Public opinion analysis of university's "closed management policy" under the improvement of epidemic situation based on topic mining and emotional computing

Motong Tian[1]

1 School of information science and Engineering, Lanzhou University, Gansu, Lanzhou 730000, China

Abstract.*Under the background of the gradual improvement of epidemic control of COVID-19 in China, some universities still implement closed management policy. In view of the response and discussion of the policy by college students, a public opinion analysis method based on topic mining and emotional calculation is proposed. By using Python to capture 3424 answers of Zhihu's hot questions about University closed management, we use Chinese word segmentation processing, word frequency statistics and naive Bayesian sentiment value calculation to show college students' views and Thoughts on the policy, and then use LDA topic model, TFIDF calculation and topic word extraction to mine different aspects of public opinion. The results show that college students' emotion towards the implementation of closed management is negative, and the hot topics include faculty, students, formalism, graduation, internship, canteen, etc. This method can effectively mine the theme of public opinion events, and has certain use value.*

Keywords*: topic mining; emotional computing; closed management;*

1. Introduction

In September 2020, the epidemic situation of COVID-19 in China will obviously improve, tourist attractions will be reopened, and enterprises will resume production and work. However, many universities still implement the policy of closed management. The closed management policy of university makes college students isolated from the crowd, which has a great impact on College Students' life, study and psychology. It has aroused extensive attention and Discussion on the network based social platform. How to use computer methods and emotional mining methods to accurately identify college students' emotional state and tendency of this policy has become an important issue. This paper proposes a public opinion analysis method based on topic mining and sentiment computing.

In recent years, scholars at home and abroad have devoted themselves to the analysis of public opinion and put forward relevant methods. Public opinion of novel coronavirus pneumonia epidemic situation was analyzed by Yang Xiuzhang et[1] from visualization methods such as emotion analysis and knowledge map. Tang Xiaobo et[2] improved LDA model by taking microblog heat as the base of analysis and calculation, and thus obtained a distribution of microblog topic heat. GE Nilin[3] use the word vector model to represent the comment text, and PCA algorithm is used to reduce the dimension of the feature vector. Naive Bayes and support vector machine are selected for sentiment analysis.In the work , this paper analyzes public opinion from two aspects of topic mining and emotion calculation combined with data specificity, and improves the accuracy of emotional value by extracting abstract first and then calculating emotion by Bayesian formula.It provides strong support for the formulation and revision of relevant policies in Colleges and universities.

2. Research method

2.1 Overall flow of algorithm

This paper aims to analyze the public opinion situation of University closed management policy, and the overall flow of the algorithm is shown in Figure 1.

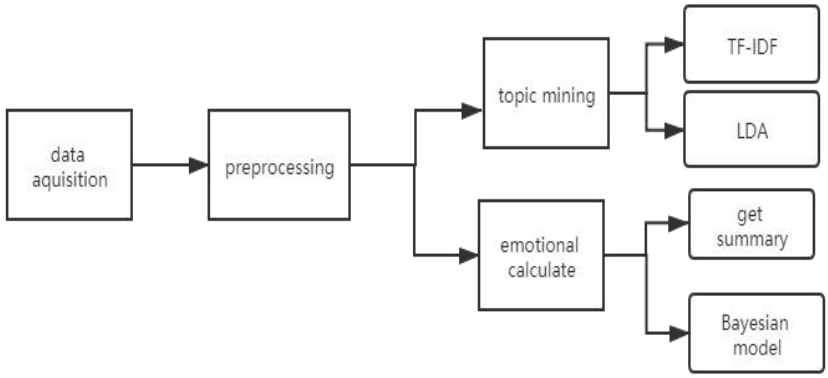

Figure 1. Overall flow of algorithm

(1) Through python, we can get more than 3400 answers to the question "will universities still have closed management after the opening of school in September?" Including the text of the answer, the ID of the respondent, the number of fans of the respondent etc.

(2) Preprocess the text, including clear pictures and word segmentation. Then the preprocessed data is stored in JSON and CSV files which are easy to process.

