Research on COVID-19 network public opinion events based on topic model and community detection algorithm

Yong Huang, Haoyu Wang*, Jinjiang Yan#, Weijing Huang
Business School, Sichuan University, Chengdu, China

ABSTRACT

COVID-19 has caused a large number of online public opinion incidents. How to timely and effectively guide the resulting network public opinion has become an urgent problem to be solved. This paper collects more than 130,000 original Weibo posts during the Wuhan "city closure" incident, and analyses the topic characteristics of the incident on the basis of user classification through topic models and community detection algorithms. It was found that during this period, the government responded quickly to the epidemic and gained public support. For different Weibo users, officially certified users mainly publish information about the epidemic and epidemic prevention measures. Personally certified users mainly forwarded and transmitted official information actively, and they also expressed their opinions and made suggestions. Non-certified users actively expressed their emotions and opinions, so they were important users that reflect public opinion.

Keywords: Topic extraction, community detection, COVID-19, public opinion analysis, Weibo

1. INTRODUCTION

In December 2019, several cases of COVID-19 were found in Wuhan. Although Wuhan was actively preventing the epidemic and announced the closure of the city in the early morning of January 23, however, according to data, COVID-19 has spread outside Hubei Province after January 20, 2020. As a major public health emergency, COVID-19 has received high attention from the public. The public has negative emotions such as sadness and panic because of concerns about the safety and health of themselves and their families. Nowadays, with the rapid development of the Internet, the negative emotions caused by COVID-19 spread rapidly through online media, causing online public opinion. The various opinions formed by such a huge number of netizens are exchanged and collided on the Internet, which further increases the complexity of network public opinion. Therefore, it is necessary to accurately guide the network public opinion, to prevent the scope of the incident from expanding, and to maintain the stability of the social and network environment.

2. LITERATURE REVIEW

2.1 Research on internet public opinion

Scholars have studied the dissemination of network public opinion from various aspects. Some scholars have studied the influencing factors of network public opinion evolution. Based on the "silence spiral" theory, Clemente et al. proposed that social participants tend to support the views of most people [1]. Therefore, in online public opinion, the main body of public opinion will spread information to make the online public opinion develop in a direction that is beneficial to them. Liu et al. Constructed a sensitive propagation model of negative network public opinion based on the two-layer network topology [2]. This model can effectively reduce the propagation speed of negative content in network public opinion and provide a new method for the research of complex network public opinion. Dong et al. proposed a dynamic model of network public opinion under the online-offline social network, which revealed the interaction mechanism between online and offline social networks, and made suggestions for network public opinion governance [3]. Zhao et al. defined the transition probability of optimism and pessimism, established an emotional contagion model, and discovered the change law of public sentiment [4]. In order to find the community structure in network public opinion, Zhang et al.

* 359426442@qq.com

john2016@126.com

Third International Conference on Computer Science and Communication Technology (ICCSCT 2022)
edited by Yingfa Lu, Changbo Cheng, Proc. of SPIE Vol. 12506, 125062R
 · 0277-786X · doi: 10.1117/12.2662583

proposed an improved Fast Louvain algorithm based on the traditional Louvain algorithm, which is superior to the traditional method in effect and efficiency [5]. Some scholars use big data technology to mine the characteristics of network public opinion. In order to study the influence of images on the spread of online public opinion, Seltzer et al. took the online public opinion of the Zika virus epidemic as an example, and analysed the spread of image public opinion on the Instagram platform [6]. Oliveira et al. used the LDA model to identify user text topics on Twitter, and used an optimized BERT pre-training model to identify the sentiment, and conducted topic-sentiment analysis on the Twitter network public opinion of COVID-19 [7]. Singh et al. used the LDA algorithm for topic modeling, and based on the NRC sentiment vocabulary, they used Syuzhet software package to perform sentiment topic analysis on the food network public opinion on Twitter [8]. Ding et al. integrated the influence characteristics of comments into the TF-IDF algorithm to mine topics, and proposed six methods of topic evolution to mine the evolution forms and laws of network public opinion [9].

