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.

2020 ASIS&T Asia-Pacific Regional Conference (Virtual Conference), December 12-13, 2020, Wuhan, China

Open Access

Runbin Xie*, Samuel Kai Wah Chu, Dickson Kak Wah Chiu, Yangshu Wang

Exploring Public Response to COVID-19 on Weibo with LDA Topic Modeling and Sentiment Analysis

https://doi.org/10.2478/dim-2020-0023
received August 15, 2020; accepted September 15, 2020.

Abstract: It is necessary and important to understand public responses to crises, including disease outbreaks. Traditionally, surveys have played an essential role in collecting public opinion, while nowadays, with the increasing popularity of social media, mining social media data serves as another popular tool in opinion mining research. To understand the public response to COVID-19 on Weibo, this research collects 719,570 Weibo posts through a web crawler and analyzes the data with text mining techniques, including Latent Dirichlet Allocation (LDA) topic modeling and sentiment analysis. It is found that, in response to the COVID-19 outbreak, people learn about COVID-19, show their support for frontline warriors, encourage each other spiritually, and, in terms of taking preventive measures, express concerns about economic and life restoration, and so on. Analysis of sentiments and semantic networks further reveals that country media, as well as influential individuals and "self-media," together contribute to the information spread of positive sentiment.

Keywords: COVID-19, Weibo, web crawling, LDA, sentiment analysis

1 Introduction

In recent decades, we have encountered several disease outbreaks such as SARS in 200 and MERS in 2012. Nowadays, we are facing another enemy: COVID-19. As of July 26, 2020, more than 16 million COVID-19 confirmed cases have been reported around the world (JHU COVID-19 Resource Center, 2020). With the first confirmed case officially reported in Wuhan, China, in late December 2019, the global COVID-19 outbreak has brought great damage to the world's normal operations for over half a year and has been, unfortunately, aggravating, as of July 2020.

During and after the disease outbreak, public opinion is commonly collected as it is useful in many ways such as improving communications in terms of public concerns, crisis management, health knowledge promotion, and so on between governments and the public (Holmes, Henrich, Hancock, & Lestou, 2009; Mollema et al., 2015). Traditionally, surveys have played an important role in gathering public opinion, and undoubtedly, they have also been applied to disease outbreak research about public opinion. Table 1 summarizes some popularly investigated themes in survey-based research of the COVID-19 outbreak.

While surveys play an irreplaceable role, there is another way of collecting public opinion about disease outbreaks: the increasing popularity of social media and the development of text analysis techniques have given rise to mining social media data. Unlike surveys that can mainly collect only a limited volume of data and may not be able to reflect the real thoughts of the public because of some influential factors during the survey process, such as the environmental distraction, subject's psychological pressure, and so on (Hridoy, Ekram, Islam, Ahmed, & Rahman, 2015), social media data, which is the collection of short texts posted online to express one's feelings at any time, are generated on a large scale. Thus, the efficient use of social media data can contribute largely to the research about public opinion.

Reported as the first storm center affected by COVID-19, China had undergone an incredibly hard time at the very beginning, encountering difficulties such as city lockdowns, lack of medical supplies, lack of hospital resources, and so on. But luckily, by July 2020, China had also made satisfactory achievement by getting the disease outbreak under control with aggressive approaches (Campbell, 2020; Clinch, 2020). Weibo, the biggest social

***Corresponding author: Runbin Xie,** University of Hong Kong, Hong Kong, China, Email: reuben88@connect.hku.hk
Samuel Kai Wah Chu, Dickson Kak Wah Chiu, Yangshu Wang, University of Hong Kong, Hong Kong, China

Table 1
Popular Themes in Survey-Based Research of COVID-19 Outbreak

Themes	Sample research
Knowledge, attitudes & practices (KAP)	Wolf et al. (2020); Zhong et al. (2020)
Psychological stress	Huang and Zhao (2020); Mazza et al. (2020); Qiu et al. (2020)
Information seeking	Ebrahim et al. (2020); Liu (2020)
Misinformation (fake news)	Greene and Murphy (2020); Motta, Stecula, and Farhart (2020)
Sensitive individuals, including front-line hospital staffs and recovered patient	Bhagavathula, Aldhaleei, Rahmani, Mahabadi, and Bandari (2020); Huang, Han, Luo, Ren, and Zhou (2020)
Attitudes towards government actions on disease control	Atchison et al. (2020)

media platform in mainland China, serves the functions of information sharing and communications in the country. Data collected from Weibo can be analyzed to understand the Chinese people's reaction to the disease outbreak and the characteristics of semantic networks, which lead to the research questions (RQs) of this research.

RQ1: What topics can be detected from Weibo posts on COVID-19?

RQ2: How do sentiments change over time, and what are the characteristics of different semantic networks?

The rest of this article is organized as follows. The second part reviews some related work in the context of this research, after which methodologies used to conduct this research are introduced, and the results are presented and discussed. Finally, conclusions are made, together with the contribution of this work to the research context and suggestions on further research directions.

2 Related Work

2.1 Disease Outbreaks on Social Media

Table 2 records some exemplary research on the disease outbreak on social media. It can be observed that social media data contribute to several types of research on disease outbreaks, including public opinion mining, sentiment analysis, semantic network analysis, disease outbreak detection, and so on; the methods applied to social media data analysis vary from manual coding to machine learning techniques; Twitter is a worldwide platform for collection and analysis of social media data, while for social media data collection from residents in mainland China, Weibo is a more popular choice.

2.2 Topic Modeling Using LDA

Latent Dirichlet Allocation (LDA) assumes that a document is generated based on a certain number of topics, and each word in the document is randomly selected from its corresponding topic vocabulary (Blei, Ng, & Jordan, 2003; Gruber, Weiss, & Rosen-Zvi, 2009). It is an excellent probabilistic model that performs well in topic modeling and has been widely applied in research (Hu et al., 2017). For example, Barua, Thomas, and Hassan (2012) utilized LDA to discover topics and topic trends from a popular Q&A website in the programming field and found that the developer community discussed a wide range of topics and discussions of different topics are interconnected. Similarly, Hu et al. (2017) studied email corpora with the LDA model.

LDA has been applied to extract topics from various kinds of corpora, including but not limited to microblogs. For example, Huang, Yang, Mahmood, and Wang (2012) applied LDA with web usage data; Xu, Zhang, and Yi (2018) and Lim and Buntine (2014) studied tweets with LDA modeling. Therefore, it is logical to follow that exploring topics from Weibo datasets with LDA topic modeling should also harvest satisfactory results.

2.3 Sentiment Analysis

Sentiment of text data mainly refers to the emotions hidden within the text. Sentiment analysis has been widely applied to a large volume of opinion mining research, including product review analysis, public response to the stock market, and so on, as valuable information can be discovered if emotions in texts are well analyzed (Bakshi, Kaur, Kaur, & Kaur, 2016). In general, sentiments are classified into three categories: positive, neutral, and

Table 2
Research on Disease Outbreaks on Social Media

Author	Disease outbreak	Social media	Methods	Findings
Mollema et al. (2015)	Measles	Twitter and others	Thematic analysis	People on Twitter cared about disease transmission, preventive actions, and vaccination; governments needed to promote vaccination acceptability.
Fung et al. (2013)	MERS-CoV & H7N9	Weibo	Statistical analysis on the number of Weibo posts	Weibo users reacted to the disease outbreak significantly, and people paid more attention to the H7N9 outbreak.
Ye, Li, Yang, and Qin (2016)	Dengue	Weibo	Analysis on the numbers of posts and spatial information	Spatially and temporally, there was a correlation between the number of posts and disease development trends.
Chew and Eysenbach (2010)	H1N1	Twitter	Manual and automated coding	Several sentiments, including confusion, humor, risk, and so on, were discovered, among which humor was the most popular sentiment.
Ye et al. (2016)	Influenza	Twitter	Modeling	A prediction model built on Twitter data could be used for influenza outbreak alerts.
Zhang et al. (2015)	H7N9	Weibo	Analysis on the number of Weibo posts and the number of new confirmed cases	There was a positive correlation between discussion and disease outbreak level, and Weibo served as a good medium to promote communications of public health.
Li et al. (2020)	COVID-19	Weibo	Machine learning algorithms	Weibo posts were classified into seven categories of situational information. Useful text features should be helpful in building an emergence response system.

negative, and there are mainly two ways of sentiment tagging (Ray & Chakrabarti, 2017).