(3) Public opinion analysis includes two modules: topic mining and sentiment calculation. In topic mining, the topic words in all answers are calculated according to TFIDF, and the topic is clustered by LDA clustering algorithm. In the process of emotion calculation, first extract the summary of each answer to make the calculation more accurate, and then use Bayesian algorithm to calculate the

emotional value of each answer.

2.2 Data acquisition and preprocessing

This paper crawls 3424 answers to the popular question "universities will be closed management after the start of school in September", and then delete useless data such as pictures. According to the crawler algorithm, the earlier the answer is crawled, the higher the exposure. The exposure ranges from 1 to 3424. The statistics of high frequency words were also carried out for all the answers. Table 1 shows the high frequency word list.

Table1 .High frequency word list.

Word	Frequency
notice	641
faculty	623
School gate	588
freedom	577
everybody	543
canteen	532
leadership	496

3. Topic mining

3.1 Analyze the theme words with tf-idf

TF-IDF is a common weighting technology for information retrieval and exploration. TF-IDF is a statistical method to evaluate the importance of a word to a document set or one of the documents in a corpus. The importance of a word increases in proportion to the number of times it appears in the document, but at the same time it decreases inversely with the frequency of its appearance in the corpus.

TF is the abbreviation of term frequency, which refers to the word frequency in the text. There are many ways to measure the frequency of a word in a document. The simplest and most effective way is to directly calculate the frequency of occurrence of a word as its TF value.

IDF is the abbreviation of inverse document frequency, which means "reverse document frequency". It is a value used to measure the common degree of a word. The calculation of this value should not be based on a single document, but should consider all the documents to be analyzed to get the result.

TF-IDF = TF*IDF

According to the order of TF-IDF from large to small, the topic words in all the answers are selected. Table 2 shows the high TF-IDF keywords.

Table 2.key words with TF-IDF

Key word	TF-IDF
school	0.1505
student	0.1163
close	0.1086
open	0.0792
management	0.0505
faculty	0.0356
canteen	0.0235
totally closed	0.0227
take out	0.0223

3.2 Topic clustering using LDA algorithm

LDA (late Dirichlet allocation) is a three-layer Bayesian topic model proposed by BLEI in 2003. It can discover the hidden topic information in the text by unsupervised learning method. The purpose is to find the implicit semantic dimension from the text by the method of unsupervised learning, that is, "topic" or "concept". LDA schematic diagram is shown in Figure 2.

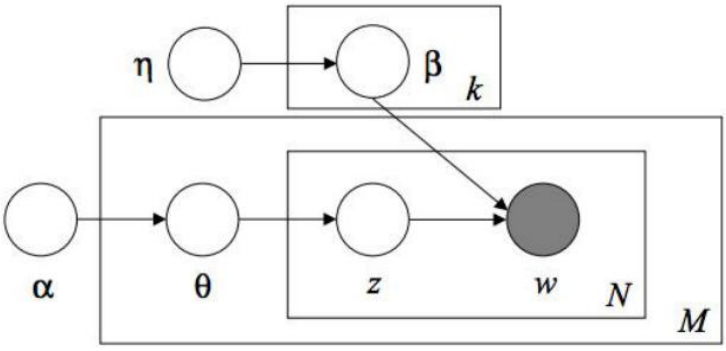

Figure 2.LDA schematic diagram

(1)There are documents, each document has words, involving a total of topics;

(2)Each document has its own topic. The topic distribution is a polynomial distribution. The parameters of the polynomial distribution obey the Dirichlet distribution, and the parameter of the Dirichlet distribution is α;

(3)Each topic has its own word distribution. The word distribution is a polynomial distribution. The parameters of the polynomial distribution obey the Dirichlet distribution, and the parameter of the Dirichlet distribution is η;

(4)For the nth word in a document D, a topic is first used from the topic distribution of the document, and then a word is used in the word distribution corresponding to the topic. This operation is repeated until all the M documents have completed the above process.

According to many experiments, it is the best to cluster topics into three categories by LDA Algorithm. Select the first 70 words under each topic, and the results are as follows.