2.2 Network public opinion research on public health emergencies

On the basis of network public opinion research, many scholars have conducted detailed research on network public opinion in public health emergencies. Taking COVID-19 as an example, Yang analyzed the network public opinion patterns and transmission characteristics of public health emergencies, and put forward suggestions for improving the efficiency of network public opinion governance during epidemic prevention and control [10]. Ma et al. analysed the emotional transmission characteristics of online public opinion in COVID-19, and put forward suggestions on how the government should guide public sentiment in the network public opinion [11]. Machuca et al. applied the Logistic regression algorithm to classify the sentiment of COVID-19 related tweets in Twitter, and the classification accuracy reached 78.5% [12].

After the outbreak of COVID-19, online public opinion research on COVID-19 has also emerged. Li et al. used Weibo data and natural language processing technology to classify information related to COVID-19 into seven situational information, thereby helping the public and authorities to identify valuable information [13]. Jena et al. analyzed the Twitter public opinion in India during COVID-19 outbreak by using natural language processing-based text mining technology, and provided a powerful NLP-based text mining framework to effectively extract topics related to COVID-19 outbreak [14].

It can be seen that the research on public opinion during COVID-19 mainly focuses on public opinion patterns and the optimization of emotions and topic algorithms. There is less empirical analysis on typical events during the epidemic, and there is a lack of research on the characteristics of Internet public opinion events classified by users. Therefore, this paper chooses to study the public opinion characteristics of the Wuhan “city closure” incident during the period of COVID-19 from the perspective of data mining on the Weibo platform, and study the focus of attention of Internet users by building topic word network communities of different types of users.

3. RESEARCH DESIGN

3.1 Design of data collection and preprocessing

In order to study the characteristics of online public opinion event during COVID-19, this paper selects a representative Wuhan “city closure” incident for empirical research. In terms of data collection, this paper is based on the Weibo platform, and uses the crawler software “Houyi Collector” to collect related Weibo.

Due to the problems of invalid characters, garbled data and “dirty data” in the Weibo text, the data collected through the collector can not be directly used for analysis. Therefore, data cleaning and other processing processes are needed to support follow-up research. This paper cleans the repeated Weibo and invalid Weibo, and removes the emojis that affect the topic mining.

3.2 Design of topic word network analysis

Classify the pre-processing event Weibo into officially certified users, personally certified users and non-certified users, and use LDA topic model to extract the topic information implicit in all Weibo, officially certified users’ Weibo, personally certified users’ Weibo and non-certified users’ Weibo. Then the extracted topics are combined, classified, and analyzed according to the user category.

Secondly, the co-occurrence characteristics are used to make undirected connections to form a topic word network of different types of users. The topic community is divided according to the constructed topic word network, and analyzed according to the user type. When dividing the network community, this paper uses the Louvain algorithm and calculates the modularity Q to evaluate the quality of the network community division.

4. EMPIRICAL ANALYSIS

4.1 Data collection and preprocessing

Since the minimum time interval of the Weibo displayed on the Weibo platform is one hour, this paper uses the "Houyi collector" to collect the original Weibo of Wuhan "city closure" incident at one hour intervals, and collects the original Weibo text for a total of 10 days from January 22, 2020 to January 31, 2020, with a total of 131,528 Weibo. After data cleaning, 107,236 effective Weibo were obtained. The number of Weibo collected and the effective rate of collected data are shown in Table 1.

Table 1. Number and efficiency of Weibo.

Event Name	Total Number of Weibo	Invalid Weibo	Repeat Weibo	Effective Weibo	Data Efficiency
Wuhan "city closure" incident	131,528	765	23,527	107,236	81.53%

4.2 Topic word network analysis

4.2.1 Word Information Analysis. In this paper, the LDA Algorithm was used to extract the topic of Weibo texts. Before topic extraction this paper determined the optimal number of topics by indicators of perplexity and consistency. When the number of topics was 10, the consistency score was the highest, and at the same time, perplexity decreased the fastest. Therefore, the number of topics selected in the topic extraction of the Wuhan "city closure" incident is 10.

The LDA model in the genism library is called to extract the topic of Weibo text of the Wuhan "city closure" incident, the number of topics is set to 10, and the first 10 topic words are output. The results are shown in Table 2.

Table 2. The extraction results of the Wuhan "city closure" incident topic.

