Research on Theme Identification of Public Opinion in Colleges and Universities

Xueyan Liu[1,a], Xuefeng Long[2,b](Corresponding Author), Hejie Chen[3,c](Corresponding Author)

liuxy@cqupt.edu.cn[a], 745640353@qq.com[b], 2585593533@qq.com[c]

School of modern post, Chongqing University of Posts and Telecommunications, Chongqing, China[1]
School of Economics and Management, Chongqing University of Posts and Telecommunications, Chongqing, China[2]
School of economics and management, Beijing Institute of Graphic Communication, Beijing, China[3]

Abstract. The 51st Statistical Report on Internet Development in China shows that the number of Internet users in China is 1.067 billion, and the Internet penetration rate reaches 75.6%, among which the percentage of Chinese Internet users among teenagers will reach 26.6%, and college students are an important group of teenage Internet users, whose knowledge, feelings, intentions and behaviors are all influenced by the Internet. Internet public opinion in colleges and universities is the focus and source of social public opinion, and is more likely to trigger discussions among netizens to form hotspots of public opinion. In this paper, based on the text data of Baidu posting in Chongqing universities, we extracted the topic words of public opinion through the LDA topic model, and constructed the topic co-occurrence network in each stage by using the Gephi social network analysis tool through the topic word co-occurrence method. The results of the study show the themes of university online public opinion at different stages and the heat changes of different themes at different time stages.

Keywords: Public Opinion in Higher Education, LDA Thematic Modeling, Theme Co-occurrence

1 Introduction

In the era of increasing level of network technology, communication and dissemination modes are becoming richer and richer, all kinds of new software are appearing one after another, and the speed of dissemination and the scope of influence of public opinion are showing unprecedented development with the rapid growth of new media. In recent years, there have been frequent emergencies of university network public opinion, such as school fire and other safety accidents; accidents when organizing outing activities; group food poisoning caused by school canteen or other food and drinks; campus bullying and sexual abuse; hot topics in society; infectious or seasonal, outbreaks of diseases and other public health events; teachers and students running away, self-injury, self-mutilation, and suicidal events triggered by various reasons; international hot events; and university students and students running away, self-injury, self-mutilation, and suicidal events. Events; international hot events; criminal offenses of school students; other natural or man-made emergencies are frequent, and colleges and universities are very prone to generating network public opinion. The emergence of these college emergencies makes us deeply realize the importance of the network to social supervision.

PMIS 2024, March 15-17, Changsha, People's Republic of China

DOI 10.4108/eai.15-3-2024.2346412

Public opinion is the sum of attitudes, views and feelings expressed by the public towards hot social events closely related to their own interests. Internet public opinion in colleges and universities assigns public opinion to a special environment, i.e. hot events in colleges and universities. The subjects of public opinion may include students, teachers, administrators, and staff[1]. Higher education online public opinion focuses more on hot issues on campus, and because a large proportion of the audience is students, it can produce faster dissemination and wider dissemination than ordinary public opinion, and the public opinion has a stronger influence. Scholars have conducted empirical studies on the dissemination of college online public opinion and the research on the response strategies to college online public opinion. Li[2] constructed a multi-event network public opinion resonance model based on the bi-stable stochastic resonance model, and analyzed the reasons for its resonance; Ling[3] designed a decision-making model for Internet users of university network public opinion based on the situational crisis communication theory; Chen[4] built a dynamics model for the evolution of university network public opinion in the mobile environment, and explored the methods for coping with the efficiency of network public opinion in the mobile environment. However, the current research is dominated by the number of theoretical studies on online public opinion in universities, and fewer scholars analyze the thematic linkage law of online and offline public opinion in universities and the thematic clustering method through empirical and quantitative studies.

2 Theoretical foundation

2.1 LDA topic model

The topic model LDA assumes that there is a total of D document in the document set, K topics, and V vocabularies (no repetition), and after inputting all the documents, after the LDA algorithm, it will get the probability distribution of each document belonging to these K topics θ_d and the probability distribution of V vocabularies β_k under each topic.

