Research on the evolution and prediction of Internet public opinion of major pandemics—Taking the COVID-19 pandemic as an example

Jing Li¹, Huimin Tang², Hongjuan Tan³

¹,²,³School of Management, Xinhua College of Sun Yat-sen University, Guangzhou Guangdong, China, 510520

Corresponding author and e-mail: Huimin Tang, mdtanghuimin@xhsysu.edu.cn

Abstract. To clarify the evolutionary law of major epidemics online public opinion is a prerequisite for targeted public opinion management. This article takes the Novel Coronavirus Pneumonias as an example and proposes a method to predict the hot spots of public opinion in epidemic. We firstly used the python web crawler technology to capture 58916 online public opinion data on the Novel Coronavirus Pneumonias of the Forum on ‘Making the Country Prosperous’ of People's Daily Online from January to March 2020, and then used the jieba thesaurus for word segmentation and keyword extraction, and performed word frequency statistics to show the hot content of the new crown epidemic network public opinion, then through the keyword association rule mining to find out the hot topics, finally through the information entropy to evaluate the hot topics, and discover potential hot topics in the future.

1. Introduction
Network public sentiment refers to the mapping of social public opinion in the Internet space formed by posting comments and opinions on online platforms such as Weibo, WeChat, and BBS forums [1]. When public sentiment develops to a certain extent which will be transformed into public consensus. The fierce language caused by public sentiment events in the society is likely to have a negative impact on netizens’ values and perceptions. Negative online public sentiment will increase the difficulty of social management. When network incidents harm the government’s image, which will have negative influence on the China’s government [2]. Therefore, the management of online public sentiment is of great significance for understanding the dynamics of people's thinking, promoting the scientific and transparent public decision-making, and maintaining social stability.

2. Major epidemics online public sentiment
At present, the research of online public sentiment at home and abroad mainly analyzes short-term emergencies, and divides the development of online public sentiment into budding-outbreak-climax-decay-disappearance [3]. The whole development cycle of online public sentiment generally ends in a relatively short period of time, which is Different with general social emergencies, major epidemics have the characteristics of strong outbreaks, strong uncertainties, difficult to control, and long response periods [4]. Correspondingly, there is a big difference between the online public sentiment of major epidemics and common social hot events, which requires special research, in-depth
3. Analysis

This article uses python web crawler technology to grab 58916 topics related to the novel coronavirus pneumonia from January 22 to March 10, 2020 from the People's Daily Online Forum. To analyze the evolution of online public sentiment, we use the jieba thesaurus first to segment topic posts and extract keywords, then perform word frequency statistics on keywords and build a high-frequency word cloud to show the hot content of the novel coronavirus pneumonia online public sentiment. We analyzed in ten-day units, divided into five time periods: late January, early February, mid-February, late February, and early March, corresponding to the construction of five keywords word clouds. Since keywords such as "epidemic" and "prevention and control" appear more frequently than others, in order to better display the content of the Network public sentiment of the epidemic, they are removed from the word cloud. The five keywords word clouds are shown in Figure 1 to Figure 5. The larger the keyword font in the figure, the higher the frequency appear. The frequency ranking of keywords from month to March is shown in Table 1 to Table 5.

**Table 1. Keywords frequency ranking in late January.**

| Rank | Keywords       |
|------|----------------|
| 1    | pneumonia      |
| 2    | coronavirus    |
| 3    | new style      |
| 4    | Wuhan          |
| 5    | win            |
| 6    | blockade       |
| 7    | Infection      |
| 8    | China          |
| 9    | guarantee      |
| 10   | anti-epidemic  |

**Figure 1.** Keywords word cloud in late January.
Table 2. Keywords frequency ranking in late February.

| Rank | Keywords              |
|------|-----------------------|
| 1    | blockade              |
| 2    | win                   |
| 3    | anti-epidemic         |
| 4    | pneumonia             |
| 5    | China                 |
| 6    | new style             |
| 7    | party member          |
| 8    | Wuhan                 |
| 9    | coronavirus           |
| 10   | assist                |

Figure 2. Keywords word cloud in late February.

