Emotional Analysis of Weibo Based on Naive Bayes: A Case Study of the Speech of "Ticket-snatching Speedup Package"

Yue Zhang[a], Xiaohuan Gao[b]

Department of Information Management Nanjing University of Science and Technology Nanjing, China

[a]1789622508@qq.com, [b]1050703217@qq.com

Keywords: Naive Bayes, Weibo, Emotional Analysis

Abstract: In this paper, we will use the naive Bayesian classification method to analyze the sentiment of the "ticket- snatching speedup package" from Weibo. Using the train browser to crawl data on Weibo, after manual labeling, word segmentation and feature selection, randomly select 80% as a training set to construct a naive Bayes classifier. After testing the test set, the correct rate of the constructed naive Bayes classifier was 71.5%. It can help to automatically discover public sentiment and public opinion trends, which is of great significance to the development of major ticketing platforms.

1. Introduction

With the rapid development of the Internet, the Internet has become an indispensable part of people's lives. According to the 43rd Statistical Report on China's Internet Development Status released by China Internet Network Information Center (CNNIC), as of December 2018, the number of Internet users in China was 829 million, and the number of new Internet users was 56.53 million in the whole year. The Internet penetration rate reached 59.6% [1]. As an emerging information publishing and social platform, Weibo has been booming since its inception and has attracted a large number of users. People can get information, make friends, post messages, record life status, share moods, and express opinions through Weibo.

A huge amount of text data has been accumulated in Weibo [2]. These massive speech messages on the Weibo platform can be divided into two categories: one is the text that describes the facts, and the other is the text that expresses the opinions. The texts that express opinions are very important. It has great practical value for commenting on products and lyrics, providing users with feedback and absorbing opinions, and improving products to meet more users. Demand; can also be used to predict whether the society supports the government's policy or the hot point of current events [3].

This paper collects the speech of "ticket-snatching speedup package" on Weibo, carries out text analysis and mining, classifies the speech of "ticket-snatching speedup package" based on naive Bayesian classifier, understands the public's emotional tendency and viewpoint, and grasps the public's emotional attitude and public opinion tendency.

2. Related Works

2.1 Emotional Analysis Model of Speech on"ticket-snatching speedup package"

Based on the characteristics of the speech text of "ticket- snatching speedup package", this paper proposes the emotion analysis model of speech on "ticket-snatching speedup package" in Fig. 1, which demonstrates the research ideas and research methods of this paper. Firstly, we use the crawler software to crawl the speech data about the "ticket- snatching speedup package" on Weibo, label the captured data artificially one by one, and then perform word segmentation, word deactivation and feature selection operations. Finally, the pre- processed training samples are trained to classify the text with the trained classifier, and the emotional orientation of the commentary text is obtained.

