ISSN: 2688-8645

Research on Online Data-driven Public Opinion Analysis Method of College Students

Shenning Song, Haiyang Wang, Haozhe Jiang, Taiyuan Wang

School of Jilin University Changchun, 130000, China

Abstract

The rapid development of the mobile Internet has promoted the explosive growth of the number of online public opinions of the student group. Real-time screening of online public opinion text content is of great significance for rationally guiding students' thinking and maintaining campus safety. This paper proposes a university network public opinion monitoring model based on LDA-LSTM, and designs and implements a network public opinion monitoring system for college students in a specific application environment. The results show that this model can efficiently obtain topic distribution and emotional polarity, and provide an effective method for university network public opinion management.

Keywords

University network public opinion; Topic model public opinion monitoring; Implicit Dirichlet distribution; Long and short-term memory network.

1. Research Background

With the rapid development of smart phones and social software, the number of netizens in our country is increasing. According to the National Bureau of Statistics, as of December 2020, the number of Internet users in our country was 989 million, and the number of mobile Internet users reached 986 million. Student netizens accounted for the largest proportion, reaching 21.0% [1]. More and more college students choose to express their positions, attitudes and opinions on Internet social platforms.

The online public opinion of universities refers to the remarks made by college students when they put forward their opinions and express their attitudes on the emergent events and hot social events around them on the online platform. Public opinion in universities has the characteristics of real-time, suddenness, diversity, and concealment, which reflect the state of mind and specific concerns of college students. College students have the characteristics of active thinking and distinctive personality, but at the same time they lack social experience, are easily bewitched and impulsive. Social emergencies, domestic and foreign focal issues, or topics closely related to the college students themselves are very easy to attract the attention of college students [2]. The resulting negative events happen from time to time. Therefore, the use of large-scale public opinion text analysis and monitoring of colleges and universities network public opinion will be of great significance.

2. Related Research

Internet public opinion monitoring refers to the entire process of using a large amount of text corpus information generated by Internet users, using computer technology to collect and process, and form visual charts, reports and other analysis results. Natural Language Processing [3] technology is the core technology, which includes two major tasks: topic extraction and sentiment analysis .

ISSN: 2688-8645

Chinese text topic mining is mainly unsupervised learning. The Chinese corpus is segmented into words and converted into matrix vectors for quantitative calculation and analysis. There are mainly the following three methods in text representation. The first is text representation based on space vector model, the second is text representation based on topic model, and the third is Representation based on neural network model.

Sentiment analysis mainly includes three tasks: sentiment information extraction, sentiment information classification, and sentiment analysis evaluation and resource construction. The core task divides text emotions into three categories [4]: negative, positive, and neutral. Natural language sentiment analysis, also known as text orientation analysis, originated from human analysis of words with emotional color. In 2002, Pang et al. used naive Bayes algorithm, support vector machine algorithm, and maximum entropy algorithm for text orientation Perform analysis [5]. Jumadi et al. used the support vector machine algorithm to conduct data mining based on a large number of Twitter opinions, and conducted text sentiment analysis. The final accuracy rate is as high as 78.75%.

At present, most public opinion analysis systems at home and abroad are oriented to the government. Among them, the public opinion analysis system based on the dynamic mechanism model researched and developed by Lemos Blois is the best public opinion analysis system [6]. There is currently no mature public opinion analysis system for colleges and universities in China

3. Related Technology

3.1. Distributed Crawler

A web crawler is a program that automatically extracts web content and can automatically download resources on the web. Quickly download the required text corpus by simulating clicking. In this research, distributed crawlers are used to deploy crawler programs on several different servers and crawl information concurrently, which greatly improves the efficiency of program operation.

3.2. Doc2vec Word Vector Model

In the Doc2vec model [7], a certain column of matrix D represents the unique vector of each sentence of text. A certain column of the matrix W represents the unique vector of each word, sliding sampling in each sentence of the text, each time the same fixed length words are selected, one of the words is used as the prediction, and the others are used as the input words. The word vector corresponding to the input word and the sentence vector corresponding to the sentence text are used as the input of the input layer. The sentence vector of the Paragraph vector can be regarded as another word vector. It has been continuously trained in many times.