Table 3. Keywords frequency ranking in mid-February.

| Rank | Key words              |
|------|------------------------|
| 1    | anti-epidemic          |
| 2    | blockade               |
| 3    | win                    |
| 4    | pneumonia              |
| 5    | China                  |
| 6    | coronavirus            |
| 7    | resume work            |
| 8    | Wuhan                  |
| 9    | enterprise             |
| 10   | assist                 |

Figure 3. Keywords word cloud in mid-February.

Table 4. Keywords frequency ranking in late February.

| Rank | Keywords        |
|------|-----------------|
| 1    | anti-epidemic   |
| 2    | win             |
| 3    | blockade        |
| 4    | China           |
| 5    | supervise       |
| 6    | pneumonia       |
| 7    | coronavirus     |
| 8    | resume work     |
| 9    | precise         |
| 10   | develop         |

Figure 4. Keywords word cloud in late February.
Table 5. Keywords frequency ranking in early March.

| Rank | Keywords             |
|------|----------------------|
| 1    | anti-epidemic        |
| 2    | China                |
| 3    | coronavirus          |
| 4    | pneumonia            |
| 5    | american             |
| 6    | resume work          |
| 7    | develop              |
| 8    | WuHan                |
| 9    | science              |
| 10   | party member         |

Figure 5. Keywords word cloud in early March.

The word cloud in Figure 1 and the word frequency ranking in Table 1 reflect the early stage of the outbreak of the Novel Coronavirus Pneumonias (Late January) The focus of public attention. It can be seen from the figure that in the early stage of the epidemic, in addition to the themes of fighting the epidemic and winning the battle against the epidemic, the knowledge of the Novel Coronavirus Pneumonias also attracted great attention from netizens. In addition, the state's deployment and response measures, the dispatch and guarantee of medical supplies, and the support and donations from all walks of life are also hot topics discussed by netizens.

The word frequency ranking in Figure 2 and Table 2 reflects the hot content of the epidemic public opinion in early February. Seeing from Table 1 that the content of the keywords is similar to Figure 1. Topics such as how to fight the epidemic and how party members play their roles have received widespread public attention.

Figure 3 and 4 and Table 3 and 4 reflect the hot topics of public opinion about the Novel Coronavirus Pneumonias epidemic in mid-to-late February. Entering mid-February, my country's epidemic control has achieved initial results, and the number of newly diagnosed patients per day has been declining, entering a three-digit range. It can be seen from the chart that in addition to the themes of fighting the epidemic, not forgetting the original aspirations of party members and cadres, and practicing missions, the overall promotion of the resumption of work, production as well as the gradual restoration of economic, social order have become the new focus of public attention.

The word cloud of Figure 5 and the word frequency ranking of Table 5 reflect the hot topics of the epidemic that netizens paid attention to in early March. Entering the first ten days of March, the domestic epidemic was basically under control, and the number of newly diagnosed people every day entered a double-digit range. It can be seen from Figure1 that similar to the situation in mid-to-late February, topics such as domestic fight against the epidemic and resumption of work and production are hot discussion among netizens. In addition, epidemics in foreign country are gradually breaking out has also aroused public concern.

Based on the analysis of the above 5 word cloud contents, it can be seen that, unlike the general online public opinion hotspots that fade in a short time, the entire process of major epidemics with the development of the epidemic has the following three characteristics:

First one is the continuity of the theme. From the keyword cloud of Figures 1 to 5, it can be seen that the subject terms such as anti-epidemic and winning the battle are prominent in the five word clouds, indicating that these themes have been throughout all stages of the development of the epidemic. Hot topics of public concern and discussion. In addition, the performance and responsibility of party members and cadres has always been a topic of discussion among netizens. All this shows that the online public opinion of the major epidemic is continuous in the theme.
The second is the diversity of content. The diversity of content refers to the richness of public opinion about the epidemic. Judging from the content of the five word clouds, netizens discussed many topics during the epidemic, from anti-epidemic measures to front-line situations, from material supplies to social donations, from leadership dynamics to people’s lives, from resumption of work and production to economic and social development, from the responsibilities of party members and cadres to the role of volunteers, from domestic to foreign conditions etc, which involves all aspects of fighting the epidemic.

