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.
Analysis of spatiotemporal characteristics of big data on social media sentiment with COVID-19 epidemic topics

Bangren Zhu, Xinqi Zheng, Haiyan Liu, Jiayang Li, Peipei Wang

Abstract
COVID-19 blocked Wuhan in China, which was sealed off on Chinese New Year’s Eve. During this period, the research on the relevant topics of COVID-19 and emotional expressions published on social media can provide decision support for the management and control of large-scale public health events. The research assisted the analysis of microblog text topics with the help of the LDA model, and obtained 8 topics (“origin”, “host”, “organization”, “quarantine measures”, “role models”, “education”, “economic”, “rumor”) and 28 interactive topics. Obtain data through crawler tools, with the help of big data technology, social media topics and emotional change characteristics are analyzed from spatiotemporal perspectives. The results show that: (1) “Double peaks” feature appears in the epidemic topic search curve. Weibo on the topic of the epidemic gradually reduced after January 24. However, the proportion of epidemic topic searches has gradually increased, and a “double peaks” phenomenon appeared within a week; (2) The topic changes with time and the fluctuation of the topic discussion rate gradually weakens. The number of texts on different topics and interactive topics changes with time. At the same time, the discussion rate of epidemic topics gradually weakens; (3) The political and economic center is an area where social media is highly concerned. The areas formed by Beijing, Shanghai, Guangdong, Sichuan and Hubei have published more microblog texts. The spatial division of the number of Weibo social media texts has a high correlation with the economic zone division; (4) The existence of the topic of “rumor” will enable people to have more communication and discussion. The interactive topics of “rumors” always have higher topic popularity and low emotion text expressions. Through the analysis of media information, it helps relevant decision makers to grasp social media topics from spatiotemporal characteristics, so that relevant departments can accurately grasp the public’s subjective ideas and emotional expressions, and provide decision support for macro-control response strategies and measures and risk communication.

1. Introduction
With the development of COVID-19 (Corona Virus Disease 2019), the number of infected people worldwide has exceeded 8.9 million, and the epidemic has become the most serious public health event affecting humankind in the 21st century. In the course of human development, infectious disease has always been a life health problem that can’t be ignored and has caused a certain impact on people’s physiology and psychology. In December 2019, China’s Wuhan Health Commission reported 27 cases of pneumonia of unknown cause [1]. On February 11, 2020, the World Health Organization officially named the new coronavirus COVID-19 in Geneva. Since the outbreak of COVID-19, the global panic caused by the epidemic has escalated. In the confrontation process of epidemic spread and control, real-time social media has become an important way to obtain people’s self-perception and information communication [2–7]. The period of occurrence and evolution of the epidemic happens to be the Chinese New Year (January 25). It is valuable to study social media topics and sentiment during this period has special value, which also provides typical cases for similar studies.

After the emergency happened, it triggered a complex network of public opinion. At this time, the positive role of social media in society will be limited by rumors, violence, and other special factors. If it is not actively guided, it may cause a crisis of online public opinion, so it is necessary to analyze the online public opinion [8–11]. By analyzing the text obtained by real-time media to obtain people’s subconscious mind’s opinions and emotions, so as
to mine people’s true cognitive ideas [12–15]. According to the analysis of emotional changes, we can understand which topics will have a positive and negative impact on emotions during emotional changes, and provide some social media emotional guidance opinions for relevant departments.

With the widespread popularity of social media, the unique advantages of social media will provide the public with a rapid and convenient platform to obtain and communicate information, thereby improving people’s ability to respond to emergencies [16]. Latent Dirichlet Allocation (LDA) is one of the most powerful technologies in social media semantic mining. The LDA topic model has been widely applied, including semantic analysis of social media, topic extraction of articles, unordered text mining, etc. [17]. Research shows that LDA has obvious advantages in topics extraction and semantic mining [18,19]. For example, the subjective expression of customers to the hotel was extracted to analyze the key factors affecting customer satisfaction [20]. Combined with the LDA model, a case studied on the New York Times report on nuclear technology, which proved that LDA has the characteristics of fast speed in analyzing large-scale texts [21]. Analysis of microblog dynamic information of China’s dengue fever based on LDA. Meanwhile, the spatial aggregation characteristics and temporal evolution of dengue fever cases were explored in combination with spatial analysis [18]. The topic was extracted from tweets about drugs, and a new method of seasonal influenza monitoring was mined based on LDA [22].

