A Sentiment Analysis Method Based on Emoticons and Sentiment Words

Baolin Gao[1], Zhiguo Zhou[1,*], Mingxue Zou[1] and Chunyan Deng[2]

[1]College of Computer Science and Information Technology, Northeast Normal University, Changchun130117, Jilin, China

[2]College of Computer Science and Technology, Jilin University, Changchun130012, Jilin, China

{gaobl893,zhouzg281,zoumx630}@nenu.edu.cn, dengcy@jlu.edu.cn

Keywords: Emoticons; Sentiment words; Sentiment classification; Weibo

Abstract. Weibo, a Twitter-like online social network in China. The statistical analysis indicates that, most weibo users will use the emoticons when they release tweets. And it is more intuitive to observe the user's sentiment attitude through these emoticons. The purpose of this paper is dividing the weibo text into four categories: happy, sad, angry and disgust, according to the expression of emotion and attitude. This paper uses the Naive Bayes classification method, in the division on the training set, collect 165 commonly used emoticons and 200 sentiment words.According to the emotion intensity that the emoticons and emotion words expressed, give various weights. By computing the emoticons and sentiment words' weighted average value, determine the classification. At the same time, use the category balance method to keep the proportion of four categories balance in the training set, to prevent classifier deviation. Through the experiments, the classification accuracy can reach 78.89%. Indicate that this method can get a better result in weibo text sentiment classification.

Introduction

With the constant development of social networks, people are more willing to express their views through weibo and blog community, comment on the hot spots. So that through the weibo, blog and product evaluation and so on to understand the social network user's emotional tendency has been widely concerned by the academic community. Sentiment analysis based on weibo data is a challenging task, in recent years, has sparked research interest of scholars [1].

At present, the main research methods of sentiment analysis are some traditional algorithms based on machine learning. Such as SVM, information entropy, CRF and so on. These methods are summed up in 3 categories: supervised learning, unsupervised learning and semi supervised learning. And most of the current researches based on supervised learning have achieved good results [2]. However, supervised learning relies on a large number of manually labeled data. Makes the system based on supervised learning need to pay a high labeling cost.

At present, weibo sentiment analysis, have already achieved some results. Zhang Chenggong [3] etc. and Wang Yong [4] etc. use the method based on the polarity lexicon to divide the weibo text into positive and negative. Zhao Jichang [5] etc. use the emoticons as the basis for dividing the training set and use Naive Bayes method to classify the sentiment into four categories. Li Dong [6] etc. use statistical analysis methods to classify the sentiment of weibo text.

In this paper, we mainly use the improved training set divide method. Assign values for a small amount of emoticons and sentiment words, according to the weighted average value of the emoticons and sentiment words in weibo text to divide the training set. Using Naive Bayes method for sentiment classification. As the experiment proves, using this method can significantly improve the accuracy of classification.

Data and Methods

Data Acquisition and Preprocessing. All data are getting from Sina weibo. Through the interface provided by the official, 200 weibo texts can be collected every time. We collected a total of about 100000 weibo texts. Because of there are a lot of garbage weibo texts, such as advertising, etc. So filter out this part of the data. In addition, there are links and other contents in some weibo texts, and this paper does not take this part into account, therefore also filter out. At last, get 68446 valid weibo texts.

Dataset Partition. First, according to the statistical results of the emoticons appearing in about 100000 weibo texts, we have selected 165 emoticons which are the most frequently used. According to the emotional attitude and intensity that the emoticons expressed, give them different values, range of -10 to 10, take integer. -10 represents the most negative emotions, 10 represents the most positive emotions. At the same time, 200 sentiment words with the highest frequency of daily use are selected from the dictionary HowNet [7], 100 positive and 100 negative. Use the same method to give different values to the 200 sentiment words. In the course of assignment, in order to avoid the subjective influence, there are 5 people doing this work. Finally, integrated the views of 5 people.

On the division of the training set, the method of calculating the weighted average value is used. Use Eq.1 to calculate the weighted average value of all the emoticons and sentiment words that appear in weibo text. n represents how many kinds of emoticons and sentiment words in a weibo text. W_i represents the weight of each emoticon and the sentiment word. C_i represents the times that the same emoticon and the sentiment word appear in a weibo text. Then divide the dataset by the average value S. For S, if S<-6.5, then classified as angry, if $-6.5 \leq S < -3.5$, then classified as sad, if $-3.5 \leq S \leq 0$, then classified as disgust, if S>0, then classified as happy. The final results are shown in Table 1.

$$S = \frac{\sum_{i=1}^{n} W_i C_i}{\sum_{i=1}^{n} C_i} \tag{1}$$

Table 1. Classification result

Emotion type	count
sad	8940
happy	27267
angry	3654
disgust	28605

Experimental Analysis

In order to avoid the small proportion of the classification at a disadvantage, so as to affect the accuracy of the classifier. Use the category balance method [8] to balance the number of weibo texts for each category of the training set. Finally, from the divided dataset randomly select data, each category has 3654 weibo texts, as experimental data.