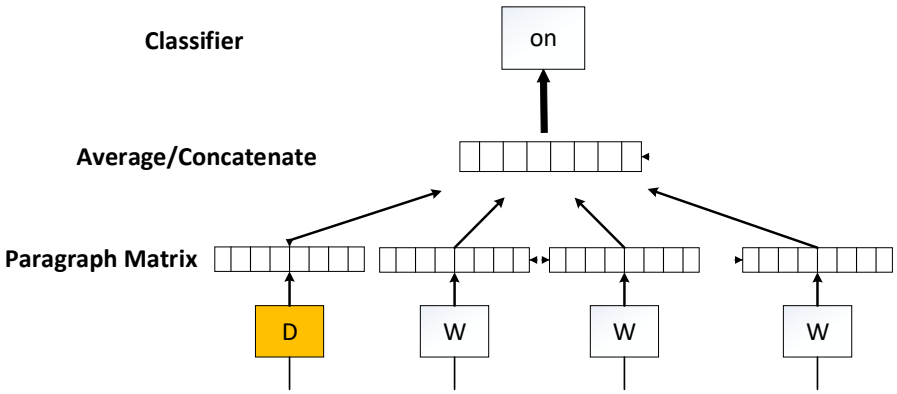

Figure 1. Doc2vec Framework

ISSN: 2688-8645

3.3. LDA Topic Model Based on Part-of-Speech Filtering

The Latent Dirichlet Allocation (LDA) model [8] is an unsupervised probabilistic topic model proposed by David Blei, Andrew Ng and Michael I. Jordan in 2003. It is a typical bag of words (Bag Of Words, BOW) model, which is composed of three levels of document, topic and word, and is also called a three-level Bayesian probability model. LDA describes the relationship between topics and words in a global way. It expresses text as a vector in an invisible semantic space. The model divides different documents into probability distributions of different topics, and different topics are composed of probability distributions of different words. The potential association between each word in the document and its corresponding topic is established. The probability distribution between the document and the topic and the topic and the word conforms to the polynomial distribution. The correlation can be judged by the value of the probability distribution.

The LDA model has two main processes, as shown in Figure 2. First, for the m -th document in the corpus, sample the Dirichlet distribution $\vec{\alpha}$ to generate the topic distribution $\overrightarrow{\theta_m}$ of the m -th document, and generate the topic $Z_{m,n}$ of the n -th word of the m -th document from the obtained topic distribution $\overrightarrow{\theta_m}$. Second, sample from the Dirichlet distribution $\vec{\beta}$ to generate the word distribution $\overrightarrow{\varphi_k}$ corresponding to the topic $Z_{m,n}$ of the n th word in the m document, and sample from the polynomial distribution $\overrightarrow{\varphi_{z_{m,n}}}$ of words to generate the word $W_{i,j}$. Repeat the above two steps to get the words, until all the words in the document are generated. In the process of generating words, the topic distribution and the distribution of topic words are constantly modified.

Traditional LDA is oriented to words in all corpora, and words with different parts of speech may cause the readability of the topics extracted by LDA to be quite different. Therefore, an LDA topic model based on part-of-speech filtering is proposed to mark the words after segmentation in the corpus, and shield the poorly readable words such as adjective words and adverbs, and retain nouns that are good for topic extraction.

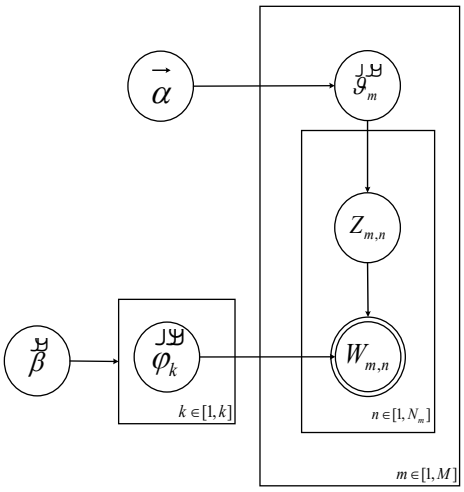

Figure 2. LDA Model

3.4. Sentiment Analysis Model Based on LSTM Neural Network

LSTM (Long Short-Term Memory) stands for Long Short-Term Memory in Chinese. It is an event neural recurrent network with long-term memory capabilities. The network structure of LSTM contains one or more units that can forget and remember. The emergence of LSTM overcomes the problem of weight loss in the back propagation of traditional recurrent neural networks (RNN) over time. It replaces the hidden nodes of RNN with memory cells, and protects or

ISSN: 2688-8645

controls the neural network through the gate structure The node state of the hidden layer forms a closed loop between the hidden layers. At the same time, the weight of the hidden layer is responsible for controlling the scheduling memory, and the state of the hidden layer is used as the memory state at the moment to participate in the next prediction.