Lexicon-based sentiment tagging analyzes the words in a sentence and harvests the overall score by adding up the scores of each word, for which sentiment dictionaries are used (Bhonde, Bhagwat, Ingulkar, & Pande, 2015). This method was widely applied in the sentiment tagging of social media posts. For example, Ray and Chakrabarti (2017) used the R language to tag Twitter posts for product reviews with lexicon vocabularies and completed sentiment analysis at three levels: document, sentence, and aspect; Pérez-Pérez, Pérez-Rodríguez, Fdez-Riverola, and Lourenço (2019) used a lexicon-based tagger to analyze tweets' sentiments under each topic discovered in the Human Bowel Disease community.

Machine learning approaches to sentiment tagging are also gaining popularity. For example, Chen and Sokolova (2018) adopted an unsupervised approach to cluster sentiments of clinical discharge summaries with word embeddings generated from Word2Vec and Doc2Vec models and compared the results. Salathé and Khandelwal (2011) compared three machine learning techniques in terms of sentiment classification performance and adopted both naive Bayes and maximum entropy algorithms for the supervised classification of vaccination sentiments.

3 Methodologies

3.1 Methodological Framework

Figure 1 shows the methodological framework of this work. This research starts with data collection of Weibo posts with a user-simulation-like web crawler. Posts are then collected and processed to remove text noises and stop words, after which text segmentation is also performed. Then topic modeling is conducted using the LDA model with the cleaned dataset. To answer the second research question, lexicon-based sentiment analysis is conducted on data, including the number of posts, network information, and so on, to reveal sentiment trends and discover the characteristics of semantic networks.

3.2 Data Collection

This work examines the public opinion related to COVID-19 on Weibo, for which theme-related posts must be collected. Both web crawlers and APIs are nowadays widely used technologies to collect data from social media, including Weibo (Liu & Hu, 2019; Liu, Wu, Wang, & Li, 2014). Although Weibo API, being the official gateway

to collect Weibo data, is easy to use, there are many constraints such as only providing a limited number of posts and so on (Zeng, Zheng, Chen, & Yu, 2014). To avoid such inconvenience, this work uses selenium to develop a simulation-based web crawler using Python, so as to satisfy the data collection task. The web crawler simulates human logins and searches Weibo posts based on given keywords. The HTML of webpages is then collected and parsed to get the posted Weibo content and relevant information such as username and so on.

To fulfill the research purposes, this work uses six keywords selected from Hu, Huang, Chen, and Mao (2020), as listed in Table 3. Hot posts on each day over the period from January 1 to June 30, 2020, are collected.

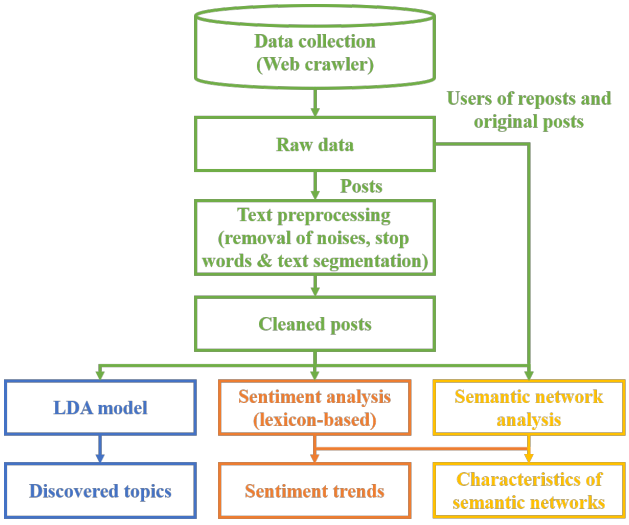

Figure 1. Methodological framework of the proposed research

3.3 Data Preprocessing

3.3.1 Removing Noises

Text noises are generally removed in the text mining research, which benefits the experimental outcomes (Celardo, Iezzi, & Vichi, 2016). In the context of this research, text noises include emoji codes, punctuation marks, symbols, non-Chinese words, and so on. To remove text noises, the "re" regular expression package in Python is adopted to remove all non-Chinese components in the dataset.

3.3.2 Stop Word List

In Chinese text, there are many "meaningless" words like "我" (I/me), "你" (you), "了" (have done something), and so on, which are normally removed in information retrieval and text mining tasks so as to improve the experimental outcomes (Zou, Wang, Deng, & Han, 2006). There are also some Chinese stop word lists built by university NLP labs and companies. To construct a stop word list, this work integrates three public stop word lists (Baidu stop word, SCU stop word, and HIT stop word) that are widely used (Xie et al., 2019), together with some domain stop words such as "转发微博" (repost) that appear frequently in most of the posts collected.

3.3.3 Chinese Text Segmentation

Unlike English text, the Chinese text needs to be segmented for analysis tasks. In terms of Chinese text segmentation tools, the "jieba" package in Python is widely used and has many advantages, such as adding customized words (Day & Lee, 2016; Peng, Liou, Chang, & Lee, 2015). In this research, the "jieba" package is adopted to perform the Chinese text segmentation task.

Table 3
Selected Keywords for Data Collection

Keyword	Translation
病毒	Virus
病例	(Confirmed/suspicious) case
肺炎	Pneumonia
新冠	COVID
新型冠状	Coronavirus
疫情	Disease outbreak

3.4 Topic Modeling

As mentioned above, LDA is a powerful tool in topic modeling. LDA can extract a given number of topics from a corpus that contains a certain number of documents. This research applies LDA to extract a certain number of topics from the cleaned dataset with the Python package "gensim." In terms of the determination of the number of topics, both perplexity and coherence scores are taken into consideration. While the former measures how well the model is generated from the corpus (the lower the better), the latter measures the sentence similarity of each topic in the dataset (the higher the better) (Blei et al., 2003; Xie, Qin, & Zhu, 2018). After the optimal model is determined,

