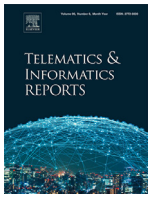
Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Telematics and Informatics Reports 8 (2022) 100016

Contents lists available at ScienceDirect

Telematics and Informatics Reports

journal homepage: www.elsevier.com/locate/teler



COVID-19 vaccine sensing: Sentiment analysis and subject distillation from twitter data

Han Xu[a,b,*], Ruixin Liu[a], Ziling Luo[a], Minghua Xu[a,b]

[a] School of Journalism and Information Communication, Huazhong University of Science and Technology Wuhan, Hubei 430074, China
[b] Philosophy and Social Science Laboratory of Big Data and National Communication Strategy, Ministry of Education, China

ARTICLE INFO

Keywords:
COVID-19 vaccine
Public opinion
Sentiment analysis
LDA
Topic analysis

ABSTRACT

The COVID-19 outbreak a pandemic, which poses a serious threat to global public health and result in a tsunami of online social media. Individuals frequently express their views, opinions and emotions about the events of the pandemic on Twitter, Facebook, etc. Many researches try to analyze the sentiment of the COVID-19-related content from these social networks. However, they have rarely focused on the vaccine. In this paper, we study the COVID-19 vaccine topic from Twitter. Specifically, all the tweets related to COVID-19 vaccine from December 15th, 2020 to December 31st, 2021 are collected by using the Twitter API, then the unsupervised learning VADER model is used to judge the emotion categories (positive, neutral, negative) and calculate the sentiment value of the dataset. After calculating the number of topics, Latent Dirichlet Allocation (LDA) model is used to extract topics and keywords. We find that people had different sentiments between Chinese vaccine and those in other countries, and the sentiment value might be affected by the number of daily news cases and deaths, and the nature of key issues in the communication network, as well as revealing the intensity and evolution of 10 topics of major public concern, and provides insights into vaccine trust.

1. Introduction

Coronaviruses (CoV) were first identified in the mid-1960s, which are known to cause colds and more serious illnesses such as Middle East Respiratory Syndrome (MERS) and Severe Acute Respiratory Syndrome (SARS) [1]. The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and infected person were first recognized and reported in Wuhan, Hubei Province, China in December 2019 [2]. This virus is a new coronavirus strain that has never been found before in humans. It is the seventh coronavirus known to infect humans, which is highly concealed, also proved more infectious than the SARS-CoV outbreak in Asia in 2003 [3,4]. In February 2020, the World Health Organization (WHO) named the SARS-CoV-2 caused disease COrona VIrus Disease 2019 (COVID-19). The source of this infectious disease, so named for its viral nature, has yet to be determined, and it quickly spread to every continent less than two months after it was first discovered [5]. On March 11st 2020, WHO declared the COVID-19 outbreak a pandemic [6], which is posing a serious threat to global public health. As with other pandemics, the spread of COVID-19 has increased exponentially. This epidemic has infected more than 487.26 million people worldwide until April 1st, 2022, of which 6,140,485 have been taken lives, according to Johns Hopkins Coronavirus Resource Center.[1]

In these early days of the outbreaks, in the absence of effective vaccine and specific therapeutic drug, many countries have implemented non-pharmaceutical interventions (NPIs), such as restricting movement of people [7]. Although these measures have been shown to be effective in curbing the spread of the virus, the availability of a safe and effective vaccine has been well-recognized by scientific experts as an additional tool to contribute to the long-term control of the pandemic [8]. All COVID-19 vaccines are designed to teach the bodys immune system to safely recognize and block the virus. Scientific institutions and manufacturers started to work on the development of COVID-19 vaccine once the pathogen was isolated and the first genomic sequence was completed.

According to data released by WHO, as of December 31st 2020, more than 200 additional vaccine candidates were in development globally, of which 47 were in clinical development, and 9 vaccines from the United States, China, the United Kingdom and Russia have entered Phase III clinical trials.[2] And the number is increasing. As shown in Table 1, these vaccines come in a variety of species currently.

[2] https://www.who.int/emergencies/diseases/novel-coronavirus-2019/covid-19-vaccines

* Corresponding author at: School of Journalism and Information Communication, Huazhong University of Science and Technology Wuhan, Hubei 430074, China.
E-mail addresses: xuh@hust.edu.cn (H. Xu), 1026855771@qq.com (R. Liu), 2797361734@qq.com (Z. Luo), xuminghua@hust.edu.cn (M. Xu).

[1] https://coronavirus.jhu.edu/map.html

https://doi.org/10.1016/j.teler.2022.100016
Received 31 March 2022; Received in revised form 16 July 2022; Accepted 5 October 2022

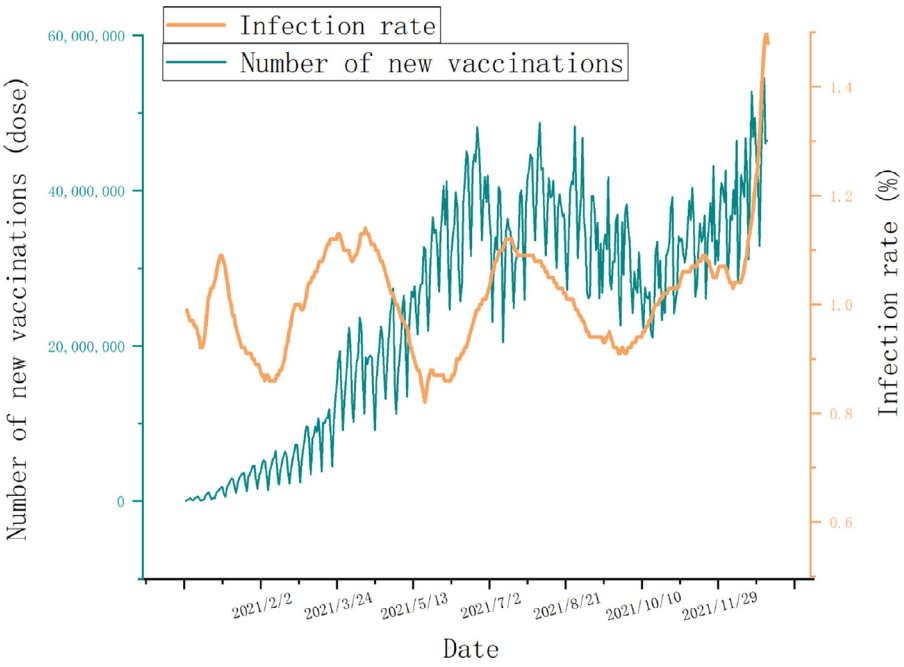


Fig. 1. The number of new vaccination and infection rate.

Table 1
The species of the COVID-19 vaccine .

Platform	Candidate vaccines
Protein subunit (PS)	51
Viral Vector non-replicating (VVnr)	21
DNA	16
Inactivated Virus (IV)	21
RNA	28
Viral Vector replicating (VVr)	4
VVr + Antigen Presenting Cell	2
Virus Like Particle (VLP)	6
Live Attenuated Virus (LAV)	2
VVnr + Antigen Presenting Cell	1
Bacterial Antigen-spore expression vector	1

The UK was the first country to officially launch vaccination campaign, using the vaccine made by BioNTech and Pfizer. The United States announced on December 14th, 2020 that it would officially begin vaccinating its citizens against COVID-19. At the same time, China began vaccinating priority populations with the Chinese-made vaccine. As of May 31st 2021, the global total of COVID-19 vaccination exceeded 40.35 million doses, with the number reaching 2.147 billion one year later. However, this number is still small compared to the total population of the world, which may be limited by the speed of vaccine production and peoples distrust of vaccines. As Fig. 1 shows, infection rates are rising. It also shows that as the virus continues to mutate, the effectiveness of vaccines is being challenged, which to some extent also affects people's attitudes towards vaccines.

Vaccine confidence is an increasingly crucial global public health issue [9]. The public's sentiments will affect their behavior, trust in vaccines will help us realize the urgent needs of achieving community protection and herd immunity against COVID-19 in the pandemic. But recent surveys indicated that a sizable proportion of the U.S. population either do not plan to or are unsure about getting vaccinated against COVID-19 [10,11]. This understandable distrust may stem from the unusually rapid speed of vaccine development. Previously, it took an average of 10 years to develop a vaccine, with the fastest mumps vaccine taking four years [12,13]. Furthermore, a survey by YouGov found that the public has different levels of trust and emotion for vaccines developed in different countries. This phenomenon may depend on people's perception of the country's image, racism and xenophobia [14]. The vast majority of the first vaccines approved for clinical use were developed in developed western countries. It is worth mentioning that this includes the developing countries of the Eastern world, China. According to Joshua Cooper Ramo, an American scholar, one of the biggest strategic threats to China today lies in its national image. The world, especially in the West, has stereotyped China to this day. Therefore, this study takes China's vaccine-related tweets as one of the objects, not only because of China's special position in the field of COVID-19 vaccine research and development, but also because China is one of the most controversial countries in the world's public opinion field, so it is of great value to study the public opinion and sentiment related to China.

People's attitudes and opinions about vaccine are not static. Their sentiments can spread through different kinds of social medias in various contexts between people in frequent close contact [15]. The outbreak of the COVID-19 resulted with a tsunami of online social media due to the implementation of NPIs measures such as regional lockdowns and social distancing [16]. Online social media has become a key platform for the public to collect information and express opinions, and individuals can express different views, opinions and emotions about the events of the COVID-19 pandemic on it [17]. Moreover, Sentiment in contents from different social media platforms might diffuse in the corresponding subsequent comments or replies and form emotional contagion [18]. Within days of onset of the COVID-19 outbreak, the emotional contagion, digitally enabled, metastasized faster than SARS-CoV-2 itself. WHO has referred this informational contagion that breeds fear and panic a "coronavirus infodemic", which can undermine trust in vaccine so much as to render them moot, then have some catastrophic effects on control and outcomes of pandemic [19,20]. Within the online social media that provide the primary platforms for emotional contagion, Twitter, which has 166 million monetized active users, needs to be highlighted.