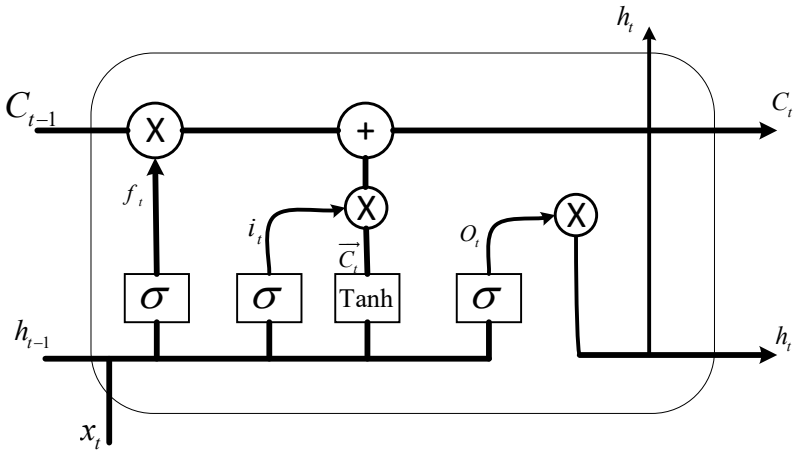

Figure 3. LSTM Model

The cell of LSTM includes four components: forget gate, input gate, memory gate and output gate. These components can significantly improve the ability of recurrent neural networks to process long sequence data.

(1) Forgetting gate, which determines what information to discard from the cell. The gate reads the previous hidden state and the current input word, and outputs a set of weights between [0,1] to the previous cell state, which means to retain or discard , As shown in formula (1).

$$f_t = \sigma(W_f \bullet [h_{t-1}, x_t] + b_f) \tag{1}$$

(2) The input gate determines how much new information is added to the cell. First, it is generated by the Sigmoid layer, which calculates the relatively more important words, and at the same time generates alternative update information through the layer. Then, the Sigmoid layer and the Tanh layer are combined. The status of the cell is updated. Finally, add the output of the forgetting gate to it to obtain a new candidate value. As shown in formulas (2) ~ (4).

$$i_t = \sigma(W_i \bullet [h_{t-1}, x_t] + b_i) \tag{2}$$

$$C_t = \tanh(W_c \bullet [h_{t-1}, x_t] + b_c) \tag{3}$$

$$C_t = f_t * c_{t-1} + i_t * C_t \tag{4}$$

(3) Calculate the memory gate and select the important information that needs to be memorized. The state of the hidden layer at the previous moment and the input word at the current moment are used as input. Output memory gate value and temporary cell state. As shown in formulas (5) ~ (6).

$$i_t = \sigma(W_i \bullet [h_{t-1}, x_t] + b_i) \tag{5}$$

$$C_t = \tanh(W_c \bullet [h_{t-1}, x_t] + b_c) \tag{6}$$

ISSN: 2688-8645

(4) The output gate determines the value of the cell output. First, the Sigmoid layer is used to determine which part of the cell state is output. Then, pass the cell state to Tanh for processing and multiply it with the output of the Sigmoid layer, and finally get the next hidden state. As shown in formula (7) ~ formula (8).

$$O_t = (W_o \bullet [h_{t-1}, x_t] + b_o) \tag{7}$$

$$h_t = O_t * \tanh(C_t) \tag{8}$$

It can be seen that the LSTM model overcomes the short-text data sparseness of the general model, and can memorize context information, learn text features better, and can handle multiple short text corpus well.

Due to its own characteristics, the LSTM model is particularly suitable for modeling time series data. It can capture longer-distance dependencies and unite contextual words, but it is one-way when faced with more fine-grained classification tasks and shorter corpora. The LSTM cannot encode back-to-front information. The forward LSTM and the backward LSTM are combined to form a two-way long and short-term memory cyclic neural network[9](Bi-LSTM), which can better capture the two-way semantic dependence.

4. Empirical Analysis

4.1. Data Collection and Preprocessing

QQ confession wall is one of the important network platforms for college students to conduct social activities, and it is also the main tool for college students to obtain information, express opinions, and speak out in response to various events. The content published by Jilin University's QQ confession wall (QQ number: 3061374768) was selected as the data source. A total of 1,420 corpora from September 2020 to January 2021 are obtained through distributed crawlers.