UserID	User	Contents	PostTime	LikeNum	TransNum	Comment	RepostFlag	OrginID	Org	OrginContents
2.7E+09	Lei	:你不会看时间差吗？欧洲什么时候爆发的？就这段时	04月01日	1	0	0	0	NA	NA	NA
2.95E+09	72m	<U+00A0>原图<U+00A0>	04月01日	NA	NA	2693	1	rmrb	人民	【#你好，明天#】欧美疫情形势严峻，可近来却有个
5.66E+09	今日	:身先士卒？#英国首相新冠肺炎检测呈阳性# 英国首相	04月01日	0	0	0	0	NA	NA	NA
9034	南空	转发理由:新冠诚可怕~健康价更高~若为皇冠故~二者	04月01日	1	0	0	1	entpapa	新浪	#查尔斯王子发视频谈患病感受# 1日，英国王室发布
3.17E+09	才不	<U+00A0>原图<U+00A0>	04月01日	NA	NA	42	1	liyinghai	李英	这是我手机里存的两张图，是武汉封城之初，恐慌造成
2.44E+09	清风	转发理由:/@ 配音演员吴磊同盟会:/@ 食不食毛毛:小	04月01日	0	0	0	1	shlingsh	领声	#抗疫行动# 抗击疫情原创公益宣传曲《病毒byebye》
2.44E+09	m in	转发理由:持久战打的苦，病毒害人啊！<U+00A0><U	04月01日	0	0	0	1	chinanet	中国	【#河南郑县村庄社区重新封闭管理# 近期曾查出2名无
1.98E+09	陈震	#美国拒绝进口中国KN95口罩#疫情之中更可怕的恐惧	04月01日	11	3	6	0	NA	NA	NA
1.42E+09	Scor	:全世界这次犯的最大的错误是用以往处理流行病的经	04月01日	0	0	0	0	NA	NA	NA
6.91E+09	是块	转发理由:这货要是在放他自己国家估计早就被打了，	04月01日	0	0	0	1	8E+07	泰安	#青岛外国人插队# 转自网友，坐标青岛：今天来检测
6.26E+09	里Kl	转发理由:/@ 政委灿荣:把国门守好，把闭门羹留给病	04月01日	0	0	0	1	qjk007	千钧	中国不养巨婴！拉萨广电，顶你！
2.72E+09	liarf	#广州护士被外国人攻击#@ 人民日报，党媒发生吧，	04月01日	0	0	0	0	NA	NA	NA
5.23E+09	wu ji	:哪是因为没有及时发现和检测和用有效药，走进急诊	04月01日	0	0	0	0	NA	NA	NA
7.27E+09	用户	NA	04月01日	0	0	0	0	NA	NA	NA
7.27E+09	达摩	转发理由:放松<U+00A0><U+00A0>	04月01日	0	0	0	1	6E+09	优必	还在玩游戏？别人家孩子都开始做游戏了<U+2757>2
7.36E+09	不曾	#抗击新型肺炎我们在行动#	04月01日	0	0	0	0	NA	NA	NA
6.33E+09	原号	转发理由:/@ 竹山吃山竹:大胆点，这新闻就算不配个	04月01日	0	0	0	1	RTrussia	今日	1日，俄罗斯军机承载口罩及其它新冠病毒救援物资飞
2.74E+09	小末	<U+00A0>原图<U+00A0>	04月01日	NA	NA	751	1	xodn	赵盛	俺早晨说什么来着？德特里克堡的军事生物武器专家先
6.01E+09	归零	#官方回应核酸检测外籍人员插队#如果我们不解决外	04月01日	4	0	2	0	NA	NA	NA
5.11E+09	qws	:病毒到来时，我喜欢你哦e	04月01日	0	0	0	0	NA	NA	NA
1.98E+09	未来	<U+00A0>原图<U+00A0>	04月01日	NA	NA	319	1	6E+08	指溶	扭腰市长说→→ 我们以为这病毒（COVID—19）只攻

Figure 2. A screenshot of some collected raw data

another Python package, "pyLDAvis," is adopted to visualize the topic extraction results, with a coordinate graph to show the distribution of topics and lists of the top 30 most salient words in each topic. Topic labeling is completed manually, based on the given salient words.

3.5 Sentiment Analysis

3.5.1 Lexicon-based Sentiment Tagging

The sentiment tagging task in this research adopts a lexicon-based approach recorded in https://www.cnblogs.com/qiaoyanlin/p/6891437.html. In short, the process of calculating the sentiment score of a post mainly contains four steps: (1) sentences in a post are split based on the punctuation marks; (2) each sentence is then segmented and meaningful words remain; (3) each remaining word (including negation words that might reverse the sentiment of a sentence) provides its sentiment and/or weight score based on the lexical dictionaries; and (4) the overall score of the post is calculated based on the scores of each sentence, which are calculated from the scores of each word.

3.5.2 Statistics and Semantic Network Analysis

After sentiment tagging, descriptive statistics of the results are described. Time series analysis is then performed to discover interesting phenomena from the number of positive and negative posts over time. To take a closer look at the sentiments of public opinion, positive and negative semantic networks are constructed to identify important roles in and the characteristics of respective networks, for which network visualization and statistical analysis are performed.

4 Results and Discussion

4.1 Data Collected

Figure 2 is a screenshot of some collected raw data. In total, 719,570 posts are collected over the period from January 1 to June 30, 2020, based on the given keywords. After data cleaning is performed, which includes removing duplicate posts, blanks, and "NA" that come probably from parsing failure, only 374,225 posts remain.

The dataset is then further processed with Chinese text segmentation and removing stop words, after which it is ready for text mining analysis. Figure 3 visualizes the top 500 frequent terms with a word cloud, and Table 4 records the top 50 frequent words in the dataset, from which some interesting preliminary findings can be observed: (1) undoubtedly, term frequencies of each keyword selected for data collection rank top in the vocabulary and (2) discussions of COVID-19 on Weibo might cover a wide range of topics, including local disease outbreak, international relations, prevention measures, global pandemic, hospital staff, vaccination, and so on, which are in line with some previous research findings on public opinions about disease outbreak (Jalloh et al., 2017; Nickell et al., 2004; Rubin, Amlôt, Page, & Wessely, 2009). The preliminary findings can be further confirmed with the following analysis.

Figure 3. Word cloud of top 500 terms in the cleaned dataset

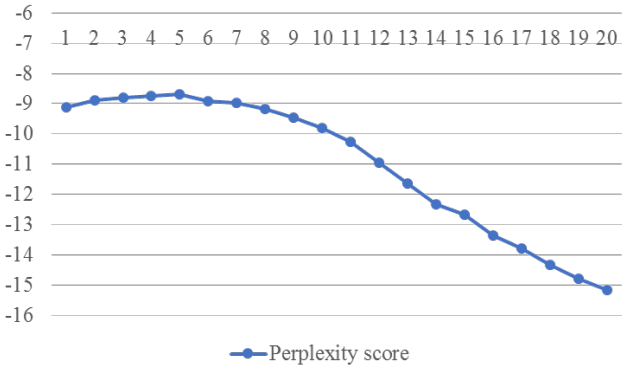

Figure 4. Perplexity scores of models under different settings of number of topics

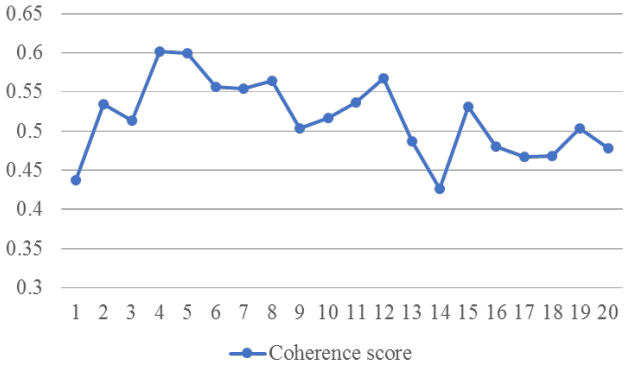

Figure 5. Coherence scores of models under different settings of number of topics

4.2 Topic Modeling

4.2.1 Perplexity and Coherence Scores of LDA Models

Figure 4 and 5, respectively, show the perplexity and coherence scores of models trained under different settings, namely, different numbers of topics. Overall, with an increasing number of topics set for model training, a smaller perplexity score is gained, that is, the model performs better in predicting the samples. Notably, the setting of a larger number of topics may lead to model overfitting issues, and thus, it needs to be careful in deciding the number of topics for training the model. Another measurement for deciding the number of topics is to compare the coherence scores of different models. In this case, coherence scores fluctuate with an increasing number of topics, and peaks are observed when numbers of topics for model training are set as 4, 8, 12, 15, respectively. Based on the perplexity and coherence scores, first, it can be concluded that the number of topics set for LDA training should be 12 or 15. By observing the keywords generated for each topic, the number of topics is finally determined as 12, because it is easier to code the topics from the given keywords.

