

# Generalizing factors of COVID-19 vaccine attitudes in different regions: A summary generation and topic modeling approach

DIGITAL HEALTH  
Volume 9: 1–16  
© The Author(s) 2023  
Article reuse guidelines:  
sagepub.com/journals-permissions  
DOI: 10.1177/20552076231188852  
journals.sagepub.com/home/dhj



Yang Liu<sup>1</sup> , Jiale Shi<sup>2</sup>, Chenxu Zhao<sup>2</sup> and Chengzhi Zhang<sup>3</sup>

## Abstract

**Objective:** The goal of this study is to use summary generation and topic modeling to identify factors contributing to vaccine attitudes for three different vaccine brands, with the aim of generalizing these factors across different regions.

**Methods:** A total of 5562 tweets about three vaccine brands (Sinovac, AstraZeneca, and Pfizer) were collected from 14 December 2020 to 30 December 2021. BERTopic clustering is used to group the tweets into topics, and then contrastive learning (CL) is adopted to generate summaries of each topic. The main content of each topic is generalized into three factors that contribute to vaccine attitudes: vaccine-related factors, health system-related factors, and individual social attributes.

**Results:** BERTopic clustering outperforms Latent Dirichlet Allocation clustering in our analysis. It can also be found that using CL for summary generation helped to better model the topics, particularly at the center-point of the clustering. Our model identifies three main factors contributing to vaccine attitudes that are consistent across different regions.

**Conclusions:** Our study demonstrates the effectiveness of deep learning methods for identifying factors contributing to vaccine attitudes in different regions. By determining these factors, policymakers and medical institutions can develop more effective strategies for addressing concerns related to the vaccination process.

## Keywords

Social media, vaccine attitudes, summary generation, topic modeling, different regions

Submission date: 11 March 2023; Acceptance date: 26 June 2023

## Introduction

The response to the COVID-19 pandemic has severely disrupted the daily lives of people and organizations worldwide<sup>1</sup>; governments are making the vaccine available to all in the hope of curbing the spread of the disease and ending the pandemic.<sup>2</sup> Given the severity of COVID-19, a vaccine must be developed. Despite significant efforts in clinical trials and the remarkable success of the vaccine in terms of safety and efficacy,<sup>3</sup> a substantial obstacle to achieving global vaccination coverage is vaccine attitudes.<sup>4</sup> Besides, the World Health Organization (WHO) has identified vaccine hesitancy as one of the top 10 threats to global health care in 2020. Thus, it is crucial to investigate the factor of the vaccination process.<sup>5</sup> In particular, identifying positions that may lead to factors about vaccination, that several adverse events on social media relate to vaccine

safety and can undermine vaccine confidence.<sup>6</sup> As a result, analyzing the attitudes to vaccine hesitancy reflected by users on social media is important.

Social media has become an enormous source of data due to its rapid information dissemination, and many people express their opinions through social media platforms.<sup>7</sup>

<sup>1</sup>School of Information Management, Wuhan University, Wuhan, China

<sup>2</sup>School of Computer Science, Wuhan University, Wuhan, China

<sup>3</sup>Department of Information Management, Nanjing University of Science & Technology, Nanjing, China

### Corresponding author:

Yang Liu, School of Information Management, Wuhan University, Wuhan 430072, China.

Email: yang.liu27@whu.edu.cn



Text mining on social media is a viable option for analyzing public opinion.<sup>8</sup> Due to the rapid development of artificial intelligence and natural language processing (NLP),<sup>9</sup> this technology can address analyzing public opinion. Meanwhile, Internet users need to read a minimum number of words to adapt to today's fast-paced life and glean the thrust of an article. Users can deploy automatic text summarization techniques to precise the main contents,<sup>10</sup> saving reading time, and improving the efficiency of information uptake. Hence, the tweet-oriented text auto-summarization model proposed in the present research is of great significance.

