Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Public attention about COVID-19 on social media: An investigation based on data mining and text analysis

Keke Hou a, Tingting Hou b,*, Lili Cai c

a School of Health Sciences, Xinhua College of Sun Yat-sen University, Guangzhou, China
b School of Information Technology and Management, University of International Business and Economics, Beijing, China
c School of Biomedical Engineering, Xinhua College of Sun Yat-sen University, Guangzhou, China

ARTICLE INFO

Keywords:
COVID-19
Public attention level
Topics analysis
Correlation analysis
Sentiment analysis

ABSTRACT

The COVID-19 epidemic is influencing global population. Social media has become important platforms to acquire and exchange information during the outbreak of COVID-19. This study explores public attention on social media. Popular Weibo texts related to COVID-19 with “coronavirus” and “pneumonia” as the keywords during December 27, 2019 and May 31, 2020 were collected in our study for public attention analysis. By combining data mining and text analysis, the public attention level trend in different stages were presented. Then a correlation analysis between public attention level and COVID-19 related cases number, topic analysis, and sentiment analysis were conducted. Significant positive correlation between public attention level and COVID-19 related cases number was identified. Based on Latent Dirichlet Allocation model, topic extraction was implemented in different stages and 41 topics were identified totally. For a comprehensive understanding of public emotions, sentiment analysis was performed. This study provides valuable lessons for public response to COVID-19.

1. Introduction

Clusters of pneumonia cases of unknown etiology emerged from 9 December 2019 in Wuhan City, Hubei Province, China (Jung et al., 2020) and were confirmed to be infected by a novel coronavirus. Because of the quick spread of COVID-19, Wuhan city was closed on 23 January, 2020 to avoid the outbreak of the novel coronavirus. Two hospitals (i.e., Huoshenshan hospital and Leishenshan hospital) were built within a short time to aim at the treatment of COVID-19 related cases. Other local governments also took measures to reduce population flow. With the containment measures implemented in China, the new cases have been reduced by more than 90% (Remuzzi & Remuzzi, 2020).

However, the spread of COVID-19 has become a global threat. As of September 16, 2020, coronavirus disease 2019 (COVID-19) has been reported with 29,444,198 cases, 931,321 deaths, covering 216 countries, areas or territories. The diffusion of COVID-19 has become unstoppable and is affecting the world’s population.

The spread of COVID-19 has changed people’s life and attracted great attention (Han, Wang, Zhang, & Wang, 2020). Except for the daily updates of confirmed cases and deaths released by official departments to improve public’s attention of prevention and protection (Bao, Sun, Meng, Shi, & Lu, 2020), publics and self-media also release opinion related to COVID-19 and attract a large number of retweeting, comments, and thump-up. Social media exposure has become an important channel for publics to actively express their opinions and attention about coronavirus disease.

With the support of mobile and web-based technologies, social media create high interaction among communities (Kietzmann, Hermkens, McCarthy, & Silvestre, 2011). Social media has become important platforms to acquire and exchange information in different situations, especially on an unprecedented scale during the outbreak of COVID-19 (L. Li et al., 2020). Existing research recognizes the critical role played by public attention in the prevention of coronavirus disease (Gao et al., 2020; Zhao, Cheng, Yu, & Xu, 2020). It is necessary to clarify the relationship between public concerns and the development of COVID-19. In addition, the change of public opinions could imply public participation that is crucial for the prevention of coronavirus disease (Han et al., 2020).

The development of data analysis in social network by virtue of machine learning and text mining enables analysis of public opinions in
social media with the ongoing break of COVID-19. Han et al. (2020) developed a topic extraction and classification model and identified seven topics and 13 sub-topics related to COVID-19 in the early stage. L. Li et al. (2020) used Weibo data and natural language processing techniques and identified seven categories of situational information related to COVID-19. Qin et al. (2020) used the lagged series of social media search indexes to predict new suspected COVID-19 case numbers. These studies provided the knowledge of public attention about COVID-19 on social media and technological feasibility of social media data mining and analysis.

This study aims to investigate the public attention on social media from December 27, 2019 to May 31, 2020 in China. According to the White paper of the Chinese People’s Government, China’s fight against COVID-19 was divided into five stages. Therefore the data was divided into five stages in our study. This study explored the trend of public attention level related to COVID-19 popular Weibo. Then a topic extraction and classification model was built to identify different topics of COVID-19 related Weibo texts in different stages. Furthermore, sentiment analysis using API sentiment orientation analysis interface developed by Baidu platform was performed.

2. Materials and methods

2.1. Data source and pre-processing

Sina Weibo (also referred to as Weibo) is the most popular microblogging site and the biggest portal Internet sites in China (Kim, Lee, Shin, & Yang, 2017), was selected to be the social media that we acquired the texts. Users can access Weibo through various mobile terminals, such as PC and mobile phone, and realize instant information sharing, communication and interaction in the form of multimedia such as text, pictures and videos. Weibo had more than 500 million monthly active users and 220 million daily active users in 2019. Popular Weibo texts is a list of texts through comprehensively calculating retweeting number, comments number and thump-up number of each text, combined with the frequency of retweeting and comments within a certain period of time. Popular Weibo texts pertain to the objective data which are calculated by the operator of Weibo. The popularity of Weibo texts can reflect the focus of attention and participation of users. Therefore, popular Weibo texts were acquired as the data source of this paper. The popular Weibo is the content flow calculated according to the user’s interest, the popularity of the microblog, the timeliness, and other dimensions, which represents the current problem widely concerned by netizens.