Table 4
Top 50 Words in the Cleaned Dataset

No.	Term	Translation	Frequency	No.	Term	Translation	Frequency	No.	Term	Translation	Frequency
1	疫情	Disease outbreak	153,855	18	国家	Country	23,856	35	输入	Import	15,642
2	肺炎	Pneumonia	126,294	19	隔离	Quarantine	23,311	36	疫苗	Vaccine	15,560
3	新冠	COVID	98,978	20	法院	Court	23,168	37	治疗	Treatment	15,506
4	病例	Case	97,597	21	工作	Work	22,905	38	真的	Genuine	15,478
5	病毒	Virus	71,299	22	希望	Hope	22,836	39	人员	Staff	14,722
6	新型	Novel	66, 986	23	死亡	Death	21,772	40	情况	Situation	14,243
7	确诊	Infected	65,884	24	医院	Hospital	21,209	41	湖北	Hubei	13,338
8	冠状病毒	Coronavirus	60,636	25	检测	Test	21,195	42	北京	Beijing	13,018
9	美国	America	52,716	26	加油	Add oil	20,350	43	健康	Health	13,005
10	中国	China	50,935	27	全球	Globe	19,263	44	报告	Report	12,769
11	武汉	Wuhan	45,959	28	时间	Time	19,247	45	期间	Period	12,223
12	感染	Infected	38,556	29	累计	Accumulate	18,367	46	境外	abroad	12,162
13	防控	Disease control	33,447	30	抗击	Fight	17,996	47	出院	Discharged	11,678
14	视频	Video	32,489	31	人民	Folk	17,381	48	医生	Doctor	1,111
15	患者	Patients	31,036	32	新闻	News	17,000	49	世界	World	11,558
16	新增	New case	28,674	33	发现	Discover	16,799	50	雨花区	Yuhua district	11,391
17	口罩	Face mask	23,942	34	全国	Nation	15,650				

4.2.2 Discovered Topics

Figure 6 is a visualization of the selected LDA model mentioned above. While the left panel shows the distribution of each topic, the word list on the right panel shows the top 30 most salient terms of the selected topic, in which the blue bar shows the overall term frequency in the dataset, and the red bar represents the estimated term frequency in the selected topic. Table 5 records the discovered topics, each with 10 representative words selected from the top 30 most salient terms.

It can be seen from Figure 6 that though there are some overlapped areas, in general, topics extracted are evenly distributed. From Table 5, it can be inferred that users on Weibo discuss a wide range of topics, among which 10 topics are related to the COVID-19 outbreak while 2 topics on "Law" and "People," seem irrelevant to the research context. Seven topics, including "Fight the virus together," "Knowledge," "Assistance," "Prevention," "Treatment," "Global pandemic," and "Stay at home" are directly related to the disease outbreak, while three topics, "Economics," "Study," and "Celebrity and charity" could be regarded as topics derived from COVID-19, because they are either fields affected by the disease outbreak or tightly associated with it. From the results, it is interesting to know that people encourage each other in terms of fighting the disease, share disease-related knowledge such as prevention and transmission, pay attention to global disease outbreak development, and also discuss the affected life together, which are consistent with some findings of previous research (Chung, He, & Zeng, 2015; Corley, Cook, Mikler, & Singh, 2010; Lazard, Scheinfeld, Bernhardt, Wilcox, & Suran, 2015; Signorini, 2014).

4.3 Sentiment Tagging Results and Trend Analysis

After data preprocessing and text segmentation, 207,323 posts in total remained for sentiment tagging, with 79,861 positive posts and 33,049 negative posts, while the rest were all tagged as neutral posts (sentiment score is 0). Figure 7 shows the trends of the number of positive and negative posts over the data collection period. It can be observed that the number of positive posts exceed that of the negative ones over the whole period.

