon with clear sentiment but not enough to decide the sentence sentiment. 0 represent emoticon without any emotional tendency. The composition of this emoticon lexicon is shown in Table 2.

Table 2: Emoticon Sentiment Lexicon

Polarity	Number
-10	12
10	22
-1	89
1	85
0	85

4.3 The Lexicon-based classifier

Like many Lexicon-based methods, we simply calculate the score of a Weibo sentence by adding up the scores of each sentiment words appeared in the sentence. For restricted resource task, only NTUSD Lexicon is used. Our Emoticon Lexicon is added to the Lexicon in unrestricted resource task.

5 Machine Learning-Based Approaches

Support Vector Machine (SVM)(Cortes and Vapnik, 1995) is used as the statistical classifier. We use a rich feature set to build the model.

5.1 Features

5.1.1 Linguistic Features

In this part, different linguistic features are considered. For the choice of n-gram, we only consider n=1 (refer to as unigram) and n=2 (refer to as bigram) due to the limited size of training corpus. For other linguistic features, we also extract character-bigram and TFIDF features from the training dataset. In section 5.2 we will discuss ways to select these features.

5.1.2 Weibo-Based Features

Apart from linguistic features, we also extract a series of Weibo-based features shown as follows:

- **Textlength**. It is believed that long Weibos tend to contain more sentiment terms, and thus are more likely to be non-neutral in sentiment.

- **Hashtag**. We consider Hashtags ("#") because they usually include topic information for a Weibo. The number of Hashtags are extracted in our experiment.

- **Punctuation**. We assume that punctuation is relevant to Weibo sentiment. We extract the number of four commonly used punctuation as features: period(""), comma(""), exclamation ("") and question ("").

- **Emoticon**. Based on the pre-constructed dictionary, we extract the number of positive and negative emoticons respectively for a Weibo, forming a 2-dimension feature vector.

- **POS**. It is spontaneous that a Weibo's sentiment can be reflected in the Part-Of-Speech (POS) features. In this paper the number of nouns, adjectives, verbs and adverbs are extracted, forming a 4-dimension feature vector.

- **URL**. The contents in the url link may be relevant to the content of the Weibo and the sentiment polarity. The number of URLs is extracted in our model.

- **ATSign**. ATSigns ("@") associate a Weibo with other people, and prior knowledge of those people may affect the Weibo sentiment. The number of ATSigns is extracted in our model.

Table 3: results for feature selection

features	neuF	posF	negF
all	0.6096	0.1975	0.3194
-Unigram	0.6091	0.1941	0.3258
-Bigram	0.6444	0.2083	0.3296
-Character-Bigram	0.7099	0.2210	0.3410
-TFIDF	0.6313	0.2105	0.3358
-textLength	0.5969	0.1905	0.3459
-#	0.6096	0.1979	0.3182
-punctuation	0.6159	0.1949	0.3282
-Emoticon	0.6134	0.1987	0.3194
-POS	0.5482	0.178	0.3346
-URL	0.5987	0.1934	0.3194
-@	0.5995	0.1837	0.3580

5.2 Feature Selection

The feature selection method is inspired by (Mohammd et al., 2013). For a detailed description, we first experiment on all aforementioned features, and then in turn kick out every feature and repeat the experiment. To make a fair comparison, in each experiment a five-fold cross validation method is proposed on the training set, and we average the F-parameters for negative, neutral and positive labels over the five sub-experiments to measure the performance of the feature combination. For the SVM training setup, we use linear kernel and default parameter. The results for the feature selection experiments are shown in Table 3.

From Table 3, the elimination of Bigram, Character-Bigram and TFIDF bring about increased performance, the elimination of POS leads to decreased performance, while the elimination of other features does not influence much of the performance. Therefore, we choose Unigram as the only linguistic feature, and remain all the Weibo-based features.

6 Model-Fusion Framework

Our final system is set up by merging the two models discussed in Chapter 4 and 5 respectively. The merging method is shown as follows. For a Weibo w, we have

$$decisionValue(w) = \lambda C_{dic}(w) + (1 - \lambda)C_{svm}(w) \quad (1)$$

where $C_{dic}(w)$ and $C_{svm}(w)$ are the classification results for the lexicon-based system and the machine learning-based system, and $\lambda \in [0, 1]$ is the linear combination parameter. The computed

decision value $decisionValue(w)$ is a real-valued number in [-1,1], based on which we obtain the final sentiment polarity as the output of our model-fusion framework:

$$sentiment(w) = \begin{cases} 1 & decisionValue(w){>}{=}0.5 \\ 0 & |decisionValue(w)|{<}0.5 \\ -1 & decisionValue(w){<}{=}\text{-}0.5 \end{cases}$$

7 Experiments

7.1 Experimental Setup

The model-fusion framework is adopted on both restricted and unrestricted requirements, but the parameter choices are slightly different for these two cases.