Although there have been many studies on the sentiment analysis of COVID-19 [21–24], few researchers have focused on the vaccine. In addition, they emphasized the emotional changes of online social media messages without considering their comparison. In this context, we combined emotion analysis with subject analysis to study the content

and sentiment of tweets related to COVID-19 vaccine. In our work, we mainly discussed three questions.

- **Q1**: Whether people have different sentiment on Chinese vaccine than other countries' vaccine?
- **Q2**: Do COVID-19 data, such as the number of daily news cases and deaths, affect people's attitudes towards vaccine?
- **Q3**: At different times, the discussion of COVID-19 vaccine focused on what topics, how they were related, and how people expressed themselves.

This article is organized as follows. In Section 2, we review the relevant studies while Section 3 introduces the methods. The Section 4 details our research results and answers the above questions mainly through emotion analysis and theme analysis of data. Finally, conclusions are made in Section 5.

2. Related work

During the pandemic, the study of COVID-19-related information on social media platforms mainly focused on two aspects, that is, content and communication structure.

Structural analysis is to use social network analysis method to study the interaction pattern and attribute characteristics among participants in the process of information dissemination. Paola et al. [25] through the social network analysis of three key communications on Twitter, it was found that the dialogue network around COVID-19 topic is highly dispersed and loosely connected, which will hinder the successful dissemination of public health information in the network. Ahmed et al. [26] fetched data from the #5GCoronavirus hashtag on Twitter and used social network graphs to cluster influential users. They found that the two largest network structures include an isolates group and a broadcast group and there is a lack of an authority figure to actively combat misinformation.

There is a lot of analysis about content, and the core of such research can be summarized as topic discovery, tracking and sentiment orientation analysis. Kaur et al. [22] used IBM Watson Tome Analyzer to extract and analyze a total of 16,138 tweets about COVID-19 and found that the number of negative tweets outnumbered the number of neutral and positive tweets. Similar conclusions have been reached by Singh et al. [27]. Boon et al. [28] highlighted that the number of tweets about COVID-19 containing negative emotions is more than 50% higher than the positive emotions.

In addition, some scholars distinguished between groups. Imran et al. [29] used deep learning classifiers to study differences between the reactions towards the pandemic in different cultures. Azzam et al. [30] compared the interactions between people of different occupations or specialities on Twitter, as well as the subject matter and emotional orientation of the tweets.

Some researches focused on topic classification related to COVID-19. Wang et al. [23] adopted TF-IDF model to summarize the topics of posts which on Sina Weibo related to the COVID-19 and used unsupervised BERT model to classify sentiment categories. Using machine learning, Sear et al. [8] found that the online anti-vax community is developing a more diverse and hence more broadly accommodating discussion about COVID-19 than the pro-vax community. Cotfa et al. [31] collected the dynamics of opinion on Twitter regarding COVID-19 vaccination and analyzed it to find that the majority of tweets were neutral, while the number of in favor tweets overpasses the number of against tweets. Rafi et al. [24] used ResNet-50 CNN and attention-based LSTM networks to analysis the sentiment of the images associated with COVID-19 from Instagram. Yin et al. [32] used natural language processing to analyze popular topics and the emotional polarity of netizens in different countries towards different vaccine brands. It can be seen that the combination of sentiment analysis and topic clustering is widely used for analysing COVID-19 related texts in social media.

Table 2
Set of keywords used to fetch tweets.

Set 1	China's COVID-19 vaccine, Chinese COVID-19 vaccine Sinopharm COVID-19 vaccine Sinovac COVID-19 vaccine Sinopharm vaccine, Sinovac vaccine
Set 2	COVID-19 vaccine

3. Methodology

3.1. Data collecting and cleaning

The official COVID-19 vaccination campaign, which began in the United Kingdom on December 8th, 2020, expanded to other countries around the world. We use the tweets whose language attributes were marked as English posted from December 15th, 2020 to December 31st, 2021 for our analysis. By using Twitter API, two sets of data are captured. The first set of data are tweets related to China's COVID-19 vaccine, and the second are tweets about all COVID-19 vaccines (excluding tweets included in set 1). Since most netizens used companies' abbreviations to refer to the vaccine directly, we ended up fetching the tweets based on the keywords in Table 2. The original data included 6 fields: *ID, nickname, number of fans, tweet text, date and comment number*.

Duplicate tweets are discarded to ensure the quality of the dataset. In addition, to prevent the sentiment analysis tool from misjudging the emotional overlay of retweets to the original tweets, we also removed all retweets and only studied the original tweets. Then, the dataset is preprocessed by converting emoji and hashtag into corresponding words and preserving them, removing a series of non-English symbols such as web addresses and Twitter IDs, converting punctuation marks to Spaces and all English letters to lowercase, and filtering out stop words. After preprocessing, we chose to use Natural Language Toolkit (NLTK) for word segmentation. Thus, the high-frequency vocabulary and content focus in the text can be found.

3.2. Sentiment analysis

We used the VADER (Valence Aware Dictionary for sentiment Reasoning) model [33], an unsupervised learning method for sentiment analysis. This model outperforms individual human raters, even the machine learning algorithm may not have more significantly accurate than it. Moreover, it performs exceptionally well in the social media domain, especially in the microblog-like contexts. By using Eq. (1), the type of the tweet's sentiment TS and the tweet's compound score CS can be calculated.

$$TS(CS) = \begin{cases} \text{positive} & CS \geq 0.05 \\ \text{negative} & CS \leq -0.05 \\ \text{neutral} & otherwise \end{cases} \tag{1}$$

In our work, the model identifies and calculates the polarity and emotion of the text, classifies the sentiment of tweets into positive, neutral, or negative classes, and calculates the overall emotional value of the tweets, which is denoted as compound and fluctuates within the interval is [−1,1]. If the compound value is in the range of [−1,0), it shows that the sentiment of this twitter is negative, otherwise, it is positive, except for 0 (neutral).

3.3. Topic analysis

Topic extraction can be divided into statistical feature-based, semantic-based and topic-based extraction methods. To study the topics and evolution of public opinion about COVID-19 vaccines, we use the Topic extraction based LDA model for analysis. This model is a word bag model of unsupervised machine learning technology based on probability graph model. It believes that each text can contain multiple different

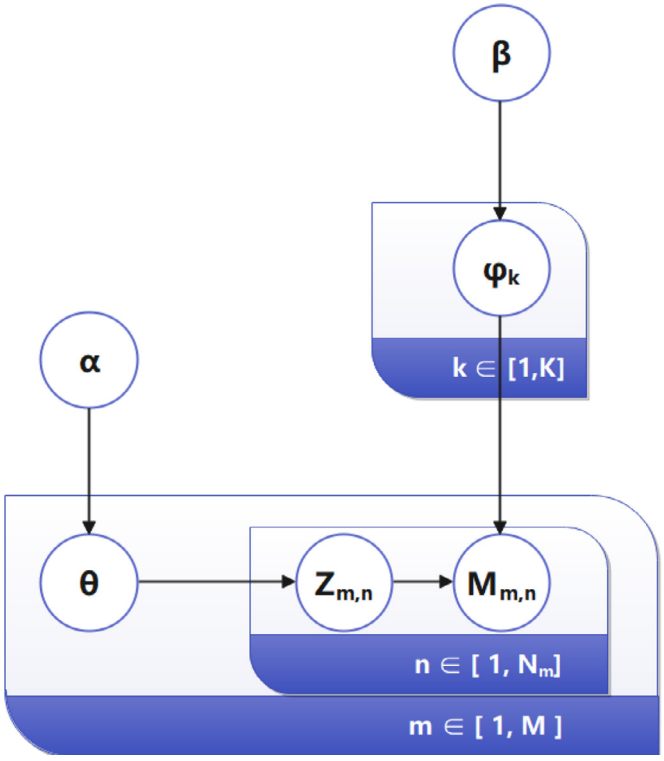


Fig. 2. LDA model.

themes, and each theme also contains multiple different words, which are mixed to form themes with a certain probability. It is a probability generation model proposed by Blei et al. [34] on the basis of Dirichlet process to solve the problem of the increase and change of the number of parameters to be estimated in PLSA model by introducing Dirichlet prior distribution.

Its main operation principle is shown as Fig. 2

Suppose there are k topics in m texts, each text has its own topic distribution and obeys the Dirichlet distribution with parameter α, and each topic has its own topic distribution and obeys the Dirichlet distribution with parameter β. For each topic in each document, there is a corresponding topic, and its probability graph model is shown in Fig. 2. Here θ represents the topic, the text i represents d_i, and the topic is displayed as $\theta_i = (\theta_{i1}, \theta_{i2}, \theta_{ik})$. If φ is used to represent the word distribution corresponding to the k topic, the word distribution of the k topic can be expressed as $\varphi_k = (\varphi_{k1}, \varphi_{k2}, \varphi_{kj})$. $Z_{M,N}$ represents the nth topic in the document m. By selecting φ_k from $Z_{M,N}$, the required observation value $W_{M,N}$ can finally be obtained, which represents the nth term in document m.