In the corresponding automatic summarization technology, the research also incorporates contrastive learning (CL),<sup>11</sup> evolving from extractive summarization to generative summarization. Our work's main contribution involves applying a combination of NLP methods to vaccines developed in different countries,<sup>12</sup> which discovers the factors related to vaccine attitudes. Collecting and monitoring data from short texts are superimposed for summary generation, such as Twitter.<sup>13–15</sup> According to the literature,<sup>16</sup> three factors (vaccine-related, health system-related, and individual societal attributes) of vaccine attitudes are defined. These three factors are utilized to cluster the topics of three different brands. This can also provide various perspectives on vaccines that can benefit public health organizations in identifying factors that contribute to vaccine confidence in people.

This study community used Twitter primarily to collect data on COVID-19 vaccination. The data are based on three different brands (Sinovac, Pfizer, and AstraZeneca) of vaccine and sought to identify topics related to factors towards vaccination<sup>14</sup> in the different regions corresponding to the brand. These results best reveal the reasons for vaccine attitudes in each region. To adapt to today's fast-paced life, Internet users need to read a minimum number of words and glean the thrust of an article, therefore, the tweet-oriented text auto-summarization model proposed in this paper is important.<sup>17</sup> This study aimed to detect the main themes expressed by users of the three vaccines when expressing their positions and attitudes towards COVID-19 vaccination. These themes point to the main reasons for vaccine attitudes in China, the UK, and the USA. This could help local experts and governments better understand the factors influencing the public to vaccinate (the reasons for young people's attitudes thereto). Furthermore, our work can provide key decision-makers with the right tools to promote vaccination campaigns more broadly.

## Literature review

### *Analysis of COVID-19 vaccine attitudes on social media*

With the development and promotion of vaccines, many researchers have conducted research on social media

discussions related to attitudes towards COVID-19 vaccines.<sup>18</sup> Previous literature mostly used sentiment analysis in machine learning,<sup>19</sup> Today, the main purpose of sentiment analysis is to understand people's attitudes towards this epidemic on social media.<sup>20</sup> Li et al.<sup>21</sup> analyzed the behavior of Americans and Chinese on different social media platforms during the COVID-19 pandemic, and the results of their sentiment analysis showed significant differences in the attitudes of the people of the two countries. Most people had confidence in controlling the spread of the virus. Zhou et al.<sup>22</sup> extracted tweets from the pandemic period, analyzed the emotional dynamics of Australians, and observed the temporal changes in emotions. Yin et al.<sup>23</sup> analyzed the sentiment of tweets related to COVID-19. They found that the positive tweets slightly outnumbered the negative ones, and provided examples of tweets with similar focuses, such as securing one's home and the people who died from the COVID-19.

Moreover, with the development and promotion of vaccines, many researchers have launched research work related to COVID-19 vaccines on social media. Kwok et al.<sup>24</sup> extracted COVID-19 vaccine-related topics and emotions from Twitter in Australia, and found that two-thirds of the tweets expressed positive opinions, while the rest expressed negative opinions. Lyu et al.<sup>25</sup> identified the emotions and topics surrounding COVID-19 vaccines over a long period of time on social media, with the aim of better understanding the emotions that may affect public immunity. Bonnevie et al.<sup>26</sup> quantified the opposition to vaccines during the COVID-19 pandemic in the USA. They found that the number of people opposing vaccines on Twitter had increased significantly. The above literatures are all about analyzing vaccine-related discussions, which ignores the public's attitude towards vaccines, especially the discussion of different regions.

### *Analysis of COVID-19 vaccine attitudes using machine learning*

The COVID-19 pandemic has demonstrated the need for advanced technology to address emergency situations.<sup>3</sup> In order to address such situations, machine learning techniques can provide effective support for public health authorities and supplement traditional applications.<sup>18</sup> Tavoschi et al.<sup>27</sup> monitored public opinion on vaccine uptake in Italy using support vector machines. Bar-Lev et al.<sup>28</sup> used several machine learning methods (logistic regression, random forest, neural network, and linear regression) to assess how online content about vaccine uptake affects vaccine hesitancy. They found that hesitancy is associated with more social media traffic for most vaccine uptakes, and social media traffic improves the performance of most models. Piedrahita et al.<sup>29</sup> achieved an 85% classification accuracy using lexicon analysis and support vector machine (SVM), which found that the percentage of

neutral tweets is decreasing, while the ratio of positive and negative tweets is increasing over time. Yuan et al.<sup>30</sup> compared five machine learning algorithms for tweet classification, SVM provided the best accuracy.