This study collected the popular Weibo texts related to COVID-19 with “coronavirus” and “pneumonia” as the keywords with timestamps from 00:00 on December 27, 2019 to 24:00 on May 31, 2020. The current study extracted the following information of the popular Weibo texts related to COVID-19: user ID, timestamp (post time of messages), text, the amount of comments, retweeting, and thump-up. A total of 153,303 popular Weibo texts were obtained.

The original microblog messages might include interfering information. Therefore, the data was pre-processed before analysis. First, some original popular Weibo texts label themselves related to COVID-19, but the actual content was irrelevant to COVID-19. After filtering and deleting this kind of data, 153,300 popular Weibo texts were finally obtained. Furthermore, the original microblog messages also contain punctuation marks, emoticons, whitespace, tags and other interfering information. Regular expressions operations in Python were used to eliminate the noise of the original data, improve the efficiency of word segmentation, and extract the topics.

2.2. Method

As discussed earlier, China’s fight against COVID-19 was divided into five stages. In the first stage (from December 27, 2019 to January 19, 2020), cases of pneumonia of unknown cause have been detected in Wuhan, Hubei province. Prompt action was taken by the Chinese government to carry out etiological and epidemiological investigations. In the second stage (from January 20 to February 20, 2020), the number of new cases had risen rapidly in China and Wuhan was shut down. The epidemic situation has attracted global attention. In the third stage (from February 21 to March 17, 2020), the number of new cases in China had been gradually decreasing. The epidemic situation in the United States and European countries had become serious. In the fourth stage (from March 18 to April 28, 2020), the spread of COVID-19 in Wuhan had been basically stopped. The number of global confirmed cases of COVID-19 has exceeded 3 million. In the fifth stage (from April 29 to May 31, 2020), epidemic prevention and control had become normal in China. While the global epidemic situation still remained in a critical condition. Based on LDA model and topic extraction, we have conducted a series of analysis.

2.2.1. LDA model and topic extraction

Latent Dirichlet Allocation (LDA) is a dominant method used to collaborative filtering, text classification, and document modeling (Blei, Ng, & Jordan, 2003). All topics learned in LDA model are from a data corpus which means that the number of topics is learned from the input data (Gao et al., 2020). As a three-level hierarchical Bayesian model, each item of a collection is modeled as “a finite mixture over an underlying set of topics” and each topic is also modeled as “an infinite mixture over an underlying set of topic probabilities” (Blei et al., 2003).

LDA model assumes that documents contain random mixtures of latent topics, which are characterized by a distribution of words in the document (Blei et al., 2003). When selecting topic distribution and word distribution, the LDA model introduces Dirichlet prior parameters. The model solves the problem that the number of parameters in LSA (latent semantic analysis) and PLSA (probabilistic latent semantic analysis) models increases with the increase of training documents when generating topics, and alleviates the problem of model overfitting. As a bag-of-word model, LDA does not consider the grammatical structure of sentences or the sequence between words, so it is suitable for corpus processing with large amount of data (Tsai, 2012). LDA adopts efficient probabilistic inference algorithm to process large-scale data, which is widely applied and makes topic model become a research focal area in the field of Internet text, such as Twitter, blog and microblog.

LDA model was applied for topic extraction and classification. According to the White paper of the Chinese People’s Government, China’s fight against COVID-19 was divided into five stages. Based on these five stages, this paper extracted and analyzed the topics of popular Weibo data sets in different stages of COVID-19. After repeated experiments, the optimal number of topics in each stage was determined, and then similar topics were combined to obtain the final topic classification. Each popular Weibo text was assigned to corresponding topics based on its relevance to each topic. Fig. 1 presents the process of topic extraction in our study.

2.2.2. Spearman correlation

Spearman correlation (also referred to as Spearman’s rank correlation coefficient) is the most widely used correlation statistic and nonparametric measure of rank correlation between two variables (Corder & Foreman, 2014). Spearman correlation is suitable for the analysis of both continuous and discrete ordinal variables (Lehman, O’Rourke, Hatcher, & Stepanski, 2005). For a bivariate sample (Xi, Yi), 1 ≤ i ≤ n, the n raw scores of Xi, Yi are converted to rgXi and rgYi through fractional ranking (de Winter, Gosling, & Potter, 2016). Eq. (1) presents the calculating method.

\[ r_s = \rho_{rgX,rgY} = \frac{cov(rgX, rgY)}{\sigma_{rgX}\sigma_{rgY}} \]  

(1)
2.2.3. Sentiment analysis

Baidu AI Open Platform is an artificial intelligence service platform. It provides more than 120 subdivisions of scenarios-based capabilities and solutions, including a series of capabilities such as voice, face recognition, text recognition, fine density image recognition, vertical image recognition, natural language of video, and knowledge map of processing. The current study used the API sentiment orientation analysis interface developed by Baidu platform for the sentiment analysis of popular Weibo content. Baidu sentiment analysis application programming interface (API) is a corpus containing vast contexts. It could be used to report the emotional score and determine the emotional polarity category (neutral, positive, negative) of context with subjective information (Y. Li et al., 2020), and help callers understand user needs, analyze hot topics and crisis public opinion monitoring. The developer accesses the API interface according to APPID, API Key and Secret Key, and uses each text as the URL request interface for emotional analysis and returns the analysis results.