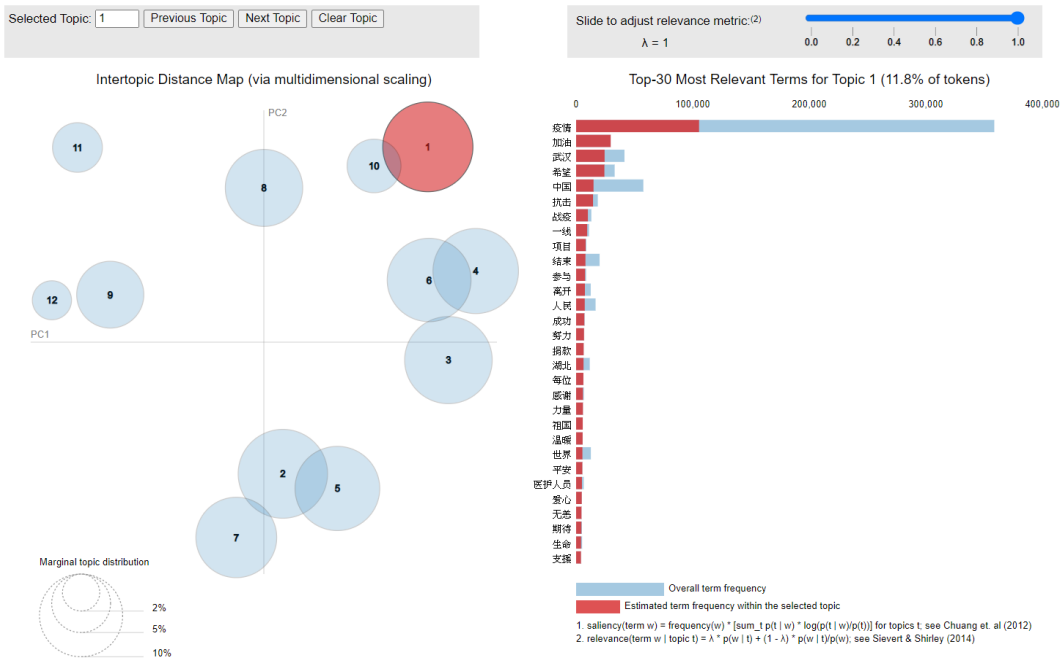

Figure 6. *LDA model visualization*

Table 5
Top 30 Most Salient Terms of Each Topic and Topic Coding Results

Topic ID	Topic Label	10 representative words selected from the top 30 most salient words
1	Fight the virus together	**力量** (power), **加油** (add oil), 人民 (folk), **希望** (hope), **中国** (China), 抗击 (fight), **战疫** (fight the virus), **一线** (front line), 成功 (success), 努力 (work hard)
2	Knowledge	研究 (research), 专家 (expert), 原因 (reason), 科普 (popular science), **传播** (transmission), 疾病 (disease), **科学** (science), 预防 (prevention), 考试 (exam), 发现 (discover)
3	Assistance	工作 (work), 奋战 (fight), 归来 (come back), 全国 (nation), **万众一心** (united), **现场** (on site), **关注** (pay attention to), 驰援 (support), 新闻 (news), 情况 (situation)
4	Economics	全球 (globe), 经济 (economics), 影响 (influence), **基金会** (fund), **世界** (world), 国际 (internationality), **控制** (control), 合作 (cooperation), **社会** (society), 市场 (market)
5	Global pandemic	伊朗 (Iran), 病例 (case), 英国 (UK), 疫情 (disease outbreak), 新增 (new case), 日本 (Japan), 意大利 (Italy), 累计 (accumulate), 告急 (urgent), 境外 (abroad)
6	Prevention	疫情 (disease outbreak), 口罩 (face mask), 宣传 (promotion), 防护 (prevention), 做好 (do well in), 防控 (prevent), 防疫 (fight the virus), 风险 (risk), 健康 (health), 措施 (measure)
7	Treatment	隔离 (quarantine), 出院 (discharged), 无症状 (no symptoms), 医学观察 (medical observation), 密切接触 (close contact), 治愈 (cure), 发热 (fever), 重症 (severely ill), 确诊 (confirmed affection), 治疗 (treatment)
8	Stay at home	期间 (period), 生活 (life), 这次 (this time), 开心 (happy), 回家 (go back home), 在家 (at home), 喜欢 (like), 事情 (thing), 朋友 (friend), 家里 (home)
9	Law	法院 (court), 法官 (judge), 人民 (folk), 违法 (breach the law), 案件 (case), 被告 (defendant), 证据 (proof), 录音 (audio recording), 依法 (according to the law), 真相 (truth)
10	Study	学校 (school), 孩子 (children), 开学 (start school), 学生 (student), 入学 (start school), 大学 (university), 专业 (major), 家长 (parents), 作业 (homework), 高考 (college entrance examination)
11	Celebrity and charity	蔡徐坤 (Xukun Cai), 杨紫 (Zi Yang), 公益 (charity), 超话 (super topic), 粉丝 (fan), 微光 (dawn), 慈善 (charity), 肖战 (Zhan Xiao), 守护 (protect), 打赢 (defeat)
12	People	直播 (live stream), 儿子 (son), 弟弟 (younger brother), 奶奶 (grandmother), 放假 (vocation), 老公 (husband), 爱豆 (idol), 价值 (value), 打榜 (boost popularity), 遇见 (meet)

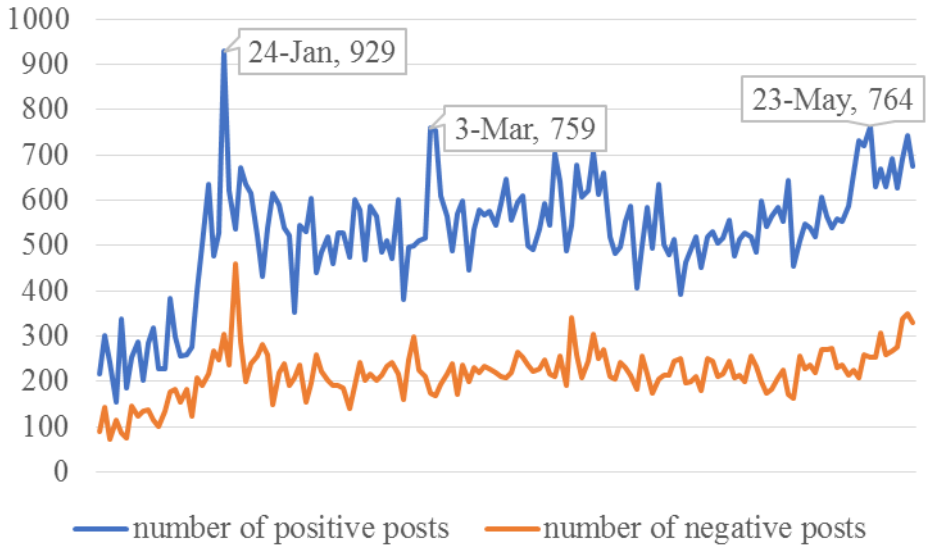

Figure 7. Number of positive and negative posts over time

Table 6
Top 10 Frequent Terms Extracted from Posts at Each Peak

No.	Peak 1			Peak 2			Peak 3		
	Top term	Translation	Freq.	Top term	Translation	Freq.	Top term	Translation	Freq.
1	希望	Hope	470	保护	Protect	767	疫情	Disease outbreak	533
2	武汉	Wuhan	388	疫情	Disease outbreak	374	病例	Case	395
3	疫情	Disease outbreak	320	希望	Hope	364	新冠	COVID-19	287
4	肺炎	Pneumonia	260	打赢	Defeat	351	病毒	Coronavirus	263
5	加油	Add oil	235	这场	This	350	肺炎	Pneumonia	193
6	新型	Novel	204	战疫	Fight the virus	344	确诊	Confirmed Affection	189
7	平安	Safe and sound	179	抗病毒	Fight the virus	313	中国	China	185
8	冠状病毒	Coronavirus	173	信息	Information	306	疫苗	Vaccine	173
9	病毒	Virus	157	好孩子	Good kids	288	累计	Accumulate	148
10	一年	1 year	154	去伪存真	Eliminate the false and retain the true	263	希望	Hope	141

Specifically, in terms of positive posts, three peaks can be observed, as also marked in Figure 7. The top 10 most frequent terms extracted from posts at each peak are listed in Table 6. The first peak is recorded on January 24, the day of the Chinese New Year's Eve, and right after the lockdown of Wuhan, which took place on January 23. From the top 10 terms, it can be inferred that the posts are mostly related to wishes for the coming New Year as well as the COVID-19 outbreak in Wuhan. The second peak comes on March 3, when the central government announced the preliminary success in fighting COVID-19, for which relevant words, for example, 保护 (protect), can also be observed. The third peak comes on May 23, when the success of the phase 1 vaccine trial was announced.