LDA has strong topic mining ability and is an unsupervised model[5], so it is considered that it does not rely on training samples, there is no domain transfer problem, and it has good domain adaptability. The LDA model is a fully Bayesian probabilistic graphical model, and the inference of the parameters needs to infer the posterior distribution of the parameters, so it uses the Gibbs sampling algorithm to estimate the parameters of the model. As Fig. 1 shows the schematic diagram of LDA probabilistic graphical model, which portrays the generation process of the whole long text dataset:

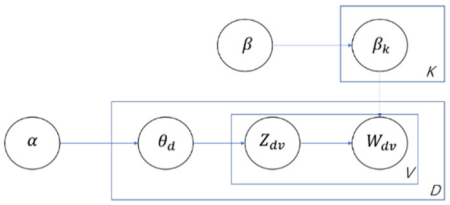

Fig. 1 Schematic of the LDA probabilistic model

1. For each document $d = 1, L, D$: sampling a distribution of document topics$\theta_d \sim D\ ir(\alpha)$;

2. For each topic $k = 1, L, K$: sampling a topic vocabulary distribution $\beta_k \sim D\,ir(\eta)$;

3. For each word in the text: $w = 1, L, V$;

(1) Sample a topic label $Z_{dv} \sim M\,ult(\theta_d)$;

(2) Sample a word under the topic vocabulary distribution of the topic label $W_{dv} \sim M\,ult(\beta_{Z_{dv}})$.

2.2 Theme word co-occurrence analysis

The co-occurrence analysis method is widely used in knowledge text network research, which mainly utilizes text keywords and their co-occurrence relationships to construct a keyword-word co-occurrence matrix, which can be used to carry out thematic hotspot and evolution research. This method assumes that in the corpus, if the frequency of two words appearing together in the same document is greater, then the correlation between these two words is also greater. The basic principle of the word co-occurrence analysis method is to construct the keyword word co-occurrence matrix by counting the number of times any two of all the keywords co-occur in a document and using this value[6].

3 Theme Analysis of Public Opinion in Colleges and Universities

3.1 Selection of reaserch cases and data crawling

In this research, we chose Chongqing colleges and universities as the research object to study the text data during the epidemic in Chongqing. Baidu college posting bar serves as a gathering place for college personnel to voice their opinions online, so the Baidu posting bars of major colleges and universities in Chongqing were selected as the data source for this study. The Baidu postings of 9 representative colleges and universities in Chongqing were selected as data sources. Using python to crawl the Baidu posting bar text data of 9 Chongqing colleges and universities 30714, each data including title, content, reply, time, and user information. According to the opening of the fall semester of 2023 in Chongqing colleges and universities, the official silent management of Chongqing during the epidemic, the lifting of the silent management of Chongqing, the preliminary examination of the 23rd National Master's Degree Admission Examination and the 23rd National Master's Degree Admission Resumption, the five time nodes of crawling the text data are divided into four phases.

3.2 Data preprocessing

Data preprocessing is the pre-processing of the analyzed source text data, which is not only to ensure the reliability of the text data but also to help improve the accuracy of the final analysis results. First of all, to remove invalid data, in the collected network text data often contain some text data with low or no research value. If we ignore the existence of this part of the invalid data, not to deal with it will certainly affect the results of our research. Therefore, we need to carry out basic processing of such data first. Invalid comment data mainly includes the following three types of data:First, repeated comment data. The comment data of the bar user may appear that the user completely agrees with the views of a user, and then completely copy and paste the user's comments, for this kind of data can be retained only one of them. The second category, the comment length is less or the content is too short comment data. Characterize this type of

data as worthless and delete it by setting the minimum comment length. The third category, only one word repeated comment data, this type of data is of low research value, so it is not analyzed.

4 Topic Mining of Public Opinion in Colleges and Universities

Constructing topic co-occurrence network firstly needs to extract the high-frequency topic words in the text through the LDA topic model, and construct the topic word-topic word co-occurrence matrix through the document-topic word co-occurrence. Then input the co-occurrence matrix into Gephi. When importing the co-occurrence matrix into Gephi, it should be set as an undirected edge. After importing the co-occurrence matrix, calculate the average path length (near centrality) of the topic-word network through the statistics toolbar in Gephi, calculate it and perform topic community detection through the modular community detection algorithm that comes with Gephi. Set the nodes in the Appearance toolbar as community distribution ordering and the edges as degree distribution ordering for topic co-occurrence network visualisation. The nodes in the graph represent different topic words, and the lines between the nodes represent the co-occurrence relationship of different topic words, the thicker and deeper the lines are, the deeper the degree of co-occurrence between the topic words is. The words belonging to different clusters are represented in different colours by setting the node size, colour and using the community detection function, and then, the graph is made more intuitive by using the Fruchterman Reingold layout.

