Emotional Analysis of Public Opinions in Colleges and Universities: Based on Naive Bayesian Classification Method

Qingjia Wang¹, Kun Liu¹* and Kun Ma¹
¹ School of Information Science and Engineering, University of Jinan, Jinan, Shandong, 250022, PR China
* liukun@ujn.edu.cn

Abstract. In this paper, we use the emotion dictionary to process and express the public opinion texts of colleges and universities. We construct the public opinion texts of colleges and universities sentiment emotional classifier based on Naive Bayes theory, and then judge the emotional tendency of public opinion texts of colleges and universities. Experiments show that the method has the characteristics of fast classification and high accuracy, and is suitable for the public opinion system of colleges and universities. The method facilitates the university management's interpretation and mastery of the public opinions related to the school in the network, promotes and guides the positive energy in the communication network, corrects the vulgar and biased content, prevents the forwarding and dissemination of inappropriate speech, and establishes a positive campus environment among the teachers and students of the whole school.

1. Introduction
The Internet has become the main way of information dissemination in the 21st century, but at the same time it has two sides, it can spread positive energy and positive content, and also spread some negative and misleading speech. Colleges and universities are the initial place for higher education, and it plays a pivotal role. As college students are gradually becoming the main force and target of online public opinions, more and more colleges and universities have become the high-risk places for online public opinions emergencies. Therefore, the emotional analysis of college public opinions is of great significance. Through the emotional analysis of college public opinions, it guides and controls the public opinions of colleges and universities, so that school leaders and student workers can timely understand students' views and attitudes toward the event, keep an eye on the development of events and monitor the development trend of public opinions in colleges and universities, provide early warning of speeches that endanger school safety, guide social opinion orientation and provide guarantee for the healthy and stable development of society.

2. Research

2.1. Research on public opinions in colleges and universities
The psychological status of college students in a period of time and space is reflected by the enthusiasm of colleges and universities. At present, many researchers have made certain research results in colleges public opinions, which has led to the rapid development of college public opinion platforms. The amount of data generated on campus every day is in PB, which puts some pressure on the students' mental state and the organization of college administrators. So Wang J F used the idea of
big data to construct a college public opinion platform in 2017 [1]. Jumadi et al. conducted research on public opinions data mining and sentiment classification based on SVM [2]. Zhao Y Q proposed a complex proxy network model for college public opinion systems [3]. Through complex adaptive system theory and complex network analysis methods, the impact of future public opinions can be evolved in this model to help managers to better guide the direction of public opinions and the way students think about problems in colleges and universities, and provide powerful practical guidance for human progress. Many well-known scholars at home and abroad have carried out certain research on the public opinions of colleges and universities. Through the mining and preliminary analysis of the public opinions data of colleges and universities, many guiding analysis models for the network public opinions of colleges and universities have been put forward. This includes building a data analysis platform or using relevant theories to monitor public opinions in colleges and universities. These are just excavations and preliminary analysis of university public opinions data.

2.2. Research on text sentiment analysis

Text sentiment analysis is a key algorithm in the field of data mining and machine learning, which is one of the research hotspots in the field of natural language processing (NLP). In recent years, with the continuous deepening of people's emotional analysis of online texts, the research scope of text emotions is expanding and gradually occupying an important position in the academic world, and many domestic and foreign scholars have long studied this aspect. The text orientation analysis method based on machine learning produces text data that satisfies a certain format by labeling the tendency of some texts, then uses these marked text as training data, constructs an emotion classifier by using the classification method of machine learning and data mining. Finally, the method uses the classifier to classify the test data, and calculates the emotional tendency and emotional intensity of the test data. Pang et al. used Native Bayesian Algorithm, Support Vector Machines Algorithm, and Maximum Entropy Algorithm for experimental verifications and comparative experiments on text emotions as early as 2002. The end result is that the effects of the three algorithms are not much different [4]. Mullen et al. used the Support Vector Machines method to establish a text classification model and perform sentiment analysis on texts [5]. There are still many scholars who choose different features and machine learning algorithms to study, and the sources of the research texts are more diversified. Gammon et al. choose customer feedback texts as the research object [6], and Li et al. choose sentence context and emotion transfer words as the research object [7]. In addition, some scholars have proposed to measure the semantic tendency of emotional words in the text, obtain the average value of the semantic directional value of all words and finally obtain the tendency metric of the overall text.