3. Results of public attention level

The public attention level in this study was as follows: (Daily) public attention level = (Daily) comments number of popular Weibo + (Daily) retweeting number of popular Weibo + (Daily) thump-up number of popular Weibo.

The results of public attention level trend from December 27, 2019 to May 31, 2020 are presented in Fig. 2. Public attention level related to COVID-19 shows a curve of fluctuation. Fig. 3 presents detailed daily public attention level trends in different stages. Fig. 3a shows the original time series of popular Weibo in the first stage, including 918 popular Weibo texts (i.e., from December 27, 2019 to January 19, 2020). Split by day, it shows that the highest point of public attention in the first stage was on January 19, 2020 and the curve shows an upward trend. Two peaks on January 11, 2020 and January 16, 2020 were identified. On January 11, 2020, netizens’ search was related to the theme “novel coronavirus caused unidentified pneumonia in Wuhan and experts urge prevention of super transmission in Hong Kong” and “students in the Chinese University of Hong Kong wear face masks to protect themselves from disease”. On January 16, 2020, the main topics related to COVID-19 on Weibo were “the first case of novel Coronavirus was confirmed in Japan”, “World Health Organization officially named pneumonia virus outbreak in Wuhan as 2019-nCoV”, and “the University of Hong Kong has made rapid tests for the virus”. On January 19, 2020, a fair amount of discussion related to COVID-19 symptom, transmission, treatment appeared.

Fig. 3b presents the original time series of popular Weibo in the
The curve shows two obvious peak points on January 24, 2020 and January 31, 2020. On January 24, 2020, the main topics related to COVID-19 on Weibo were “lockdown in Wuhan”, “Wuhan face on New Year’s Eve”, “Wuhan version of Xiaotangshan began to build” and “doctors from all over the country rushed to Hubei for help”. On January 31, 2020, netizens’ search on Weibo was related to the theme “COVID-19 update”, “fight against COVID-19”, and “Huoshenshan hospital and Leishenshan hospital”.

Fig. 3c shows the original time series of popular Weibo in the third stage, including 34,770 popular Weibo texts (i.e., from February 21 to March 17, 2020). The curve shows a trend of fluctuation and two peaks on February 27, and March 11, 2020. The popular topics related to COVID-19 on February 27 included “505 cases were found in South Korea within a day”, “the start of school will continue to be delayed”, and “COVID-19 update”. On March 11, 2020, netizens’ search on Weibo was related to the theme “Zhong Nanshan shares Chinese experience in English”, “global epidemic situation”, and “Zhengzhou confirmed an imported case”.

Fig. 3d presents the original time series of popular Weibo in the fourth stage, including 47,681 popular Weibo texts (i.e., from March 18 to April 28, 2020). There are three obvious peaks on March 24, April 1, and April 11, 2020. The whole curve shows a trend of downward trend. On March 24, 2020, the main topics related to COVID-19 on Weibo included “epidemic situation in Italy”, “India declared a 21-day national lockdown”, and “imported cases”. On April 1, 2020, public attention was related to the imported cases. On April 11, 2020, netizens’ search on Weibo was related to the theme “there are more than 500,000 confirmed cases of COVID-19 in the United States”, “two people from a Beijing family of five who were traveling abroad at a high incidence of COVID-19 were confirmed”, and “the first COVID-19 lecture between Doctors in China and Africa”.

Fig. 3e presents the original time series of popular Weibo in the fifth stage, including 27,795 popular Weibo texts (i.e., from April 29 to May 31, 2020). The curve shows two obvious peaks on May 14 and May 16, 2020 and a downward trend. On May 14, 2020, the main topics related to COVID-19 on Weibo included “110 COVID-19 vaccines are under development”, and “the time for resuming classes in Shandong province is determined”. On May 16, 2020, public attention on Weibo was related to the theme “WHO: More than 300,000 deaths from COVID-19”.

Fig. 3. The trend of public attention level related to COVID-19 popular Weibo in different stages.
4. Correlation analysis

4.1. Correlation analysis between public attention level and COVID-19 related cases

The Spearman correlation between public attention level and new confirmed cases number, new suspected cases number, and new deaths number in provinces. As discussed before, public attention was operated as the sum of comments, retweeting, and thump-up number. The data analysis of Spearman correlation is conducted on Statistical Package for the Social Sciences (SPSS24.0), a professional statistical analysis tools. The data of new confirmed cases, new suspected cases, and new deaths are from mainland China. The current study included data for 157 days (i.e. during December 27, 2019 and May 31, 2020) for further analysis.

