# Research on the text sentiment classification about the social hot events on Weibo

Fulian Yin[1] ,Beibei Zhang[2,*] ,Pei Su[3] ,Juanfang Chai[4]
1,2,3. College of Information Engineering, Communication University of China
Beijing, China
4. Shanghai Electromechanical Engineering Institute
Shanghai, China
yinfulian@cuc.edu.cn, *shatanshibei@163.com, supeiCynthia@163.com, chaijuanfang@163.com

***Abstract*—The public comments on the social hot events on Weibo has attracted lots of attentions in recent years. To remedy the shortage of sentiment analysis about typical events on Weibo, the classification method based on sentiment dictionary is put forward in this paper, and the accuracy rate is close to 50%. This paper also proposed a sentiment classification method based on Naive Bayesian in order to improve the accuracy. We use TF as the weight of feature words and chi-square test to extract features of each category. The accuracy rate is commonly raised to more than 70%, it can reach to 80% when analyzing text of typical events.**

***Keywords—feature extraction; Naive Bayes; sentiment classification; sentiment dictionary***

## I. INTRODUCTION

In recent years, more people tend to express their views and opinions about social hot events on Weibo. To make full use of these data to analyze and integrate the public attitudes and emotion tendency has become a significant role in the field of analyzing and controling the public opinions. As a result, text sentiment analysis on Weibo is becoming a new hot research spot. At present, the emotional analysis is applied in the comments of hotel, movie, product, news and other fields [1],[2],[3]. There are lots of research achievements for text sentiment classification, but most of them are aimed to classify the whole content on Weibo. There are few sentiment research results for specific events. Compared with the traditional long text, the Weibo short text is limited by the number of words, which is characterized by sparse feature, short content and easy expression. In social hot event of different natures, people also express their attitudes in different styles, which makes it difficult to guarantee the text classification effect when applied previous efficient methods to new specific event [4].

Sentiment analysis is also called opinion mining. It is a process that extracts the relevant information of public attitudes from the subjective text. It belongs to the text classification. The research method of emotion analysis can be divided into three categories: the method based on the emotion dictionary and rules, the method based on machine learning and the method based on semantic analysis. The sentiment dictionary and rules based method depends on the quality of sentiment dictionary. The domestic and foreign scholars have done a lot of researches. Kamps used Word Net synonymous structure graph to compute semantic distance between neologisms and benchmark words and come to their tendency [5]. Zhu Yanlan, Min Jin used semantic similarity method to calculate new words' appraisive degree and judge whether a word positive or negative based on HowNet [6]. The method based on machine learning include the famous algorithms such as Naive Bayes [1], KNN [7], Support Vector Machine [8], Maximum Entropy [9] and so on. Xu Jun's experiment showed that the method of machine learning can achieve good classification results in text sentiment classification [3]. The method based on semantic analysis is mainly shallow semantic parsing. Tumey application of semantic orientation method to the automotive, banking, film and other places of the service of the comments were emotional analysis [2]. Mullen and others proposed the potential information of different sources, such as emotional phrases, adjective, subject related words, etc., as the feature, used SVM method to classify the text and the accuracy rate is greatly improved [10].

This paper studied the sentiment classification about public comments of the social hot event on Weibo. First, we put forward the sentiment dictionary based classification method and the accuracy rate is close to 50%. In order to improve the accuracy, we also proposed a sentiment classification method based on Naive Bayesian. We use TF as the feature weight and chi-square test to extract features of each class. The accuracy rate is raised to more than 70%. It's better than present method aiming to the whole unified text on Weibo and has remedied the shortage of sentiment analysis aiming to typical events on Weibo.