The third is the complexity of response. On the one hand, the diversified content of the epidemic public opinion needs to be dealt with in a targeted manner, making the response to the epidemic public opinion more complicated than a single public opinion hot event. On the other hand, some sudden individual hot spots during the development of the epidemic have received a lot of attention in a short period of time, formed a large scale, and increased the uncertainty of public opinion response. For example, in the Dr. Li Wenliang event that occurred in early February, the number of readings and comments on related topic posts far exceeded the average number of forums, becoming a focus topic, causing a huge impact on the Internet, and posing greater challenges to public opinion response and guidance.

4. Hot spot forecast of internet public opinion on major epidemic

Major epidemics are the focus of public attention, covering a wide range of areas and having great influence. Therefore, it is necessary and important to predict online public opinion hotspots of major epidemics and take corresponding measures to guide them. The above analysis has pointed out that the hot topics of online public opinion on major epidemics are continuous. In other words, certain themes in one stage are still hot topics of public concern in the next stage. This provides us with ideas for predicting the hot spots of online public opinion. We take the topic posts in late January as an example, and find out the hot topics by mining the association rules of keywords. Association rules are one of the main methods of data mining, used to discover the correlation between things from a large amount of data. The results of association rule analysis generally appear in the form of rules that include relevant items, which appear at the same time with a relatively high frequency. Therefore, association rules can be used to mine keyword combinations with a higher frequency of co-occurrence. Commonly used indicators to judge the effectiveness of association rules include support, confidence, and promotion [6]. Table 6 illustrates these three indicators.

**Table 6.** Common indicators of association rules.

| Rules     | Description                                      | Formula                                      |
|-----------|--------------------------------------------------|----------------------------------------------|
| Support   | Probability of item X and item Y appearing at the same time | \( \text{Support}(X \rightarrow Y) = P(X \cup Y) \) |
| Confidence| Probability of project Y is included in a transaction that includes project X | \( \text{Confidence}(X \rightarrow Y) = P(Y | X) \) |
| Promotion | The degree of influence of the emergence of project X on the emergence of project Y | \( \text{lift}(X \rightarrow Y) = \frac{P(Y | X)}{P(Y)} \) |

**Table 7.** Association rule mining results.

| Rules                                      | Support(%) | Confidence(%) | Promotion |
|--------------------------------------------|------------|---------------|-----------|
| Original intention => mission              | 0.54       | 79.49         | 111.36    |
| Flying high => party flag                  | 0.85       | 100.00        | 77.62     |
| finical => service                        | 0.68       | 100.00        | 42.24     |
| Discipline Inspection and Supervision => cadre | 0.56       | 78.05         | 26.37     |
| contain => spread                         | 0.92       | 75.71         | 45.30     |
| win => blockade                           | 5.34       | 72.41         | 9.97      |

5
Due to the large number of topic posts, we set the threshold of support at 0.5% and the threshold of confidence at 60% to ensure that the rules obtained are credible. The results of mining topic posts in late January using association rules are shown in Table 7 (rules with the same content have been eliminated, and "novel coronavirus pneumonia" is used as a term in the analysis).

It can be seen from Table 7 that the lift of all the rules is bigger than 1, indicating that these rules are meaningful [7]. After using it to mine high-frequency keyword combinations, we evaluate the amount of information of these keyword combinations. This article introduces information entropy as an evaluation index. The concept of information entropy was proposed by Shannon and used to quantify information to reflect the complexity or richness of the information contained in the target object. The high information entropy of a keyword means that the information contained in the keyword has diversity [8]. The information entropy of the keyword $i$ is calculated by the following formula:

$$\text{Entropy} = -\sum_j p_{ji} \log_2 (p_{ji})$$

In the formula, $j$ represents a keyword different with $i$, and $p_{ji}$ represents the quotient of the number of topic posts containing both keywords $i$ and $j$ and the number of subject posts containing keyword $i$. The entropy of a keyword combination is measured by the average value of the entropy of each keyword contained in the combination, which is used to reflect the richness of the information contained in the combination. The greater the entropy of a keyword combination, the greater the amount of information contained in the combination, and the greater the value of continuing discussions. Therefore, this article uses entropy to filter out the keyword combinations with greater re-discussion value and regards them as hot spots that will continue to appear in the next stage. The entropy of the 11 association rules in Table 7 is shown in Table 8.