Naive Bayes algorithm [9] is used as a classification method. For a weibo text T, use NLPIR [10] for word segmentation. NLPIR is a Chinese word segmentation system developed by the Chinese Academy of Sciences. After the word segmentation, remove stop words, punctuation and other useless information, we got the words sequence $\{Word_i\}$ of T, where $Word_i$ is a word and i is its position in T. From the labeled training set, we could obtain the $Word_i$'s prior probability of belonging to the emotional type is

$$P(Word_i \mid Type_j) = \frac{n^{Type_j}(Word_i)+1}{\sum_{j=1}^{4}(n^{Type_j}(Word_i)+1)} \quad (2)$$

Where j = 1,2,3 or 4, $n^{Type_j}(Word_i)$ is the times that $Word_i$ appears in all the weibo texts in the category $Type_j$ and Laplace smoothing is used to avoid the problem of zero probability. Then, for an unlabeled weibo text t with words sequence $\{Word_i\}$, its category could be obtained as

$$Type^*(t) = \arg\max_j P(Type_j)\prod_i P(Word_i \mid Type_j) \quad (3)$$

Where $P(Type_j)$ is the prior probability of $Type_j$. In this paper, $Type_j$ is 0.25.

The experiment of this paper uses 5-fold Cross-Validation model. The experimental data is divided into 5 equal parts, 4 parts as training set, and 1 part as test set. For each part, the proportion of the four categories is equal. The results of the experiment are shown in Table 2.

Table 2. Experimental Results

	Test set	angry	sad	disgust	happy	Precision
Exp 1	2924	680	653	387	618	79.95%
Exp 2	2924	665	647	382	621	79.17%
Exp 3	2924	660	668	386	607	79.37%
Exp 4	2924	620	658	360	636	77.77%
Exp 5	2923	648	635	393	610	78.20%

Conclusion

In this paper, we use the naive Bayes algorithm to classify the collected weibo data, and use weighted average method to calculate the weighted average value of emoticons and sentiment words on the division of training set. And use the category balance method to balance the proportion of data in each category of the training set. In the experiment, the 5-fold Cross-Validation model is used to test the method of this paper.

Experimental results show that, using the method proposed in this paper can get a better result for weibo text sentiment classification. The average accuracy can reach 78.89%.

Acknowledgements

This work is supported by the Foundation for Science Front and Crossing Discipline of Jilin University, the Integration of RIA and Panoramic virtural field outcrop Geological Information System. (Chunyan Deng, #450060521045).

This thesis is supported by the teaching development fund of Northeast Normal University. The project name: Reform and construction of the algorithms and program practice course oriented around problem-solving (15B1XZJ014).

References

[1] B. Pang, L. Lee, Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales, Proceedings of the ACL, 2005, pp.115-124.

[2] H.F. Tang, S.B. Tan, X.Q. Cheng, Research on Sentiment Classification of Chinese Reviews Based on Supervised Machine Learning Techniques, Journal of Chinese Information Processing, Vol. 21 (2007) No. 6, p. 88-94. (In Chinese)

[3] C. G. Zhang, P. Y. Liu, Z. Z. Zhu and M. Fang, A sentiment analysis method based on a polarity lexicon, Journal of Shandong University(Natural Science), vol.47 (2012) No.3, pp.47-50. (In Chinese)

[4] Y. Wang, X. Q. Lü, L. C. Ji, and S. B. Xiao, Sentiment classification for Chinese microblogging based on polarity lexicons, Computer Applications and Software, vol.31 (2014) No.1, pp.34-37. (In Chinese)

[5] J. C. Zhao, L. Dong, J. J. Wu, and K. Xu, MoodLens: an emoticon-based sentiment analysis system for Chinese tweets in Weibo, *ACM SIGKDD conference on Knowledge Discovery and Data Mining*(Beijing, China, August 12-16, 2012). pp. 1528-1531.

[6] L. Dong, F. R. Wei, S. J. Liu, M. Zhou and K. Xu, A Statistical Parsing Framework for Sentiment Classification, Computational Linguistics, vol.41 (2015), Issue 2, pp. 293-336.

[7] HowNet on http://www.keenage.com

[8] Q. R. Zhang, L. Zhang, S. B. Dong, and J. H. Tan, Effects of category distribution in training set on text categorization, Journal of Tsinghua University(Science and Technology), vol.45 (2005) No. S1, pp.1802-1805. (In Chinese)

[9] Z. M. Liu, L. Liu, Empirical study of sentiment classification for Chinese microblog based on machine learning, Computer Engineering and Applications, vol.48 (2012) No.1, pp.1-4. (In Chinese)

[10] NLPIR on http://ictclas.nlpir.or