Topic	Topic Word
Topic 0	Hope, COVID-19, Facial Mask, At home, Go out, Eat, Look, Think, Spring Festival, New year
Topic 1	Epidemic prevention and control, epidemic, Prevention and control, Work, Personnel, Assure, Measure, Emergency, Epidemic prevention, COVID-19
Topic 2	Link, Web page, Coronavirus disease, COVID-19, Epidemic, News, Real time, Map, Post, Renew, Article
Topic 3	Hospital, Patient, Infect, Treat, COVID-19, Detect, Expert, Medical, Rescue, Clinical
Topic 4	COVID-19, Virus, Drug, Research, Vaccine, WHO, Wuhan Institute of Virology, SARS, Research and development, Graduate School
Topic 5	Facial mask, Protection, Prevention, Suggest, Disinfect, Healthy, Wash hands frequently, Wear mask, Spread, Symptom
Topic 6	China, Wuhan, International, US, Country, Japan, WHO, Medical Team, PHEIC, Impact
Topic 7	Material, Market, Information, Mask, Price, Social, Business, according to law, Supervised, Sale
Topic 8	Fight, Wuhan, Come on, First line, China, Win, Medical staff, Hope, Love, Unity
Topic 9	Example, Day, Month, Infect, Case, Patient, New addition, Report, Discharged from hospital, Grand total

After identification, it can be found that although the 10 topics extracted by the LDA algorithm all describe different contents, according to the topic description and the semantics of the topic words, the contents of some topics can be classified as a type of information. Therefore, the above topics are grouped into four topics of “epidemic information”, “scientific popularization advice”, “epidemic prevention measures”, and “popular dynamics”. The specific categories are shown in Table 3.

Table 3. Four major topic categories.

Topic category	Service properties
Epidemic information	Topic 2, Topic 4, Topic 6,Topic 9
Scientific popularization advice	Topic 5
Epidemic prevention measures	Topic 1, Topic 3, Topic 7
Popular dynamics	Topic 0, Topic 8

The LDA topic model generates probabilities of different topics for each Weibo. Each Weibo is classified according to the maximum probability, and the proportion of Weibo on various topics of officially certified users, personally certified users and non- certified users is counted, as shown in Figure 1.

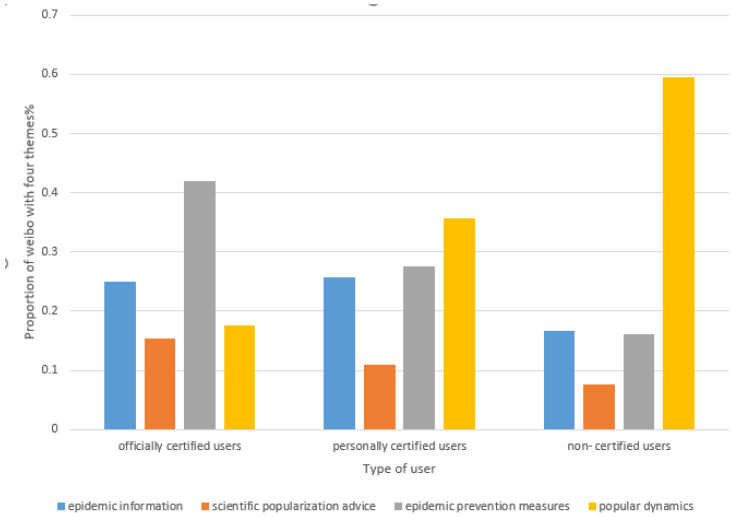

Figure 1. Weibo on the four topics of different types of users.

It can be seen from Figure 1 that scientific popularization advice account for the least among the three types of users. Among the Weibo posts published by officially certified users, epidemic prevention measures and epidemic information account for a relatively high proportion, which shows that the government attaches great importance to the epidemic and publishes relevant information in a timely manner. Personally certified users’ Weibo accounts for each topic are relatively balanced. They will not only publish information about the epidemic, but also express their views and feelings. Different topics will reflect different characteristics of public opinion, and people's livelihood has high public opinion value because it reflects the public’s living conditions and real feelings during the Wuhan "city closure" incident. Among Weibo published by different types of users, Weibo of different topics account for different proportions, so the topics of Weibo of different types of users are discussed.