Based on LDA topic discovery and topic network community detection to get the topic word social network as shown in Fig. Figure 2 represents the topic word social network graph of college posting during T1. The 20 topic word nodes in the graph have 4 different colors, each color represents different topic word communities, "Semester", "Work", "Exam", "Grades", "Graduation" for the same subject matter community, the same subject matter community represents a college posting theme, this kind of theme is mainly related to the college students on their own learning, academics, life and work concerns and worries, this kind of theme can be summarized as "Employment issues". "Materials" "Questions" "Counseling" "Graduate School" "Information ""Postgraduate" is the same theme word community, the theme involves the examination information and examination materials for help, this kind of theme can be summarized as "examination problems". "Counselor" "Game" "Epidemic" "Nucleic Acid" "Dormitory ", "Freshmen", and "Cafeteria" are the same theme word community, which involves online games, campus epidemic discussions, nucleic acid testing complaints, and campus epidemic life concerns, and these themes can be summarized as "Epidemic Issues". "Campus" and "Learning" are communities with the same subject line, and the theme involves campus learning issues and various discussions encountered on campus, which can be summarized as "Learning Issues".

Figure 2 represents the topic word social network diagram of college posting during T2, the topic word social network in the diagram is divided into 4 communities, "Company", "Graduation", "Work" "Thesis" "Credits" for the same theme word society, the theme is mainly related to the epidemic of employment concerns, the trouble of graduation and the necessary requirements of the consultation, the theme can be summarized as "Employment problems ". "Epidemic" "Dormitory" "Nucleic acid" "Cafeteria" "Campus ", "Freshmen" and "School closure" are the same thematic word community. This theme mainly concerns the learning and

living difficulties brought by the epidemic to college students, their complaints about the closure of schools and dormitories, as well as their evaluation of the cafeteria food during the epidemic. The theme can be summarized as "epidemic problem". "English", "Information", "Results", "Materials", "Examination", and "Questions" are the same theme word community, which is mainly related to the concern of university personnel about the examination policy of the year, the consultation and discussion on the examination issue, and the discussion on the impact of the epidemic on the examination. This theme can be summarized as "Exam Questions". "Going home" and "Unsealing" are the same theme word community, and the theme mainly involves the discussion of going home and unsealing of the epidemic. This theme can be summarized as "Unsealing Issues".

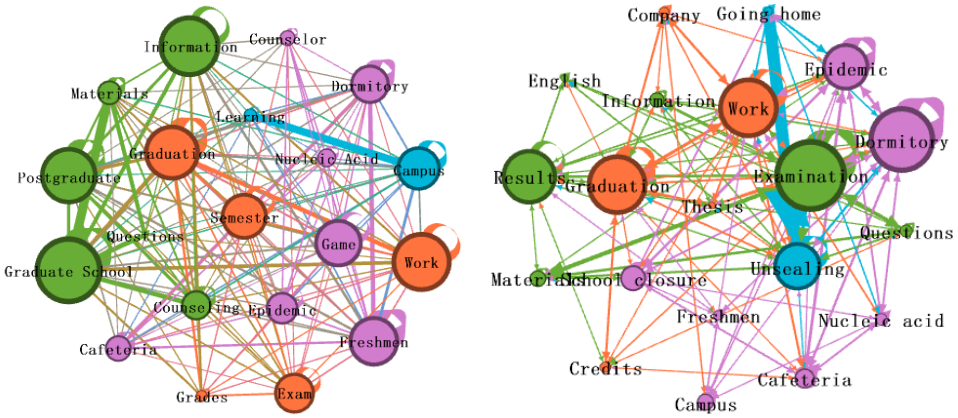

Fig.2 T1 and T2 Thematic Social Network Mapping

Figure 3 represents the social network diagram of college postings during T3, which is divided into four communities, "Examination", "Professional", " Postgraduate", "Math" and "Problems" are the same community. Mathematics" and "Real Questions" are the same theme word community, which mainly involves the discussion of various problems of the examination, and the theme can be summarized as "examination problems". "Work" "Graduation" "Grades" "Nucleic Acid" "Companies " "Thesis" is the same theme word community, the theme mainly involves the discussion of job problems, the discussion of graduation problems, and the discussion of nucleic acid problems The theme can be summarized as "Employment problems". "Exams" "Going home" "Starting school" "Renting a room" "Winter vacation ""Freshmen" is the same theme word community, the theme mainly involves the discussion of the impact of unsealing home and returning to school on learning life, the theme can be summarized as "Learning problems". "Cafeteria", "Dormitory" and "Epidemic" are the same theme word network, and the theme mainly involves the discussion of various problems of the epidemic and the campus life after the unsealing. "Epidemic Issues".