Although these studies have made gratifying progress, for COVID-19, and in the Chinese New Year of population migration, the online expression of people’s concern about the epidemic may be significantly different from other epidemics. In order to obtain the public’s social media expressions during this period, the LDA model was chosen to construct text topics and emotion recognition. With the help of crawler to obtain social media data. This study is to analyze the change of topics and mood during the epidemic from the perspective of time and space, as well as to answer the following research questions:

RQ1: What are the main topics of concern during the epidemic?
RQ2: What is the relationship between high-profile topics?
RQ3: What are the characteristics of social media topics and emotions changing with time under the background of major events?
RQ4: What are the characteristics of the spatial distribution of topics and emotions?

2. Methods

2.1. LDA model

As an unsupervised machine learning technology, LDA uses the bag-of-words method to identify the topic information hidden in large-scale document sets or corpora. The principle is to project data in a low dimension. The projection points of each type of data to be as close as possible, and the distance between different types of data centers to be as far as possible. The purpose of the LDA algorithm is to infer potential topics and build a comprehensive corpus [23]. In LDA, a document is a set of words, and there is no order between words. A document can contain multiple topics, and each word in the document is generated by one of the topics. LDA can distribute the topic of each document in the document set in the form of probability distribution [21].

The application of LDA is based on three important classifications: corpus, documents in corpora, words in documents. There is a nested relationship among these three classes [23]. Each document represents a probability distribution composed of some topics, and each theme represents a probability distribution composed of many words. Based on LDA, we can obtain the vocabulary of the topic and the frequency of vocabulary occurrence. Document topic statistics show the probability that the text is associated with each topic in the original set topic [16]. Finally, the semantic expression of the text was produced. The data processing flow is depicted in Fig. 1.

2.2. Data source and preprocessing

The data for this study was obtained from Qingbo Big Data Agency (http://www.gsdata.cn/). We defined “new crown” as the search keyword. With the help of python toolkit, crawled the data searched with the keyword “new crown” from January 24, 2020 to February 25, 2020. The data includes the publisher, title, publishing area, publishing time, likes, and comment segments. The data was acquired and stored according to the time.

In order to improve the representativeness of text data, we removed short text. The “re” toolkit in Python was used to remove special characters from text. In order to ensure the validity of the analysis statistics, we deleted the data with NA value in the “Title” field. Because part of the text is longer, we only retained Weibo with a length of less than 400 Chinese characters as the analysis sample, and obtained 1,858,288 microblog data.

The “jieba” package in Python was used to segment Weibo text. We have limited the part of speech of words, including 7 categories (“n”, “nr”, “ns”, “nt”, “eng”, “v”, “d”). The “Gensim” package in Python was used to implement the LDA model. Through more than 10 tests and some reference studies [16,24,25], we set 10 themes every day, and each theme contains 15 words.

2.3. Text sentiment

Baidu provides NLP platform (https://ai.baidu.com/tech/nlp) for sentiment analysis of text. We selected the “positive_prob” field returned by the API (Application Programming Interface) as the sentiment score. The score is in the range of 0 to 1. The higher the value, the higher the passion of emotion.

2.4. Matching rules

According to the vocabulary of daily topic extracted by LDA, the microblog text related to COVID-19 into 8 categories of topics as our analysis objects was summarized. The vocabulary contained in each category has been listed in Table 1. If the microblog text contains vocabulary of two topics, it is worth believing that the microblog also belongs to the interactive theme category. Finally, we got 28 interactive topics.