Table 2 presents the correlation between public attention level and COVID-19 related cases number. A significant positive relationship between these two variables was identified. The Spearman correlation between public attention level and new confirmed cases number is 0.823 with significant statistical significance ($p < 0.001$). The Spearman correlation of 0.873 at significant statistical significance ($p < 0.001$) indicates a significant relationship between public attention level and new suspected cases number. A significant relationship ($p < 0.001$) with correlation 0.804 is also found between public attention level and new deaths number. To further investigate if the magnitude of correlation analysis would be different across stages, further analysis was conducted. Due to the limited sample sizes distributed in each stage, the data of the first and second stage, as well as the third and fourth stages, was combined to investigate the change of correlation. Because in the fifth stage, the new cases number has a small quantity, the separate analysis of the fifth stage was not included. As presented in Table 2, the correlation coefficient between public attention level and new confirmed cases number in the first and second stages is obviously bigger than the value in the third and fourth stages. Table 3 presents that each COVID-19 related cases number positively relates to the other cases number.

To analyze the correlations among the three indicators of public attention level and test the correlation between each indicator and other outcomes, further correlation analysis was performed. Table 4 shows that daily popular Weibo texts positively relates to public attention level, new confirmed cases number, new suspected cases number, and new deaths number. As presented in Table 5, the Spearman correlation between (daily) comments number and (daily) retweeting number, between (daily) comments number and (daily) thump-up number, and between (daily) retweeting number and (daily) thump-up number are 0.928, 0.975, and 0.921 respectively with significant statistical significance ($p < 0.001$). In addition, (daily) retweeting number, (daily) thump-up number, and (daily) comments number positively relates to three new cases number respectively.

Table 2
Correlations between public attention level and COVID-19 related cases number.

| Variables                  | New confirmed cases number | New suspected cases number | New deaths number |
|----------------------------|---------------------------|---------------------------|-------------------|
| Public attention level     | 0.823***                  | 0.873***                  | 0.804***          |
| Number (N)                 | 157                       |                           |                   |
| Public attention level (the first and second stages) | 0.790***                  | 0.850***                  | 0.903***          |
| Number (N)                 | 56                        |                           |                   |
| New confirmed cases number |                           |                           |                   |
| New suspected cases number |                           |                           |                   |
| New deaths number          |                           |                           |                   |

Note: ***$p < 0.001$, **$p < 0.01$ (2-tailed).

Table 3
Correlations among three COVID-19 related cases number.

| Variables                  | New confirmed cases number | New suspected cases number | New deaths number |
|----------------------------|---------------------------|---------------------------|-------------------|
| New confirmed cases number | 0.905***                  |                           |                   |
| New confirmed cases number |                           | 0.850***                  |                   |
| New suspected cases number |                           |                           | 0.903***          |
| Number (N)                 | 157                       |                           |                   |

Note: ***$p < 0.001$ (2-tailed).

Table 4
Correlations between daily popular Weibo texts number and other variables.

| Variables                  | Public attention level | New confirmed cases number | New suspected cases number | New deaths number |
|----------------------------|------------------------|-----------------------------|-----------------------------|-------------------|
| Public attention level     | 0.792***               |                             |                             |                   |
| New confirmed cases number | 0.765***               |                             |                             |                   |
| New suspected cases number | 0.820***               |                             |                             |                   |
| New deaths number          | 0.795***               |                             |                             |                   |
| Number (N)                 | 157                    |                             |                             |                   |

Note: ***$p < 0.001$ (2-tailed).

4.2. Correlation analysis between cases and lag series of public attention level

The correlation between COVID-19 related cases number and lag series of social media attention could exist (Qin et al., 2020). Therefore, the correlation analysis between COVID-19 related cases number and the lagged series of public attention level was presented in Table 6. The three COVID-19 related cases number in our study indicates the new cases number emerged, not the cumulative quantity. A significant
6.43% respectively. In the first stage, the route of COVID-19 transmission, symptoms and controllability.

Table 6

Correlations between COVID-19 related cases number and lag value of public attention level.

| Variables          | New confirmed cases number | New suspected cases number | New deaths number |
|--------------------|-----------------------------|---------------------------|------------------|
| Lag 0 day          | 0.823***                    | 0.873***                  | 0.804***         |
| Lag 1 day          | 0.837***                    | 0.845***                  | 0.801***         |
| Lag 2 day          | 0.832***                    | 0.829***                  | 0.800***         |
| Lag 3 day          | 0.832***                    | 0.809***                  | 0.777***         |
| Lag 4 day          | 0.807***                    | 0.772***                  | 0.767***         |
| Lag 5 day          | 0.778***                    | 0.746***                  | 0.753***         |
| Lag 6 day          | 0.760***                    | 0.722***                  | 0.747***         |
| Number (N)         | 157                         |                           |                  |

Note: ***p < 0.001 (2-tailed).

positive relationship between the lag days of public attention level and COVID-19 related cases number (i.e., new confirmed cases number, new suspected cases number, and new deaths number) was found, which revealed that discussion of COVID-19 related cases number would continue for several days.

5. Topic analysis

5.1. Topic analysis in the first stage: December 27, 2019 - January 19, 2020

LDA model is used in our study for topic extraction and classification. As discussed earlier, the topic analysis was divided into five stages. Table 7 shows the COVID-19 related topics and relative weight in the first stage (December 27, 2019 - January 19, 2020). Seven topics in the first stage were identified. The first frequent topic, “Human-to-human transmission in Wuhan”, accounted for 35.08% of all topics. “Epidemic situation in Wuhan” and “cases have been detected in Japan”, accounting for 18.19% and 13.4% respectively, were the second and the third most frequent topics. The next four COVID-19 related topics were “cases have been detected in Thailand”, “unidentified coronavirus found in Wuhan”, “novel coronavirus was preliminarily determined controllable”, and “novel coronavirus symptoms”, at 9.26%, 8.93%, 8.71%, and 6.43% respectively. In the first stage, the route of COVID-19 transmission and other information is not clear. Public attention is focused on the transmission route, symptoms and controllability.