4.2.2 Analysis of Topic Word Network Community. In order to explore the differences in Weibo topics of different types of users during the Wuhan “city closure” incident, and to discuss the correlation between the topic words, based on the topic extraction of the Wuhan “city closure”, the topic extraction was performed and 20 topic words were obtained respectively, and the co-occurrence relationship between the topic words was used to form word pairs to construct an

undirected network graph. The topic words were imported into Gephi software, and the Louvain algorithm was used for community detection, and the series of Figures 2-5 were obtained. Among them, officially certified users, personal certified users and non-certified users have the topic words community modularity of 0.457, 0.485 and 0.537 respectively, all of which have good community structure characteristics.

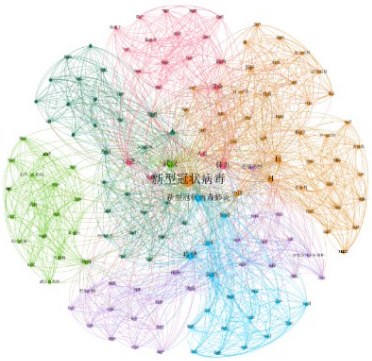

Figure 2. All users topic word community network.

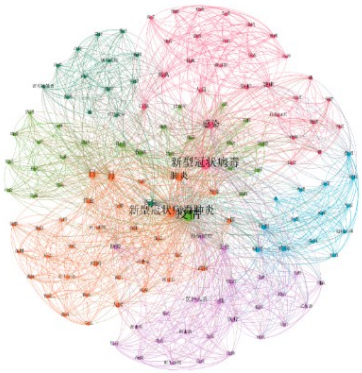

Figure 3. Officially certified users topic word community network.

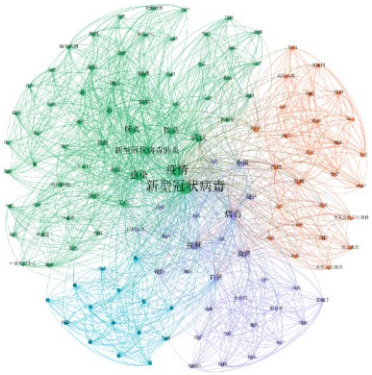

Figure 4. Personally certified users topic word community network.

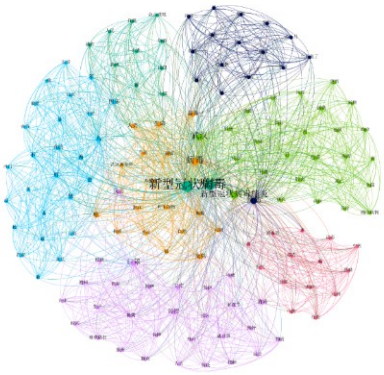

Figure 5. Non-certified users topic word community network.

It can be seen from Figure 2 that the community analysis merges Topic 0 and 8, Topic 1 and 7, Topic 3, 6 and 9 through the co-occurrence of the topic words. In the network diagram, there are more topic word communities belonging to the topics of epidemic information and epidemic prevention measures, and fewer topic word communities belonging to the topics of people's livelihood dynamics and popular science advice. It can be seen from Figure 3 that the topic words of officially certified users are divided into 6 communities. Most of the topic word communities belong to the topic of epidemic information and epidemic prevention measures, and the classification of the communities is more detailed. The word community network reflects the fact that officially certified users focus on publishing the latest news about the epidemic and the measures taken to prevent it. At the same time, they also publish information on popular science and advice about the epidemic, but lack public opinion guidance. It can be seen from Figure 4 that the topic words of personally certified users are divided into four simple communities. Among them, the topic words belonging to the epidemic information are relatively large and highly concentrated, which reflects the high repetition of the discussion content of the personally certified users on the epidemic information, which shows that the personally certified users are more willing to spread the Weibo of the officially certified users. They also publish some suggestions and popular science content summarized by themselves to help other users. It can be seen from Figure 5 that the topic words of non-certified users are divided into 7 communities. Most of the topic word networks of non-certified users belong to the dynamic topics of people's livelihood, and express rich public opinion. The topic word community included in the topic of public opinion information reflects the main content that netizens pay attention to, that is, the data and traceability of the epidemic. Non- certified users also actively spread advice and popular science information to help other users.

