Multimedia Tools and Applications (2023) 82:25811–25827
https://doi.org/10.1007/s11042-023-14916-x

Text recognition and analysis of network public opinion focus events of a major epidemic: a case study of "COVID-19" in Sina Microblogs

HeLin Wei[1,2] · Chenying Hai[3] · Donglu Shan[1] · Bei Lyu[4,5,6] · Xiulai Wang[7]

Received: 14 June 2021 / Revised: 12 February 2023 / Accepted: 14 February 2023 /
Published online: 7 March 2023

Abstract

Identifying and analyzing the public's opinion of focal events during a major epidemic can help the government grasp the vicissitudes of network public opinion in a timely manner and provide the appropriate responses. Taking the COVID-19 epidemic as an example, this study begins by using Python-selenium to capture the original text and comment data related to COVID-19 from Sina Microblog's CCTV News from Jan. 19, 2020, to Feb. 20, 2020. The study subsequently uses a manual interpretation method to classify the Weibo content and analyzes the shifting focus phenomena of network public opinion based on the moving average method. Next, the study uses an enhances TF-IDF to extract keywords from the Weibo comment and uses the keywords to construct a word co-occurrence network. The results show that during the epidemic, the network public opinion focus shifted significantly over time. With the progression of the epidemic, the focus of network public opinion diversified, and various categories stabilized. Compared to simple keyword and text classification recognition focus problems, the proposed

✉ Chenying Hai
893046060@qq.com

✉ Bei Lyu
peter1983123@hotmail.com

[1] School of Business, Guangxi University, Nanning, China

[2] Key Laboratory of Interdisciplinary Science of Statistics and Management, Education Department of Guangxi, Guangxi University, Nanning, China

[3] School of Management, Huazhong University of Science and Technology, Wuhan, China

[4] School of Economics and Management, Huaibei Normal University, Huaibei, China

[5] Panyapiwat Institute of management, Nonthaburi, Thailand

[6] Leeds University Business School, University of Leeds, Leeds, UK

[7] Institute of Big Data on Talents, Nanjing University of Information Science and Technology, Nanjing, China

model, which is highly accurate, identified multiple network public opinion focus problems and described the core contradictions of the different focus problems.

Keywords COVID-19 · Keyword co-occurrence · Major epidemic · Network public opinion

1 Introduction

The SARS incident, the Japanese nuclear radiation incident, the Australian wildfire and other incidents threatening the lives and health of people occur frequently and have attracted immense attention in recent years. New media platforms such as Instagram, Facebook and TikTok provide a broad variety of media channels for people to follow current events, obtain information and post comments, so network public opinion has the characteristics of massive content, scattered users, diversity of arguments and so on [1]. Public opinion refers to the collection of cognitions, attitudes and emotions produced by people due to the occurrence and development of a social situation [2]. Network public opinion refers to the sum of peoples' cognitions, attitudes, and emotions about social situations published on social media platforms with the help of mobile phones, computers and other devices [3]. As the epidemic is characterized by a wide impact range, long duration, strong uncertainty and great harm [4], its emergence could lead to the sudden appearance of a large number of information sources with low controllability and uneven quality throughout all the media platforms, which could easily cause a network public opinion crisis and have a negative impact on society. Additionally, due to the influence of factors such as information asymmetry and the decline of the public's cognitive abilities, a network public opinion crisis caused by the epidemic presents significant challenges to the government's information governance. Therefore, it is essential to control the development of network public opinion quickly and effectively.

Although the government attaches great importance to the supervision of network public opinion, the government's ability to deal with the network public opinion that emerges from major epidemics still needs to be improved. Once a major epidemic occurs, if the network public opinion is difficult to guide or control, the social and economic situation is likely to become chaotic. The key to guiding or controlling the development of network public opinion is to identify and understand the focus of the public's attention over time. Typically, network public opinion will continue to ferment as a situation develops, and the internet also tends to accelerate the evolution of network public opinion, indicating that the public's opinion also changes rapidly. Because a major epidemic will go through various development stages (e.g., emergence, outbreak, and decline), in this paper we investigate: 1) what specific issues are the focal points of network public opinion at different times during an epidemic; 2) how are the focal points are identified and analyzed; and 3) what trends exist in the relevant issues. These are the issues that must be resolved to improve the government's ability to govern network public opinion.

The research on the network public opinion management of past major social events is a rich source of information, and big data and artificial intelligence technology has been used to deeply explore numerous issues, such as the public opinion communication modes [5], the mechanisms used for public opinion diffusion [6] and public opinion monitoring and early warnings [7]. In terms of algorithms for public opinion analysis, most of the existing studies are based on probabilistic statistical models or natural language processing models, such as, the incremental hierarchical clustering algorithm [8], K-nearest classification algorithm [9], etc. However, these methods focus on the frequency and occurrence time of public opinion related information, pay more attention to the efficiency of public opinion analysis, but ignore its accuracy. There are still many problems in the management of network public opinion. For example, the public opinion presented by an algorithm may deviate from the real public opinion, and the complexity of a public sentiment analysis could lead to misjudgments. Therefore, it is imperative to find an optimization method that accurately identifies and analyzes the network public opinion of major epidemics. This study proposes a method to solve this problem; text analysis of comments on social media posts. The analysis method is a process that uses the moving average method, TF-IDF keyword extraction and a keyword co-occurrence network analysis. This paper takes network public opinion related to the topic of "COVID-19" (COVID-19 for short) on the Sina Microblog platform (https://weibo.com/), as an example, to verify this method and analyzes the network public opinion focus trend changes in different time periods during the epidemic, as well as the specific problems occurring in the different time periods. This article not only enriches the application of the co-occurrence network analysis method in the research of public opinion governance but also provides references for the government to identify and analyze the focus of the network public opinion of major epidemics.