Table 7

The COVID-19 related topics in the first stage (December 27, 2019 - January 19, 2020).

| Topic name                                           | Rate (%) | LDA keywords                  |
|------------------------------------------------------|----------|-------------------------------|
| 1 Human-to-human transmission in Wuhan              | 35.08    | Coronavirus, novel, Wuhan, human-to-human, pneumonia, cases, detection, infect |
| 2 Epidemic situation in Wuhan                        | 18.19    | Cases, pneumonia, coronavirus, infect, detection, hospital discharge, newly increased, Wuhan |
| 3 Cases have been detected in Japan                  | 13.4     | Coronavirus, novel, Japan, pneumonia, cases, patients, detection, infect |
| 4 Cases have been detected in Thailand               | 9.26     | novel, cases, Thailand, coronavirus, find, infect, pneumonia |
| 5 Unidentified coronavirus found in Wuhan             | 8.93     | Coronavirus, pneumonia, patients, cases, find, infect, Wuhan, unknown |
| 6 Novel coronavirus was preliminarily determined controllable | 8.71 | Pneumonia, coronavirus, unknown, epidemic, novel, preliminarily, controllable, determined |
| 7 Novel coronavirus symptoms                          | 6.43     | Pneumonia, cases, Wuhan city, coronavirus, ages, fever, detection, patients |

5.2. Topic analysis in the second stage: January 20–February 20, 2020

The COVID-19 related topics and relative weight in the second stage (during January 20 and February 20, 2020) are presented in Table 8. The results show that “treatment condition”, “epidemic prevention and control”, and “notification of epidemic situation” emerged to be the top three topics in the second stage, accounting for 25.04%, 16.56%, and 13.12% respectively. The next three most frequent topics were “global attention on the epidemic in China”, “the medical team supported Wuhan”, and “novel coronavirus scientific research” comprised 11.69%, 11.62%, and 8.74% respectively. “Front-line clinical staff”, “medical resources”, and “route of transmission” emerged to be the last three topics, at 6.84%, 3.22%, and 3.18%. In this stage, initial containment of the epidemic was achieved. Medical resources were concentrated in Wuhan. Public attention is related to prevention and control and medical staff support.

Table 8

The COVID-19 related topics in the second stage (January 20–February 20, 2020).

| Topic name                                           | Rate (%) | LDA keywords                  |
|------------------------------------------------------|----------|-------------------------------|
| 1 Treatment condition                                | 25.04    | Clinic, treatment, pneumonia, quarantine, detection, treat and cure, nucleic acid, traditional Chinese medicine |
| 2 Epidemic prevention and control                    | 16.56    | Epidemic, prevention and control, achieve, epidemic prevention, community, face masks, wear, alcohol |
| 3 Notification of epidemic situation                  | 13.12    | Cases, confirmed, newly increased, hospital discharge, accumulation, novel, cured, remain in hospital for observation |
| 4 Global attention on the epidemic in China          | 11.69    | Epidemic, China, coronavirus, World Health Organization, spread, global, international |
| 5 The medical team supported Wuhan                   | 11.62    | China, epidemic, Wuhan, Hubei, fight, medical team, support, urgency |
| 6 Novel coronavirus scientific research              | 8.74     | Coronavirus, novel, research, academician, laboratory, Zhong Nanshan, SARS, vaccine |
| 7 Front-line clinical staff                          | 6.84     | Epidemic, doctor, nurses, pneumonia, medical worker, hospitals, front-line, family |
| 8 Medical resources                                  | 3.22     | Detection test kits, reagent, hospital, novel, coronavirus, supplies, detection, medical |
| 9 Route of transmission                              | 3.18     | Transmission, disinfection, contact, aerosol, droplet, ventilation, face masks, wash hands |
epidemic situation in Wuhan.

5.3. Topic analysis in the third stage: February 21—March 17, 2020

Table 9 presents the COVID-19 related topics in the third stage (during February 21 and March 17, 2020). The rates of topics in this stage were dispersive. “Notification of epidemic situation”, “fight against epidemic”, “prevention of overseas imported cases”, and “the spread of epidemic” accounted for 15.01%, 14.41%, 13.82%, and 13.39% of all topics respectively. The following six topics, “epidemic in other countries”, “epidemic situation in Wuhan”, “epidemic situation in the United States”, “epidemic situation in Germany”, “epidemic situation in Korea” and “epidemic situation in Italy”, comprised 12.72%, 8.64%, 7.39%, 5.61%, 4.89%, and 4.10% respectively. In the third stage, the number of new cases has gradually dropped to single digits. However, epidemic has spread globally. Public attention is focused on domestic prevention of epidemic and the global situation.