(0,0.023 * "close" + 0.014 * "close" + 0.014 * "open school" + 0.014 * "start school" + 0.013 * "management" + 0.010 * "no" + 0.008 * "one" + 0.006 * "epidemic" + 0.006 * "college students" + 0.006 * "now" + 0.006 * "already" + 0.005 * "returning to school" + 0.005 * "notice" + 0.005 * "questions" + 0.005 * "faculty" + 0.004 * "teachers" + 0.004 * "dormitory" + 0.004 * "dormitory" + 0.004 * "possible" + 0.004 * "possible" + 0.004 * "possible" + 0.004 * "possible" + 0.004 * "possible" + 0.004 * "possible" + really "+ 0.004 *" University "+ 0.004 *" in and out "+ 0.004 *" out " +The University + 0.004 * "University" + 0.004 * "canteen" + 0.004 * "won't" + 0.004 * "won't" + 0.004 * "can't" + 0.004 * "school gate" + 0.004 * "know" + 0.003 * "know" + 0.003 * "practice" + 0.003 * "situation" + 0.003 * "semester" + 0.003 * "need" + 0.003 * "Hope" + 0.003 * "Hope" + 0.003 * "infection" + 0.003 * "many" + 0.003 * "one" + 0.003 * "fully enclosed" + 0.003 * "enclosed" + 0.003 * "enclosed" + 0.002 * "freedom" + 0.002 * "risk" + 0.002 * "risk" + 0.002 * "*" positive "+ 0.002 *" time "+0. 002 * "campus" + 0.002 * "this" + 0.002 * "this" + 0.002 * "leadership" + 0.002 * "formalism" + 0.002 * "always" + 0.002 * "always" + 0.002 * "feel" + 0.002 * "feel" + 0.002 * "complete" + 0.002 * "class" + 0.002 * "isolation" + 0.002 * "requirements" + 0.002 * "seal school" + 0.002 * "every day" + 0.002 * "don't let" + 0.002 * "don't" + 0.002 * "work" + 0.002 * "random" + 0.002 * "feel" + 0.002 * "feel" + 0.002 * "complete" + 0.002 * "class" + 0.002 * "actually" + 0.002* "only" + 0.002 * "in and out" + 0.002 * "should" + 0.002 * "life" + 0.002 * "living" + 0.002 * "coordinates" + 0.002 * "Mask" + 0.002 * "place" + 0.002 * "people" + 0.002 * "go home" + 0.002 * "together" + 0.002 * "application" + 0.002 * "don't want" + 0.001 * "go back" + 0.001 * "study" + 0.001 * "answer" + 0.001 * "home" + 0.001 * "home" + 0.001 * "policy" + 0.001 * "region" + 0.001 * "region" + 0.001 * "unseal" + 0.001 * "continue")