2 Review of related research

In past major emergencies, scholars primarily conducted research in the fields related to the dissemination of network public opinion and the monitoring and control of network public opinion. For example, some scholars took the "Ebola virus" event as an example to study whether the dissemination of picture information of public health emergencies on different social media platforms would produce different effects[4]. In addition, a positive correlation was found between the number of Twitter posts and the number of virus cases during the H1N1 pandemic [10]. In terms of the monitoring and control of network public opinion, there are scholars studied the trends and characteristics of public concerns during the Ebola, Seca and Middle East respiratory syndrome epidemics [11], then found the response and handling of media governance is a key factor influencing the development of public opinion on public health emergencies [12]. Most of the studies on the dissemination and monitoring of network public opinion were performed from objective perspectives, investigating the generation mechanisms, transmission channels and evolution mechanisms of network public opinion. Few scholars have studied changes in the focus of public opinion events from the public's subjective perspective and analyzed the changes to identify and discover the focus of public opinion issues. The duration of an epidemic is often extensive, and its impact on public health, work and study and the country's economic development will occur sequentially. Therefore, the focus of public attention during an epidemic is likely to change over time on the internet, and there may be specific trends in the public opinion focus. Based on the public's subjective perspective, a timely understanding of the changing trends of public opinion at different times during the epidemic, and the specific issues that are the focus of the public, are important in allowing the government to guide network public opinion, control information and maintain social order.

3 Research design

The purpose of text stereo recognition of major epidemic network public opinion focus events is to understand the dynamic changing trends of the public's opinion focus and the specific types of focus events. To achieve this goal, a text stereo recognition model of the major epidemic network public opinion focus events, as shown in Fig. 1, is constructed. The proposed model can be divided into the following three parts: 1) Data acquisition, where basic microblog information is gathered from a micro-index platform by crawlers; 2) data processing, which uses a manual interpretation of the Weibo content, calculation of the sliding average of the number of reposts, word segmentation and cleaning of the comment content, extraction of keywords with TF-IDF and construction of a word cooccurrence network; and 3) data analysis, which includes classifying the Weibo content and analyzing the public opinion trends.

Compared with the evolutionary game model [13], the logistic regression model [14], default model [15] and other models used in previous studies, the main contribution of this study is combining the existing methods to design a more accurate and fast text analysis model for public opinion recognition and analysis. The public opinion identification and analysis model in this study is not an algorithmic innovation but a process optimization. By mining the internal relationship between words in the public opinion text corpus and using the internal logical relationships between the texts to clean up, refine, summarize and analyze the public opinion text, managers can quickly and efficiently grasp the public opinion development trends of major events, understand the public's hot issues of concern and find countermeasures.

The traditional way to grasp the dynamics of public opinion through text analysis is to extract keywords from all relevant content in social media to identify the key information in the public opinion text [16]. As we know, the keyword is usually a single word or a phrase, it can not express limited information or reflect the internal relationship between the same or similar topics, moreover, if the same keyword appears in different topics or is combined with other different keywords, it may represent different views [17]. Therefore, traditional methods are difficult to accurately grasp the different demands and rich emotions of the public, which will reduce the accuracy of public opinion analysis. In addition, there are many theme analysis

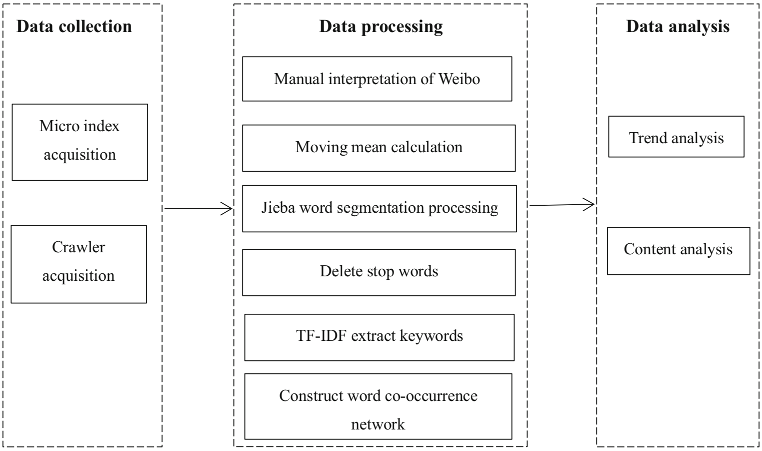

Fig. 1 Text stereo recognition model of the focal events of major epidemics on the Internet

models, one of which is a theme model that introduces potential semantic dimensions to obtain document themes through unsupervised learning of text, represented by the classical Latent Dirichlet Allocation (LDA) model, with a series of improved models derived from it [18]. These models can adapt to large-scale text and have better semantic representation characteristics, but they also have some drawbacks, such as limited semantic information recognized, low theme recognition and difficult to interpret in context. The other one is term clustering that establishes some association rules, such as theme mining based on terms co-occurrence relationship analysis [18]. The advantage is that it is easy to use and feasible, but the drawback is that the meaning of the obtained clustered themes is more ambiguous.

Since the content of public opinion on major epidemics is complex and closely related to various contents, and there are many semantic information and special contexts, this study chooses to combine TF-IDF, co-occurrence analysis and manual recognition to perform theme mining of online public opinion, which ensures that the identified themes are comprehensive and accurate. And these methods are applied to bibliometric analysis [19], consumer purchase behavior prediction [20], higher educational institutions evaluation [21], and other fields, which can ensure the feasibility of the methods and the validity of the results.

For research data, we selected the comment contents of posts on the mainstream social media platform in China-Sina Microblog for text analysis. As the expression of comments is generally unrestricted, the public can fully express their personal opinions and emotions, including the opinion interaction between the public, so it is an excellent data resource for analyzing public opinion trends.

3.1 Data collection

This study mainly discusses two issues, the first is the analysis of the development trend of internet public opinion, and the second is the identification and analysis of the focus events at each stage of public opinion development. Next, it uses the public opinion of COVID-19 as expressed in Sina Microblog as an example. The data source for Question 1 is the micro index of Sina Microblog, which is an index obtained by comprehensively weighting indicators such as the number of mentions, readings and interactions in Weibo related to a specific keyword, which accurately describes the popularity of a discussion in Weibo about an event that the keyword belongs to, along with the network public opinion trend. The data source for Question 2 is the basic information in Sina Microblog related to COVID-19, including text body content, comment content, number of reposts, user ID and posting time. The methods and procedures used to obtain the two data sources are described below.