For restricted results, the parameter λ is set to be 1, which means only lexicon-based system is adopted. Since only two different results can be submitted, we submit the results by considering I) main body of the Weibo only and II)main body and forward chains.

For unrestricted results, we combine the provided training corpus with extra training dataset to train the SVM classifier for machine learning-based system. For lexicon-based system, the main body only is considered, but the emoticon lexicon is incorporated. The two results were generated by setting the fusion parameter λ as 1(lexicon-based only) and 0.5 respectively.

7.2 Results and Discussions

For each subtask (restricted and unrestricted), the better performed system is automatically chosen from the two submitted results, and performance and rank are returned. The results for our system is shown in table 4.

The results show that our system generally ranks in the middle of the 13 teams who participated in the evaluation, which proves the effectiveness of our system. Since our system is targeted on improving F-values, and most Weibos are of neutral sentiment for both training and testing corpus, more non-neutral labels will be generated but with low accuracy. Therefore, our system is unsatisfactory in overall accuracy and precision, but rather competitive in terms of recall and F-values.

It is further revealed that the other system has consistently higher F-values than the accepted system for both tasks. This means that the abandoned system generates more non-neutral polari-

Table 4: Performance and ranks of our system in evaluation.

	value	best	rank
U-ACC	**0.6351**	**0.8535**	**11**
U-pre+	0.1212	0.5880	10
U-rec+	0.1788	0.6203	6
U-F1+	**0.1445**	**0.6039**	**7**
U-2-F1+	**0.2108**	**0.6039**	**5**
U-pre-	0.3412	0.7917	9
U-rec-	0.3954	0.6175	5
U-F1-	**0.3663**	**0.6938**	**6**
U-2-F1-	**0.4096**	**0.6938**	**5**
ACC	**0.6489**	**0.8357**	**9**
pre+	0.0988	0.6258	10
rec+	0.0946	0.5139	9
F1+	**0.0967**	**0.5643**	**10**
2-F1+	**0.1480**	**0.5643**	**7**
pre-	0.3320	0.8232	9
rec-	0.3767	0.4671	4
F1-	**0.3530**	**0.5960**	**5**
2-F1-	**0.3805**	**0.5960**	**4**

Note: Words that start with "U-" stand for unrestricted situation. Words that end with "+" and "-" stand for results on positive and negative polarities. Words that contains "2" refer to the other submitted system. The highlighted values correspond to key measures in the evaluation.

ties, resulting in higher F-values for both positive and negative class, but the overall accuracy is also lower than the recorded system, so it is neglected automatically by the evaluation system.

Nevertheless, the system still needs further improvements. The topic information is not considered, which is a major drawback for our system. The author believes that it will probably be an improvement to discover the topic-specific knowledge using some unsupervised methods prior to the whole system. These knowledge can not only be somehow incorporated into the lexicon-based approach, but be treated as extra features for the machine learning-based system.

8 Conclusion

In this paper, a combined system is proposed on the task of topic-based Chinese Weibo sentiment analysis. It conducts a linear combination between a lexicon-based sentiment classification system and an SVM sentiment classifier. The evaluation results demonstrates the advantageous performance in measures of recall and F-values for non-neutral sentences.

References

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *LREC*, volume 10, pages 2200–2204.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine Learning*, 20:273–297.

Zhengdong Dong. 2000. Introduction to hownet. `http://www.keenage.com`.

Lun-Wei Ku and Hsin-Hsi Chen. 2007. Mining opinions from the web: Beyond relevance retrieval. *Journal of the American Society for Information Science and Technology*, 58(12):1838–1850.

Saif M Mohammd, Svetlana Kirichenko, and Xiaodan Zhu. 2013. NRC-Canada: Building the State-of-the-Art in Sentiment Analysis of Tweets. *Second Joint Conference on Lexical and Computational Semantics*, 2:321–327.

Tony Mullen and Nigel Collier. 2004. Sentiment analysis using support vector machines with diverse information sources. In *EMNLP*, volume 4, pages 412–418.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307.

Jiajun Yan, David B Bracewell, Fuji Ren, and Shingo Kuroiwa. 2008. The creation of a chinese emotion ontology based on hownet. *Engineering Letters*, 16(1):166–171.

Hua-Ping Zhang, Hong-Kui Yu, De-Yi Xiong, and Qun Liu. 2003. Hhmm-based chinese lexical analyzer ictclas. In *Proceedings of the second SIGHAN workshop on Chinese language processing-Volume 17*, pages 184–187. Association for Computational Linguistics.