3. Research method

The college sensation sentiment analysis method proposed in this paper is mainly to construct a Naive Bayesian sentiment emotional classifier. First, web crawlers are used to obtain public opinion data of universities. After obtaining the data, the text corpus preprocessing is performed, including word segmentation, deleting the stop words, reading the sentiment dictionary to obtain the emotional words in this article, and calculating the feature weights. After the text corpus preprocessing is completed, the text classifier is constructed by the Naive Bayesian algorithm. The basic idea of the construction is based on the Bayesian formula. First, the training corpus is trained to obtain the prior probability of the text category and the posterior conditional probability of the feature attribute. Then, the Bayesian formula is used to calculate the probability that the text to be classified belongs to all categories, and finally the text to be classified is divided into classes with higher probability values.

Naive Bayesian is a probability-based learning algorithm that is based on the prior probability of hypotheses. The probability of observing different features is given under the given assumptions. To the sentimental tendency of the public opinion text \( d = \{w_1, w_2, \ldots, w_n\} \) of the university to be classified belongs to \( C = \{C_{\text{Positive}}, C_{\text{Negative}}, C_{\text{Neutral}}\} \), considering the weight of the feature words, in the
case where the features are independent of each other, the classification algorithm formula is shown in formula 1.

\[
\text{category} = \arg \max_{c_i \in C} \left\{ \prod_{k=1}^{n} P(w_k, c_i)^{w_k(w_k)} \right\}
\]  

(1)

Where \( P(c_i) \) is the prior probability of the category \( c_i \), and the pre-estimate based on training expectations that have been manually labeled correctly, that is the ratio of positive text, negative text, and neutral text to the total text, respectively, is calculated as formula 2:

\[
P(c_i) = \frac{d(c_i)}{\sum_{c \in C} d(c)}
\]  

(2)

Where \( d(c_i) \) is the number of texts belonging to \( c_i \).

\( P(w_k, c_i) \) in the formula 1 is a posterior probability that the feature word \( w_k \) appears in the category \( c_i \), and \( w_k(w_k) \) is the weight of the word \( w_k \) in the text to be classified. This paper uses the sum of the weights of the feature word \( w_k \) in the text belonging to the category \( c_i \) divided by the weights of all the words of the category \( c_i \). If a feature word that does not appear in the training corpus appears in the text to be classified, \( P(w_k, c_i) \) will be 0. In order to avoid this situation, using Laplace transform, \( P(w_k, c_i) \) is calculated as shown in formula 3:

\[
P(w_k, c_i) = \frac{\text{weight}(w_k, c_i) + 1}{\sum_{k=1}^{n} \text{weight}(w_k, c_i) + V}
\]  

(3)

4. Experiment and analysis

4.1. Data set

The data set contains pieces of 1000 data, including positive and negative college public opinion data 500 each, each experiment selects 50 positive and negative data as test corpus, and the remaining 900 data as training corpus. Experiments were conducted by two feature selection methods, one is to use emotion dictionary to select emotional words, and the other is TF-IDF.

4.2. Standards for experimental evaluation

The experiment used the current widely used accuracy rate, recall rate and F1 measure to evaluate the experimental results. TP indicates the number of positive texts classified to positive categories, FN indicates the number of positive texts classified to negative categories, FP indicates the number of negative texts classified to positive categories, and TN indicates the number of negative texts classified to negative categories. The three indicators are calculated as follows:

\[
P_{pos} = TP/(TP+FP) \quad (4)
\]

\[
R_{pos} = TP/(TP+FN) \quad (5)
\]

\[
P_{neg} = TN/(FN+TN) \quad (6)
\]

\[
R_{neg} = TN/(FP+TN) \quad (7)
\]

\[
F_1 = 2PR/(P+R) \quad (8)
\]

\( P_{pos} \) and \( P_{neg} \) are the accuracy rates, and \( R_{pos} \) and \( R_{neg} \) are the recall rates. \( P_{pos} \) reflects the proportion of positive corpus classified by the classifier into positive corpus, and \( P_{neg} \) reflects the proportion of negative corpus classified by the classifier into negative corpus. \( R_{pos} \) reflects the proportion of positive corpus that is correctly classified to the total positive corpus, and \( R_{neg} \) reflects the proportion of negative corpus that is correctly classified to the total negative corpus.