5.4. Topic analysis in the fourth stage: March 18–April 28, 2020

Table 10 reveals the COVID-19 related topics in the fourth stage (during March 18 and April 28, 2020). The top three topics were “epidemic situation in the United States”, “notification of epidemic situation”, and “global response to the epidemic”, accounting for 23.22%, 22.28%, and 16.87% of all topics. The following three topics, “epidemic prevention and control”, “epidemic situation in Japan”, and “viral vaccine”, comprised 10.11%, 8.03%, and 6.64% respectively. “Enterprises resumed work and production”, “the reinstatement of schools”, and “medical resources” accounted for 5.80%, 3.81%, and 3.24% respectively. In the fourth stage, decisive results were achieved in the fight with COVID-19 in China. Work and production were resumed gradually. The United States was still in the outbreak of epidemic. Therefore, public attention in this stage is focused on epidemic situation in the United States and domestic resumed work and production.

Table 9
The COVID-19 related topics in the third stage (February 21–March 17, 2020).

| Topic name | Rate (%) | LDA keywords |
|------------|----------|--------------|
| 1 Notification of epidemic situation | 15.01 | Cases, confirmed, newly increased, pneumonia, cured, hospital discharge, death, medical observation |
| 2 Fight against epidemic | 14.41 | Epidemic, pneumonia, novel coronavirus, fight, anti-epidemic, nation, news, prevention and control |
| 3 Prevention of overseas imported cases | 13.82 | Epidemic, prevention and control, overseas, people, detection, measures, enter the country, nucleic acid testing |
| 4 The spread of epidemic | 13.39 | Epidemic, China, the United States, global, novel coronavirus, nation, pneumonia, spread, Europe |
| 5 Epidemic in other countries | 12.72 | Novel coronavirus, infection, pneumonia, Iran, Britain, Spain, Ministry of Health, detection |
| 6 Epidemic situation in Wuhan | 8.64 | Novel coronavirus, pneumonia, Wuhan, virus, News, Research, Zhong Nanshan, specialists |
| 7 Epidemic situation in the United States | 7.39 | the United States, President, novel coronavirus, face masks, epidemic, pneumonia, state of emergency, infection |
| 8 Epidemic situation in Germany | 5.61 | Novel coronavirus, pneumonia, Germany, competition, sport, epidemic, World Health Organization |
| 9 Epidemic situation in Korea | 4.89 | Korea, novel coronavirus, pneumonia, confirmed, infection, Daegu, daily, church |
| 10 Epidemic situation in Italy | 4.10 | Confirmed, pneumonia, Germany, cases, novel coronavirus, quarantine, cases, first case |

5.5. Topic analysis in the fifth stage: April 29–May 31, 2020

Table 11 presents the COVID-19 related topics in the fifth stage (during April 29 and May 31, 2020). The results show that “epidemic prevention and control” emerged to be the first frequent topic and accounted for 32.74% of all topics. The following three topics were “epidemic situation in the United States”, “notification of epidemic situation”, and “economic impact of the epidemic”, accounting for 22.84%, 18.59%, and 16.64% respectively. “The students resumed their studies” and “Epidemic situation in Brazil” then comprised 5.77% and 3.42% respectively. In the fifth stage, fight with COVID-19 has become a normalcy in China. The global epidemic was still a concern. Therefore, public attention is focused on epidemic prevention and control and global epidemic situation.

Table 10
The COVID-19 related topics in the fourth stage (March 18–April 28, 2020).

| Topic name | Rate (%) | LDA keywords |
|------------|----------|--------------|
| 1 Epidemic situation in the United States | 23.22 | The United States, coronavirus, deaths, number of people, virus, confirmed, ten thousand, infection |
| 2 Notification of epidemic situation | 22.28 | Confirmed, newly increased, accumulation, imported, overseas, hospital discharge, medical observation, death |
| 3 Global response to the epidemic | 16.87 | Epidemic, China, global, response, provide, economics, international, organization |
| 4 Epidemic prevention and control | 10.11 | Novel, coronavirus, fight, prevention and control, health, detection, close, temporary |
| 5 Epidemic situation in Japan | 8.03 | Coronavirus, Japan, novel, epidemic, affect, Olympic Games, World Health Organization |
| 6 Viral vaccine | 6.64 | Coronavirus, vaccine, novel, research, treatment, antibody, clinic, clinical test |
| 7 Enterprises resumed work and production | 5.80 | Epidemic, prevention and control, enterprises, resumed work, measures, aggregation, release, resumed production |
| 8 The reinstatement of schools | 3.81 | Face masks, school reopen, school, prevention and control, students, grades, achieve, prevention |
| 9 Medical resources | 3.24 | Britain, China, detection test kits, medical, reagent, detection, breathing machine, products |

Table 11
The COVID-19 related topics in the fifth stage (April 29–May 31, 2020).

| Topic name | Rate (%) | LDA keywords |
|------------|----------|--------------|
| 1 Epidemic prevention and control | 32.74 | Epidemic, prevention and control, work, detection, achieve, normalcy, place, fever |
| 2 Epidemic situation in the United States | 22.84 | the United States, novel coronavirus, pneumonia, death, President, time, ten thousand, epidemic |
| 3 Notification of epidemic situation | 18.59 | Confirmed, newly increased, accumulation, imported, overseas, report, hospital discharge, asymptomatic |
| 4 Economic impact of the epidemic | 16.64 | Epidemic, China, pneumonia, economics, affect, global, Japan, tourism |
| 5 The students resumed their studies | 5.77 | School, students, resume classes, resume studies, back to school, prevention and control, time, school reopens |
| 6 Epidemic situation in Brazil | 3.42 | Cases, confirmed, Brazil, death, accumulation, newly increased, ten thousands of cases, one day |
6. Sentiment analysis