According to the above analysis, the topic word communities presented by the three types of users reflect their own Weibo characteristics. Officially certified users' Weibo focuses on the release of various information and measures. Personally certified users will actively forward Weibo of officially certified users and publish their own views and feelings. Weibo of non-certified users focuses on the expression of mood and life status. When conducting public opinion guidance and control, the public opinion of personally certified users and non-certified users should be analyzed emphatically. Personally certified users have been screened by Weibo and have a certain fan base. Personally certified users should be appropriately used to respond to public opinion to help guide the development of public opinion.

5. CONCLUSION

This paper uses the Weibo platform as the data source to study the topic characteristics of the Wuhan "city closure" incident during COVID-19. The significance of the research results is mainly reflected in the following aspects:

- After the outbreak of COVID-19, the government took epidemic prevention measures in time and released various information through social media, which received positive public discussions and played a positive role in guiding online public opinion.

- The content released by officially certified users during public opinion events is mainly to announce epidemic information. The content released by personally certified users during public opinion events is similar to that of officially certified users, and they also post a small amount of comments and opinions on Weibo. non-certified users are an important group of users who express public opinions, this type of users will express their emotions due to the situation of the epidemic, and will also express their views on the epidemic.

- Suggestions for public opinion guidance: When an epidemic occurs, the government should respond quickly, reduce the negative impact, and release information to the public in time. After a public opinion event occurs, different types of users will have their own public opinion characteristics. Therefore, a public opinion management system should be established according to the type of users, and special individual users should be fully utilized to guide and control them in a timely manner.

There are still shortcomings in this paper, such as the long duration of COVID-19, which can increase the research on the evolution of public opinion over a long time span. At the same time, the public opinion research on COVID-19 should not be limited to the Wuhan "city closure" incident, and multiple network public opinion events can be studied.

REFERENCES

[1] Clemente, M. and Roulet, T., “Public opinion as a source of deinstitutionalization: A ‘spiral of silence’ approach,” Academy of Managenebt Review, 40(1), 96-114(2015).

[2] Liu, X., Tang, T. and He, D., “Double-layer network negative public opinion information propagation modeling based on continuous-time Markov Chain,” The Computer Journal, 64(9), 1315-1325(2020).

[3] Dong, Y., Ding, Z., Chiclana, F., et al., “Dynamics of public opinions in an online and offline social network,” IEEE Transactions on Big Data, 7(4), 610-618(2017).

[4] Zhao, L., Wang, J., Huang, R., et al., “Sentiment contagion in complex networks,” Physica A Statistical Mechanics & Its Applications, 394(2), 17-23(2014).

[5] Zhang, J., Fei, J., Song, X., et al., “An improved Louvain algorithm for community detection,” Mathematical Problems in Engineering, 1-14(2021).

[6] Seltzer, E. K., Horst-Martz, E., Lu, M., et al., “Public sentiment and discourse about Zika virus on Instagram,” Public Health, 150, 170-175(2017).

[7] Oliveira, F. B., Haque, A., Mougouei, D., et al., Investigating the emotional response to COVID-19 News on Twitter: a topic modelling and emotion classification approach, IEEE Access, 10, 16883-16897(2022).

[8] Singh, A. and Glinska-Newes, A., “Modeling the public attitude towards organic foods: A big data and text mining approach,” Journal of Big Data, 9(1), 1-21(2022).

[9] Ding, S., Liu, X. and Li, Z., “Research on the evolution of internet public opinion hot topics based on the influence of comments,” Modern Intelligence, 41(8), 87-97(2021).

[10] Yang, L., “Network public opinion governance in major public health emergencies,” News Research Guide, 11(14), 40-41(2020).

[11] Ma, D., “The dissemination mechanism of public sentiment in public opinion emergencies,” Journalism and Communication, (5), 7-8(2022).
[12] Machuca, C. R., Gallardo, C. and Toasa, R. M., “Twitter sentiment analysis on Coronavirus: Machine learning approach,” 2020 International Symposium on Automation, Information and Computing (ISAIC 2020), 12-20(2021).
[13] Li, L., Zhang, Q., Wang, X., et al., “Characterizing the propagation of situational information in social media during COVID-19 epidemic: A case study on Weibo,” IEEE Transactions on Computational Social Systems, (99), 1-7(2020).
[14] Jena, R. K. and Goswami, R., “Understanding peoples’ sentiment during different phases of COVID-19 lockdown in India: A text mining approach,” International Journal of Business Analytics, 8(4), 52-68(2021).