The data acquisition method of the micro index is the official data analysis tool “Micro Index” of Sina Microblog. The data acquisition steps are as follows. First, we selected search keywords, including words such as “new crown virus”, “Wuhan epidemic” and “COVID-19”; however, no data were found. Then, considering that in the early stage of the epidemic, the patient numbers were small, most patients appeared in Wuhan, and no major epidemics had yet occurred, we considered that the COVID-19 epidemic primarily appeared in the public eye as “pneumonia”. Therefore, the index trend of “pneumonia” represents the network public opinion trend of COVID-19. During the outbreak, the number of patients increased sharply, and the epidemic spread across the country. The word “pandemic” then became synonymous with COVID-19 incidences in the eyes of the public. Finally, at this stage, the index trend of “epidemic” was used to represent the COVID-19 network and the public opinion trend. Thus, we searched for “epidemic + pneumonia” as a keyword, which comprehensively describes the micro-index of COVID-19. The results of this search are shown in Fig. 2. Second, we selected the search time period. Because the index trend analysis time period stipulated by the "micro-index" system is "1 hour", "24 hours", "30 days" and "90 days", combined with the COVID-19 network public opinion, the earliest occurence appeared in mid-December 2019, and the number of days from mid-December 2019 to February 20, 2020, was below 90 days and more than 24 hours, so we selected "30 days" as the search time period. Using these two steps, the micro index map shown in Fig. 2 was created.

The basic information on Sina Microblog related to COVID-19 was obtained through the Sina Microblog page. The data acquisition steps used in this study are as follows. First, we determined the information source. Because CCTV News is the official Weibo of CCTV News, which has some authority and is more popular when compared to other Weibo websites, the official Weibo of CCTV News was thus selected as the information source. Second, we determined the release date of Weibo. Fig. 2 shows that from Jan. 20, 2020, to Jan. 23, 2020, the micro-index of "pneumonia" increased markedly, and a clear network public opinion formed. Additionally, from Jan. 23, 2020, to Feb. 20, 2020, the micro-index of the epidemic fluctuated, and the strength of the network public opinion peaked. These results show that between Jan. 20, 2020, and Feb. 20, 2020, the public's attention to COVID-19 reached and sustained a relatively high level. Additionally, these results indicate that related popular topics and their derivative topics spread quickly and had a large influence; thus, information may be amplified or distorted, which can easily lead to a public opinion crisis. Thus, it is believed that analyzing the changes in the focus of public opinion during this period and understanding the issues that the public is truly concerned about plays an important role in the precise control of network public opinion. Third, we obtained data. Python-selenium was used to capture the basic information from Weibo related to COVID-19 that was released by CCTV News from Jan. 20, 2020 to Feb. 20, 2020.

3.2 Data processing

First, a manual interpretation was used to classify the content of the Weibo text. Manual interpretation refers to researchers' manual interpretation of the analyzed content. This method is widely used in text analysis, such as when deleting irrelevant text and classifying text [22]. In this study, we used this method to classify the content of the Weibo text. The operational steps for a manual interpretation are roughly as follows: 1) In Excel, the crawled Weibo text content is arranged in descending order according to the number of "likes" obtained. The one with the highest number of "likes" is in the front. If the number of likes is the same, it is arranged in descending order according to the release time, and the latest one is in the front; 2) Browse the content of the Weibo text, set the classification standards according to the events

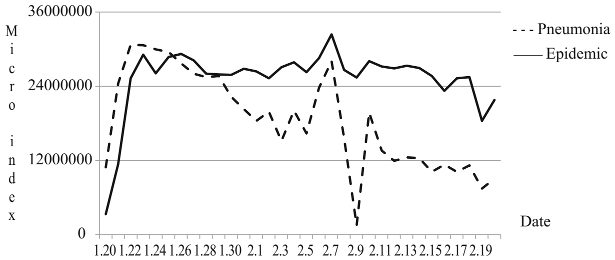

Fig. 2 "Pneumonia + pneumonia" micro index chart

covered by the content, and then set the category name according to the classification standards; 3) Interpret the content of the Weibo text and make a preliminary classification; 4) Improve the classification standard according to the classification situation (the final category name and classification standard are shown in Table 1); and 5) Classify the content of the Weibo text again.

Second, the moving average method was used to calculate the moving average of the number of reposts of each type of microblog. According to previous studies, compared to the number of likes and comments, the number of Weibo reposts is more important for public opinion monitoring [23]; thus, an analysis of the changes in the number of reposts of microblogs on different topics determines the development path of network public opinion. Concurrently, a moving average can be calculated from the data of adjacent periods to reduce the impact of accidental change factors on the trends [24]; thus, the sliding average is used to process the data and eliminate accidental fluctuations in the forwarding number (e.g., a popular event occurred under a certain category on a certain day). Additionally, to prevent differences in the number of microblogs on the same topic in a single day, three microblogs on the same topic on the same day are randomly selected to eliminate the influence of an excessive number of microblogs of a certain type on a certain day.

Third, we created a line chart of the sliding average of the number of Weibo reposts and divided the development of network public opinion according to the change in the sliding average. The sliding averages of the number of reposts of different types of microblogs have numerical differences in the different time periods, which can describe the degree of public attention to the events reflected in each type of microblog at the different times, which describes the changes in the focus of network public opinion at different times. When the

Table 1 Weibo content classification and standard description of each category

Classification	Classification standard
Epidemic progress	Changes in the situation of confirmed, suspected, confirmed, dead, cured, etc. at home and abroad; looking for people who have taken the same means of transportation as the suspected or confirmed case; the public's attitude towards the epidemic.
Popular science	Measures to prevent infection; knowledge of the transmission route, mortality, pathology, and naming of the new coronavirus; knowledge of home observation and care by oneself; methods of screening feverish persons at stations; methods of psychological counseling.
Prevention policies and measures	To control the epidemic, governments at all levels and governments have formulated policies and measures in transportation, education, economy, medical care, entertainment, life, and work.
Official response and action	Official explanations and responses of government departments, hospital leaders and government officials on issues related to the epidemic, prevention and control work, medical assistance, violations of laws and regulations, diplomacy, and rumors; field inspections and supervision conducted by leaders.
Frontline personnel situation	The infection of front-line personnel; the working and living conditions of front-line personnel.
Supplies	Supply and demand of protective equipment; announcement of material donation; public announcement of receiving material; disclosure of material management and usage; material quality; medical conditions.
Special event	Special cases of illness, cure, and death; deeds and sacrifices of relevant staff; donors of blood plasma; other incidents (such as incidents related to games and software).
Treatment measures	Specific treatment methods; diagnosis and treatment plan; drug development and testing; pathological research.

sliding mean polylines of different types of Weibo reposts intersect, the focus of network public opinion shifts.