3. Results

3.1. Results of topic analysis

As shown in Table 2, “origin” is the topic with the largest number of microblogs (11.86%), with “role models” topic (0.6182) having the highest average popularity and “rumor” topic (0.2018) having the lowest average. In addition to expressing thoughts and needs directly by publishing the Weibo text, people will also express their attention to topics through likes and comments. We also calculated the rate of discussion for each topic by summing the total number of comments and likes per topic divided by the sum of the total number of posts per topic. The topic of “host” had the highest discussion rate (19.70). Judging from the rate of topic discussion, there are three main areas of concern: (1) The appeal of intermediate host for the spread of the epidemic and the prohibition of the use of wildlife; (2) Wide spread of the epidemic; (3) The situation in the place where the outbreak occurred.
Table 1
Associated terms for each topic.

| Topic          | Selected Search Terms                                                                 |
|----------------|----------------------------------------------------------------------------------------|
| Origin         | ["Hubei", "Wuhan"]                                                                    |
| Host           | ["game", "wild animals", "bat", "pangolin", "intermediate host"]                     |
| Organization   | ["Health Committee", "Health Department", "Red Cross", "charity", "community"]        |
| Quarantine measures | ["in and out", "seal off", "prohibition", "separaten", "sealed type", "online", "remote", "mask"] |
| Role models    | ["front line", "angel", "white", "protector", "doctor", "assistance", "Li Lanjuan", "Zhong Nanshan", "Li Wenhong"] |
| Education      | ["Ministry of Education", "school", "teachers", "students", "teach", "internet courses"] |
| Economic       | ["resume work", "resume production", "economic", "company", "enterprise", "factory", "financial", "business", "work"] |
| Rumor          | ["rumor", "start a rumor", "fake news", "Don't believe the rumors", "Don't spread rumors"] |

Table 2
Individual topic data and percentage.

| Topic          | Count  | Likes  | Comments | Emotion | Rate (n = 1,858,288) | Discussion rate |
|----------------|--------|--------|----------|---------|---------------------|-----------------|
| Origin         | 220,331| 3,063,935 | 438,126 | 0.5360 | 11.86%              | 15.89           |
| Host           | 27,609 | 511,181  | 32,631  | 0.2513 | 1.49%               | 19.70           |
| Organization   | 60,433 | 604,980  | 149,738 | 0.5442 | 3.25%               | 12.49           |
| Quarantine measures | 179,619 | 2,662,811 | 369,663 | 0.4311 | 9.67%               | 16.88           |
| Role Models    | 146,482 | 1,871,704 | 275,695 | 0.6182 | 7.88%               | 14.66           |
| Education      | 35,520  | 303,604  | 45,383  | 0.5682 | 1.81%               | 9.83            |
| Economic       | 118,686 | 842,949  | 129,663 | 0.5436 | 6.39%               | 8.19            |
| Rumor          | 20,386  | 221,133  | 40,791  | 0.2018 | 1.10%               | 13.29           |

Table 3 shows that the “origin & model” contains the most data (2.07%), which is consistent with the results of the single topic analysis. Although there are a lot of microblogs on the topic of “origin & model”, the topic discussion rate (28.29) is not the highest. The topic discussion proportions were sorted in descending order, and 4 of the top 10 interactive topics were interactive with “rumor” topics: “economic & rumor” (143.69), “organization & rumor” (88.52), “quarantine measures & rumor” (45.78), and “organization & rumor” (35.88). This shows that the existence of “rumor” will enable people to have more communication and discussions.

We sorted the sentiment values in ascending order. Seven of the top 10 low-emotional topics were related to rumor. According to the analysis of the topic popularity, there are always high topic popularity and low emotion text expressions in the interactive topics of “rumor”.