## II. SENTIMENT ANALYSIS OF WEIBO HOT EVENTS BASED ON SENTIMENT DICTIONARY

This paper put forward the classification method based on sentiment dictionary to classify the comments about hot events on Weibo. We use the emotional ontology thesaurus of Dalian University of Technology as a basic emotion dictionary, with 1102 colloquial network words we summarized, such as "nima" "shangbuqi" and so on, up to 28848 words marked with their part of speech, as a basis for sentiment classification. The emotional classification process of hot events on Weibo is as follows:

**978-1-4673-9613-4/16/$31.00 ©2016 IEEE**

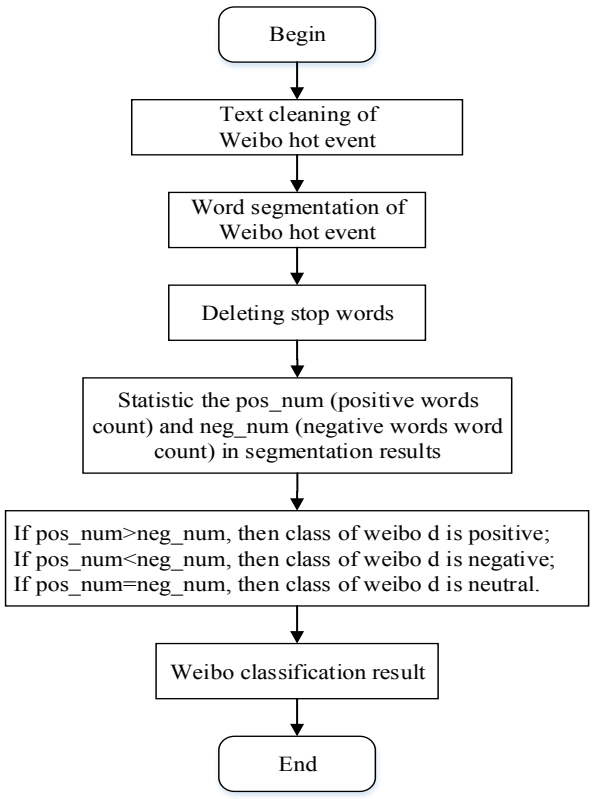


Fig.1. The text emotional classification process of hot events on Weibo based on sentiment dictionary.

Step1: Weibo text cleaning. Remove the Weibo nickname after @, address, source and common phrases which is meaningless on Weibo, such as "|Tian An Men", "I'm in XX", "from the network", "from XX". And remove the URL, number and punctuation.

Step2: Chinese words segmentation. This paper uses the library Rwordseg in software R to segment the text about the hot events on Weibo. R provides the library Rwordseg to segment Chinese words. It refers to the Chinese word segmentation tool ansj developed by Sun Jian. It is based on the Chinese Academy of Sciences of the ICTCLAS and can be used for personal name identification, toponymy identification, organization identification, multi class part of speech tagging, keyword extraction, fingerprint extraction and other fields. It also supports industry dictionary and user defined dictionary. Accuracy of word segmentation and the convenience of a custom dictionary or the operation efficiency performance better.

Step3: Remove the stop words. There are many stop words in Chinese, such as "at", "in", "also", "of", "it", "for". These kinds of words appear frequently and they're meaningless, so we should remove them to reduce the computing time.

Step4: We matched the segmentation result with the Weibo after pre-processing. Statistic the positive words count and negative words count existing in emotional thesaurus and the higher is the Weibo's emotional tendency.

## III. SENTIMENT ANALYSIS OF WEIBO HOT EVENTS BASED ON NAIVE BAYES

The performance of method based on emotional dictionary severely relies on the quality of the dictionary, so the classification performance is far lower than the accuracy of the current general text classification. The popular language in the Internet is spoken. One word always has more than one meaning. Combined with the tone of rich expression such as irony, rhetorical question, the performance of method based on emotional dictionary is reasonably bad. This paper selects the famous algorithm Naive Bayesian as classification method in many machine learning methods, and the classification accuracy generally reaches more than 70%, with the classification results significantly improved.

Text sentiment classification process for Weibo hot events is as follows. Firstly, clean and express the Weibo text of specific hot events. After feature extraction, training classifier, classifier performance evaluation, apply it to the unknown category Weibo text finally.

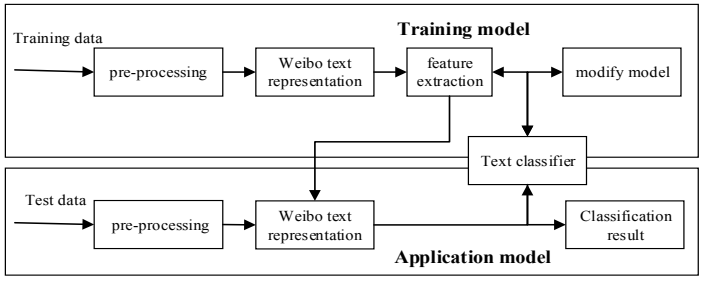


Fig.2. Naive Bayes based sentiment classification process for Weibo hot events.