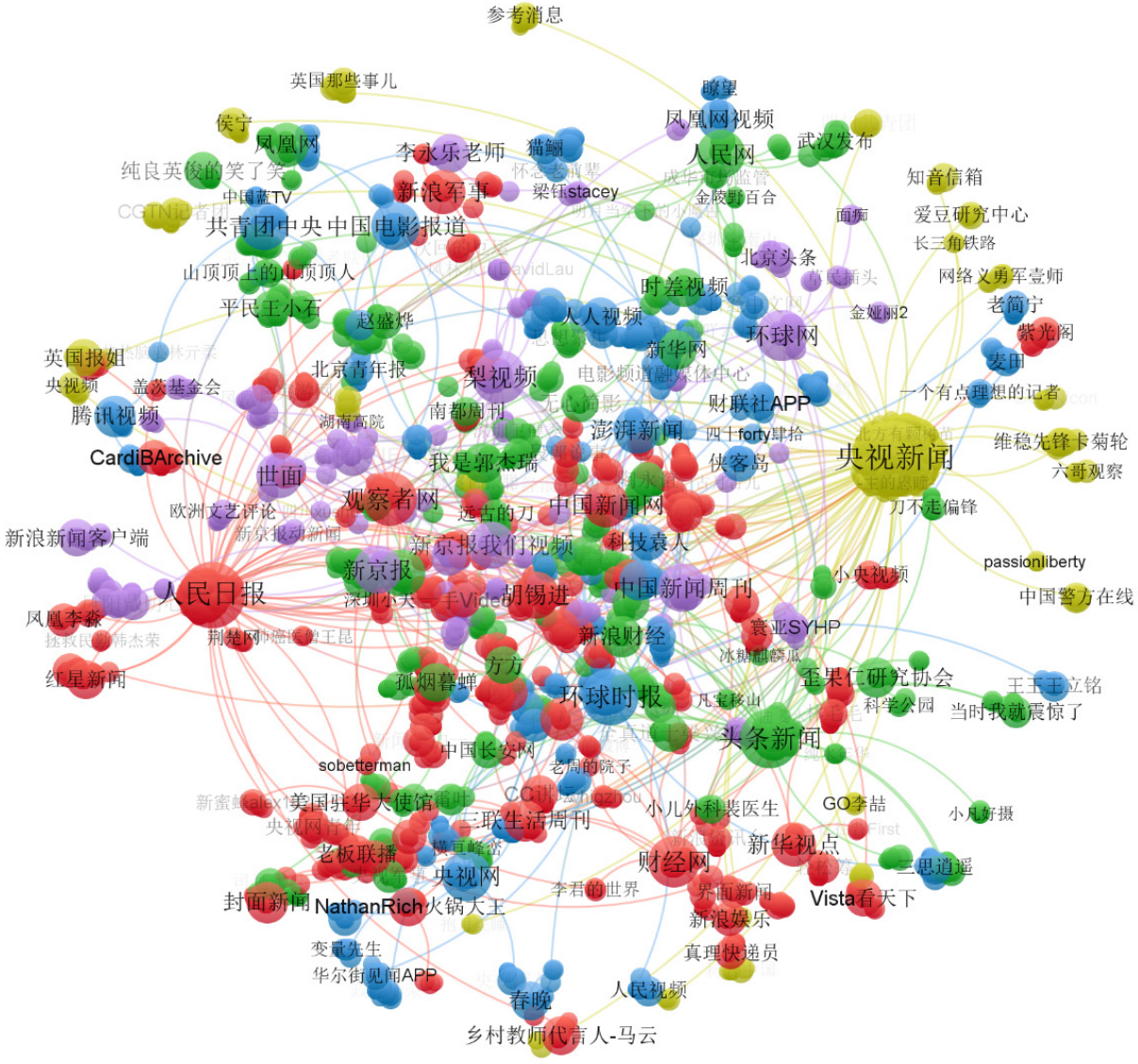

Figure 8. Visualization of semantic network of positive sentiment

4.4 Semantic Network Analysis

Figures 8 and 9 record the visualization of semantic networks of the positive and negative sentiments, respectively. In the network visualization, the color of each modularity is different from one another, and the size of a node stands for its interaction frequency: the more interaction it has, the bigger size of node it is. In both semantic networks, it could be easily observed that country media, including 人民日报 (*People's Daily*), 央视新闻 (*CCTV News*), 环球时报 (*Global Times*), and so on are leading the discussions, and basically, each of them forms a relatively independent community. As for the differences, it could be seen that there are more medium-sized nodes around each leading actor in the semantic network of positive sentiment, which means that mainstream media and influential KOL (key opinion leader), including entrepreneurs such as 乡村教师代言人-马云(Jack Ma), 胡锡进 (Hu Xijin), self-media such as 英国那些事 (Things in the UK) and so on, play an important role in leading the information spread and discussions of positive sentiment, while in the semantic network of negative sentiment, the discussions, also led by country media, are more scattered.

To take a closer look, the reposting frequencies of each node are calculated for each semantic network and then normalized between 0 and 1 for comparison purposes. Quartiles of normalized data are shown in

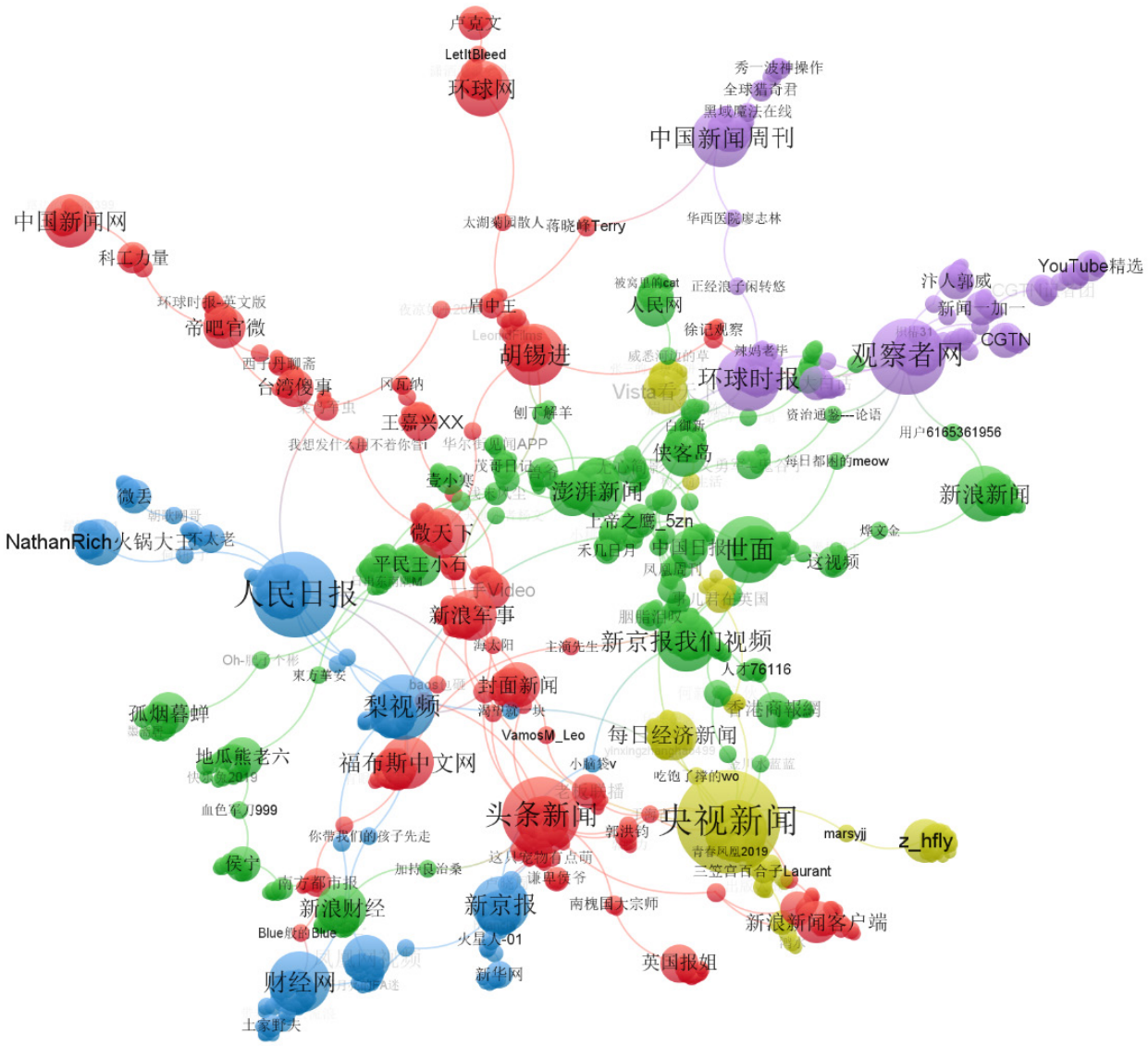

Figure 9. Visualization of semantic network of negative sentiment

Table 7
Quartiles of Normalized Reposting Frequencies of Each Semantic Network

Sentiment	Min	Q1	Median	Q3	Max
Positive	0.000	0.002	0.013	0.109	1.000
Negative	0.000	0.000	0.000	0.002	1.000

Table 7, from which it can be seen that Q1, median, and Q3 of the reposting frequencies of the positive semantic network are all greater than those of the negative semantic network, meaning that there are more influential nodes in the semantic network of positive sentiment, which are consistent with the previous findings.