The above literatures all use machine learning, but due to factors such as poor classification effect and generalization ability of machine learning,<sup>31</sup> deep learning provides a feasible solution.<sup>32</sup> Hussain et al.<sup>33</sup> used a deep learning BERT model to analyze the public sentiment towards COVID-19 in the UK and the USA. Shahid et al.<sup>34</sup> used deep learning to predict the number of confirmed cases, deaths, and fatality rates in 10 countries. The experimental results showed that the Bi-LSTM model demonstrated stronger robustness and achieved higher prediction accuracy. Devaraj et al.<sup>35</sup> used a multivariate LSTM model to predict cumulative confirmed cases and deaths in COVID-19 cases, which also showed higher accuracy in their algorithm. Shastri et al.<sup>36</sup> conducted a comparative analysis of deep learning methods for predicting COVID-19 cases in the USA and India one month in advance, and the experimental results showed that the LSTM model performed better in predicting COVID-19 compared to the other two models. All the above literature used deep learning for classification or prediction, without using deep learning for topic clustering.

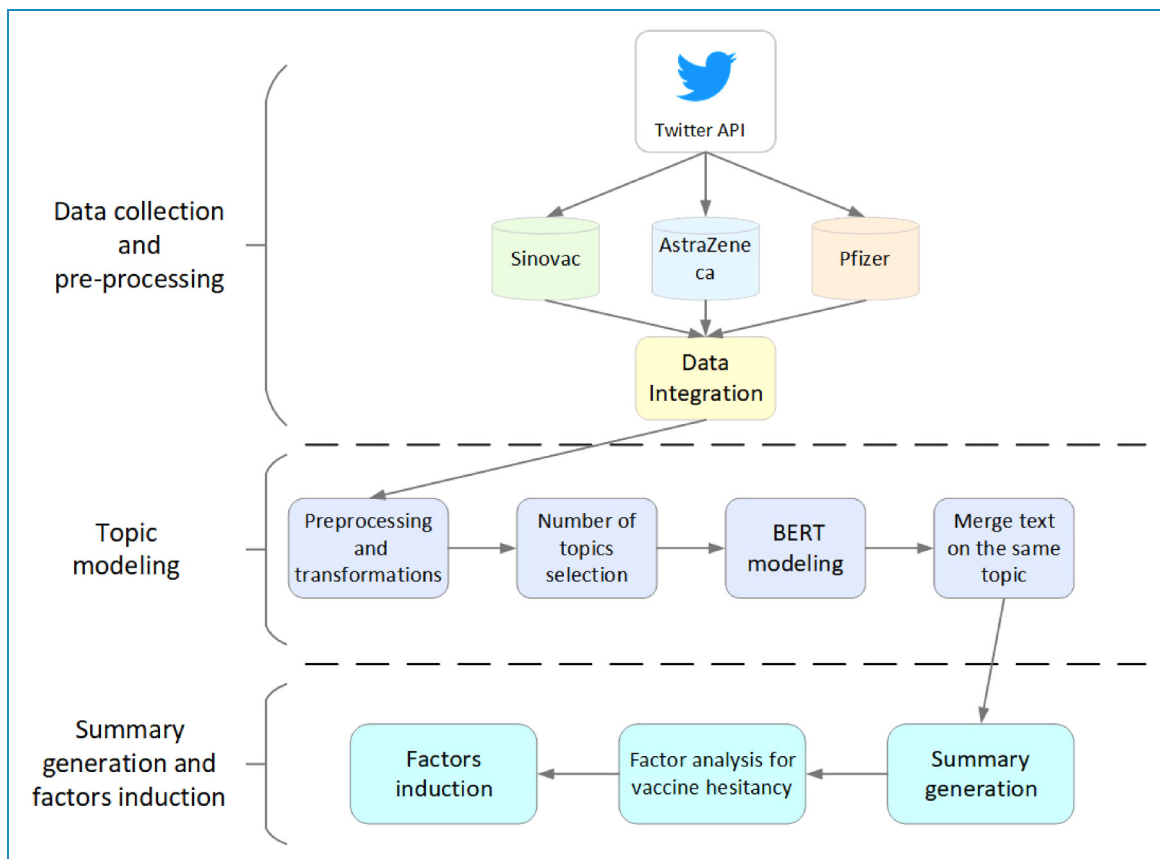
Furthermore, the summary generation technology based on deep learning also helps the topics to be better clustered. Therefore, this paper adopts text generation by CL and topic modeling by BERTopic.