(1)Weibo text segmentation about hot events

Segment relevant text about Weibo hot event.

(2)Text representation for Weibo hot event

The current text representation model on Weibo is mainly vector space model (VSM) proposed by Gerard Salton and McGill. It's basic idea is simplifing the Weibo text as a vector of feature item weight: $(w_1, w_2, ..., w_n)$, The $w_i$ is the weights of feature item $i$, the features are generally selected out of words and the weight is represented by word frequency.

(3)The text feature extraction of Weibo hot event

In the text classification of Weibo hot events, the size of the dictionary is up to thousands or more, which can't always distinguish one class from another, so we should select some representative features from the large number of features and do not affect the classification effect, that is feature extraction. Commonly the methods of feature extraction include mutual information, chi-square test, information gain, etc. In this paper,

we choose the chi-square test to extract the features. Through judging whether the term t is independent of class c, we can judge whether the term t has a characteristic function of class c. Finally we select the appropriate threshold to make the selected feature set make the classification performance best. A two-way contingency table for word t and category c is constructed and the $\chi^2$ is calculated as:

$$\chi^2(t,c_i)=\frac{N\times(AD-BC)^2}{(A+C)\times(B+D)\times(A+B)\times(C+D)} \quad (1)$$

The A means the frequency of documents in which $t$ and $c$ co-occur, B and C the frequency when either $t$ or $c$ occurs, D the frequency when neither $c$ nor $t$ occurs, and N is the total number of documents. If $c$ and $t$ are independent of each other, $\chi^2(t,c_i)$ has a value of zero.

(4)Text classification algorithm of Weibo hot event

Naive Bayes classifier is a simple yet surprisingly accurate technique in spite of the wrong independence assumption. Naive Bayes estimates the probability that Weibo d is belong to category $c_j$ through computing the priori probability and conditional probability. The algorithm's steps are as follows:

Step 1: Calculate the priori probability $P(c_j)$ for each category according to the Weibo training set D and category set $C=\{positive,negative\}$ .

$$P(c_j)=\frac{Doc(c_j)}{\sum_{c_j\in C}Doc(c_j)} \quad (2)$$

$Doc(c_j)$ is the Weibo count of category $c_j$ .

Step 2: Calculate all the conditional probability $P(t_i\mid c_j)$ the feature is belong to one category.

$$P(t_i\mid c_j)=\frac{1+TF(t_i,c_j)}{|V|+\sum_{i=1}^{|V|}TF(t_i,c_j)} \quad (3)$$

|V| is the feature count of text, $TF(t_i,c_j)$ is the frequency of feature $t_i$ in category $c_j$ . The above formula is processed with data smoothing to avoid the $P(t_i\mid c_j)$ value of zero.

Step 3: Calculate the posterior probability that Weibo d is belong to each category $c_j$ , select the highest as the category of Weibo $d$ .

$$P(c_j\mid d)=\arg\max\{P(c_j)\prod_{i=1}^{n}P(t_i\mid c_j)^{wt(t_i)}\},c_j\in C \quad (4)$$

The $t_i$ is the feature $i$ of Weibo $d$ . $P(c_j\mid d)$ is the posterior probability that Weibo $d$ is belong to each category $c_j$ , $wt(t_i)$ is the weighting of word $t_i$ in Weibo $d$ and $wt(t_i)$=1 when selecting BOOL weighting.