Fourth, we conducted a word co-occurrence network analysis on the data's content. We randomly selected one Weibo of each type published on any day in each time period. Due to a limitation of the Sina Microblog system, we can only obtain 300 comments per Weibo at most. Therefore, the comments of each Weibo are sorted according to their popularity (the number of likes of a Weibo comment), and the top 300 comments are selected for text analysis. As the comments analyzed have the highest number of likes among all the comments, that is, those that are widely of interest and agreed upon by the public, which means they are highly representative of public opinion. The remaining comments are relatively less important, and they are not analyzed, which has a relatively small impact on the final results. When analyzing the comments, Jieba word segmentation was first used to segment the comment content. Second, we treated a comment as a document and used the TF-IDF method for calculation. The TF-IDF method is often used to evaluate the criticality of words in a text database [25]. In this study, the method was used to extract keywords in the content of the Weibo comments. Numbers, prepositions, pronouns and certain adverbs are less helpful in identifying the subject of the review; thus, they are used as stop words. Finally, the extracted keywords were used to construct a word co-occurrence network and perform a word co-occurrence network analysis. Specifically, each comment is traversed, and if two keywords appear in the same Weibo, it is considered to be a keyword co-occurrence; a keyword co-occurrence network is constructed, and a visualized result is obtained.

Many low-quality words appearing in the word co-occurrence network will affect the analysis of the network; however, because the term frequency and inverted document frequency in the TF-IDF method are only first-order statistics in the text, the impact on the second-order distribution feature between the words is not linear; thus, the keywords extracted by TF-IDF alone cannot fully describe their importance in the word co-occurrence network. Typically, if a word co-occurs with other words only a few times, the importance of the word is lower. The sum of the number of times that a word co-occurs with other words is the degree centrality of the word, which is also known as the degree of the word; the degree centrality of the word co-occurrence network describes its importance from a certain level. However, there are specific vocabularies in a real situation. Although they are related to many other vocabularies, they tend to be connected with certain words with low degree centrality. Therefore, filtering nodes in a network using only degree centrality is not reliable. The K-core algorithm is an algorithm that can roughly divide nodes. First, the algorithm removes words of degree 1 (i.e., words with only one edge) from the network, which will cause a new batch of words to appear in the network. For the vocabulary of 1, there will only be nodes of degree 2 in the network after several iterations, which yields the 1-core structure in the network. Then, we remove the nodes of degree 2 in the network. After many iterations, a 2-core structure is formed. The advantage of the k-core method is that it can distinguish nodes with false high degrees of centrality in the network so that the word relationship structure in the stripped network is clearer, and the connection is closer. After filtering the weights of the words and edges, a topology-based clustering algorithm is used to cluster the word co-occurrence network. The nodes classified into the same category are words that describe the same topic. The community partition algorithm based on a complex network is an unsupervised clustering algorithm that is based on the network topology. The algorithm for community division is rich, and the algorithm used in this paper is the fast-unfolding algorithm. Compared to other community discovery algorithms, this algorithm achieves rapid computational speeds, high recognition accuracies and stable results.

4 Empirical analysis

4.1 Analysis of the transfer trend of network public opinion at different times during the epidemic

In the Weibo data published by the official account of CCTV News from Jan. 20, 2020 to Feb. 20, 2020, a total of 977 microblogs related to COVID-19 were captured and passed through the manual interpretation method to divide the body content of these microblogs into eight categories, as shown in Table 1. Because not all the topics are covered by the microblogs published every day, and if there are differences in the distribution of the number of microblogs for each type of topic, the ability to respond to the shift of public opinion focus will worsen. Therefore, it is necessary to select the microblog with a closer release date to compare all types of Weibo with their sliding averages. According to statistics, the three types of microblogs – "prevention policies and measures", "epidemic progress" and "official responses and actions" – have the most days and similar dates, while other types of microblogs have fewer days and large differences in the dates. We thus select only three types of Weibo to calculate the sliding average of the number of reposts, and the sliding average is shown in Fig. 3. Among these types, the abscissa represents the ordinal number of days starting from Jan. 20, 2020, and the ordinate represents the degree of public attention given to microblogs on different topics.

Based on Fig. 3, the following conclusions can be drawn. First, the sliding averages of the three types of Weibo reposts moved upward and then downward, which shows that the public's attention to these three types of events is not static. Second, the public pays different amounts of attention to the three types of events in different time periods. As shown in Fig. 3, the intersection of the sliding averages of the three types of Weibo reposts roughly divides the period from Jan.20, 2020 to Feb.20, 2020 into five time phases, specifically, Jan.20,2020-Jan.25,2020 is Phase 1; Jan.26, 2020-Jan.28, 2020 is Phase 2; Jan.29, 2020-Feb.3, 2020 is Phase 3; Feb.4, 2020-Feb.9, 2020 is Phase 4; Feb.10, 2020 to Feb.20, 2020 is Phase 5. This result shows that the focus of network public opinion on these three types of events shifted in different time periods. Specifically, in Phase 1, the public's attention followed these trends: epidemic > prevention policies and measures > official responses and actions. In Phase 2, the public's attention followed these trends: epidemic > official responses, and actions > policies and measures for prevention and control. In Phase 3, the public's attention followed these trends: official response and action > epidemic progress > prevention policies and measures. In Phase 4, the public's attention followed these trends: epidemic progress>control policies and measures > official responses and actions. In Phase 5, the public's attention followed these

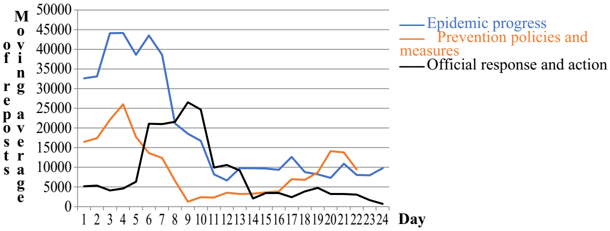

Fig. 3 Trend of the number of Weibo reposts over time

trends: prevention policies and measures > epidemic progress > official responses and actions. In the early stage of public opinion development, the public primarily focused on "epidemic progress" and "official responses and actions", while later, the focus shifted to "epidemic progress" and "prevention policies and measures." The reason for these results can be attributed to the epidemic being concentrated with many suspected cases in Wuhan in the early stage of the epidemic, where the new coronavirus was identified as easily transmitted and causing great harm. Information on the progress of the epidemic, relevant knowledge, and prevention measures were summarized by the government. As the epidemic rapidly spread across the country, affecting transportation, the economy, education and other aspects of life, the government continuously used various policies and measures to prevent and control the epidemic. Therefore, the public's attention shifted accordingly.