## Methods

Three branded tweets about the COVID-19 vaccine were first collected. This research combined three datasets and applied two topic-modeling approaches (LDA and BERTopic) to ascertain reasons for underlying vaccine attitudes. Moreover, we revert to the original text based on the topic model, merge the texts of the same topic to generate a summary, and finally perform factorial induction (Figure 1). Each of the step-by-step vignettes is described in detail below:

### Data sources

COVID-19 vaccine-related tweets containing various predefined hashtags from 14 December 2020 (after the approval of the world's first COVID-19 vaccine) to 30 December 2021 were collected. These datasets include #CovidVaccine, # AstraZeneca, #Sinovac, and #Pfizer: 41,524 tweets were collected.



**Figure 1.** Topic modeling and summary generation pipeline.

Tweets with vaccine attitudes are also manually annotated. There are two medical experts and six graduate students, and the group is composed of annotations. Those comments are discarded, and the two experts also have different opinions about those reviews with inconsistent labels. After the students complete the annotation, the experts will review all the content and correct any erroneous annotations.

Pre-processing steps for the text include removing forwards, URLs, and punctuation, converting emojis to words, removing stop words, and stemming. After data pre-processing, the dataset contains 5562 tweets, consistent with the literature.<sup>13,14</sup>

Although the user’s account comment date is helpful for practical analysis of topic timeliness, we do not need this information at the topic-modeling stage. Therefore, we removed this information, only retaining the annotated text. Meanwhile, although pictures and emojis are helpful for sentiment analysis,<sup>37</sup> we only adopt text and leave the multimodal data to future work since the objective of the work is to analyze comments in English. Then, the data are separated by vaccine brand (Table 1).

### Topic modeling

In recent years, the Latent Dirichlet Allocation (LDA)<sup>38</sup> has been a standard method to identify the latent Dirichlet distribution in large-scale corpora. However, when dealing with short texts, it is easy to have the problem that the frequency of words cannot be the basis of the topic. The LDA model follows the bag-of-words assumption, which ignores the correlation between words. The Bidirectional Encoder Representations from Transformers (BERT)<sup>39</sup> topic model for the topic clustering of self-represented content. Thus, BERTopic clustering is used.

BERTopic relies on the attention mechanism model, using masked language model capture word and sentence-level representations, adopting noise-reducing self-encoding for model training. It can better adapt to the downstream tasks of NLP. In addition, BERTopic can be better based on short texts (less than 512 words in length) at the sentence or paragraph level processing tasks. Compared with static word embedding methods such as word2vec,<sup>40</sup>

BERT’s dynamic word embedding can also better understand the meaning of the sentence semantics.