(5)Evaluation Index

Evaluation index of sentiment classification refers to the results of manual classification. The commonly used index includes precision, recall and micro_average (F1)。

$$\text{Precision}=\frac{\sum_{c_j\in C}True(c_j)}{\sum_{c_j\in C}Doc(c_j)} \quad (5)$$

$$\text{Recall}=\frac{\sum_{c_j\in C}True(c_j)}{\sum_{c_j\in C}\text{Re}sponse(c_j)} \quad (6)$$

$$F_1=\frac{2\times recall\times precision}{recall+precision}\times 100\% \quad (7)$$

$True(c_j)$ is the count of Weibo which is classified as $c_j$ correctly. Response( $c_j$ ) is the count of Weibo which is classified as $c_j$ . *Precision* and *Recall* reflect two aspects of sentiment classification performance and vary in relative effect, so $F_1$ is a comprehensive index.

## IV. EXPERIMENT RESULTS

We used the directional network information acquisition system to collect a number of relevant Weibo text about the Weibo hot event and selected the "Malaysia Airlines plane lost contact" and "2014 APEC" as cases to study. We randomly selected 2000 subjective Weibo out of the 2 events, marked emotional category and clean the text data, 3/4 text for training set and the rest for test set. Then we designed 3 experiments to prove the feature threshold's influence on Naive Bayes classification performance, the classification performance contrast of the two methods and the classification performance contrast of different events. The computer for experiments was 8G, 64 bit operating system, and 1.8GHz.

### *A. Feature threshold's influence on Naive Bayes classification*

Weibo text "Malaysia Airlines aircraft lost contact" was selected as the research object. After Chinese word segmentation and word frequency statistics, transform the result into Document Word Frequency Matrix. Then use chi-square test to extract feature words. Reduce the dimensions of Document Word Frequency Matrix according to the feature words and regard it as the input of the Naive Bayes classifier. Lastly, train the classification model in the classifier, and evaluate the performance by the test set.

(1)Feature selection results

The following chart is partial results from the feature words extracted by the method of chi-square test. It can be seen that the feature words extracted in this event do have a function to distinguish between categories and reduce feature dimension.

TABLE I. The feature words extracted by the method of chi-square test

| positive word | chi_positive | negative word | chi_negative |
|---|---|---|---|
| hope | 31.887 | event | 15.331 |
| safe | 16.688 | government | 7.319 |
| pray | 13.33 | hope | 7.084 |
| airliner | 11.812 | crash | 6.841 |
| bless | 8.797 | safe | 5.635 |
| family | 4.393 | family | 4.215 |
| back | 2.83 | bless | 3.856 |
| lose | 2.707 | pray | 3.779 |
| home | 2.042 | airliner | 2.944 |
| information | 1.97 | abduction | 2.837 |
| announce | 1.896 | survey | 2.666 |
| crash | 1.763 | lose | 2.517 |
| satellite | 1.666 | official | 2.379 |
| reveal | 1.663 | kidnap | 2.173 |
| life | 1.486 | truth | 1.078 |

(2)Feature threshold's influence on performance of Naive Bayes classification

The classification performance is calculated by cross validation method and get the accurate classification performance by 5 cross validation in this paper. When the threshold is different, the performance of the classifier is as follows:

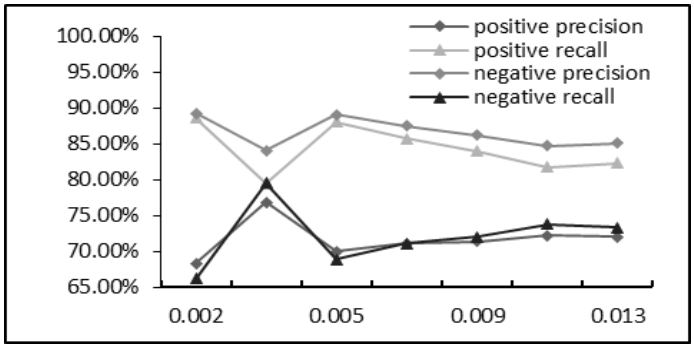


Fig.3. Feature threshold's influence on precision and recall.

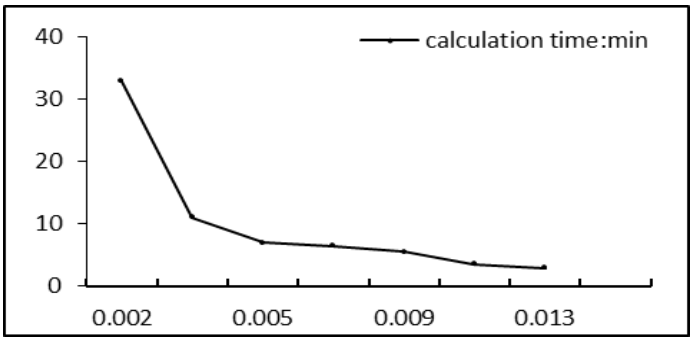


Fig.4. Feature threshold's influence on calculation time.