In order to further explore the evolution of public opinion since the outbreak of COVID-19, this paper introduced an emotional analysis of popular Weibo content. The data collection and pre-processing have been discussed before. The sentiment attitude is divided according to the returned score of Baidu sentiment analysis API. The outputting results are between 0 and 1. The scores less than 0.45, between 0.45 and 0.55, greater than 0.55, are regarded as negative, neutral, and positive emotions. The closer it is close to 1, the more positive the emotion is, while the closer it is close to 0, the more negative the emotion is. In order to analyze the emotional evolution process, the average value of the emotional analysis results of the popular Weibo every day was calculated. To investigate the relationships between sentiment and public attention level, the correlation analysis between the average sentiment across days and public attention level. As presented in Table 12, the Spearman correlation between sentiment score and public attention level is 0.562 with significant statistical significance ($p < 0.001$).

Fig. 4 presents the overall trend of sentiment analysis related to COVID-19 popular Weibo. As presented in Fig. 4, netizens’ attitude toward COVID-19 can be divided into three periods. The first period is from December 27, 2019 to January 20, 2020. Within this range, netizens’ emotions fluctuated greatly, with positive and negative emotions alternating. The second period is from January 21 to April 05, 2020. In this period, the netizens’ emotion changed from negative to positive side, and gradually rose and stabilized in positive emotion. The third period is from April 06 to May 31, 2020, during which the netizens’ emotions performed small fluctuations and stabilized at the critical point of positive and negative sides.

During the initial period of the epidemic, there were fewer popular Weibo about COVID-19. Table 7 shows that public attention was focused on the topic of “Human-to-human transmission in Wuhan” and “unidentified coronavirus found in Wuhan”. The cognition of novel coronavirus was not clear and the information was not enough. Therefore, netizens were in a period of intense anxiety with turbulent emotion. During the second period, netizens had more understanding and cognition of novel coronavirus and the epidemic situation and the epidemic information was more transparent. Netizens were in a period of active defense and showed positive emotions. With the gradual improvement of the epidemic in Wuhan and the whole country, the emotion of netizens has stabilized in a positive state. During the third period, the sentiment shows small fluctuation and stabilizes in neutral mood. Based on Fig. 2 and the results of topic analysis, public attention related to COVID-19 has reduced and epidemic prevention has become part of daily life. Therefore, public emotion was relatively calm, stabilizing in the neutral mood with less fluctuation.

7. Discussion and conclusions

7.1. Conclusions

This paper presents a comprehensive of analysis about social media data related to COVID-19. All popular Weibo texts during December 27, 2019 and May 31, 2020 were collected for further analysis. The curve of public attention level presents an upward trend during the initial period of the epidemic, a fluctuation trend during the middle period, and a downward trend during the final period. The peak points appeared in different stages related to different themes. In order to further understand the public attention, correlation analysis, topic analysis, and sentiment analysis were performed.

Second, correlation analysis found significant positive correlations between public attention level and three new cases number which indicates that public attention would increase with more new COVID-19 related cases. In addition, the positive correlations among three COVID-19 related cases number, as well as correlations between daily popular Weibo texts number and other variables were reported. Then positive correlation between COVID-19 related cases number and lagged series of public attention level indicated that the influence of COVID-19 related cases number could continue for several days.

Third, COVID-19 related topics in different stages were identified in topic analysis. The results show that the topics are related with COVID-19 situation of China and global epidemic. The current study obtained 41 topics, with seven, nine, ten, nine, and six topics in different stages respectively. Furthermore, the topic analysis concretely presents the specific contents of public attention in different stages.

Finally, the sentiment orientation analysis presents that the sentiment could be divided into three phases. Netizens’ attitude toward COVID-19 in three periods is deeply interconnected with topic analysis and present epidemic situation. The sentiment in the three phases respectively correspond to the initial stage of epidemic during which people were confused, the middle stage during which people had certain knowledge about epidemic prevention, and the final stage, during which people were confident and accustomed to the prevention.

7.2. Strengths and limitations

As one of the first countries to suffer the ravages of the epidemic, the research on the public attention related to COVID-19 epidemic has certain theoretical and practical significance. First, by visualizing the evolution process of public attention related to COVID-19 epidemic, the changes of people’s attitudes and emotions toward the epidemic at different stages were obviously displayed. This paper provides a comprehensive understanding of public attention by combing data mining, text analysis, correlation analysis, topic analysis and sentiment analysis. Second, this study provides insights for decision makers to grasp public attention trends and respond to public concerns for timely and accurate decision making. This work could improve response capacity and provide reference for scientific prediction of epidemic trends. Compared to previous studies (Han et al., 2020; Qin et al., 2020), this work provides a long-term data analysis from December 27, 2019 to May 31, 2020 and tried to analyze public attention combing prevailing situation. Third, this study provides important and useful insights into public attention research in other countries. Popular Weibo texts were selected to analyze the public attention level. Popular Weibo texts could show more representativeness in comparison with all Weibo texts. This study provides a base for future research in the operationalization of public attention level.