Figure 3 represents the social network map of college posting topic terms during T4, which is divided into four communities, "Education", "Discipline", "Dormitory", "Cafeteria", "Graduate students", "Examination" "Exam questions", "Campus network" is the same Theme word community, the theme mainly involves the discussion of various issues related to the examination, can be summarized as "Examination issues". "Grades", "Exams", "Semester",

"Start of school", "Credits", "Sports" "Repeat" is the same theme word community, the theme is mainly related to the discussion of college life and the discussion of academic performance, the theme can be summarized as "Study problems". "Graduation," "Job," and "Company" are communities of the same theme word, and the theme mainly involves discussion of graduation and job issues, which can be summarized as "Employment issues". "Volunteer" "Transfer" is the same theme word community, the theme mainly involves the discussion of the examination volunteer problem and the examination review discussion, the theme can be summarized as "transfer problem".

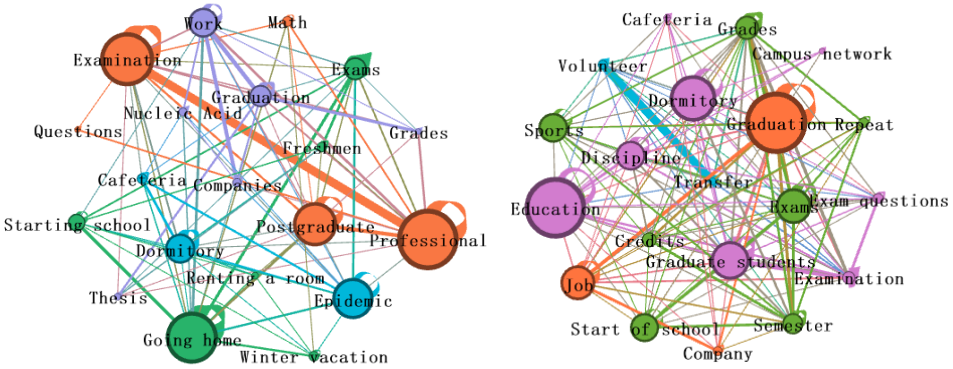

Fig. 3 T3 and T4 Thematic Social Network Mapping

5 Conclusion and discussion

In this study, the LDA topic model is used to obtain the topic words to construct the topic word co-occurrence matrix, and the social network analysis tool is used to construct the topic network of each stage of public opinion in Chongqing colleges and universities, which can help college and university administrators to identify the topic of public opinion and the direction of each stage of public opinion through the visualization of the topic network. Meanwhile, in the face of the complex and changeable network public opinion situation, universities can start to establish university public opinion monitoring through the identification of themes, optimize the construction of university network content, and set up online public opinion topics, so as to establish the emergency management mechanism of university network public opinion. Universities should closely combine with the actual situation of their own schools to start to establish the emergency management mechanism of network public opinion.

Acknowledgments. This research was supported by the National Social Science Foundation of China (Grant No.20BSH076) and the Opening Foundation of Center for Capital Commercial Industry (JD-KFKT-2020-006).

Reference

[1] Mancini C, Pickett T J, Call C, et al.Sexual Assault in the Ivory Tower: Public Opinion on University Accountability and Mandatory Reporting[J].Sexual Abuse: A Journal of Research and Treatment,2019,31(3):344-365.

[2] Li, Z Y.: Simulation of Internet Public Opinion Resonance Phenomenon in Colleges and Universities and Research on Response Strategies [J]. Journal of Intelligence,2019,38(12):107-113.

[3] Ling, Feng, W P.: Research on Emergency Response to Online Public Opinion in Colleges and Universities Based on SOAR Modeling [J]. Information Science,2019,37(09):145-152.

[4] Chen, W L.: Research on Internet Public Opinion Response Strategies of Universities in Mobile Environment Based on System Dynamics [J]. Journal of Modern Information,2018,38(04):118-123+176.

[5] Muzumdar P, Kurian G, Basyal P G .: A Latent Dirichlet Allocation (LDA) Semantic Text Analytics Approach to Explore Topical Features in Charity Crowdfunding Campaigns[J].Asian Journal of Economics, Business and Accounting,2024,24(1):1-10.

[6] WANG, Song, Lu.: A study on the application of co-occurrence analysis in textual knowledge mining [J]. Journal of Library Science in China,2007,(02):59-64.