The research shows that LDA model can reduce the dimension of text and avoid the dimension disaster in the mass text data. In addition, it is effective in text mining, especially in short text processing. Therefore, LDA model has been widely used in the fields of economy, education, book information management and network public opinion [35]. On the basis of preprocessing the data, we calculated the right number of topics, then extracted the keywords contained in each topic, calculated the intensity and the evolution of importance.

4. COVID-19 vaccine sensing

4.1. Dataset

A total of 5,272,745 tweets were analyzed in the study. After capturing and cleaning the data according to the keywords in Table 2 Set 1, 136,154 valid tweets can be obtained. These data will be used as the first dataset for our analysis. 5,254,184 valid tweets related to the keywords in Table 2 Set 2 can be obtained, and the data overlaps with the first dataset can be deleted. Finally, it can be determined that there are 5,136,591 samples in the second dataset. It is worth noting that, through preliminary research, we found that netizens may use expressions such as “Sinopharm vaccine”, “Sinovac vaccine” when discussing the COVID-19 vaccine produced in China. While such situations also occur when people discuss vaccines produced in other countries, they are minuscule

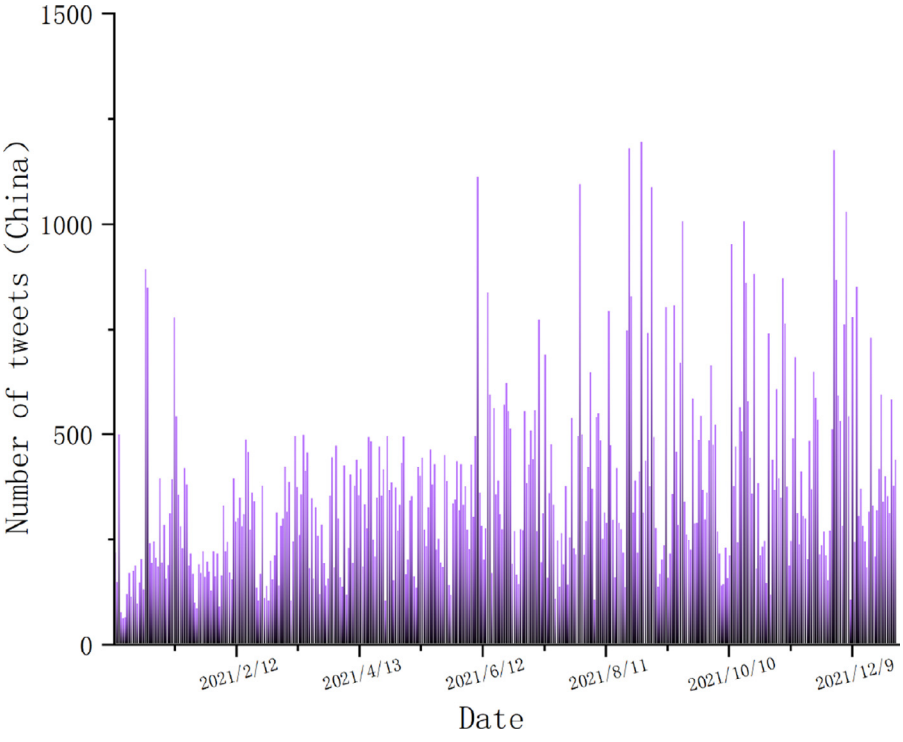


Fig. 3. Sample sizes for the first dataset.

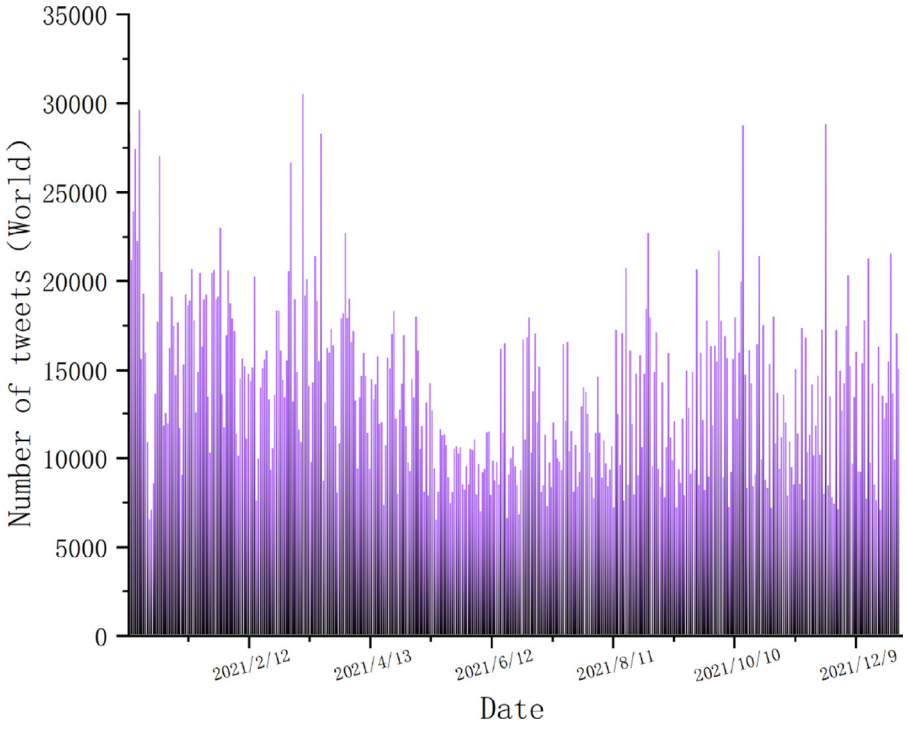


Fig. 4. Sample sizes for the second dataset.

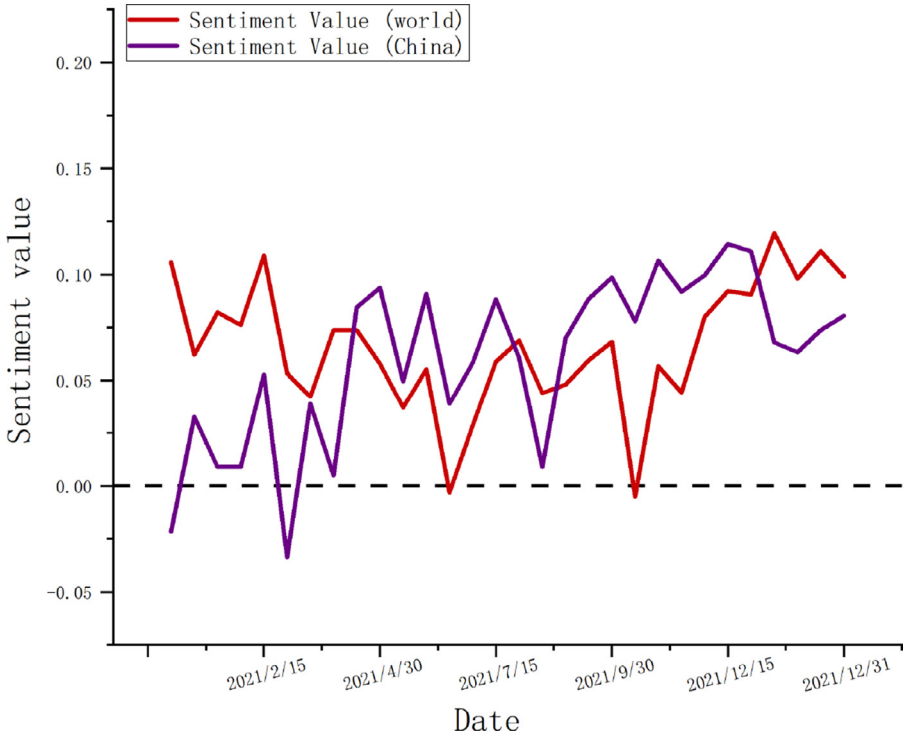
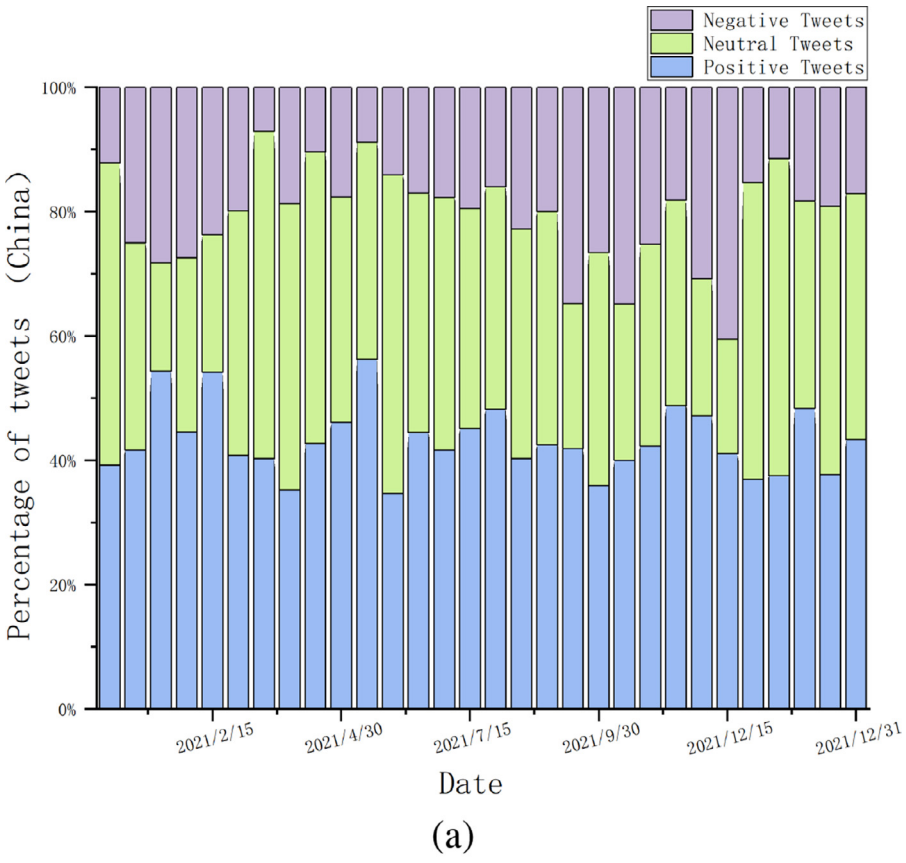