Additionally, the weighted moving averages of the number of forward shifts of the three types of microblogs in the five stages were calculated. The weighted average of "epidemic progress" was 19,109.80, that of "prevention policies and measures" was 9817.16, and that of "action" was 8426.98. These results show that from Jan. 20, 2020, to Feb. 20, 2020, the public's attention followed the progression of epidemic > prevention policies and measures > official responses and actions. These results can be attributed to the impact of the epidemic on the public throughout the epidemic, which made the public continue to consider the progress of the epidemic. Additionally, the prevention policies and measures were closely related to everyone's daily life, and the public is typically concerned about these types of topics. However, the official responses and actions also involve issues that are distant from the public's awareness, such as diplomacy and official accountability, and the public's attention to them is inconsistent. Thus, it is believed that during a major epidemic, the focus of network public opinion changes over time. The government thus must adjust the direction of public opinion guidance in different stages of the development of network public opinion and rationally allocate the information management and control resources according to the public's attention to each type of event. Taking COVID-19 as an example, the overall focus of network public opinion governance should have been "epidemic progress", followed by "prevention policies and measures" and then "official responses and actions". In terms of the different stages, the focus of network public opinion governance in the early stage of the epidemic should have been "epidemic progress" and "official responses and actions", while in the early stage of the epidemic, the focus of network public opinion governance should have shifted to "epidemic progress" and "prevention policies and measures".

4.2 Internet public opinion focus issues at different times during the epidemic

The content of comments on Weibo can accurately describe issues of public concern. Therefore, this study applies a word co-occurrence network analysis to the content of comments on Weibo to clarify the focus of public attention at different times during the epidemic. The specific method randomly selects one day in each of the five stages and then randomly selects one Weibo on each of the three themes of "prevention policies and measures", "epidemic progress" and "official response and actions" released on that day. This process gathers all the relevant comment content and then allows for a word cooccurrence network analysis to be performed on each Weibo. Word co-occurrence network analysis can describe the second-order distribution characteristics between the keywords and uses a modular observation standard. The effective value of modularity is typically between 0.3–0.7, where the larger the value, the clearer the community structure. After these steps, dates were randomly selected from Jan. 21, 2020, Jan. 27, 2020, Jan. 29, 2020, Feb. 5, 2020, and Feb. 14, 2020. Two Weibo comments were controlled, and their comments could not be obtained; thus, a total of 13 Weibo comments were finally obtained. Then, word co-occurrence network analysis was performed on these 13 microblogs, and 13-word co-occurrence network diagrams were drawn. Keywords detected as the same event in each diagram were rendered with the same color, and then each type was rendered, and the events were classified. Due to space reasons, only the most modular word cooccurrence network diagram and its event classification are shown in this study, as shown in Fig. 4. Finally, the word co-occurrence network analysis results of 13 Weibo comments are shown in Table 2.

The following explains the results of the word co-occurrence network analysis. First, according to the modularity values in Table 2, the overall modularity of the review content tends to increase over time, and the larger the modularity value is, the clearer the community structure. During the epidemic, the issues of public concern gradually became clearer and more concentrated over time, because as the epidemic continued to ferment, various problems became increasingly prominent; thus, the public's attention to various problems gradually became clearer and more concentrated. Second, via inductive analysis with the 13-word co-occurrence network diagrams, the content of "comments involving problem classification" in Table 2 (i.e., the types of problems that the public is concerned about) are determined. Table 2 shows that the classification of the question involved in the comment is not completely consistent with the classification of the original blog. Although the content of Weibo only involves a certain type of issue, public comments are not limited to discussing the content of the extracted Weibo but may discuss other types of issues because different types of issues are

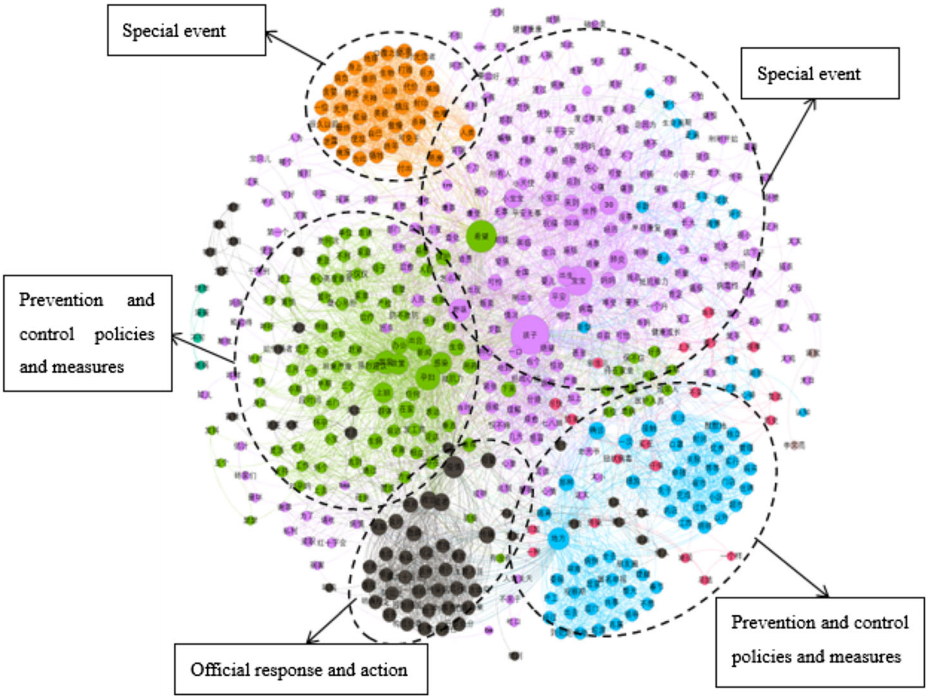
Special event
Special event
Prevention and control policies and measures
Official response and action
Prevention and control policies and measures