3.2. Results of temporal-level analysis

In order to explore the changing characteristics of topics and emotions in the period, the research period was subdivided into
daily as a unit of temporal level. As shown in Fig. 2(a), after entering the New Year, the number of microblog posts gradually decreased. After January 28th, the number of microblogs began to rise, and the peak of the epidemic topic data curve appeared around February 10. As shown in Fig. 2(b), the topic discussion rate also showed a downward trend after the New Year, and continued until January 31. The topic discussion rate fluctuated greatly from February 5 to February 13, and also fluctuated greatly from February 17 to February 25, but the volatility declined significantly. As shown in Fig. 2(c), the emotion gradually showed an increasing trend. There was no abnormal fluctuation or reduction in the time interval. It can be seen that the emotion is relatively flat.

(1) Single topic

As shown in Fig. 3, the number of microblogs began to increase after 7:00 a.m. and began to decrease after 10:00 p.m. As a result of the establishment of some embargo measures to prevent outbreaks of the epidemic, people generally reduced their travel. Therefore, a large number of microblogs were published from 7:00 a.m. to 12:00 p.m. But it can still be seen that there were more microblog posts at 10:00 a.m. and 4:00 p.m. As shown in Fig. 4(a), the topic of “origin” has long been ahead of other topics. However, after February 23, there was a decline and people were more inclined to publish texts on topics related to “quarantine measures” and “economic”. As shown in Fig. 4(b), the discussion rate on some dates is relatively high (February 6, February 12, and February 20). There are two major fluctuations in the interactive topic, including “host”.

It can be seen in Fig. 4(c) that the mood of the epidemic topic also changes with time. From January 24 to February 4, people had higher emotional expressions in the texts on the topic of “organizations” and “role models”. From February 2 to February 25, the topics of “education” and “role models” alternately became the topics with the highest emotional value. Weibo is time-sensitive; it will cause emotions to change with events in the period. But from the perspective of the emotional curve, the topics of “host” and “rumor” are in a low emotional state throughout the period.

(2) Interactive topic

We chose the most popular interactive topic of the day as the research object. During the period from January 24 to January 30, the “origin & quarantine measures” released a larger amount of data, and from January 31 to February 25, the “origin & role models” released more data. The topic of “role models” in the interactive topic will bring more positive emotional expression, which has been reflected in the single topic analysis at the temporal level. Among the 33-day high-level interactive topics, a total of 23 days contain “rumor” and there is a phenomenon of depression.

3.3. Spatial-level analysis

The research area is subdivided into the province as a unit. Although the epidemic has swept through all provinces and cities across the country, due to the differences in social conditions and the severity of the epidemic, there are some spatial heterogeneity in the topic of the epidemic. As can be seen in Fig. 5, the number of topic microblogs published in central China and coastal areas is higher. The area is formed around Hubei with Beijing, Shanghai, Guangdong and Sichuan as its borders. As shown in Fig. 6, the areas with high topic discussion rate are mainly distributed in areas with convenient external contact.

(1) Single topic

Most regions still pay more attention to “origin”. The discussion rate of popular topics in most areas of the country is between 11 and 40, and various provinces and cities also present different forms of regional interactive topics. People in Beijing call for an end to the consumption of wildlife. This appeal led more people to participate in the discussion on the topic of “host”. It is similar in Hebei and Chongqing. “Origin” is the topic with the highest topic discussion rate in Hubei Province. As far as emotions are
Fig. 2. General topic —— data volume (a), discussion rate (b), and emotion expression (c).

Fig. 3. Number of microblogs published by time nodes.
concerned, the topic with the highest emotions in most regions is "role models", while the topic with the lowest emotions is "rumor".

(2) Interactive topic

The areas with the largest number of microblogs are mainly concentrated in the eastern part of the Chinese mainland. Although the theme with the largest amount of data in each province is "origin & role models", the number varies. There are more microblogs in Beijing and Hubei than other places. In terms of topic discussion rates, Beijing ("host & quarantine measures", 1,391.35), Chongqing ("origin & host", 537.55), Yunnan ("origin & quarantine measures", 501.78), Zhejiang ("organization & rumor", 150.68), Gansu ("host & role models", 102.86) have higher discussion rates.