| Keyword combination | entropy     |
|---------------------|-------------|
| original intention, mission | 18.35       |
| Flying high, the party flag | 19.42       |
| finacal, service      | 11.73       |
| Discipline Inspection and Supervision, leader | 12.54       |
| contain, spread       | 15.57       |
| win, blockade         | 18.27       |
| novel coronavirus, infection | 17.84       |
| people, life, security | 15.38       |
| Party organization, party member | 14.72       |
| novel coronavirus, meeting, pneumonia | 17.68       |
| novel coronavirus, pneumonia, leader, infection | 16.09       |

The average value of the information entropy of the 11 keyword combinations in Table 8 is 16.14. We regard the five keyword combinations with higher information entropy as hot-spot combinations, that is, the theme reflected by these keyword combinations will remain in the next stage. It is a hot spot of public concern. In order to test the effectiveness of the analysis in this paper, we have counted the frequency of these 5 keyword combinations in early February. The results are shown in Table 9. It can be seen from Table 9 that the keyword combinations selected in this article still appeared with a higher frequency in early February. To win the fight against the epidemic, party members did not
forget their original aspirations and fulfilled their mission. Topics such as the situation and response to the new crown pneumonia epidemic are still hot topics of lively discussion on the Internet. This proves the effectiveness of the method proposed in this paper.

**Table 9.** Keywords combination frequency of occurrence.

| Keywords combination                      | Frequency of occurrence |
|------------------------------------------|-------------------------|
| original intention, mission              | 91                      |
| flying high, party flag                  | 99                      |
| win, blockade                            | 860                     |
| novel coronavirus, infection              | 305                     |
| novel coronavirus, meeting, pneumonia    | 0.57                    |

In order to test the robustness of the method proposed in this paper, we also use association rules to conduct keyword mining on the topic posts in late February, and verify with the topic posts in early March. The results of association rule mining are shown in Table 10.

**Table 10.** Results of keyword association rule mining in late February.

| Rules                          | Support(%) | Confidence(%) | Promotion | Information entropy |
|-------------------------------|------------|---------------|-----------|---------------------|
| people battle => people       | 0.54       | 100.00        | 34.51     | 19.86               |
| original intention => mission | 1.06       | 66.35         | 48.75     | 17.62               |
| practice => mission           | 0.54       | 80.46         | 59.12     | 13.55               |
| resume production => resume work | 3.07     | 97.57         | 22.43     | 17.43               |
| win => blockade                | 4.89       | 70.45         | 10.26     | 17.65               |
| pioneer, model function => party member | 0.51   | 95.71         | 17.44     | 12.99               |
| novel coronavirus, meeting => pneumonia | 0.57  | 98.68         | 17.35     | 22.22               |
| developement, overall planning => economic society | 1.12 | 87.50         | 40.15     | 18.69               |
| Accurate, scientific, prudent => supervise | 1.72 | 91.84         | 15.72     | 12.81               |

According to the above ideas, we counted the frequency of occurrence of keyword combinations that have an average information entropy in the first ten days of March. The results are shown in Table 11. It can be seen from Table 11 that the keyword combinations we selected in late February still appeared with a higher frequency in early March. To win the fight against the epidemic and the People’s War, Communist Party members did not forget their original aspirations, fulfilled their mission, resumed production and work, and restored social order. Topics about the restoration of social order remained the focus of netizens’ attention in early March. Therefore, the method proposed in this paper has stability.