Fig. 4 Word co-occurrence network diagram (partial)

Table 2 Word co-occurrence network analysis results of Weibo comment content

Period	Weibo Date	Original blog classification	Comments related to problem classification	Modularity
Stage one	January 21, 2020	Prevention and control policies and measures	Prevention and control policies and measures Popular science	0.468
		Official response and action	Official response and action Frontline personnel situation Popular science Prevention and control policies and measures	0.468
		Epidemic progress	Prevention and control policies and measures Popular science	0.403
Stage two	January 27, 2020	Prevention and control policies and measures	Prevention and control policies and measures Supplies	0.421
		Epidemic progress	Epidemic progress Supplies	0.408
Stage three	January 29, 2020	Prevention and control policies and measures	Prevention and control policies and measures Popular science	0.408
		Official response and action	Prevention and control policies and measures Official response and action	0.576
		Epidemic progress	Supplies Epidemic progress Prevention and control policies and measures	0.555
Stage four	February 5, 2020	Official response and action	Supplies Prevention and control policies and measures Epidemic progress	0.576
		Epidemic progress	Special event Official response and action Prevention and control policies and measures	0.595
		Prevention and control policies and measures	Special event Epidemic progress Prevention and control policies and measures	0.442
Stage five	February 14, 2020	Official response and action	Prevention and control policies and measures Epidemic progress Frontline personnel situation Treatment measures	0.568
		Epidemic progress	Official response and action Frontline personnel situation Supplies Epidemic progress	0.498

often related and have a large impact on the public. Third, the issues in the comment contents in different periods were different; thus, the focus of public opinion was different.

Through the analysis of the keyword co-occurrence network, combined with the real the COVID-19 epidemic situation, the public did not know about COVID-19 in Phase 1, and the epidemic developed rapidly and was harmful. Therefore, in Phase 1, the public primarily considered COVID-19, the relevant knowledge and the corresponding prevention policies and measures. The content of the comments primarily related to “Relevant departments use text messages, broadcasts, news and other indicators to promote knowledge of the epidemic”, “Out-of-office protection measures”, “Comparison of COVID-19 and SARS”, “People who are vulnerable to infection”, “The cost of treatment is paid by the state”, “Refund policy” and other issues. In Phase 2, the public began to buy protective equipment such as masks and disinfectants after they had a certain understanding of COVID-19. Hubei became the hardest-hit area, and the number of confirmed diagnoses in various provinces soared. These problems led to medical treatments and a shortage of supplies. Then, many domestic and foreign institutions, enterprises and people from all walks of life donated money and materials to the disaster area. Concurrently, poor merchants sold inferior and high-priced protective products to make a fortune. Therefore, in Phase 2, issues such as the supply and demand of materials, quality, price and distribution attracted much attention. Comment contents were thus primarily related to issues such as “resumption of work and school opening”, “not sufficient hospital beds”, “cannot buy masks” and “lack of supplies in hospitals”. In Phase 3, due to the end of the Spring Festival holiday, many migrant workers were required to return to work. Thus, it was important to prevent and control their return trips and resumption of work. Therefore, in Phase 3, the public was more concerned about prevention and control policies and measures. Comment contents primarily focused on issues such as “protective measures on the way back to work”, “protective measures in communities”, “protective measures in public places” and “working methods after resumption of work”. In Phase 4, the epidemic was likely to have a stronger impact on economic development and was related to the livelihood of many people because the epidemic had not been controlled. Additionally, various new protective measures were continuously introduced, causing the public to pay more attention to the progress of the epidemic and the related prevention policies and measures. During Phase 4, a series of special incidents also occurred, which aroused widespread public concern. Therefore, the public was primarily concerned with the progress of the epidemic, prevention policies and measures and special events. Comment contents primarily related to “the problem of seeing a doctor for nonpneumonia patients during the epidemic”, “the unified distribution of drugs online and offline”, “the impact of the epidemic on the economy”, “Yunnan hijacks Chongqing supplies”, “Red Cross negligence”, “New-born babies are diagnosed” and other issues. In Phase 5, the medical staffs and other front-line personnel worked hard, and the risk of infection was high. As the number of infections and deaths of front-line personnel increased, the public’s situation in relation to front-line personnel received more attention. Comment contents primarily involved issues such as “recovery of e-commerce”, “allocation of medical resources”, “conditions of front-line personnel”, “integrated Chinese and Western medicine treatment”, “Jingzhou violation incidents” and “implementation of interprovincial assistance”.

Based on these results, the issues of public concern were unclear, more scattered, and fewer in the early stage of the epidemic. As the epidemic developed, issues of public concern gradually became clearer and more concentrated, the focus of network public opinion gradually changed, and the types of problems of concern also increased and diversified. Therefore, it is more efficient to conduct network public opinion control in the middle and late stages of an epidemic than in the early stage. The government should gradually strengthen the control of network public opinion as an epidemic develops. At different stages, the focus of public opinion on the internet is different, and the government should precisely manage problems that arise at each stage. Additionally, the emergence of the focus of network public opinion is consistent with the development of the epidemic. The government can predict the focus of network public opinion based on the development of the epidemic and make a network public opinion governance plan before the focus is formed in order to channel network public opinion effectively and stabilize social order.

5 Discussion

In sudden and major crisis events, social media plays an important role, can quickly convey official and key event information and can provide basic information for decision-making [26]. However, due to the large number of users of social media platforms and independent user information, there are occurrences of misleading and unreliable information [27], which can easily trigger a network public opinion crisis. Weibo is a social software platform that is frequently used by Chinese citizens. The quality of its information dissemination plays an important role in the supervision of government network public opinion, and the focus of public attention is different in different periods. In the early stage of the epidemic, the public paid more attention to the issue of "epidemic progress"; thus, the public felt panicked. Driven by their panic, the public actively looked for ways to vent their negative emotions, such as disseminating inappropriate public health information on the internet [28, 29]. In the later stage of an epidemic, the public was more concerned about the government's prevention and control measures and actions related to the epidemic. If the government does not implement effective measures, the public may lose control of their emotions, causing more serious problems related to network public opinion [11]. Taking the outbreak of COVID-19 as an example and comments on Sina Microblog as the main data source, this study conducted text analysis research on the focus events of network public opinion during the outbreak of the epidemic, established a new process of public opinion recognition and analysis, analyzed the focus of network public opinion in different periods and provided suggestions for the government to monitor network public opinion. The results of this study identify several important implications.