Duplicate data lacks practical value in analysis, so it is necessary to remove this part of data to improve data quality. Use mechanical compression to process text data. Then use the combination of word tagging and part-of-speech tagging for word segmentation. Since special vocabularies such as school buildings and common names are not included in the original dictionary, the introduction of college-related vocabulary into the word segmentation database has effectively improved the word segmentation effect. In addition, the stop word list of Harbin Institute of Technology was cited to build a dictionary, and the stop words in the text were deleted.

4.2. Parameter Estimation and Topic Extraction

After data preprocessing, the text data set has met the topic model specifications, and then read the data part of speech, filter the non-noun part of speech data and import the LDA topic model to obtain the keywords under each topic. Many parameters in the LDA model need to be set by themselves. Among them, the number of topics (Topic) is the most important. There are many calculation methods for the number of topics. According to the methods used in most references, the number of topics corresponding to the smallest degree of confusion is the best Number of topics. The formula for calculating perplexity is as follows.

$$\text{perplexity(D)}=e^{\frac{-\sum \log(p_w)}{\sum_{d=1}^{M} N_d}}$$

ISSN: 2688-8645

After determining that the number of topics required for the text corpus is 8, other parameters need to be set. First, set the minimum threshold of topic probability to 0.01 to exclude unpopular topics; the hyperparameters and are set to automatically learn prior knowledge from the corpus, and the number of iterations is set to 600, which helps the model to infer the topic distribution results to stabilize. The final model sorted out 6 topics with clear content after training.

4.3. Analysis of Emotional Tendency

In terms of model training data, this article uses open source Chinese sentiment annotation corpus, including 3000 positive, neutral, and negative corpus, divided into training set and test set at a 4:1 ratio. After hyperparameter adjustment and multiple rounds of training, the accuracy, recall, and F1 factor of the Bi-LSTM model all reached more than 90%, indicating that it can accurately predict the emotional tendency of the text. The results are shown in Table 1.

Table 1. Bi-LSTM Training result

	Precision	Recall	F1-score	support
Positive	0.92	0.92	0.92	600
Neutral	0.99	0.97	0.98	600
Negative	0.92	0.93	0.92	600
Accuracy	-	-	0.94	1800
Macro avg	0.94	0.94	0.94	1800
Weighted avg	0.94	0.94	0.94	1800

Use the trained model to predict the actual corpus in March 2021. The distribution of positive, neutral, and negative texts on the six topics is shown in Figure 4.

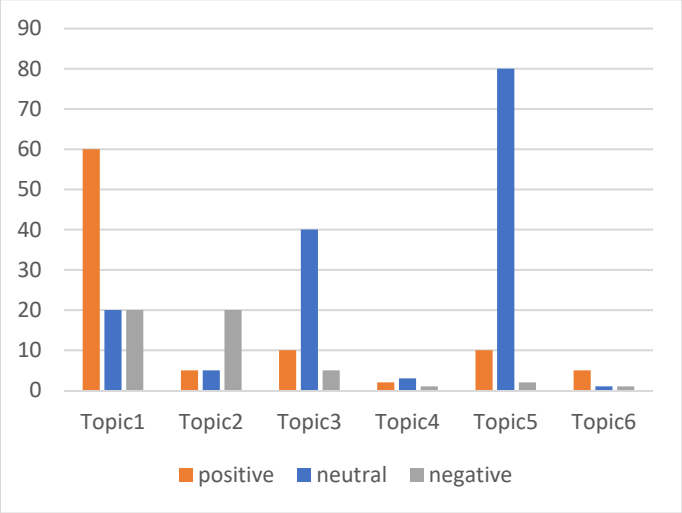

Figure 4. Emotional polarity distribution

4.4. Implementation of the Visualization System

For college management, we have built a visual public opinion platform for colleges and universities. The main roles of the system are divided into: system staff, university administrators and ordinary users. System staff have the highest operating authority and can use all the functions of the system to facilitate the management and maintenance of the system; college administrators also have the highest operating authority, can monitor real-time public opinion information and receive early warnings when public opinion crises, and at the same

ISSN: 2688-8645

time Can control the authority and identity information of ordinary users; ordinary users can only use the system to query public opinion information by keywords. User authority management makes information asymmetry between ordinary users and university administrators, which is convenient for administrators to control real-time emotional orientation and public opinion orientation. At the same time, all accounts of personnel using the system are uniformly assigned by the school, avoiding the leakage of public opinion information. The platform page is shown in Figure 5.