5 Conclusion and Future Work

Weibo serves as a social media platform for people in mainland China to share information and communicate with each other. With the help of 719,570 collected posts and application of the LDA model, a wide range of topics discussed in relation to COVID-19 on Weibo is discovered. In response to the COVID-19 outbreak, people gain knowledge about COVID-19, show their support for frontline warriors, encourage each other spiritually, and, in terms of taking preventive measures, express concerns about economic and life restoration, and so on. Sentiment analysis further reveals that country media are leading the discussions on Weibo in both semantic networks, while, specifically, country media, as well as influential

individuals and "self-media" together contribute to the information spread of positive sentiment, indicating that the government could better fulfill its role as crisis communicator through the utilization of such kind of media network.

Although there have been studies of public opinion on COVID-19 using surveys, scant studies focus on COVID-19 opinion mining based on social media data. With LDA's excellent performance in topics modeling and sentiment analysis to take a closer look at people's feelings, this work contributes to the understanding of peoples' response to COVID-19 on Weibo and may probably serve as an example of preliminary research in the application of LDA and sentiment analysis on the COVID-19 social media dataset. Further investigation of this topic can be done in different ways. One direct way is to extend the research context, such as tracing relevant posts of each topic, analyzing peaks of negative posts, revealing the relationships between positive and negative trends, and so on. Besides, identification of topic trends, correlation analysis between the number of posts and disease development trends, and so on can also lead to meaningful findings, for which a similar approach can be seen in some previous research (Fung et al., 2013; Hu et al., 2017).

References

Atchison, C., Bowman, L., Eaton, J. W., Imai, N., Redd, R., Pristera, P., Vrinten, C., & Ward, H. (2020). *Public response to UK government recommendations on COVID-19: Population survey* (Report No. 10). Retrieved from https://doi.org/10.25561/77581

Bakshi, R. K., Kaur, N., Kaur, R., & Kaur, G. (2016). Opinion mining and sentiment analysis. In M. N. Hoda (Ed.), *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)* (pp. 452–455). Piscataway, NJ: IEEE.

Barua, A., Thomas, S. W., & Hassan, A. E. (2014). What are developers talking about? An analysis of topics and trends in Stack Overflow. *Empirical Software Engineering, 19*(3), 619–654.

Bhagavathula, A. S., Aldhaleei, W. A., Rahmani, J., Mahabadi, M. A., & Bandari, D. K. (2020). Novel coronavirus (COVID-19) knowledge and perceptions: A survey of healthcare workers. *MedRxiv*, 1–15. doi:10.1101/2020.03.09.20033381

Bhonde, R., Bhagwat, B., Ingulkar, S., & Pande, A. (2015). Sentiment analysis based on dictionary approach. *International Journal of Emerging Engineering Research and Technology, 3*(1), 51–55.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research, 3*(Jan), 993–1022.

Campbell, C. (2020). China appears to have tamed a second wave of coronavirus in just 21 days with no deaths. *Time*. Retrieved from https://time.com/5862482/china-beijing-coronavirus-second-wave-covid19-xinfadi/

Celardo, L., Iezzi, D. F., & Vichi, M. (2016). *Multi-mode partitioning for text clustering to reduce dimensionality and noises. Proceedings of the 13th International Conference on Statistical Analysis of Textual Data* (pp. 181–192), Nice : Les Press de Fac Imprimeur.

Chen, Q., & Sokolova, M. (2018). Word2Vec and Doc2Vec in unsupervised sentiment analysis of clinical discharge summaries. *ArXiv Preprint.* arXiv:1805.00352.

Chew, C., & Eysenbach, G. (2010). Pandemics in the age of Twitter: Content analysis of Tweets during the 2009 H1N1 outbreak. *PLoS One , 5*(11), 1–13. doi: 10.1371/journal.pone.0014118

Chung, W., He, S., & Zeng, D. (2015). *eMood: Modeling emotion for social media analytics on Ebola disease outbreak.* Paper presented at the ICIS2015, Beijing, China.

Clinch, M. (2020). Beijing's coronavirus outbreak is under control, Chinese health expert says. *CNBC*. Retrieved from https://www.cnbc.com/2020/06/18/beijings-coronavirus-outbreak-under-control-china-health-expert-says.html

Corley, C. D., Cook, D. J., Mikler, A. R., & Singh, K. P. (2010). Text and structural data mining of influenza mentions in web and social media. *International Journal of Environmental Research and Public Health, 7*(2), 596–615.

Day, M. Y., & Lee, C. C. (2016). Deep learning for financial sentiment analysis on finance news providers. *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* (pp. 1127–1134). Piscataway, NJ: IEEE.

Ebrahim, A. H., Saif, Z. Q., Buheji, M., AlBasri, N., Al-Husaini, F. A., & Jahrami, H. (2020). COVID-19 information-seeking behavior and anxiety symptoms among parents. *Journal of Health Care and Medicine, 1*(1), 1–9.

Fung, I. C. H., Fu, K. W., Ying, Y., Schaible, B., Hao, Y., Chan, C. H., & Tse, Z. T. H. (2013). Chinese social media reaction to the MERS-CoV and avian influenza A(H7N9) outbreaks. *Infectious Diseases of Poverty, 2*(1), 1–12. doi:10.1186/2049-9957-2-31

Greene, C. M., & Murphy, G. (2020). Can fake news really change behaviour? Evidence from a study of COVID-19 misinformation. *PsyArXiv*, 1–32. doi:10.31234/osf.io/qfnm3

Gruber, A., Weiss, Y., & Rosen-Zvi, M. (2009). Hidden topic Markov models. *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*, 163–170.

Holmes, B. J., Henrich, N., Hancock, S., & Lestou, V. (2009). Communicating with the public during health crises: Experts' experiences and opinions. *Journal of Risk Research, 12*(6), 793–807. doi:10.1080/13669870802648486

Hridoy, S. A. A., Ekram, M. T., Islam, M. S., Ahmed, F., & Rahman, R. M. (2015). Localized twitter opinion mining using sentiment analysis. *Decision Analytics, 2*(1), 1–19. doi:10.1186/s40165-015-0016-4

Hu, X., Choi, K., Hao, Y., Cunningham, S. J., Lee, J. H., Laplante, A., ... & Downie, J. S. (2017). Exploring the music library association mailing list: A text mining approach. In X. Hu, S. J. Cunningham, D. Turnbull, & Z. Duan (Eds.), *Proceeding of 18th International Society for Music Information Retrieval Conference*, 302–308.

Hu, Y., Huang, H., Chen, A., & Mao, X. L. (2020). Weibo-COV: A large-scale COVID-19 social media dataset from Weibo. *ArXiv Preprint.* arXiv: 2005.09174.

Huang, B., Yang, Y., Mahmood, A., & Wang, H. (2012, August). *Microblog topic detection based on LDA model and single-pass clustering*. Paper presented at the International Conference on Rough Sets and Current Trends in Computing, Chengdu, China. doi:10.1007/978-3-642-32115-3_19

Huang, J. Z., Han, M. F., Luo, T. D., Ren, A. K., & Zhou, X. P. (2020). Mental health survey of medical staff in a tertiary infectious

disease hospital for COVID-19. *Chinese Journal of Industrial Hygiene and Occupational Diseases, 38(3)*, 192–195. doi:10.3760/cma.j.cn121094-20200219-00063

Huang, Y., & Zhao, N. (2020). Generalized anxiety disorder, depressive symptoms and sleep quality during COVID-19 outbreak in China: A Web-based cross-sectional survey. *Psychiatry Research, 288*, 1–6.