First, during the period of public opinion development, the public paid the most attention to the issue of "epidemic progress", followed by the issue of "prevention and control policies and measures" and finally the issue of "official response and action". Accordingly, during the epidemic period, the primary task of the government should be to timely disclose the real progress of the epidemic to the public. In the COVID-19 epidemic, a comparison of the sliding averages of the number of reposts on Weibo related to the three types of issues ("epidemic progress", "prevention policies and measures" and "official responses and actions") shows that in the early stage of the development of network public opinion, the public was primarily concerned about "epidemic progress" and "official responses and actions", while in the later stages of development, the focus shifted to "epidemic progress" and "prevention policies and measures." Additionally, overall, the public's attention toward the three types of issues differed. Throughout the study period, the public paid the most attention to the issue of "epidemic progress", followed by the issue of "prevention policies and measures" and finally the issue of "official responses and actions". Therefore, the government must flexibly adjust the direction of network public opinion governance at each stage according to the focus of network public opinion at each stage and rationally allocate network public opinion control resources according to the public's degree of attention to each type of event. Previous studies have also found that in public health events, the public pays attention to different focal issues at different stages, which is consistent with the research in this article [2].

Second, the fermentation of the epidemic takes time, and the government should focus on the middle and late stages of the epidemic. At the early stage of the development of the COVID-19 epidemic, the problems of public concern were not clear, the types of problems were scattered, and the number of types was small. As the epidemic developed, the issues of public concern gradually became clear and concentrated, the focus of network public opinion gradually formed, the variety of problems increased, and the focus was diversified. Therefore, the government should not focus on the governance of network public opinion in the early stage of the epidemic but should gradually strengthen the governance of network public opinion as the epidemic develops to improve governance efficiency. In addition, in the middle and late stages of the epidemic, the public will pay attention to a series of social events triggered by the epidemic in addition to the problems of the epidemic itself. It can be seen from the cluster diagram that these special events received high levels of public attention. Therefore, the government should pay attention to the handling of special events to prevent the expansion of the impact of negative events.

Third, the emergence of the network public opinion focus is consistent with the development of the epidemic. The government can predict the network public opinion focus according to the epidemic's development and make the network public opinion governance plan before the focus is formed to effectively dredge the network public opinion and stabilize the social order. In addition, COVID-19 is still occurring in the world, and public opinions are still forming. The current development of public opinion can also test the correctness and effectiveness of the public opinion identification and analysis methods proposed in this paper. For example, in the COVID-19 incident that broke out again in Shanghai, China in 2022, public opinion developed very rapidly, and its focus still showed a change trend of "epidemic progress" - "official response and action" - "prevention and control policies and measures". In the middle and late stages of the epidemic, various social events occurred frequently and public opinion continued to ferment, which caused great public pressure on the government. The government made various countermeasures to stabilize social order, which also proved that the government should focus on the middle and late stages of the epidemic. Similarly, for other countries in the world, the development trend of the public opinion focus of the COVID-19 is also consistent with that of China. These examples prove the validity of the proposed model again. These examples prove the validity of the proposed model again. In addition, through the analysis of public opinion in each stage, we can also identify and summarize certain rules and provide effective suggestions for public opinion prediction and governance in the future.

At the same time, there are some limitations in this study, and need to be further improved in future research. In terms of research methodology, this research only focused on combining existing methods and technologies to create a new public opinion analysis process, though it can effectively deal with the identification and analysis of public opinion, but lacking innovation and optimization of algorithms. Future research should pay more attention to algorithm innovation and breakthrough to make more contributions to research in the computer field.

As for research data selection, on the one hand, we only used the comment contents on the mainstream social media platform in China-Sina Microblog, and they are all in text type. So, in the future, it is necessary to select the data from more media platforms with different types, such as image content, video content, etc., to enrich the diversity of data. On the other hand additional types of major social events can be selected as research samples to verify and improve this study's method to enhance the universality of the method.

Finally, the manual interpretation method is used to summarize and classify the microblog text content and comment content, which may lead to deviations in the research results. In the future, it will be important to find more rigorous methods to reduce the existing deviations as much as possible.

Funding National Natural Science Foundation of China (Grant No. 71662004, 71872055); Key Research Base of Humanities and Social Sciences in Guangxi Universities and the Project of Guangxi Development Strategy Institute in China (Grant No. 2022GDSIYB14; 2021GDSIYB05); Guangxi philosophy and Social Science Planning Research Project (Grant No. 21FMZ050); Innovation Project of Guangxi Graduate Education (Grant No. JGY2022018).

Data availability The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Declarations

Conflict of interest We declare that we do not have any commercial or associative interest that represents a conflict of interest in connection with the work submitted.