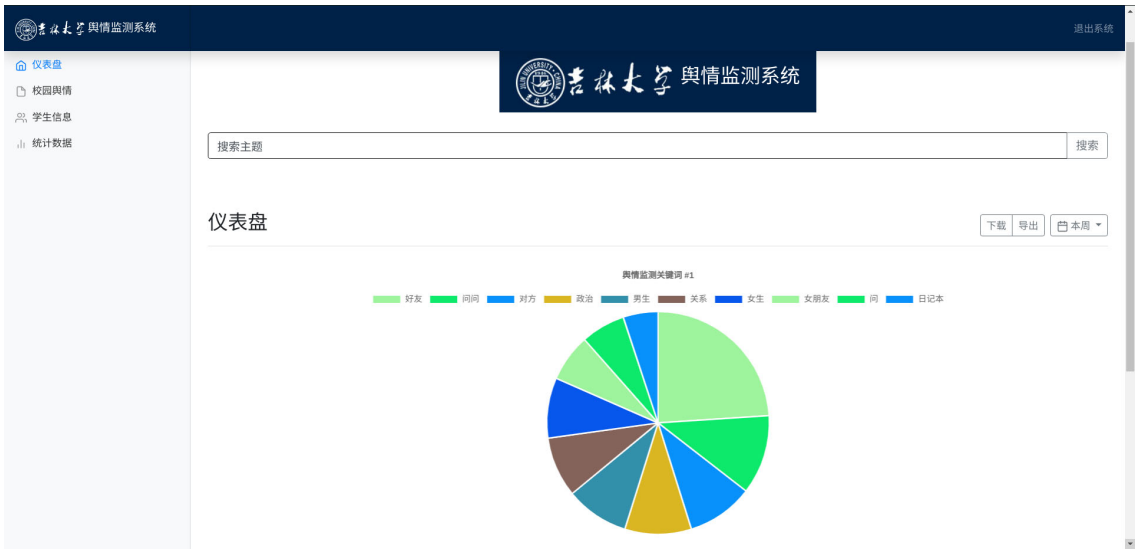

Figure 5. Visualization platform homepage

5. Concluding Remarks

This research proposes a method for monitoring public opinion in colleges and universities based on the combination of topic mining and sentiment analysis[10[, using Jilin University's QQ confession wall as the data source, combining the LDA model based on part-of-speech filtering and the BiLSTM model to mine hot topics in text data , Filter text data related to popular topics for sentiment analysis, and analyze the opinions and opinions of college students based on sentiment polarity.

The research results show that this method can dig out the hot topic information in the text, obtain the emotional distribution of each topic, understand the ideas and attitudes of college students, and help colleges and universities to grasp the public opinion, implement corresponding strategies in time, and avoid the development of public opinion. The stability of the network environment provides theoretical and methodological support.

References

[1] China Internet Network Information Center. The 47th "Statistical Report on Internet Development in China." [R].2021.

[2] Harbin Municipal Party School of the Communist Party of China Wang Haitao. "Research on the Countermeasures for the Guidance of Internet Public Opinion of College Students.".

[3] Chengqing Zong. Statistical natural language processing [M]. Tsinghua University Press.

[4] Information Retrieval Research Center, School of Computer Science and Technology, Harbin Institute of Technology Zhao Yanyan, Bing Qin, Tingliu. "Text sentiment analysis."

[5] Ensemble of keyword extraction methods and classifiers in text classification Onan;Serdar Korukoglu;Hasan Bulut Expert Systems With Application.

ISSN: 2688-8645

[6] Zhang W,Ming Z,Zhang Y,et al.The use of Dependency Relation Graph to analysis the Public opinion Systems.[C]//COLING,2012:3105-3120.

[7] Le Q V,Mikolov T.Distributed Representations of Sentences and Documents[C]//Proceedings of the 31st International Conference on Machine Learning 2014.

[8] Blei M, Ng A Y, Jordan M I. Latent Dirichlet allocation[J].Journal of Machine Learing Research, 2003, 3(4/5):993-1022.

[9] Deng Nan, Bengong Yu. Analysis of sentiment tendency of review text based on sentiment word vector and BiLSTM [J]. Application Research of Computers, 2018.35 (12): 3547-3550.

[10] Wang zhibo,Ma Long,Zhang Yangqing.A hybird document feature extraction method using latent dirichelt allocation and Word2Vec[C]//2016 IEEE First International Conference on Data Seience in Syberspace(DSC), 2016.