It can be seen that the negative text precision is higher with a value of about 80%, the negative text recall is between 70%-80%, the positive text precision is between 70%-80%, the positive text recall is above 80%, and the calculation time decreases with the increase of feature threshold. With a comprehensive consideration, the comprehensive classification performance is better when the feature threshold is 0.003. At this time, the positive and negative precision, recall and the calculation time reached a relatively good level.

### *B. Classification performance contrast of the two methods*

Select the same 500 Weibo text to data in last experiment and use the sentiment dictionary based classification method and the Naive Bayes based classification method to classify the text. Feature threshold was set as 0.003 in Naive Bayes based classification method and the feature words set includes 1139 words. The following table is the results of two methods.

TABLE II. CLASSIFICATION PERFORMANCE OF THE TWO METHODS

| Method | Positive precision | Negative precision | Positive recall |
|---|---|---|---|
| Sentiment dictionary based | 47.23% | 50.45% | 41.34% |
| Naive Bayes based | 76.80% | 79.43% | 78.09% |
| **Method** | **Negative recall** | **Positive F1** | **Negative F1** |
| Sentiment dictionary based | 37.10% | 44.09% | 42.76% |
| Naive Bayes based | 84.04% | 79.56% | 81.74% |

The results show that the performance of classification based on Naive Bayes is much higher than the classification based on sentiment dictionary, with the accuracy increased by 30%-47%, the recall increased by 29%-35% and the F1 value increased by 37%-39%. On the one hand, the imperfect emotional dictionary result to bad performance of the classification method based on emotional dictionary. On the other hand, Weibo text tends to be colloquial and the way people express their feelings are not simple to be understood by machine, such as irony, rhetorical question and so on. Naive Bayes is based on statistics. First, it extracted feature words with class differentiation according to the artificial tagging corpus, then judged the category of Weibo text according to the probability. So the performance of the classification method is better.

### C. Classification performance contrast of different events

We took two hot events "Malaysia Airlines plane lost contact" and "2014 APEC" as a contrast experiment. The feature threshold of "2014 APEC" was set 0.006 at which classification result was best. Naive Bayes classification method is used to classify the Weibo test set. The results are as follows.

TABLE III. CLASSIFICATION PERFORMANCE OF THE TWO METHODS

| Event | Positive precision | Negative precision | Positive recall |
|---|---|---|---|
| 2014 APEC | 68.84% | 64.41% | 73.25% |
| Malaysia Airlines plane lost contact | 76.80% | 79.43% | 78.09% |
| Event | Negative recall | Positive F1 | Negative F1 |
| 2014 APEC | 76.39% | 70.98% | 69.89% |
| Malaysia Airlines plane lost contact | 84.04% | 79.56% | 81.74% |

The results show that the performance of the same classification method is not exactly same in different events, but the overall performance is good. It proves that the classification method based on Naive Bayes in this paper is fairly applicable for sentiment classification of hot events on Weibo.

## V. CONCLUSIONS

This paper studied the sentiment classification about public comments of the social hot event on Weibo. First, we researched the sentiment dictionary based classification method and found the accuracy is too bad. In order to improve the accuracy, we also put forward a sentiment classification method based on Naive Bayes. We use TF as the feature weighting and chi-square test to extract features of each class. The accuracy rate is raised to more than 70%. The experiments proved that the performance of classification based on Naive Bayes is much higher than the classification based on sentiment dictionary and the performance of the same classification method is not exactly same in different events, but the overall performance is good, the accuracy rate is about 70%. Next, we will study to merge a variety of methods to improve the performance of sentiment classification.


## ACKNOWLEDGMENT

This work was supported by the Capital college students ideological and political education research subject under Grant No. BJSZ2016ZC041 and Engineering preparatory programme of Communication University of China under Grant No. 3132016XNG16.