References

1. An L, Yu C, Lin X, Du TY et al (2018) Topical evolution patterns and temporal trends of microblogs on public health emergencies: An exploratory study of Ebola on Twitter and Weibo. Online Inf Rev 42(6):821–846. https://doi.org/10.1108/OIR-04-2016-0100
2. Ataa Allah F, Grosky WI, Aboutajdine D (2007) On-line single-pass clustering based on diffusion maps. In: International Conference on Application of Natural Language to Information Systems 107–118. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-73,351-5_10
3. Bora K, Das D, Barman B, Borah P (2018) Are internet videos useful sources of information during global public health emergencies? A case study of YouTube videos during the 2015–16 Zika virus pandemic. Pathog Global Health 112:320–328. https://doi.org/10.1080/20,477,724.2018.1507784
4. Chen S, Xu Q, Buchenberger J et al (2018) Dynamics of Health Agency Response and Public Engagement in Public Health Emergency: A Case Study of CDC Tweeting Patterns During the 2016 Zika Epidemic. JMIR Public Health Surveill 4:e10827. https://doi.org/10.2196/10827
5. Chua AYK, Banerjee S (2018) Intentions to Trust and Share Online Health Rumors: An Experiment with Medical Professionals. Comput Hum Behav 87:1–9. https://doi.org/10.1016/j.chb.2018.05.021
6. Cui JH, Zhu F (2021) Analysis of Patent themes and Hot spots for the COVID-19 epidemic in China. J Modern Inf 41(11):161–169. https://doi.org/10.3969/j.issn.1008-0821.2021.11.016
7. Gil-García R, Pons-Porrata A (2010) Dynamic hierarchical algorithms for document clustering. Pattern Recogn Lett 31(6):469–477. https://doi.org/10.1016/j.patrec.2009.11.011
8. Hadi TA, Fleshler K (2016) Integrating Social Media Monitoring into Public Health Emergency Response Operations. Disaster Med Public Health Prep 10:775–780. https://doi.org/10.1017/dmp.2016.39
9. He H, Zhu N, Lyu B, Zhai SB (2023) Relationship between nurses' psychological capital and satisfaction of elderly cancer patients during the COVID-19 pandemic. Front Psychol 13:1,121,636. https://doi.org/10.3389/fpsyg.2023.1121636
10. Hong W, Shi M, Hong XJ, Pu XJ (2016) Factors Affecting the Netizen's Microblog Retweet Behavior in Food Safety Internet Public Sentiment: The Case of Shanghai Husi Incident. China Pop Resource Env 26(5):167–176. https://doi.org/10.3969/j.issn.1.002-2104.2016.05.021
11. Huang G, Li Y, Wang Q et al (2019) Automatic classification method for software vulnerability based on deep neural network. IEEE Access 2019(1):1. https://doi.org/10.1109/ACCESS.2019.2900462
12. Lee JY, Jo WK, Chun HH (2015) Long-Term Trends in Visibility and Its Relationship with Mortality, Air-Quality Index, and Meteorological Factors in Selected Areas of Korea. Aerosol Air Qual Res 15:673–681. https://doi.org/10.4209/aaqr.2014.02.0036

13. Li Y, Teng YCH (2021) Government Network Public Opinion Governance Integration and Government Information Synergy Measurement. Inf Sci 39(12):113–117. https://doi.org/10.13833/j.issn.1007-7634.2021.12.017
14. Liu Y, Wang W, Shang MS, Tang M (2016) The spread of epidemic and public opinion on complex networks and its immune-based control strategy. Complex Syst Complex Sci 13:74–83. https://doi.org/10.13306/j.1672-3813.2016.01.007
15. Ma YP, Shu XM, Shen SF et al (2014) Study on Network Public Opinion Dissemination and Coping Strategies in Large Fire Disasters. Procedia Eng 71:616–621. https://doi.org/10.1016/j.proeng.2014.04.088
16. Matheson C, Jaffray M, Ryan M, Bond CM et al (2014) Public opinion of drug treatment policy: Exploring the public's attitudes, knowledge, experience and willingness to pay for drug treatment strategies. Int J Drug Policy 25:407–415. https://doi.org/10.1016/j.drugpo.2013.11.001
17. Mccauley M, Minsky S, Viswanath K (2013) The H1N1 pandemic: media frames, stigmatization and coping. BMC Public Health 13:1116. https://doi.org/10.1186/1471-2458-13-1116
18. Qi K, Yang Z (2020) Multiscenario Evolutionary Game Analysis of Network Public Opinion Governance in Sudden Crisis. Chin J Manag Sci 28(3):59–70. https://doi.org/10.16381/j.cnki.isn1003-207x.2020.03.007
19. Schultz F, Utz S, Göritz A (2011) Is the medium the message? Perceptions of and reactions to crisis communication via twitter, blogs and traditional media. Public Relat Rev 37(1):20–27. https://doi.org/10.1016/j.pubrev.2010.12.001
20. Seltzer EK, Jean NS, Kramer-Golinkoff E et al (2015) The content of social media's shared images about Ebola: a retrospective study. Public Health 129:1273–1277. https://doi.org/10.1016/j.puhe.2015.07.025
21. Signorini A, Segre AM, Polgreen PM (2011) The use of Twitter to track levels of disease activity and public concern in the U.S. during the influenza A H1N1 pandemic. PLoS One 6:e19467. https://doi.org/10.1371/journal.pone.0019467
22. Toçoğlu, MA, Onan, A (2021) Sentiment analysis on students' evaluation of higher educational institutions. In Intelligent and Fuzzy Techniques: Smart and Innovative Solutions: Proceedings of the INFUS 2020 Conference, Istanbul, Turkey, July 21–23, 2020 (pp. 1693–1700). Springer International Publishing. https://doi.org/10.1007/978-3-030-51,156-2_197
23. Trieschnigg, D, Kraaij, W (2004) TNO Hierarchical topic detection report at TDT 2004. In Topic Detection and Tracking Workshop Report. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.418.2165&rep=rep1&type=pdf
24. Wang G, Chi Y, Liu Y, Wang Y (2019) Studies on a multidimensional public opinion network model and its topic detection algorithm. Inf Process Manag 56(3):584–608. https://doi.org/10.1016/j.ipm.2018.11.010
25. Xiao S, Tong W (2021) Prediction of user consumption behavior data based on the combined model of TF-IDF and logistic regression. J Phys Conf Ser 1757(1):012089. IOP Publishing. https://doi.org/10.1088/1742-6596/1757/1/012089
26. Xiao, J, Yang, Z, Li, Z, Chen, Z (2022) A review of social roles in green consumer behaviour. Int J Consum Stud, 1–38. https://doi.org/10.1111/ijcs.12865
27. Yang L, Lin H, Lin Y, Liu S (2016) Detection and extraction of hot topics on chinese microblogs. Cogn Comput 8(4):577–586. https://doi.org/10.1007/s12559-015-9380-6
28. Yu LAN, Li L, Dai W, Tang L (2016) Crisis Information Release Policy and Online Public Opinion Dissemination in Emergency of Hazardous Chemicals Leakage into River: A Multiagent-based Model. Manag Rev 28(8):175–185. https://doi.org/10.14120/j.cnki.cn11-5057/f.2016.08.022
29. Zhang YF, Li H, Peng LH, Chen YF (2017) An Empirical Research on Monitoring and Early Warning of Internet Public Opinion Based on Fuzzy Inference of Semantic Membership Degree. Inf Theory Practice 40: 82–89. https://doi.org/10.16353/j.cnki.1000-7490.2017.09.016

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.