4. Discussion

The Internet and social media have become a widespread, large scale and easy to use platform for real-time information dissemination. It has become an open stage for discussion, ideological expression, knowledge dissemination, emotions and sentiment sharing. Social media has the characteristic of a double-edged sword. We should make full use of its advantages to serve the public, especially when dealing with emergency situations. We selected January 24, 2020 to February 25, 2020 as the study period. Analyzing the opinions and emotions expressed by Weibo users during the outbreak of the epidemic from a spatiotemporal perspective. The temporal and spatial changes of people's discussion topics and emotions are analyzed to provide decision support for relevant departments.

4.1. "Role models" triggers high topic discussions and emotional expression

On January 29, 31 regions of China initiated a first-level response to public health emergencies. The formation of medical teams all over China to support regions with severe epidemics, "Academician Zhong Nanshan" and "Academician Li Lanjuan" frequently appeared on the topic of "role models". At the same time, some doctors who once supported the SARS epidemic have once again joined the fight against the COVID-19 epidemic. People hold a positive attitude towards this example and it is easy to resonate. Role models are a kind of spiritual carrier and embodiment, which affects people's behavior in a subtle way [26]. In the current society, it is necessary to publicize this topic, which will enhance people's determination and confidence in fighting the epidemic.

4.2. "Double peaks" on the public opinion curve

The number of epidemic topics gradually decreased after January 24. However, compared with the general topics in daily life, its proportion is increasing constantly, and there is a "double peaks" phenomenon in the public opinion curve. We got the daily hot search list of microblogs as the total number of topics for each day. Combining the topic phrases analyzed by this article and the phrases analyzed by scholars [27] as the characteristics of our
judgment whether the hot search topic belongs to the epidemic topic. It is found that the epidemic topic search volume curve showed two peaks in a week, which appeared on January 29 and February 6 respectively. The epidemic topic searches accounts for 60% of all topic searches. Since then, the proportion of popular topics has gradually decreased, but it remains above 30%. We count the public opinions of social media on some major disaster events.

During the H1N1 flu, the search volume dropped sharply within one month after the peak search volume appeared on Twitter [28]. There were two peaks of concern within one month during the H7N9 flu, but the second peak was about 50% lower than the first [29]. The public's attention to the Ebola epidemic reached its climax, and the second peak of attention appeared three months later [30]. The current COVID-19 epidemic shows different trends of concern, which have not been seen before, and is a new feature of this epidemic. As shown in Fig. 7, during the period from January 30 to February 3, although other hot topics appeared, the Spring Festival topic also received continuous attention.

4.3. The topic changes with time and the fluctuation of the topic discussion rate gradually weakens

People published a large number of tweets about “origin” in the early stage, and gradually change to “quarantine measures” and “economic” topics in the later period. Wuhan was sealed off on January 23. With the outbreak of the epidemic, every move of the city has attracted the attention of the whole nation. With the passage of time, the epidemic situation in Wuhan has been controlled, and people's attention has gradually shifted to the topics of “quarantine measures” and “economic”. This phenomenon is in line with the characteristic that online public opinion would change topics within a period of time [16]. The major media platforms have publicized the epidemic prevention measures, such as “stay at home”, “wear masks”, “keep distance”, etc. At the same time, people’s attention to the economy has gradually increased, and how to safely and orderly resume work. Production is also the focus of discussion.
4.4. People in densely populated areas and economically developed areas are more inclined to express their views on the epidemic

Beijing-Tianjin-Hebei, Yangtze River Delta, Pearl River Delta and Chengdu-Chongqing posted more microblogs. Economic concentration also exists in these areas [16]. Beijing, as the capital and political center of China, is geographically far from the outbreak area (Wuhan), but still has more topics to participate in and discuss. First of all, Beijing, as the transportation hub of the country, the impact of the changing epidemic on people may be clearly reflected on social media. Second, Beijing was greatly affected by SARS in 2002. There are still some microblogs about SARS in the topic (3.50%), and the mood was very low (0.2837). Third, the economic, political, and cultural background has created people who can obtain information in time and participate in discussions [31]. The low-mood areas are not necessarily adjacent to Hubei Province. Beijing, Shandong, Jiangsu, Shanghai, Guangdong and other regions are far away from Hubei, but the mood for the epidemic is not high. This phenomenon may be related to the unique culture of China.