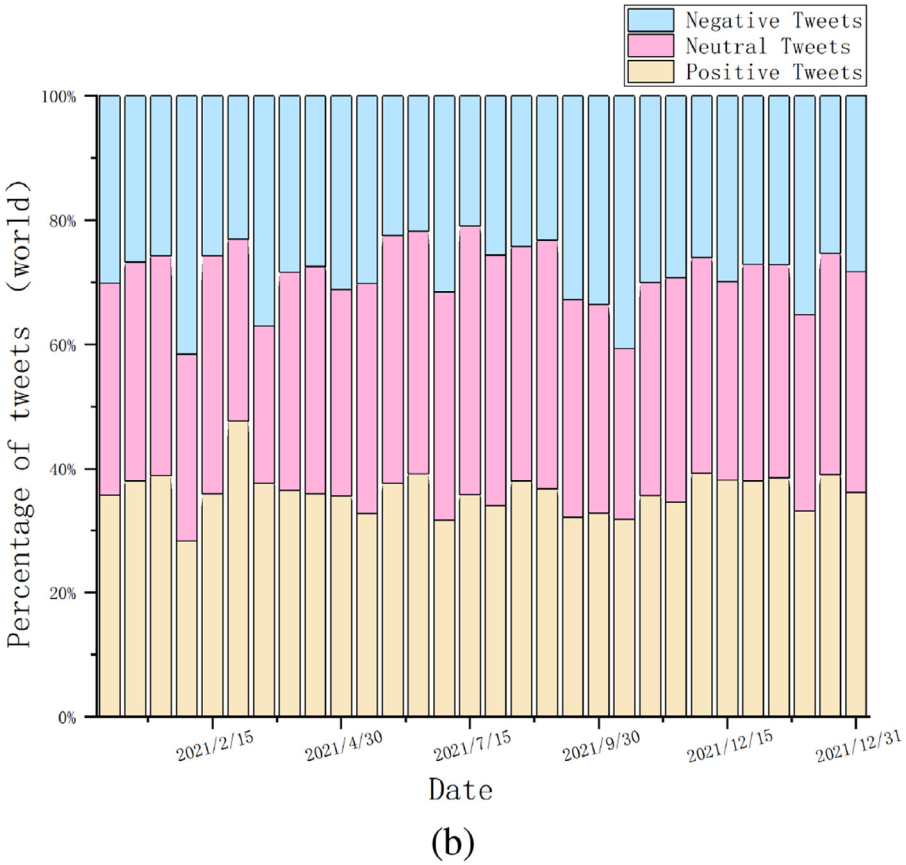
Fig. 5. The emotional value curve for the first dataset and second dataset.

in the vast number of samples discussed for COVID-19 vaccines. The number of tweets discussing the Chinese vaccine was much smaller, so when we filtered this data, we took into account this particular case.

Figs. 3 and 4 show that the number of tweets related to the Chinese COVID-19 vaccine is small and only nine days have more than 1000 tweets. The volume of the second set is huge, averaging more than 13,446 tweets a day. However, with the advancement of time and the popularization of vaccination, people were more active in discussing the Chinese vaccine on Twitter. But there has been no significant increase in discussions about vaccines overall. It can be seen that when some countries first started vaccinating against the novel coronavirus, the global buzz about the issue became apparent, and people still have a high interest in the issue for more than one year.



(a)



(b)

Fig. 6. Proportion of positive, negative and neutral tweets in the 2 sets. (a) Set 1. (b) Set 2.

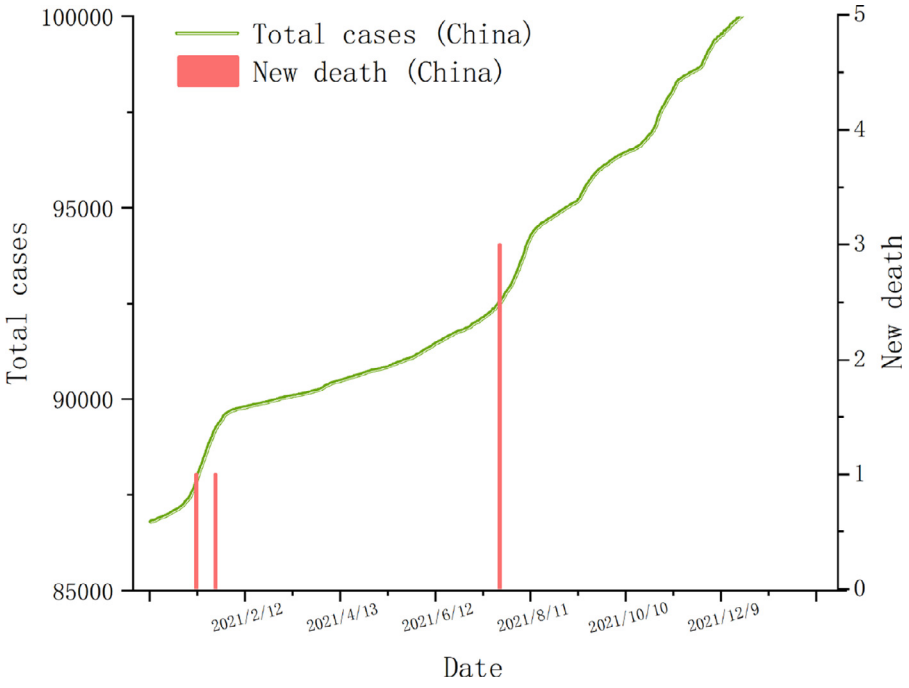


Fig. 7. Total confirmed cases, new deaths in China.

4.2. Vaccine tweets sentiment analysis

4.2.1. Differences in sentiment values between Chinese vaccine and other countries vaccine

Fig. 5 shows that for each set, the emotional value curve is mostly located above 0. It can be seen that the public sentiment to the COVID-19 vaccine is mainly positive, which can reflect people's demand for vaccines to a certain extent. People are more optimistic about vaccine development and vaccination, and trust their quality. Specifically, people's feelings about COVID-19 vaccines were higher at the beginning of the global vaccination campaign. It can be inferred that there was a high level of trust in vaccines and an expectation that vaccines could effectively contain the spread of the virus. With the rapid increase in the number of people vaccinated, there has been an increase in the number of cases of health damage as a result of vaccination, although the incidence of such problems remains low. Due to the media and netizens attention to relevant risk events on social media platforms, the possible risks of vaccination will be gradually recognized by people to a certain extent. For example, the Website of NOMA 2021 reported a total of 23 suspected deaths after vaccination on 18 January 2021. Over the next three days, 4467 tweets directly referred to the incident. And research continues to show that mutations in the virus can make vaccines less effective. On the same day, the Gupta Laboratory at the University of Cambridge confirmed that Pfizer's and BioNTech's COVID-19 vaccine was significantly less protective against COVID-19 mutant strains. Therefore, we can conclude that the increased awareness of the risks and effectiveness of vaccines may be one of the reasons for the overall decrease in netizens' emotional value of vaccines around February 15, 2021, three months after vaccination began. But the overall emotional value is still positive or neutral. The value also rose by the end of 2021, so people are still relatively positive about the effect of COVID-19 vaccine overall. But a comparison reveals a clear shift in sentiment towards Chinese vaccines. From the Fig. 5, the portion of the curve below 0 in first dataset is significantly more than in second dataset. In the early stage of vaccine promotion, people's attitude towards the rapidly emerging Chinese vaccine was obviously lower than that of vaccines produced in other countries, and their trust in Chinese vaccine was also not very adequate. However, with the witness of time, the relatively stable quality of Chinese vaccines and China's epidemic prevention policy have gradually been recognized by netizens from all over the world, and even their emotional value was once higher than that of other countries.

Table 3
Proportion of positive, negative and neutral tweets in the first set and second set.

	PP(pos)	PP(neg)	PP(neu)
Set 1	42.80%	21.12%	36.08%
Set 2	36.53%	28.28%	35.19%

As can be seen from the Fig. 6, the number of positive tweets and neutral tweets for each set both account for a large proportion. This speaks to the general trust in vaccines, but it also shows that the diversity of opinions on social media platforms is inevitable.

Table 3 shows the number and of positive, negative and neutral tweets in the two sets. We can see that people are more positive about Chinese vaccines (covering 42.80% of tweets in dataset 1) than other countries' COVID-19 vaccines (covering 36.53% of tweets in dataset 2). But people were also more likely to be negative about the Chinese vaccines (covering 36.08% of tweets in dataset 1) than about the other (covering 35.19% of tweets in dataset 2).

This also suggests that attitudes towards Chinese vaccines differ significantly, but this may be limited by sample size. Because fewer people chose to tweet about the Chinese vaccines than about vaccines in general. At the same time, there is no denying that people who pay more attention to or have a more distinct attitude towards Vaccines in China are more likely to post relevant information on Twitter.