Jalloh, M. F., Sengeh, P., Monasch, R., Jalloh, M. B., DeLuca, N., Dyson, M., . . . Bunnell, R. (2017). National survey of Ebola-related knowledge, attitudes and practices before the outbreak peak in Sierra Leone: August 2014. *BMJ Global Health, 2*(4), 1–10. doi:10.1136/bmjgh-2017-000285

JHU COVID-19 Resource Center. (2020). *COVID-19 Dashboard by the Center for Systems Science and Engineering (CSSE)*. Retrieved from https://coronavirus.jhu.edu/map.html

Lazard, A. J., Scheinfeld, E., Bernhardt, J. M., Wilcox, G. B., & Suran, M. (2015). Detecting themes of public concern: A text mining analysis of the Centers for Disease Control and Prevention's Ebola live Twitter chat. *American Journal of Infection Control, 43*(10), 1109–1111.

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.

Lim, K. W., & Buntine, W. (2014). Twitter opinion topic model: Extracting product opinions from tweets by leveraging hashtags and sentiment lexicon. *Proceedings of the 23rd ACM international conference on conference on information and knowledge management*, 1319–1328. doi:10.1145/2661829.2662005

Liu, P. L. (2020). COVID-19 information seeking on digital media and preventive behaviors: The mediation role of worry. *Cyberpsychology, Behavior, and Social Networking, 23*(10), 677–682. doi:10.1089/cyber.2020.0250

Liu, X., & Hu, W. (2019). Attention and sentiment of Chinese public toward green buildings based on Sina Weibo. *Sustainable Cities and Society, 44*, 550–558. doi: 10.1016/j.scs.2018.10.047

Liu, Y., Wu, B., Wang, B., & Li, G. (2014, August). *SDHM: A hybrid model for spammer detection in Weibo*. Paper presented at the 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014), Piscataway, NJ: IEEE. doi:10.1109/ASONAM.2014.6921699

Mazza, C., Ricci, E., Biondi, S., Colasanti, M., Ferracuti, S., Napoli, C., & Roma, P. (2020). A nationwide survey of psychological distress among Italian people during the COVID-19 pandemic: Immediate psychological responses and associated factors. *International Journal of Environmental Research and Public Health, 17*(9), 1–14. doi:10.3390/ijerph17093165

Mollema, L., Harmsen, I. A., Broekhuizen, E., Clijnk, R., De Melker, H., Paulussen, T., . . . Das, E. (2015). Disease detection or public opinion reflection? Content analysis of tweets, other social media, and online newspapers during the measles outbreak in The Netherlands in 2013. *Journal of Medical Internet Research, 17*(5), 1–12.

Motta, M., Stecula, D., & Farhart, C. (2020). How right-leaning media coverage of COVID-19 facilitated the spread of misinformation in the early stages of the pandemic in the US. *Canadian Journal of Political Science/Revue Canadienne de Science Politique, 53*(2), 335–342. doi:10.1017/S0008423920000396

Nickell, L. A., Crighton, E. J., Tracy, C. S., Al-Enazy, H., Bolaji, Y., Hanjrah, S., . . . Upshur, R. E. (2004). Psychosocial effects of SARS on hospital staff: Survey of a large tertiary care institution. *Canadian Medical Association Journal, 170(5)*, 793–798. doi:10.1503/cmaj.1031077

Peng, K. H., Liou, L. H., Chang, C. S., & Lee, D. S. (2015, October). *Predicting personality traits of Chinese users based on Facebook wall posts*. Paper presented at the 2015 24th Wireless and Optical Communication Conference (WOCC) , Piscataway, NJ: IEEE.

Pérez-Pérez, M., Pérez-Rodríguez, G., Fdez-Riverola, F., & Lourenço, A. (2019). Using twitter to understand the human bowel disease community: Exploratory analysis of key topics. *Journal of Medical Internet Research, 21*(8), 1–16.

Qiu, J., Shen, B., Zhao, M., Wang, Z., Xie, B., & Xu, Y. (2020). A nationwide survey of psychological distress among Chinese people in the COVID-19 epidemic: Implications and policy recommendations. *General Psychiatry, 33*(2), 1–3. doi:10.1136/gpsych-2020-100213

Ray, P., & Chakrabarti, A. (2017, February). *Twitter sentiment analysis for product review using lexicon method*. Paper presented at the 2017 International Conference on Data Management, Analytics and Innovation (ICDMAI). Piscataway, NJ: IEEE.

Rubin, G. J., Amlôt, R., Page, L., & Wessely, S. (2009). Public perceptions, anxiety, and behaviour change in relation to the swine flu outbreak: Cross sectional telephone survey. *BMJ (Clinical Research Ed.), 339*(jul02 3). doi:10.1136/bmj.b2651

Salathé, M., & Khandelwal, S. (2011). Assessing vaccination sentiments with online social media: Implications for infectious disease dynamics and control. *PLoS Computational Biology, 7*(10), 1–7.

Signorini, A. (2014). *Use of social media to monitor and predict outbreaks and public opinion on health topics*. (Doctoral dissertation, University of Iowa, Iowa). Retrieved from https://doi.org/10.17077/etd.841vvmuz

Wolf, M. S., Serper, M., Opsasnick, L., O'Conor , R. M., Curtis, L., Benavente, J. Y., . . . Bailey, S. C. (2020). Awareness, attitudes, and actions related to COVID-19 among adults with chronic conditions at the onset of the US outbreak : A cross-sectional survey. *Annals of Internal Medicine, 173*(2), 100–109. doi:10.7326/M20-1239

Xie, T., Qin, P., & Zhu, L. (2018). Study on the topic mining and dynamic visualization in view of LDA model. *Modern Applied Science, 13*(1), 204–213.

Xu, G., Zhang, Y., & Yi, X. (2008, December). *Modelling user behaviour for web recommendation using lda model*. Paper presented at the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, Piscataway, NJ: IEEE. doi: 10.1109/WIIAT.2008.313

Ye, X., Li, S., Yang, X., & Qin, C. (2016). Use of social media for the detection and analysis of infectious diseases in China. *International Journal of Geo-Information, 5*(9), 1–17. doi:10.3390/ijgi5090156

Zeng, Z., Zheng, X., Chen, G., & Yu, Y. (2014, December). *Spammer detection on Weibo social network*. Paper presented at the 2014 IEEE 6th International Conference on Cloud Computing Technology and Science, Piscataway, NJ: IEEE. doi: 10.1109/CloudCom.2014.14

Zhang, E. X., Yang, Y., Di Shang, R., Simons, J. J. P., Quek, B. K., Yin, X. F., . . . Tey, J. S. (2015). Leveraging social networking sites for disease surveillance and public sensing: The case of the 2013 avian influenza A (H7N9) outbreak in China. *Western Pacific Surveillance and Response Journal: WPSAR, 6*(2), 66–72.

Zhong, B. L., Luo, W., Li, H. M., Zhang, Q. Q., Liu, X. G., Li, W. T., & Li, Y. (2020). Knowledge, attitudes, and practices towards COVID-19 among Chinese residents during the rapid rise period of the COVID-19 outbreak: A quick online cross-sectional survey. *International Journal of Biological Sciences, 16*(10), 1745–1752.

Zou, F., Wang, F. L., Deng, X., & Han, S. (2006). Automatic identification of Chinese stop words. *Research on Computing Science, 18*, 151–162.