  DOI: 10.25236/icemit.2019.015

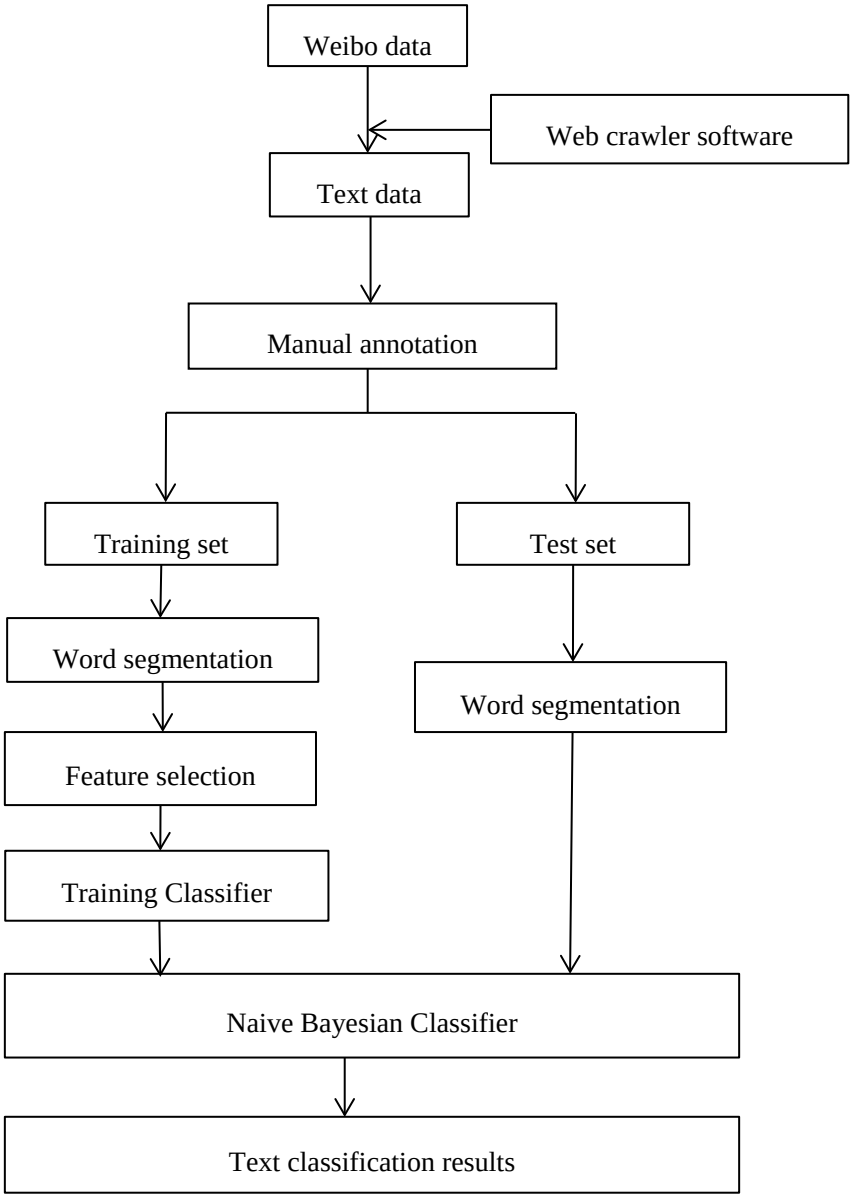

Figure 1. Emotional Analysis Model of Speech on “ticket-snatching speedup package”

2.2 Text preprocessing

2.2.1 Word segmentation

The template is used to format your paper and style the text. All margins, column widths, line spaces, and text fonts are prescribed; please do not alter them. You may note peculiarities. For example, the head margin in this template measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations.

This paper uses a software called Train Browser to write crawler script to crawl the speech text of“ticket-snatching speedup package” and construct the corpus of “ticket-snatching speedup package”. The text needs to be pre-processed before text sentiment classification. The steps are as follows:

a) Artificial annotation: The crawled corpus is labeled as three categories: positive emotion, negative emotion and neutral emotion. Among them, unrelated comments are labeled as neutral emotions.

b) Word segmentation: Chinese text only has the form of word, sentence, paragraph and so on. It cannot directly obtain the emotional words needed in the experiment. Before the experiment, word segmentation is needed. In this paper, the Jieba module in Python is used to realize Chinese word segmentation and keyword extraction. Jieba is a very concise and efficient word segmentation tool, which supports three word segmentation modes and a custom dictionary. For some proper nouns, it can be easily solved by using a custom dictionary.

c) De-stop words: remove symbols, auxiliary words, verbs and so on by importing stop words text.

The first 80% of word segmentation results were randomly selected as training set samples and the remaining 20% as test set samples. Save each data in the same row.

2.2.2 Feature selection

Common feature extraction methods include information gain method, mutual information method, CHI statistical method, etc. The CHI feature selection algorithm uses the basic idea of hypothesis test in statistics: assuming that the distribution of feature t and category ci conforms to the CHI distribution, the larger the CHI statistics, the stronger the correlation between features and categories, and the greater the contribution to categories.