4.2.2. The relationship between sentiment value and the number of daily new cases or daily new deaths

Vaccine can generally be divided into two categories, namely, preventive vaccine and therapeutic vaccine. The COVID-19 vaccine is a preventive vaccine. Experts dedicated to vaccine research generally said that vaccination of COVID-19 can increase antibodies to human body, thereby reducing the risk of infection with virus. In addition, the COVID-19 vaccine has a strong rate of severe protection, that is, the probability of severe infection after vaccination is much smaller. In this regard, we assume that the number of confirmed people and the number of deaths can reflect the effectiveness of the vaccine to a certain extent, and their

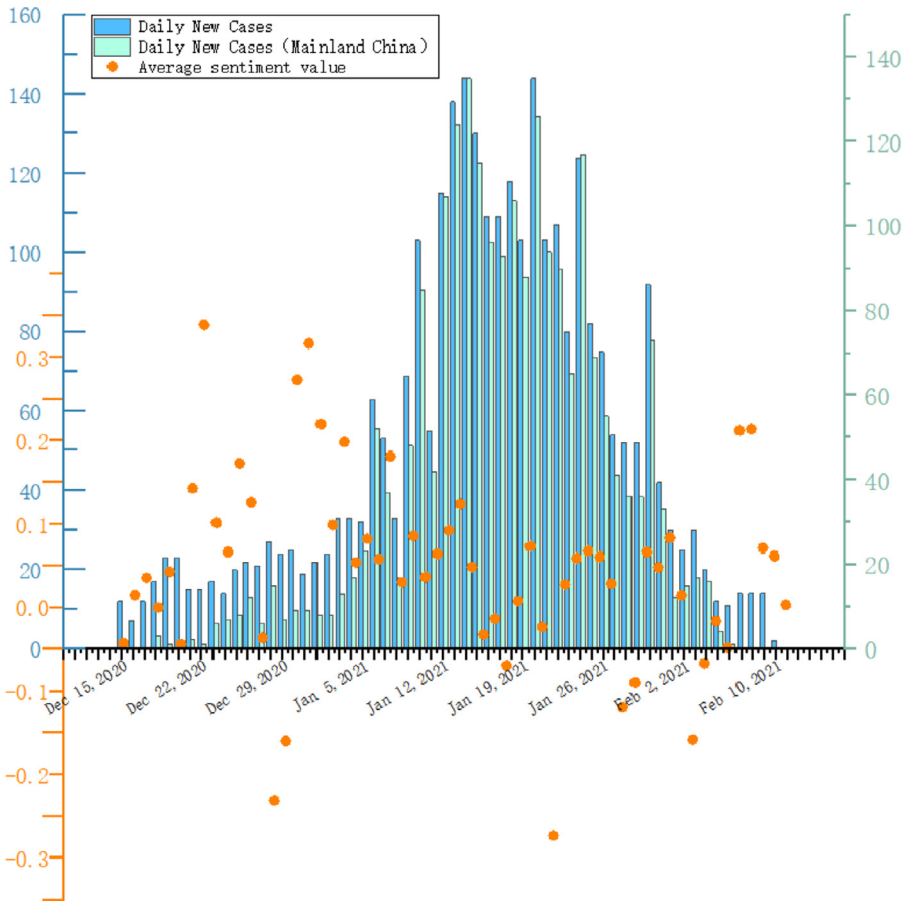


Fig. 8. Daily confirmed cases, local confirmed cases and sentiment value in China from December 15rd, 2020 to February 10rd, 2021.

changes will change the public's awareness of the vaccine to a certain extent, thereby affecting sentiment.

From Fig. 7 we can see that during the time period selected for this study, there were no deaths in China, according to WHO dataset, except for January 13rd, 2021, January 15rd, 2021 and July 23rd, 2021.

For example, in January 2021, a small epidemic broke out in Hebei Province, China, and two death case occurred. From Fig. 8 we can see that during these days, the sentiment value is greater than 0. But in the three days after the death, the average sentiment value has a significant sustained decline. Especially on January 28th, the value decreases to negative and forms the first trough. Through the word frequency statistics and word cloud analysis of the tweets on January 28th in set 1, it can be seen from Fig. 9 that although the emergence of new deaths has been over three days, there is still a certain amount of discussion on this event in Twitter. It can be inferred that the public discussion of death cases can be attributed to the strong autonomy of the content they publish, and the emergence of local cases in China will affect the public cognition on China's epidemic, to a certain extent, and then have an emotional impact on the vaccine, policy and other sub-issues.

It is difficult to visualize the relationship between sentiment value and the number of daily confirmed cases and deaths in the world in Fig. 10. Here we use Pearson correlation coefficient to analyze the relationship among the above three variables, according to (2).

$$r_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y} \tag{2}$$

The calculating results are listed in Table 4, from which we can find there was no significant correlation between sentiment value and the deaths per day. It is speculated that people's cognition of vaccine efficacy was mainly focused on the prevention of infection. However, the sentiment value is significantly negatively correlated with the number of daily new cases. When the number of newly diagnosed people increases, negative emotions will also increase. It is noteworthy that vaccination takes a long time to produce effectiveness, so the scientific of exploring the relationship between the number of confirmed cases and emotional fluctuations under the cross-section is questionable.

Table 4
Correlation analysis between sentiment value and the number of daily new cases and deaths in the world.

		Cases	Deaths
Sentiment	Pearson Correlation	−0.294*	−0.152
	Sig. (2-tailed)	0.025	0.253

*Correlation is significant at the 0.05 level(2-tailed).

4.3. Topic analysis based on LDA

4.3.1. Determine the number of topics

In order to better analyze the tweets related to the COVID-19 vaccine topic, we chose to use the LDA model to conduct topic analysis on all the data. When using the LDA model, it is critical to use scientific methods to determine the optimal number of topics. We calculate the perplexity to determine the number of topics, which is also a commonly used method at present [36]. The perplexity can be used to evaluate the applicability of the LDA topic model and to quantitatively evaluate the performance of the model [37]. The calculation method is as follows:

$$\mathrm{Perplexity}\ (D) = \exp\left(-\frac{\sum_{m=1}^{M} \log_D p(w_m)}{\sum_{m=1}^{M} N_m}\right) \tag{3}$$

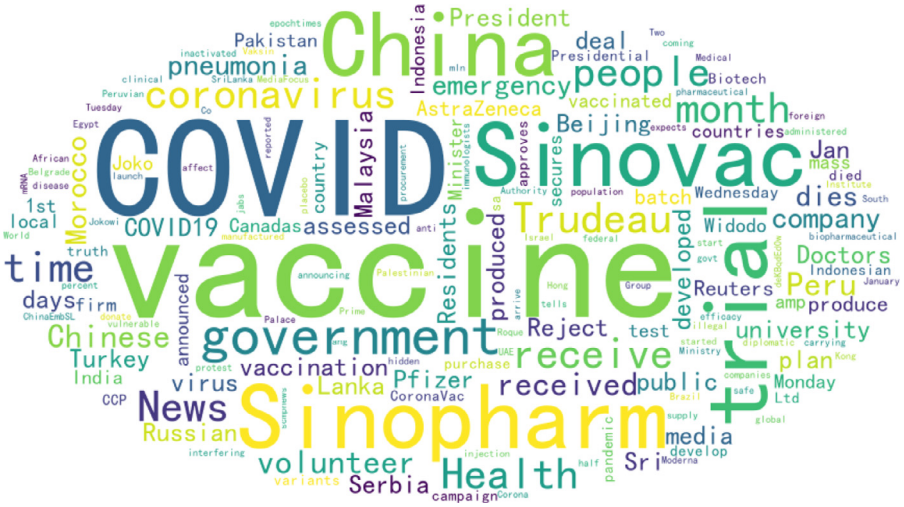


Fig. 9. In a word cloud of tweets related to COVID-19 in China on Jan. 28th, 2021, the word "dies" is more pronounced at the top right.

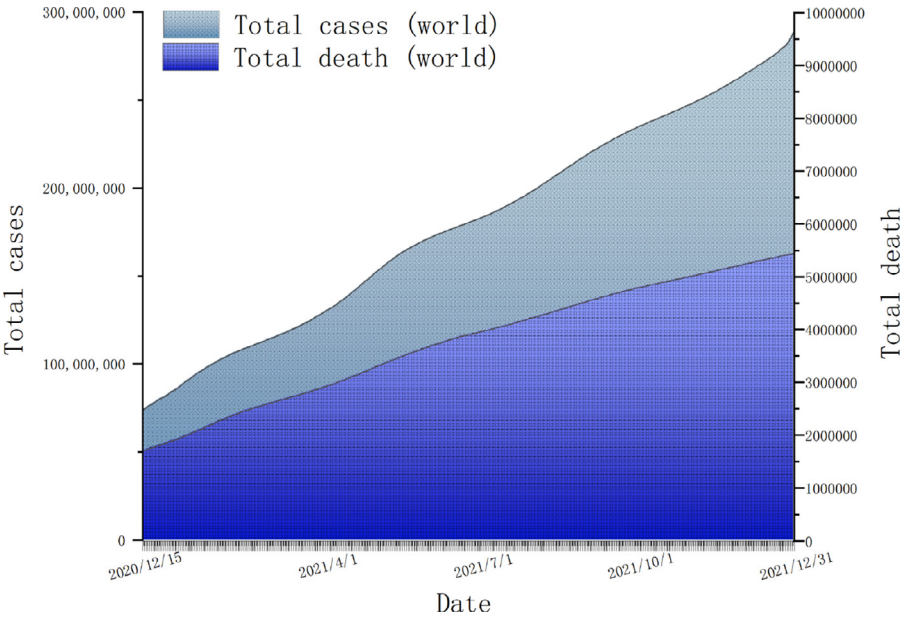


Fig. 10. Total confirmed cases, total deaths in world.

$$p(w_m) = \sum_{d} \prod_{n=1}^{T} \sum_{j=1}^{T} p(w_j \mid z_j = j) \cdot p(z_j = j \mid w_m) \cdot p(d) \tag{4}$$

Here D represents the test set in the corpus, M represents the number of documents, N_m represents the number of words held in document m, and $p(w_m)$ represents the probability of w_m generation.