(1, '0.013 * "management" + 0.011 * "management" + 0.011 * "close" + 0.008 * "start school" + 0.007 * "can't" + 0.007 * "one" + 0.007 * "one" + 0.006 * "returning to school" + 0.006 * "going out" + 0.006 * "no" + 0.005 * "epidemic situation" + 0.005 * "now" + 0.005 * "now" + 0.005 * "access" + 0.005 * "notification" + 0.005 * "dormitory" + 0.005 * "college students" + 0.004 * "colleges" + 0.004 * "really" + 0.004 * "really" + 0.004 * "already" + 0.004 * "already" + 0.004 * "already" + 0. 004 * "in and out" + 0.004 * "classmate" + 0.004 * "freedom"+ Know + 0.003 * "know" + 0.003 * "feel" + 0.003 * "know" + 0.003 * "feel" + 0.003 * "school gate" + 0.003 * "faculty" + 0.003 * "teacher" + 0.003 * "University" + 0.003 * "possible" + 0.003 * "express" + 0.003 * "semester" + 0.003 * "Mask" + 0.003 * "coordinates" + 0.003 * "take away" + 0.003 * "leadership" + 0.003 * "should" + 0.003 * "need" + 0.003 * "need + 0.003 *" campus "+ 0.003 *" campus "+ 0.003 *" University "+ 0.003 *" University "+ 0.003 *" possible "+ 0.003 *" possible "+ 0.003 *" may. 002 * "canteen" + 0.002 * "problem"+ 002 * "a lot" + 0.002 * "a lot" + 0.002 * "Hope" + 0.002 * "Hope" + 0.002 * "counselors" + 0.002 * "time" + 0.002 * "seal school" + 0.002 * "conduct" + 0.002 * "fully closed" + 0.002 * "requirements" + 0.002 * "work" + 0.002 * "prevention and control" + 0.002 * "this" + 0.002 * "application" + 0.002 * "Hope" + 0.002 * "Hope" + 0.002 * "counselors" + 0.002 * "time" + 0.002 * "seal school" + 0.002 * "seal school" + 0.002 * "conduct" + 0.002 * "conduct" + 0.002 * "work" + 0.002 * "all closed" + 002 * "isolated" + 0.002 * "random"+ Staff + 0.002 * "staff" + 0.002 * "can only" + 0.002 * "only" + 0.002 * "safety" + 0.002 * "security" + 0.002 * "any way" + 0.002 * "implementation" + 0.002 * "don't let" + 0.002 * "don't want" + 0.002 * "prevent epidemic" + 0.002 * "learn" + 0.002 * "go home" + 0.002 * "one point" + 0.002 * "feeling" + 0.002 * "one time" + 0.002 * "don't" + 0.002 * "detection" + 0.001 * "limit" + 0.001 * "allow" + 0.001 * "permission" + 0.001 * "region")

(2, '0.013 * "close" + 0.010 * "management" + 0.010 * "management" + 0.008 * "start school" + 0.007 * "epidemic" + 0.006 * "no" + 0.005 * "can not" + 0.005 * "college students" + 0.005 * "true" + 0.005 * "in and out" + 0.004 * "freedom" + 0.004 * "problem" + 0.004 * "problem" + 0.004 * "guidance" + 0.004 * "teacher" + 0.004 * "one" + 0.004 * "now" + 0.004 * "now" + 0.003 * "know" + 0.003 * "know" + 0.003 * "go out" + 0.003 * "go out" + 0.003 * "go out" + 0.003* "closure" + 0.003 * "Mask" + 0.003 * "University"+ 003 * "staff" + 0.003 * "dormitory" + 0.003 * "staff" + 0.003 * "dormitory" + 0.003 * "already" + 0.003 * "infection" + 0.003 * "canteen" + 0.003 * "fully closed" + 0.003 * "need" + 0.003 * "classmates" + 0.002 * "school gate" + 0.002 * "possible" + 0.002 * "won't" + 0.002 * "situation" + 0.002 * "take out" + 0.002 * "staff" + 0.002 * "not let" + 0.002 * "coordinate" + 0.002 *

"coordinate" + 0.002 * "this" + 0.002 * "this" + 0.002 * "risk" + 0.002 * "risk" + 0.002 * "risk" + 0.002 *. 002 * express + 0.002 * hope+ 002 * "should" + 0.002 * "should" + 0.002 * "campus" + 0.002 * "return to school" + 0.002 * "life" + 0.002 * "University" + 0.002 * "epidemic prevention" + 0.002 * "a lot" + 0.002 * "feel" + 0.002 * "feel" + 0.002 * "region" + 0.002 * "access" + 0.002 * "health" + 0.002 * "current" + 0.002 * "prevention and control" + 0.002 * "isolation" + 0.002 * "a bit" + 0.002 * "semester" + 0.002 * "every day" + 0.002 * "outside" + 0.002 * "outside" + 0.002 * "outside" + 0.002 * "outside" + 0.002 * "outside" + 0.002"notice" + 0.002 * "get up"+ 001 * "virus" + 0.001 * "virus" + 0.001 * "going out" + 0.001 * "two" + 0.001 * "two" + 0.001 * "closed loop" + 0.001 * "formalism" + 0.001 * "school" + 0.001 * "new crown" + 0.001 * "practice" + 0.001 * "every day" + 0.001 * "place" + 0.001 * "learning" + 0.001 * "don't" + 0.001 * "don't" + 0.001 * "request" + 0.001 * "direct" + 0.001 * "strict" + 0.001 * "counselors" + 0.001 * "can't" + 0.001 * "can")

On this basis, public opinion is clustered into three categories.