4.3. Results and analysis
Each evaluation criterion was used to evaluate the classification effect. Tables 1 and 2 respectively list the classification accuracy rate, recall rate, F1 measure and their arithmetic mean results of the positive corpus and negative corpus in four experiments.

Table 1. Accuracy, recall and F1 measure of classification methods characterized by emotional words

| Data set | P_pos | R_pos | F1  | P_neg | R_neg | F1  |
|----------|-------|-------|-----|-------|-------|-----|
| C1       | 0.98  | 0.98  | 0.98| 0.98  | 0.98  | 0.98|
| C2       | 0.98  | 0.98  | 0.98| 0.98  | 0.98  | 0.98|
| C3       | 0.83  | 0.98  | 0.90| 0.98  | 0.80  | 0.88|
| C4       | 0.86  | 0.84  | 0.85| 0.84  | 0.86  | 0.85|
| Average  | 0.91  | 0.95  | 0.93| 0.95  | 0.91  | 0.92|

Table 2. Accuracy, recall and F1 measure of classification methods characterized by TF-IDF

| Data set | P_pos | R_pos | F1  | P_neg | R_neg | F1  |
|----------|-------|-------|-----|-------|-------|-----|
| C1       | 0.93  | 0.86  | 0.89| 0.94  | 0.94  | 0.90|
| C2       | 0.80  | 0.86  | 0.83| 0.78  | 0.78  | 0.81|
| C3       | 0.71  | 0.90  | 0.79| 0.64  | 0.64  | 0.73|
| C4       | 0.81  | 0.52  | 0.63| 0.88  | 0.88  | 0.75|
| Average  | 0.81  | 0.79  | 0.79| 0.81  | 0.81  | 0.80|

As can be seen from Tables 1 and 2, using emotional words as feature selection, the F1 measurement on all corpora is improved compared with TF-IDF. This shows that the Naive Bayesian classification method using emotional dictionary as feature selection can improve the emotional classification effect of college public opinion information.

5. Conclusion

Experiments show that the sentiment analysis of college public opinions can be effectively realized through the combination of sentiment dictionary and Naive Bayesian algorithm, but this paper does not distinguish application fields. In the field of colleges and universities, the characteristics of each emotional word are different from other fields, so it is necessary to establish a comprehensive emotional dictionary based on the university field. It is necessary to expand the number of training corpora and improve the quality of training corpus. Based on the sentiment dictionary, the Naive Bayesian probability model of the university field is constructed and continuously revised.

References

[1] Wang, J.F. (2017) Research on Campus Network Public Opinion Mining Based on Big Data. Value Engineering, 26:196-198.
[2] Jumadi, Maylawati, D.S., Subaekib, B., et al. (2016) Opinion mining on Twitter microblogging using Support Vector Machine: Public opinion about State Islamic University of Bandung. In: International Conference on Cyber and IT Service Management. Bandung.
[3] Zhao, Y.Q. (2016) Research on Evolving Model of University Public Opinion Based on Complex Agent Network. Information Science, 34:130-134.
[4] Bo, P., Lillian, L., and Shivakumar, V. (2002) Thumbs up? Sentiment Classification using Machine Learning Techniques. In: Empirical Methods in Natural Language Processing. Philadelphia, pp: 79-86.

[5] Mullen, T., Collier, N. (2004) Sentiment Analysis using Support Vector Machines with Diverse Information Sources. In: Empirical Methods in Natural Language Processing. Barcelona, pp: 412-418.

[6] Gamon, M. (2004) Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis. In: International Conference on Computational Linguistics. Switzerland. pp: 831.

[7] Li, S.S., Lee, S., Chen, Y., et al. (2010) Sentiment classification and polarity shifting. In: International Conference on Computational Linguistics. Beijing. pp: 635-643.