Based on this, we iterated over a number of topics. The lower the degree of confusion is, the stronger the model generalization ability is. And when the change of confusion slows down within a certain range, the value can be expressed as the optimal number of topics in the data set. Fig. 11 shows the perplexity for different number of topics. It can be seen that the number of topics changes significantly at 10 and tends to be stable before that. Therefore, we can get the optimal number of topics to be 10.

4.3.2. Results of subject classification

Fig. 12 is the cloud map of data high-frequency words. It can be seen that discussions on COVID-19 vaccines focused on "Pfizer" and "Astrazeneca", while they also talked more about "mRna". This also shows that Chinese vaccines are far less popular on world social media platforms than vaccines produced in other countries, especially the United States, and inactivated vaccines are also less popular than mRna vaccines. In addition, we can know whether the virus "variant" and whether vaccines can be "effective", "free" and "protect", which also reflects people's concern about the effectiveness of vaccines against mutated viruses. At the same time, "risk" and "death" can also indicate distrust of vaccines. The frequent use of "community," "FDA," "country," and "government" reflects the netizens' interest in quarantine policies and government actions.

Table 5 shows the 10 topics and keywords derived from the classification.

Topic intensity refers to the proportion of the word frequency corresponding to a topic in the word frequency of all subject words in a certain period of time [38]. As shown in Fig. 13, during the period of this study, topic 3, 5, 6, 7 and 8 have a high intensity, that is, related themes have attracted high attention.

Of these, topic 5 received the most attention. Based on keywords such as "pfizer", "patent", "protection", "shared", "provide", and some developing countries, we can speculate that it is related to the vaccine production, vaccine assistance and vaccination effectiveness with Pfizer as the main body. The content related to Pfizer attracted constant attention, because "Pfizer" became the keyword with the highest frequency in

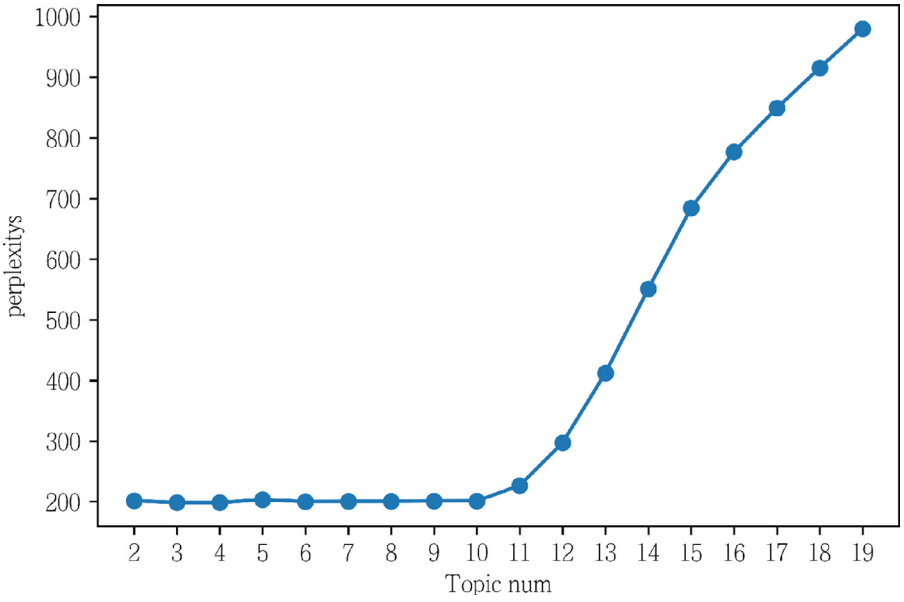


Fig. 11. The perplexity for different number of topics.

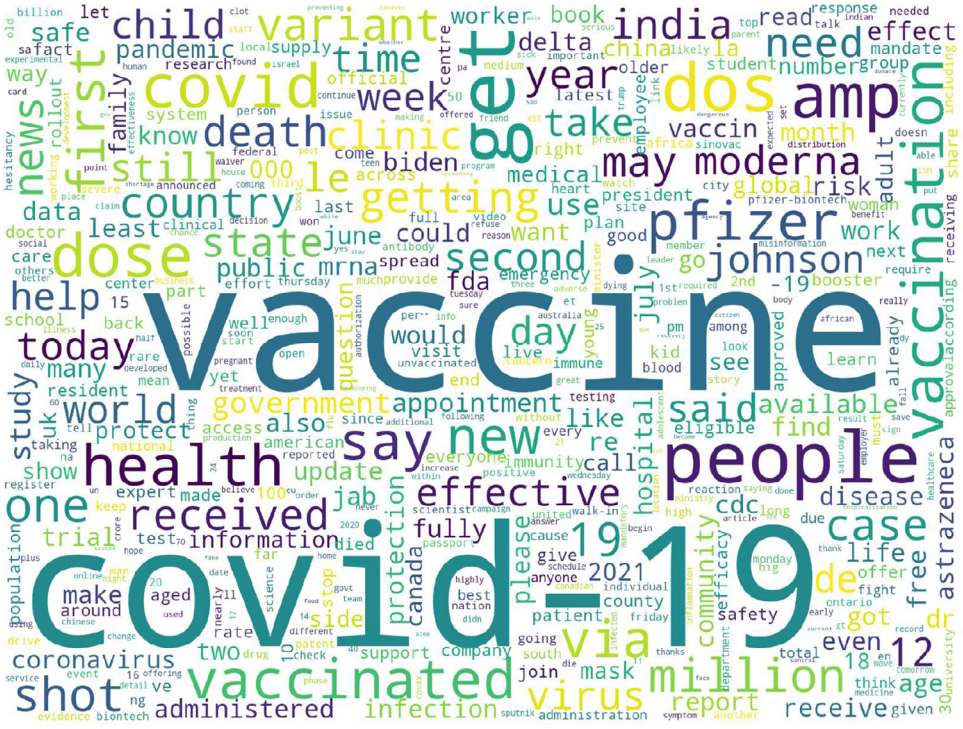


Fig. 12. The cloud map of data high-frequency words.

topic 5. As one of the first vaccines approved for marketing, Pfizer's vaccine has been concerned because of the controversy of effectiveness and adverse reactions. While paying attention to the efficacy of vaccines, the patent and technical issues of vaccine production are also highlighted by netizens. The production speed, vaccination effect and approval speed of vaccines also reveal the biological science and technology level of an enterprise or even a country. This can also reflect that the competition of science and technology, as one of the contents of the competition of comprehensive national strength, not only attracts the attention of leaders of various countries, but also the topic is sinking and widely concerned by the masses. But worldwide, the gap between countries in the level of biotechnology is very wide, and vaccine assistance programs have become the focus of widespread concern. Take Pfizer as an example, this brand of vaccine has also been announced as one of the major vaccines that the United States provides to poor countries and regions in the world, the implementation of this commitment is also a hot topic of public concern.

As the first vaccine approved for vaccination, Pfizer has gained high attention from netizens. Meanwhile, other brands of vaccines are also frequently mentioned. According to analysis, topic 3 is likely to focus on pharmaceutical companies, mainly "Astrazeneca" and Chinese vaccine companies. "Effect" and "supply" are the main concerns. As a result, Chinese vaccines have been widely compared with those produced in other countries since the beginning of the phase-in of vaccination worldwide. Based on previous research on vaccine sentiment in China, we can guess that the public's attitude towards Chinese vaccines and even Chinese

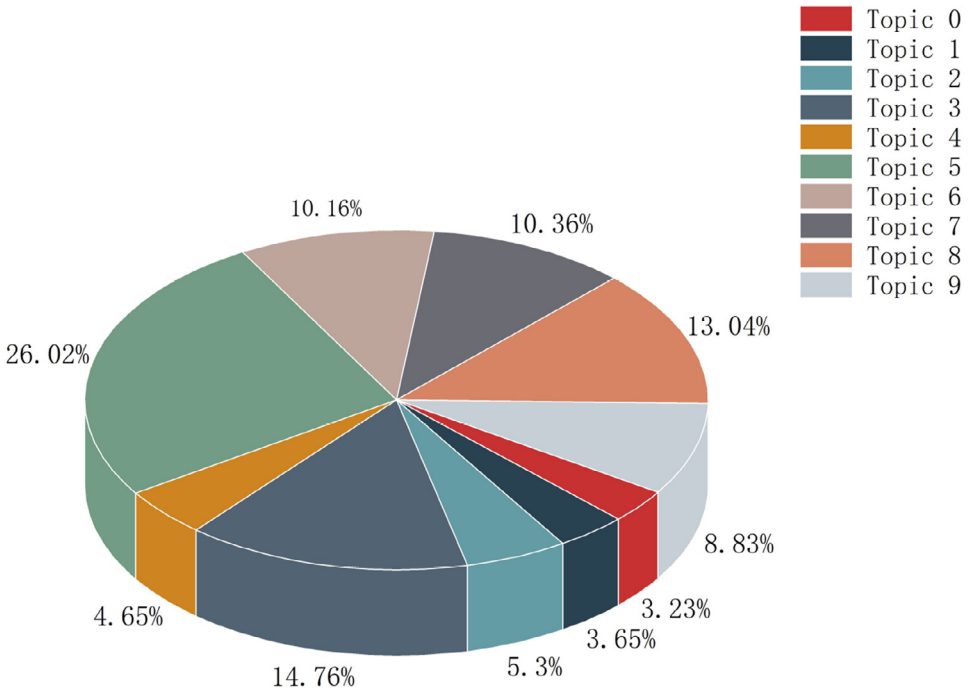


Fig. 13. The intensity of each topic.