4. Emotional computing

4.1 Get the summary

It is not good to use text directly for emotion calculation because there are too many interfering sentences. Therefore, I first extract the summary of each answer, and then calculate the emotion of the summary.

Calculate TF-IDF matrix

The weighted score of sentences is calculated as shown in Figure 3

The sentences of each answer are ranked from high to low according to the sentence weighted score, and the summary of each answer is extracted.

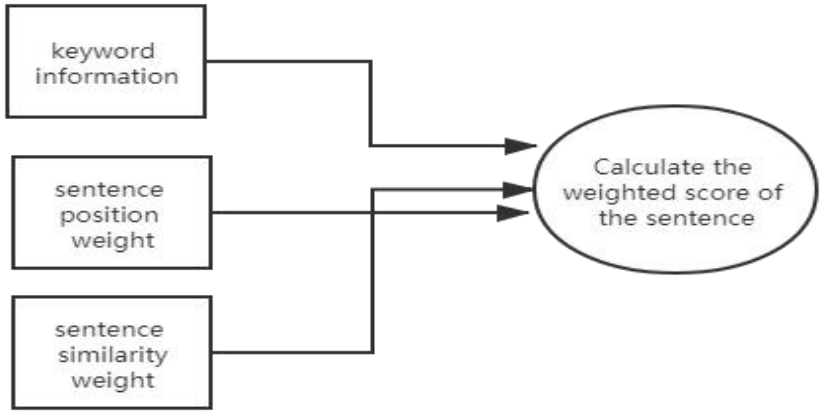

Figure3. Algorithm flow chart of calculating sentence weighted score

According to this algorithm, I can live the summary of each answer, and use these summaries to calculate the emotional value.

4.2 Emotion calculation with Bayesian algorithm

LIANG Ke[4]used to compare the accuracy of bag of words model and naive Bayesian model. SU Ying[5]combined naive Bayes with potential Dirichlet distribution.

I use Bayesian algorithm to calculate the emotion of each summary.

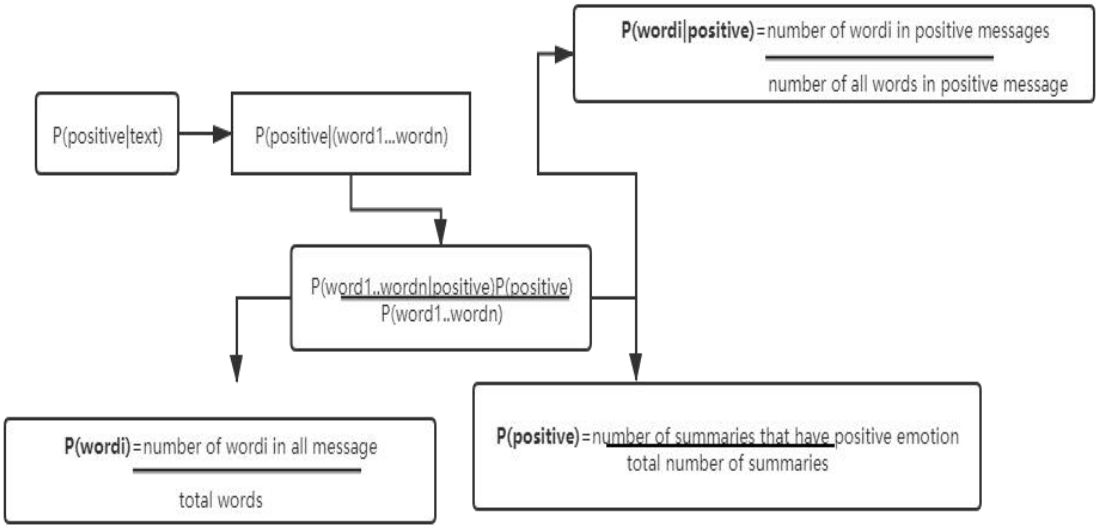

Figure 4.Bayesian algorithm

According to this principle, snownlp(a package in python) calculates that the closer the value is to 0, the more negative the emotional tendency is; the closer the value is to 1, the more positive the emotional tendency is.