This work has some limitations. Since popular Weibo texts data was used in our study, it is not possible to include personal attributes, individual characteristics. In addition, the data source of this study is Sina Weibo. Public attention data on Facebook and Twitter is not included. Future research into the comparison in different social platforms is encouraged.

| Variables                  | Minimum value | Maximum value | Mean Value | Standard deviation | Correlation coefficient |
|----------------------------|---------------|---------------|------------|--------------------|-------------------------|
| Sentiment score            | 0.072         | 0.724         | 0.535      | 0.095              | 0.562 ***               |
| Public attention level     | 0             | 22,566,518    | 2,882,824.439 | 2,854,732.627      | 0.562 ***               |

Number (N) 157.

*** p < 0.001.
Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

CRediT authorship contribution statement

Keke Hou: Conceptualization, Data curation, Formal analysis, Investigation, Resources, Software, Visualization, Writing – original draft.

Tingting Hou: Methodology, Software, Supervision, Validation, Writing – review & editing.

Lili Cai: Investigation, Validation, Writing – review & editing.

Declaration of competing interest

None.

References

Bao, Y., Sun, Y., Meng, S., Shi, J., & Lu, L. (2020). 2019-nCoV epidemic: Address mental health care to empower society. The Lancet, 395(10224), e37–e38. https://doi.org/10.1016/S0140-6736(20)30309-3.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of Machine Learning Research, 3, 993–1022. https://doi.org/10.5555/944919.944937.

Corder, G. W., & Foreman, D. I. (2014). Nonparametric statistics: A step-by-step approach. John Wiley & Sons.

Gao, J., Zheng, P., Jia, Y., Chen, H., Mao, Y., Chen, S., … Dai, J. (2020). Mental health problems and social media exposure during COVID-19 outbreak. PLoS One, 15(4), Article e0231924. https://doi.org/10.1371/journal.pone.0231924.

Han, X., Wang, J., Zhang, M., & Wang, X. (2020). Using social media to mine and analyze public opinion related to COVID-19 in China. International Journal of Environmental Research and Public Health, 17(8), 2788. https://doi.org/10.3390/ijerph17082788.

Jung, S. M., Akhmetzhanov, A. R., Hayashi, K., Linton, N. M., Yang, Y., Yuan, B., … Nishiura, H. (2020). Real-time estimation of the risk of death from novel coronavirus (COVID-19) infection: Inference using exported cases. Journal of Clinical Medicine, 9(2), 523. https://doi.org/10.3390/jcm9020523.

Kietzmann, J. H., Hermkens, K., McCarthy, I. P., & Silvestre, B. S. (2011). Social media? Get serious! Understanding the functional building blocks of social media. Business Horizons, 54(3), 241–251. https://doi.org/10.1016/j.bushor.2011.01.005.

Kim, S. E., Lee, K. Y., Shin, S. I., & Yang, S. B. (2017). Effects of tourism information quality in social media on destination image formation: The case of Sina Weibo. Information & Management, 54(6), 687–702. https://doi.org/10.1016/j.im.2017.02.009.

Lehman, A., O’Rourke, N., Hatcher, L., & Stepanski, E. J. (2005). JMP for basic univariate and multivariate statistics: A step-by-step guide. SAS Institute.

Li, Y., Gao, X., Du, M., He, R., Yang, S., & Xiong, J. (2020). What causes different sentiment classification on social network services? Evidence from Weibo with genetically modified food in China. Sustainability, 12(4), 1345. https://doi.org/10.3390/su12041345.

Li, L., Zhang, Q., Wang, X., Zhang, J., Wang, T., Gao, T. L., … Wang, F. Y. (2020). Characterizing the propagation of situational information in social media during COVID-19 epidemic: A case study on Weibo. IEEE Transactions on Computational Social Systems, 7(2), 556–562. https://doi.org/10.1109/TCSS.2020.2980007.

Qin, L., Sun, Q., Wang, Y., Wu, K. F., Chen, M., Shia, B. C., & Wu, S. Y. (2020). Prediction of number of cases of 2019 novel coronavirus (COVID-19) using social media search index. International Journal of Environmental Research and Public Health, 17(7), 2365. https://doi.org/10.3390/ijerph17072365.

Remuzzi, A., & Remuzzi, G. (2020). COVID-19 and Italy: What next? The Lancet., 395(10231), 11–17. https://doi.org/10.1016/S0140-6736(20)30627-9.

Tsal, C. F. (2012). Bag-of-words representation in image annotation: A review. International Scholarly Research Notices, 2012, 376804. https://doi.org/10.5402/2012/376804.

de Winter, J. C., Gosling, S. D., & Potter, J. (2016). Comparing the Pearson and Spearman correlation coefficients across distributions and sample sizes: A tutorial using simulations and empirical data. Psychological Methods, 21(3), 273–290. https://doi.org/10.1037/met0000079.

Zhao, Y., Cheng, S., Yu, X., & Xu, H. (2020). Chinese public’s attention to the COVID-19 epidemic on social media: Observational descriptive study. Journal of Medical Internet Research, 22(5), Article e18825. http://preprints.jmir.org/preprint/18825.