Table 5
The 10 topics and keywords.

Topic 0	Topic 1	Topic 2	Topic 3	Topic 4
use	may	vaccin	astrazeneca	health
moderna	variant	spot	effect	biden
fda	virus	eligibility	supply	public
emergency	study	line	pharma	access
expert	effective	app	immunization	worker
safety	risk	booking	register	resident
mrna	asap	delay	China	hospital
trial	disease	ang	shortage	president
company	fully	plus	saturday	school
drug	immunity	spreading	receives	bill
Topic 5	Topic 6	Topic 7	Topic 8	Topic 9
pfizer	amp	first	get	India
patent	johnson	vaccination	protect	new
protection	mass	appointment	safe	news
biontech	deal	free	via	state
older	wto	adult	doctor	case
Israel	manufacturing	available	vaccinated	death
offer	support	age	getting	day
Kenyan	global	schedule	pandemic	update
protected	waiver	clinic	help	report
shared	property	eligible	need	official

biotechnology is relatively positive, which also reflects China's improving image in the world. This also reflects the necessity of the two sets of data selected in this study.

By analyzing the high-frequency keywords of topic 8, especially "protect", "safe", "need", we can basically determine that this topic is closely related to the safety and effect of COVID-19 vaccine vaccination, and also involves vaccine assistance to a certain extent. Thus, the previously assumed problem of trust in vaccines does exist, and the possible dangers of vaccination have also been confirmed. But combined with previous sentiment analysis results and the number of vaccinations, it was clear that despite concerns about vaccine safety, people could generally be trusted to agree to be vaccinated. At the same time, the imbalance of vaccine supply around the world is still a concern and relevant enterprises, and countries should also provide humanitarian assistance to countries in need.

Topic 7 is likely to focus on conditions for COVID-19 vaccination. Because it talks about eligibility for vaccination such as age, and it talks about the price and timing of vaccination. While most countries, including China, Japan and Singapore, offer free vaccinations, citizens of some countries have to pay for them, even at a high price. Therefore, whether the vaccine can be given for free may also be the focus of netizens. Depending on the type of vaccine, more vaccines on the market require multiple doses over a period of time. At the same time, as the virus continues to mutate, more and more people began to get booster shots, and the group of vaccinations has expanded as the vaccine technology has matured. Social media platforms are a good way to know when vaccines will be administered.

It can be inferred from the keyword "deal", "manufacturing", "property", "WTO" of topic 6 that the topic is related to economy. This shows that in the context of the COVID-19 pandemic, people are concerned about economic issues as well as topics directly related to vaccines. The ravaging of the epidemic will inevitably put pressure on the growth of global economy and trade, and further impact on social development and stability. We need to understand that the longer the spread and containment of the epidemic is delayed, the more serious the damage it will do to global economic development and society. So we need to work to build trust in vaccines and further influence their willingness to be vaccinated, and to develop effective vaccines. In addition to the epidemic prevention and control, timely and effective measures should be taken to maintain economic development and social stability.

Topic evolution is the variation rule of a topic term in a time series, which can be expressed by the topic intensity in a certain period. Fig. 14 shows the heat changes of the 10 topics mentioned above during the study period.

As can be seen, topic 5 has been the most popular topic among people. It can also be concluded that Pfizer is famous for the COVID-19 vaccine, although the corresponding sound is good or bad needs further calculation and judgment. The heat of topic 3 and topic 8 increases with the progress of time, which shows that vaccine manufacturers except Pfizer begin to enter people's vision and gradually gain attention, which also reflects the tendency of globalization in people's vision. Meanwhile, the safety and efficacy of vaccination have been widely mentioned as clinical data and related studies have increased. On the other hand, the topic

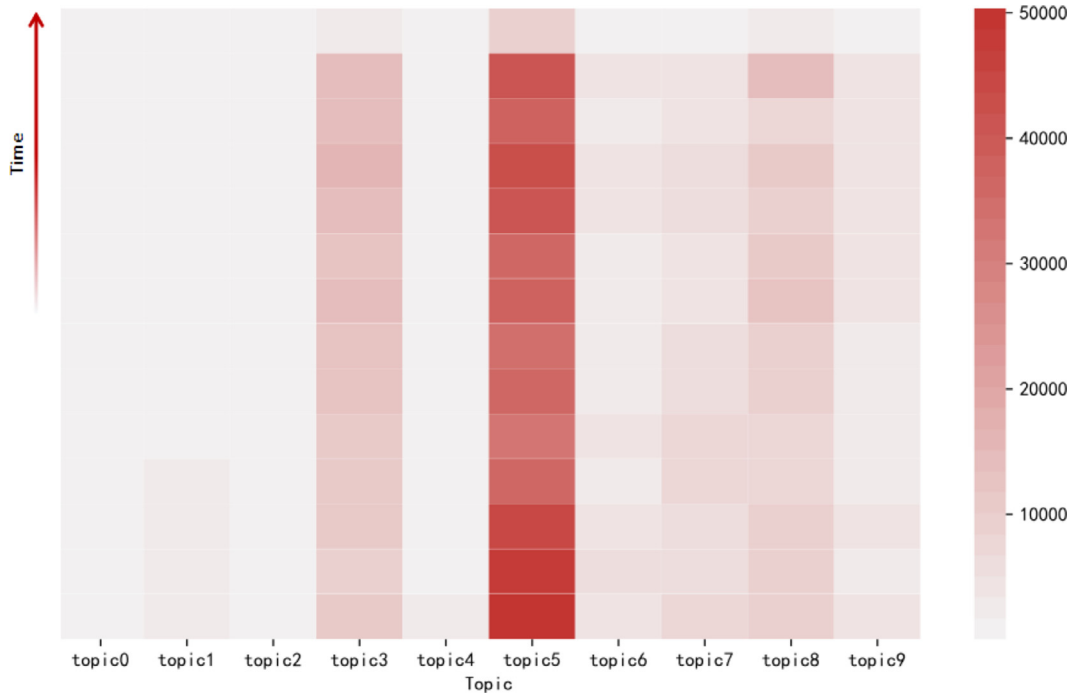


Fig. 14. The heat evolution of various topics.

1 of vaccine risk and immunity is declining. Combined with Fig. 5, it can be inferred that people were concerned about vaccine risk at the beginning, but held a positive attitude. But over time, despite their fluctuating attitudes, their feelings about the vaccine shifted to a more stable neutral range. Presumably, to some extent, people had high expectations for the vaccine's effectiveness, but the outcome was disappointing. Similarly, topic 7 is on the wane. With the popularization of vaccination and the deepening of research on the effect of vaccination, restrictions on who can be vaccinated are gradually being lifted, which may be the reason for the decreasing popularity of topic 7.

It is noteworthy that through the extraction of themes and keywords, we find that netizens have special online social media expression habits. First, abbreviations such as "FDA", "EST", "ANG", "WTO" are widely used. Secondly, internet slang like "ASAP" has gained popularity. This also reflects that network terms aim to express the maximum amount of information in the most effective and direct way, and the principle of saving effort is the starting point of their formation.

5. Conclusion

The information dissemination during this epidemic shows the characteristics of multiple subjects of discourse, the interweaving of rational and emotional cognition of discourse texts, and the complexity of topic content. How to effectively conduct emotional guidance and disseminate scientific information on social platforms is very important. Based on the VADER model, this paper conducts a sentiment analysis and social network analysis on the online social media platform Twitter. We found that people have different sentiments between Chinese vaccine and those in other countries. Public sentiment towards COVID-19 vaccine may be influenced by the number of new cases and new deaths. At the same time, through the theme analysis of the LDA model, we also get the topics that people are widely concerned about in the COVID-19 vaccine issue, which also helps us to further understand the focus of public concern, Then it can help us to effectively guide the public's attention and emotions in the COVID-19 vaccine issues communication network. In order to boost public trust in COVID-19 vaccines and further promote the process of epidemic prevention and control.

In addition, this article reflects the current situation of the international public opinion field under the issue of COVID-19 vaccine, indicating that soft power such as biotechnology is increasingly becoming an important part of building national image, and concluding that international assistance and humanitarianism are the common spiritual orientation of mankind. It may also help to further enrich our understanding of current theories on international relations, national image and cultural identity.

However, only studying the content on Twitter may lead to bias, and the accuracy of the model needs to be optimized. In our future work, we need to consider mining multilingual online data from more social network platforms, and it is also necessary to improve the accuracy of sentiment analysis model and topic clustering model.

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.

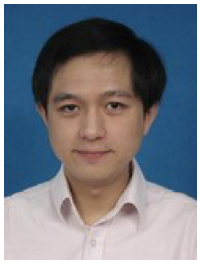
Acknowledgments

This work was supported by the Huazhong University of Science and Technology Special Funds for Development of Humanities and Social Sciences, the Open Funding Project of the State Key Laboratory of Communication Content Cognition (Grant no. 20G03), and the Open Research Project of the State Key Laboratory of Media Convergence and Communication, Communication University of China (Grant no. SKLMCC2021KF010).