## REFERENCES


[1] P. Waila, Marisha, V.K. Singh, M.K. Singh, "Evaluating Machine Learning and Unsupervised Semantic Orientation approaches for sentiment analysis of textual reviews", in *Proc.* Computational Intelligence & Computing Research (ICCIC), India, Dec. 2012, pp. 1 – 6.

[2] P.D. Turney, M.L. Littman, "Unsupervised learning of semantic oriental on from a hundred-billion-word corpus" Tech. Rep. ERB-1094, National Research Council Canada, Institute for Information Technology, 2002, pp. 359-364.

[3] V.K. Singh, R. Piryani, A. Uddin, P. Waila, "Sentiment analysis of Movie reviews and Blog posts", in *Proc.* Advance Computing Conference (IACC), Ghaziabad, India, Feb.2013, pp.893 - 898 .

[4] J. Lin, A. Yang, Y. Yong, "Classification of Microblog Sentiment Based on Naïve Bayesian", Computer Engineering and Science, 2012. vol.34, No.9, pp.160-165.

[5] J. Kamps, M. Marx, R. J. Mokken, "Using WordNet to measure semantic orientations of adjectives" in *Proc.* Language Resources and Evaluation. Lisbon, May. 2004, pp. 1115- 1118.

[6] Y. Zhu, J. Min, Y. Zhou, "Lexical semantic computation based on HowNet, Journal of Chinese Information Processing", 2006. 20(1):14-20.

[7] B. Yao, F. Li, P. Kumar, "K nearest neighbor queries and kNN-Joins in large relational databases (almost) for free", in *Proc.* Data Engineering (ICDE), Long Beach, USA, Mar. 2010, pp. 4 – 15.

[8] G. Valentini, T. G. Dietterich, "Bias-variance analysis of support vector machines for the development of SVM-based ensemble methods". Machine Learning Research, Jul.2004, vol.5, pp. 725-775.

[9] S. M. Kim, E. Hovy, "Extracting opinions, opinion holders, and topics expressed in online news media text" in *Proc.* the ACL Workshop on Sentiment and Subjectivity in Text, Morristown, 2006, pp.1-8.

[10] T. Mullen, N. Collier. "Sentiment analysis using support vector machines with diverse information sources" in *Proc.* the Conference on Empirical Methods in Natural Language Processing. Barcelona, Spain, Jul. 2004, pp. 412-418.


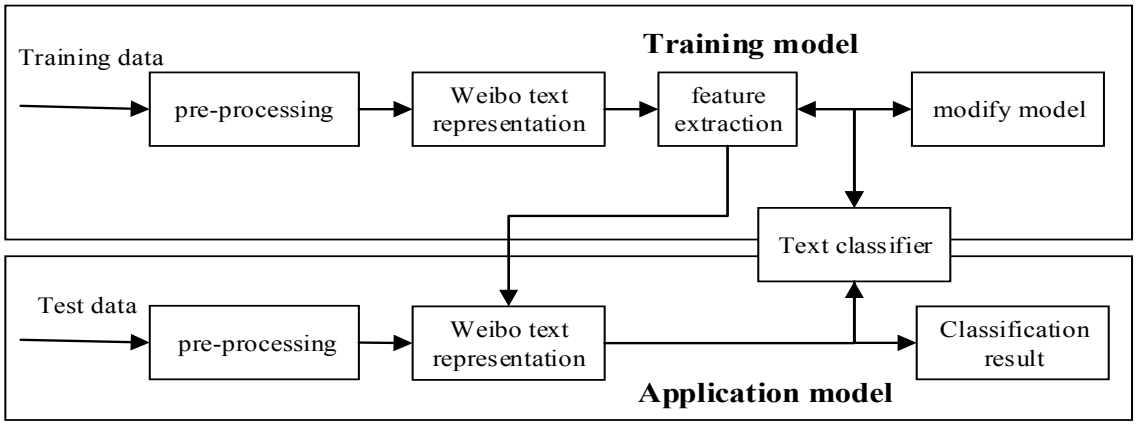


Fig.2. Naive Bayes based sentiment classification process for Weibo hot events.