The implementation steps are as follows:

- Statistics the total number of documents (N) in the sample set.
- The number of times that feature words t and ci appear together (A), the number of times that feature words to appear but ci does not appear (B), the number of times that category ci appears but it does not appear (C), and the number of times that feature words t and ci do not appear (D).
- Calculate the chi-square value of each word, and the formula (1) is as follows:

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

- The maximum value of feature t is taken as its global CHI statistic. Formula (2) is as follows:

$$\mathrm{CHI}_{\max}(t)=\max_{i=1}^{|c|}\{CHI(t,c_i)\} \tag{2}$$

3. Constructing classifiers

3.1 Summary of Naive Bayesian Algorithms

Naive Bayesian is a simple and effective classification method, which obeys the "Bayesian hypothesis", that is, the features of text are independent of each other [5]. The classification process is as follows:

- Sample data were collected and labeled manually. Let $X=\{x_1,x_2,...,x_n\}$ be a text data of "ticket-snatching speedup package"and x_i be a characteristic word of X ; category $Y=\{y_1,y_2,...,y_n\}$.
- Calculate the conditional probability of feature word x_i in category Y .That is $p(y_1|X), p(y_2|X),..., p(y_n|X)$.
- If $p(y_i|X)=\max\{p(y_1|X), p(y_2|X),..., p(y_n|X)\}$, where $i<n$, then the category of X belongs to y_i.

According to the Bayesian hypothesis $p(y_i|X)=\frac{p(X|y_i)p(y_i)}{p(X)}$, the probability is calculated by formula (3):

$$p(y_i|X)=\frac{p(X|y_i)p(y_i)}{p(X)} \tag{3}$$

• In order to maximize $p(y_i|X)$, we only need to maximize $p(X|y_i)p(y_i)$.In view of the independence of each feature word, it can be calculated according to the following formula (4):

$$p(X|y_i) = p(x_1|y_i)p(x_2|y_i)...p(x_n|y_i) \quad (4)$$

• Among them, $p(x_j|y_i)$is estimated by training samples, and the calculation formula is shown in (5):

$$p(x_j|y_i) = \frac{s_{ij}}{s_i}; \text{of which } j = 1,2,...,n \quad (5)$$

• The formulas for calculating the prior probability of category y_i are as follows (6):

$$p(y_i) = \frac{s_i}{s} \quad (6)$$

• In formula (5) and formula (6), s_{ij} is the sample number of characteristic word x_j and class y_i, s_i is the sample number of class y_i and s is the total sample number.

• After the above training stage, the text classifier is obtained. Firstly, the microblog text to be classified is represented by the eigenvector $X = \{x_1, x_2,..., x_n\}$, and then the main work is to calculate the values of $p(X|y_i)$ and $p(y_i)$ according to the above formulas. If and only if $p(y_i|X) = \max\{p(y_1|X), p(y_2|X),..., p(y_n|X)\}$ the text is divided into Class y_i [6].

In addition, considering the characteristics of sparse data, we adopt Laplacian smoothing for the features that do not appear, that is, assuming that all features appear at least once, to avoid the situation where the probability of the features that do not appear is zero and the whole conditional probability is zero. For example, suppose there are three categories in text categorization:, and. In a given training sample, the probability of observation counts of a word in each category is 0, 990, 10, the probability of is 0, 0.99, 0.01. For these three quantities, the Laplace smoothing method is used as follows: 1/1003 = 0.001, 991/1003 = 0.988, 11/1003 = 0.011 [7].

3.2 Design of Naive Bayesian Classifier Algorithms

The design idea of Naive Bayesian classifier algorithm is as follows:

Step 1: Data preparation. The results of feature selection are read into the list, each word is an element of the list; the contents of Naive Bayesian training set are read into the dictionary in the following formats: {‘1’:{‘word’: word frequency,‘ word’: word frequency, ‘word’: word frequency..., attribute:‘ ’},‘2’: {‘word’: word frequency,‘ word’: word frequency..., attribute:‘ ’}}.