On this basis, the emotional value of each answer was calculated.

When crawling data, I crawl from the first answer people see to the last answer they see, that is, crawling according to the order of exposure. According to this, I identifier each answer. The higher the exposure of the first crawled answer is, the greater the number will be (that is, the number of the first crawled answer's identifier is equal to the number of all the answers, and the identifier of the last crawled answer is 1). I made a scatter plot to record the emotional value and exposure of the responses, and selected the first 1000 responses with the lowest emotional value. The y-axis represents the exposure, and the x-axis represents the emotional value.

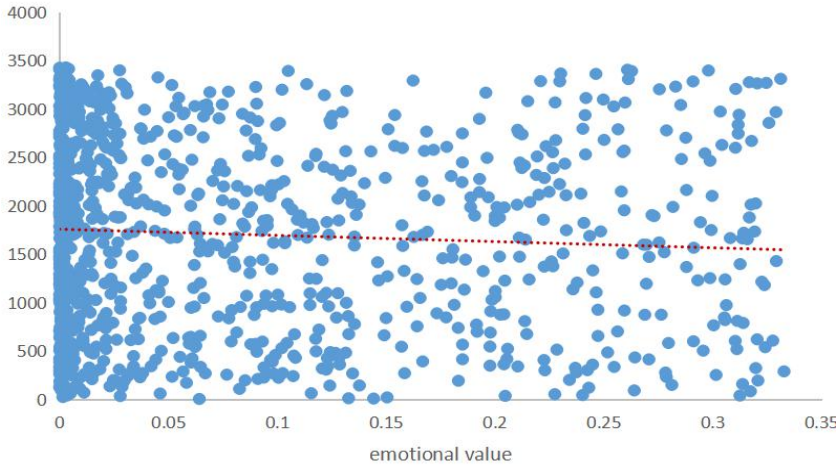

Figure 5. Visualization of exposure and emotional value

We can see that no matter in any exposure, there are more points focused on the right side(low emotional value).

5. Conclusion

Aiming at the problem that it is difficult to analyze the topic and emotion of public opinion topic in closed university management, this paper proposes a method combining topic mining and emotion calculation. Firstly, we use Python to capture 3424 answers to Zhihu hot issues, and then use data preprocessing, Bayesian classifier, TFIDF topic extraction, LDA topic model and other means to analyze public opinion. The experimental results show that the proposed method can effectively identify the keywords in the answer to hot questions and cluster public opinion effectively, and can also calculate the emotional value to analyze the public opinion. Finally, the word cloud and scatter diagram are used to visualize the direction of public opinion. The experimental results show that the public opinion emotion of University closed management policy is negative. It involves such hot topics as closure, management, faculty, internship, graduation and formalism. The method of this paper can effectively mine the theme of public opinion events, and provide ideas for the formulation and adjustment of relevant policies.

References

[1]Yang Xiuzhang, Wu Shuai, Xia Huan, Yu Xiaomin (2020) Public opinion analysis of COVID-19 epidemic using topic mining and emotion analysis. Computer Era, pp.31-36.
[2] Tang Xiaobo, Xiang Kun (2014)Hotspot Mining Based on LDA Model and Microblog Heat. Library and Information Service, pp.58-63.
[3]GE Nilin,FAN Jiajia(2020)Sentiment Analysis of Reviews Based on Naive Bayes and Support Vector Machine.Computer & Digital Engineering,pp1700-1704.
[4]LIANG Ke， LI Jian， CHEN Yingxue， LIU Zhigang(2019)Text emotional classification and realization based on Naive Bayes.Intelligent Computer and Applications,pp150-157.
[5]SU Ying，ZHANG Yong，HU Po，TU Xinhui(2016)Sentiment analysis research based on combination of naive Bayes and latent Dirichlet allocation. Journal of Computer Applications.pp1613-1618.