4.5. There is always a high rate of topic discussion and low emotion in textual expression in interactive topics that contain “rumor”

This research found that there are usually more topics to discuss in interactive topics including “rumor”. The existence of “rumor” can enable people to have more communication and discussion, and most of the low-emotional topics are related to “rumor”. It can be seen that the interactive topics of “rumor” always have higher topic popularity and low emotional text expression, which will have a negative psychological impact on the prevention and management of the epidemic to a certain extent. This situation may be caused by the uncertainty of the development of major
epidemics, reflecting people’s emotional reactions. A similar phenomenon is borne out in related studies.\(^1\)

5. Conclusion

The purpose of this study is to analyze the public opinions of social media on the topic of COVID-19 after the closure of Wuhan and the spring festival in China. We analyzed single topic and interactive topic from the perspective of temporal and spatial.

(1) The topic of “origin” and “quarantine measures” accounted for 21% of the total sample. This shows that the government’s scientific report on the origin of the epidemic will help to stabilize public opinion in the early stages of the epidemic. The discussion of “economic and rumor” is more intensive. Relevant departments should focus on controlling the spread of rumors and the economy in social media, and timely contain them to prevent further proliferation.

(2) The “double peaks” appeared in the epidemic topic curve. Several hot topics and Chinese New Year topics have led to two peaks in the search volume curve of popular topics within one week. After that, the epidemic topic search volume showed a significant downward trend, but the epidemic topic search volume accounted for more than 30% of the total topic search volume.

(3) The topic gradually shifted from the epidemic itself to the potential impact of the epidemic over time, and continued to receive attention for a long time. People gradually shifting the topic from “origin” to potential topics such as “economic”. Potential topics have been concerned for a long time since the beginning of the New Year. This shows that the prevention and control of the epidemic situation should be done well in a long period of time.

(4) The political and economic center is a high-profile area of the epidemic network. With Hubei Province as the center, Beijing-Tianjin-Hebei, the Yangtze River Delta, the Pearl River Delta, and Chengdu-Chongqing posted more microblogs. This is highly relevant to the division of economic regions. It is suggested that the government should strengthen public opinion response and prevention in and control of cities with better economic conditions.

(5) “Rumor” would enable people to have more communication and discussion. “rumor” will attract more attention. There are more exchanges on the topic of “rumor”, which makes it spread faster than other topics. At the same time, “rumor” can also cause people to experience low emotions.

However, this study also has some limitations. Firstly, the location information contained in the data information is only used at the provincial level, so there is still room for improvement in spatial analysis. Secondly, the data of this study does not collect statistical data on the gender and age of microblog users, therefore, when analyzing microblogs, some significant effects of gender and age are not reflected. Finally, we only obtained data from Sina Weibo, but for people who did not use Sina Weibo to express their opinions, we could not collect their topic focus and topic emotions, so we should be more cautious about the generalization of results. With the popularization of COVID-19, we should expand our data volume to provide more comprehensive public opinion prevention and control responses for relevant departments.

Author contributions

B.Z and X.Z. developed the original idea. All authors designed this study. B.Z collected and analyzed the data and established the model and wrote the first paper. All authors read and approved the final manuscript. X.Z. and H.L. provided project funding support.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This study was supported by The National Natural Science Foundation of China (grant no. 71673256), and the Fundamental Research Funds for the Central Universities (Grant No. 2652020001, No. 2652020002).