Step 2: Feature representation. According to the results of the first step, each document of each class is represented by features and weighted by word frequency.

Step 3: Training naive Bayesian classifier. According to the principle of Naive Bayesian algorithm and Laplacian smoothing, the probability matrix is obtained.

Step 4: Calculate the accuracy of Naive Bayesian classifier. According to the principle of Naive Bayes and the result of the third step, the accuracy of Naive Bayes classifier is calculated. Equations

4. Analysis of Experimental Results

4.1 Experimental Data

This article uses the train browser to crawl 5823 comments on the "ticket grabbing acceleration package" on micro-blog. Through manual annotation, it is found that 492 of them contain positive emotions, 1554 contain negative emotions and 3777 contain neutral emotions. The author partitioned the tagged data and removed the stop words. Eighty percent of the segmentation results were

randomly selected as the training word set of Naive Bayesian classifier and the remaining 20 percent as the test set of Naive Bayesian classifier. The training vocabulary set is used to train the emotional classifier which belongs to the speech of "ticket grabbing acceleration package". The test set is used to test the accuracy of the emotional classifier.

4.2 Test results analysis of the algorithm

In the testing stage of Naive Bayesian classifier, two main tests are carried out: one is the accuracy of Naive Bayesian classifier; the other is to automatically judge whether the emotion of this sentence is positive, negative or neutral by randomly giving a sentence about “ticket grabbing acceleration package”. The experimental results show that the accuracy of Naive Bayesian classifier is 71.5%. The process of testing is to randomly give a sentence about "ticket-grabbing acceleration package", such as “I have no feeling for ticket-grabbing acceleration package”. The author judges that this sentence is a neutral emotional comment. The test result given by the naive Bayesian classifier constructed in this paper is “test comment belongs to neutral emotional comment”, which is the same as the author's judgment. It is concluded that the effect of this classifier is very good.

5. Summary And Prospect

In this paper, a method of emotional analysis is presented. Taking the speech data of “ticket-snatching speedup package” as the research object, an emotional classifier is constructed, which is exclusive to the speech of “ticket-snatching speedup package”. Experiments show that Naive Bayesian classifier can achieve better classification results. The disadvantage of this paper is that the collected data sets are not large and the amount of data varies greatly among different types, and the invalid comments are not removed, which affects the accuracy of the classifier. Therefore, the number of training sets can be continuously expanded in the later analysis to improve the accuracy of judgment. In addition, other machine learning algorithms can be used to classify it.

Acknowledgements

Research on Media Information Effect in Financial Markets: Saliency, Attention Distribution and Time Preference, SJCX19_0046.

References

[1] Xu Jianzhong, Zhu Jun, Zhao Rui, Zhang Liang, He Liang and Li Jiao Jiao. “Emotional analysis of space micro-blog based on SVM algorithm”. Information security research, 2017, vol.3,pp.1129-1133

[2] Liu Xule, He Yanxiang. “Emotional analysis of micro-blog based on multi-features”. Computer Engineering, 2017, vol. 12, pp. 160 - 164.

[3] Yang Chao, Feng Shi, Wang Daling, et al. “Network Public Opinion Tendency Analysis Based on Emotional Dictionary Expansion Technology”. Minicomputer System, 2016, vol. 04.

[4] Meena A, Prabhakar T V. Sentence Level Sentiment Analysis in the Presence of Conjuncts Using Linguistic Analysis. 2007.

[5] Zhao Gang, Xu Zan. “Emotional Analysis Model of Commodity Review Based on Machine Learning”. Information Security Research, 2017.vol.02.

[6] Jiao Feng. “Analysis of emotional inclination of hotel reviews based on Naive Bayes”. Modern Computer (Professional Edition), 2018.vol.20, pp. 45 - 49.

[7] Wang Fei, Liu Yunfei. “Emotional Classification Research Based on Commodity Evaluation of E-commerce Platform”. Information System Engineering, 2017 .vol.09, pp. 115 - 116.