**Table 11.** Keyword combination frequency in early March.

| Keywords combination                      | Frequency of occurrence |
|------------------------------------------|-------------------------|
| people battle, people                    | 48                      |
| original intention, mission              | 51                      |
| Resume production                         | 259                     |
| win, blockade                             | 209                     |
| novel coronavirus, meeting, pneumonia    | 30                      |
| developement, overall planning, economic society | 169               |
5. Conclusions
Taking the novel coronavirus epidemic as an example, this article conducts keyword extraction and text mining on the data of the People's Daily Online Power Forum from late January to early March 2020. On this basis, it analyzes the evolution of major epidemic online public opinion and discovers three characteristics of public opinion of major epidemic networks, just as following: the continuity of the theme, the diversity of the content and the complexity of response. Based on the continuity of the topic of online public opinion, this article introduces association rules and information entropy and proposes a method to predict the hot spots of online public opinion, which can provide a reference for the guidance and management of online public opinion for major epidemics.

The Internet public opinion of a major epidemic is the manifestation of public opinions and emotions on the Internet in a special period. In addition to the characteristics of general Internet public opinion such as immediacy, interaction, concealment, and emotionality [9], it also has its particularity. Based on the three important characteristics of the online public opinion of major epidemics discovered in this article, we propose the following response strategies: Firstly, based on the continuity of public opinion themes, it is necessary to unearth important themes at the beginning of the epidemic to "strengthen confidence, gather people's hearts, and warm people's hearts." In order to guide the ideology and make a good management plan, strengthen positive publicity of related topics on various online platforms such as Weibo, WeChat, online forums, news clients, etc., to guide the development of online public opinion in a positive direction, and form a good online atmosphere. Secondly, based on the diversity of public opinion content, it is necessary to establish a public opinion analysis mechanism to collect and analyze relevant public opinion information, fully grasp all kinds of information related to the epidemic, do a good job of information classification, and take targeted measures to guide different topics. Grasp the initiative to guide public opinion. Thirdly, based on the complexity of public opinion response and second point, it is necessary to establish a prediction and prediction mechanism for potential hotspots, combining artificial intelligence, 5G and other new-generation information technology construction based on large Data network public opinion analysis system [10], do well in real-time monitoring of network public opinion, discover hot spots in time, respond in time at the initial stage of the incident, follow the principles of openness and transparency, and seek truth from facts, do a good job in information disclosure, and respond to public opinions in a timely and appropriate manner, to deal with false rumors and avoid the expansion of negative effects.

Acknowledgement
Fund project: 2016 "Public Management" Construction Project of Key Discipline with Characteristics of Guangdong Province;

References
[1] Xuelin Yang, Research on prediction algorithm of network public opinion supervision based on big data [J]. (Modern Electronic Technology), 2017, 40(24):28-30.
[2] Ying Yu, Kainan, Liu, Tingting Yang etc. Research on prediction of network public opinion based on EMD.ARXG model [J]. (Modern Electronic Technology), 2020, 43(03):82-86.
[3] Mingde Li, Shengjun Meng, Hongbang Zhang. Research on Weibo Public Opinion Communication Mode
[4] Process-based analysis [J]. (Information Magazine), 2014, 33(02):120-127.
[5] Di Xu, Research on Internet Public Opinion Research and Judgment System for Major Epidemic Emergencies Based on Spatio-temporal Big Data [J]. (Information Magazine), 2020, 40(04):23-30+81.
[6] Ming Li, Leye Yao, Changes in public opinion response and guidance policies for severe natural disasters——Analysis based on the framework of the initiative alliance [J]. (Press World), 2018(09):81-87.
[7] Kim C, Lee H and Seol H, Lee C. Identifying core technologies based on technological cross-impacts: An association rule mining (ARM) and analytic network process (ANP) approach
[8] Zhou Y et al. Combining association rules mining with complex networks to monitor coupled risks [J]. (Reliability Engineering and System Safety) 2019, 186: 194-208.

[9] Chen Y S, Chang K C. Using the entropy-based patent measure to explore the influences of related and unrelated technological diversification upon technological competences and firm performance[J]. (Scientometrics) 2012, 90(3):825-841.

[10] Suzhen Zhou, Research on Strategies of Dealing with Network Public Opinions on Network Mass Events [J]. (Academic Forum), 2014, 37(04):112-116.

[11] Zhihua Zhan,. Internet Public Opinion Analysis System Based on Big Data [J]. (Modern Electronic Technology), 2017, 40(24):15-17.