References

[1] D. Wang, B. Hu, C. Hu, Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus-infected pneumonia in Wuhan, China, J. Am. Med. Assoc. 323 (2020) 1061–1069.
[2] Q. Li, X. Guan, P. Wu, X. Wang, et al., Early transmission dynamics in Wuhan, China, of novel coronavirus infected pneumonia, N. Engl. J. Med. 382 (13) (2020) 1199–1207.
[3] X. Hao, S. Cheng, D. Wu, Reconstruction of the full transmission dynamics of COVID-19 in Wuhan, Nature 584 (2020) 420.

[4] Y. Liu, A.A. Gayle, A. Wilder-Smith, J. Rocklv, The reproductive number of COVID-19 is higher compared to SARS coronavirus, J. Travel Med. 27 (2020) 1–4.
[5] N. Zhu, D. Zhang, W. Wang, X. Li, B. Yang, J. Song, X. Zhao, B. Huang, W. Shi, R. Lu, P. Niu, F. Zhan, X. Ma, et al., A novel coronavirus from patients with pneumonia in China, 2019, N. Engl. J. Med. 382 (2020) 727–733.
[6] C. Wang, P.W. Horby, F.G. Hayden, G.F. Gao, A novel coronavirus outbreak of global health concern, Lancet 395 (2020) 470–473.
[7] S. Lai, N.W. Ruktanonchai, L. Zhou, O. Prosper, W. Luo, J.R. Floyd, A. Wesolowski, M. Santillana, C. Zhang, X. Du, H. Yu, A.J. Tatem, Effect of non-pharmaceutical interventions to contain COVID-19 in China, Nature 585 (2020) 410–413.
[8] R.F. Sear, N. Velsquez, R. Leahy, N.J. Restrepo, S.E. Oud, N. Gabriel, Y. Lupu, N.F. Johnson, Quantifying COVID-19 content in the online health opinion war using machine learning, IEEE Access 8 (2020) 91886–91893.
[9] K.F. Brown, J.S. Kroll, M.J. Hudson, M. Ramsay, J. Green, S.J. Long, C.A. Vincent, G. Fraser, N. Sevdalis, Factors underlying parental decisions about combination childhood vaccinations including MMR: a systematic review, Vaccine 28 (2010) 4235–4248.
[10] W.-Y.S. Chou, A. Budenz, Considering emotion in COVID-19 vaccine communication: addressing vaccine hesitancy and fostering vaccine confidence, Health Commun. 35 (2020) 1718–1722.
[11] S. Li, Y. Wang, J. Xue, The impact of COVID-19 epidemic declaration on psychological consequences: a study on active Weibo users, Int. J. Environ. Res. Public Health 17 (2020).
[12] T.T. Le, Z. Andreadakis, A. Kumar, R.G. Romn, S. Tollefsen, M. Saville, S. Mayhew, The COVID-19 vaccine development landscape, Nat. Rev. Drug Discov. 19 (2020) 305–306.
[13] J.F. Modlin, W.A. Orenstein, A.D. Brandling-Bennett, Current status of mumps in the United States, J. Infect. Dis. 132 (1975) 106–109.
[14] M.O. Lwin, J. Lu, A. Sheldenkar, P.J. Schulz, W. Shin, R. Gupta, Y. Yang, Global sentiments surrounding the COVID-19 pandemic on twitter: analysis of twitter trends, JMIR Public Health Surveill. 6 (2020) e19447.
[15] S. Stefan, D.-X. Linh, Emotions and information diffusion in social media-sentiment of microblogs and sharing behavior, J. Manag. Inf. Syst. 29 (2013) 217–247.
[16] K. Chakraborty, S. Bhatia, S. Bhattacharyya, J. Platos, R. Bag, A.E. Hassanien, Sentiment analysis of COVID-19 tweets by deep learning classifiers-a study to show how popularity is affecting accuracy in social media, Appl. Soft Comput. 97 (2020).
[17] Z. Li, J. Ge, M. Yang, Vicarious traumatization in the general public, members, and non-members of medical teams aiding in COVID-19 control, Brain, Behav., Immun. 88 (2020) 916–919.
[18] S. Stieglitz, D.-X. Linh, Impact and diffusion of sentiment in political communication - an empirical analysis of political weblogs, in: European Conference on Information Systems, 2012, pp. 427–430.
[19] J. Zarocostas, How to fight an infodemic, Lancet 395 (2020) 676.
[20] H.J. Larson, The biggest pandemic risk? Viral misinformation, Nature 562 (2018) 309.
[21] H. Jelodar, Y. Wang, R. Orji, S. Huang, Deep sentiment classification and topic discovery on novel coronavirus or COVID-19 online discussions: nlp using LSTM recurrent neural network approach, IEEE J. Biomed. Health Inform. 24 (10) (2020) 2733–2742.
[22] S. Kaur, P. Kaul, P.M. Zadeh, Monitoring the dynamics of emotions during COVID-19using twitter data, Procedia Comput. Sci. 177 (2020) 423–430.
[23] T. Wang, K. Lu, K.P. Chow, Q. Zhu, COVID-19 sensing: negative sentiment analysis on social media in China via bert model, IEEE Access 8 (2020) 138162–138169.
[24] A.M. Rafi, S. Rana, R. Kaur, Q.J. Wu, P. Moradian Zadeh, Understanding global reaction to the recent outbreaks of COVID-19: insights from instagram data analysis, in: IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2020, pp. 3413–3420.
[25] P. Pascual-Ferrá, N. Alperstein, D.J. Barnett, Social network analysis of COVID-19 public discourse on twitter: implications for risk communication, 2020risc, https://www.cambridge.org/core/terms.
[26] W. Ahmed, J. Vidal-Alaball, J. Downing, F.L. Seguí, COVID-19 and the 5G conspiracy theory: social network analysis of twitter data, J. Med. Internet Res. 22 (2020) e19458.
[27] P. Singh, S. Singh, M. Sohal, Y.K. Dwivedi, K.S. Kahlon, R.S. Sawhney, Psychological fear and anxiety caused by COVID-19: insights from twitter analytics, Asian J. Psychiatry 54 (2020).
[28] S. Boon-Itt, Y. Skunkan, Public perception of the COVID-19 pandemic on twitter: sentiment analysis and topic modeling study, JMIR Public Health Surveill. 6 (2020) e21978.
[29] A.S. Imran, S.M. Daudpota, Z. Kastrati, R. Batra, Cross-cultural polarity and emotion detection using sentiment analysis and deep learning on COVID-19 related tweets, IEEE Access 8 (2020) 181074–181090.
[30] A. Mourad, A. Srour, H. Harmanani, Critical impact of social networks infodemic on defeating coronavirus COVID-19 pandemic: twitter-based study and research directions, IEEE Trans. Netw. Serv. Manag. 17 (2020) 2145–2155.
[31] L.-A. Cotfas, C. Delcea, I. Roxin, C. Ioan, D.S. Gherai, F. Tajariol, The longest month: analyzing COVID-19 vaccination opinions dynamics from tweets in the month following the first vaccine announcement, IEEE Access 9 (2021) 33203–33223.
[32] H. Yin, X. Song, S. Yang, J. Li, Sentiment analysis and topic modeling for COVID-19vaccine discussions, World Wide Web-Internet and Web Information Systems, 2022.
[33] C. Hutto, E. Gilbert, Vader: a parsimonious rule-based model for sentiment analysis of social media text, in: Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media, 2015.
[34] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent dirichlet allocation, The Journal of Machine Learning Research 3 (2003) 993–1022.
[35] T. Xiaobo, X. Kun, Hotspot mining based on LDA model and microblog heat, Libr. Inf. Serv. 58 (5) (2014) 58–63.
[36] M.J. Huang Ling, C. Chunling, Topic detection from microblogs using T-LDA and perplexity, in: 2017 24th ASIA-Pacific Software Engineering Conference Workshops (APSECW), 2017, pp. 71–77.
[37] P.P. Toral Antonio, W. Longyue, Linguistically-augmented perplexity-based data selection for language models, Comput. Speech Lang. 32 (1) (2015) 11–26.
[38] H. Laurence, A comparison of Lucene search queries evolved as text classifiers, Appl. Artif. Intell. 32 (7) (2018) 768–784.



Han Xu received the B.E. degree from Wuhan University, China, in 2005, the M.S. degree from Research Institute of Post & Telecommunication (WRI), China, in 2008 and the Ph.D. degree from Huazhong University of Science and Technology, China in 2012 respectively, all in Electrical Engineering. From 2012 to 2017, he was a senior engineer with China Ship Development and Design Center (CSDDC). Since 2017, he has been an Associate Professor with the school of journalism and information communication, Huazhong University of Science and Technology. His research interests include complex network, social network analysis and information diffusion.

Ruixin Liu received the B.A degree in Communication from Lanzhou University in 2020. She is currently working toward the master degree in the school of journalism and information communication, Huazhong University of Science and Technology, Wuhan, China. Her main research interests include social network analysis and media communication.

Ziling Luo received the B.A degree in Communication, and the B.S degree in Computer Science from Huazhong University of Science and Technology in 2020. She is currently working toward the master degree in the school of journalism and information communication, HUST, Wuhan, China. Her main research interests include social network analysis and information diffusion.

Minghua Xu received the B.A. degree from the School of Journalism and Information Communication, Huazhong University of Science and Technology (HUST), in 2002, and the M.A and Ph.D. degrees from Social Science Department, National University of Singapore, in 2005 and 2010, respectively. She is currently a professor with the School of Journalism and Information Communication, HUST. Her research interests include information communication theories and the theory and practice of information diffusion in social networks.