Herein, we adopt the pre-training and fine-tuning approach of the BERTopic model<sup>41</sup> to reduce the dimensions of the resulting vectors (preserving the most important by using uniform manifold approximation and projection (UMAP)).<sup>42</sup> Moreover, clustering the reduced-dimensional content relied on a non-parametric hierarchical clustering algorithm,<sup>43</sup> which was originally developed in a different context. Cosine similarity is applied to identify the most “representative” words/phrases in a topic. Traditionally, LDA topic modeling requires a predefined number of topics, and an algorithm that clusters the corpus around several topics. BERTopic does not need a predefined  $k$ , only fine-tuning the model is required.

### Number of topics selected

The literature<sup>44</sup> has proved that topic consistency is the most consistent measure of human understanding. Topic consistency is utilized to evaluate the optimal number of topics. The main idea of topic consistency is that if the generated topics are easy to interpret, words belonging to the same topic will co-occur more frequently in the corpus. The method of calculation of topic consistency is shown below<sup>45</sup>:

$$C_k = \sum_{m=2}^M \sum_{l=1}^m \log \frac{N(v_m^k, v_l^k) + 1}{N(v_l^k)}$$

where  $V^k = (v_1^k, v_2^k, \dots, v_m^k)$  is the word space of the top- $M$  words in topic  $k$ ,  $N(v)$  represents the number of comments containing word  $v$ , and  $N(v_1, v_2, \dots)$  is the number of reviews that contain both terms  $v_1$  and  $v_2$ .

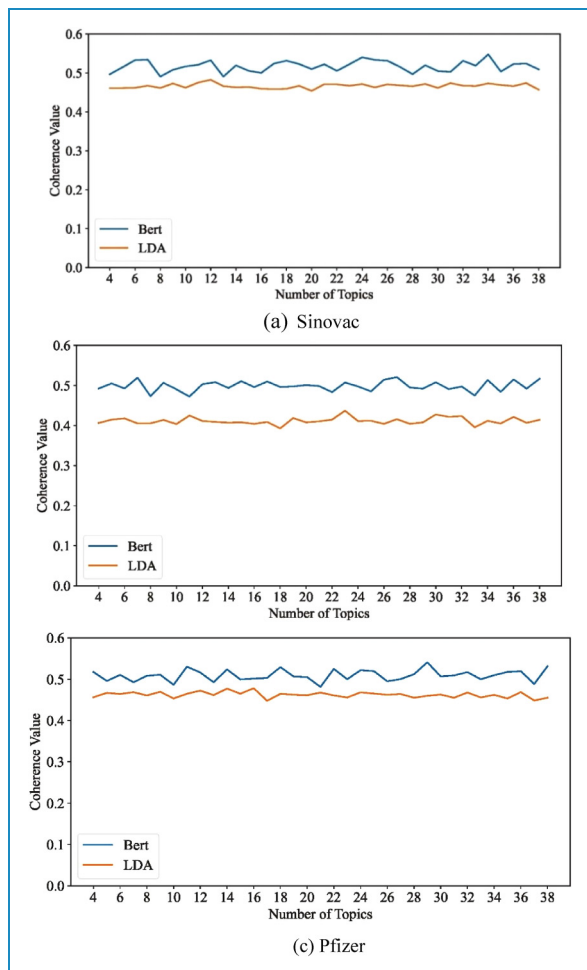
The consistency value of BERTopic under different numbers of topics is shown in Figure 2. This article also compares the consistency values of the two topic clustering results of BERTopic and LDA. The results demonstrate the consistency of LDA and BERT. The clustering performance of BERTopic is higher than that of LDA when the number of topics exceeds nine. When the peak is reached, the UK and USA brands are around 28, while Sinovac (in China) is at 35, which is different from the other three brands. American brands overlap at LDA, one at 17 and the other at 21: the interpretability clustering results become more robust when the number of themes increases, which also proves the superiority of the BERTopic method.

### Summary generation

A language representation that can be processed by a computer is obtained. Language models are used to compute arbitrary language sequences  $a_1, a_2, \dots, a_n$ . BERTopic uses a bidirectional transformer as an encoder for feature extraction, which can acquire more context information

**Table 1.** Vaccine attitudes dataset statistics.

Country	Brand	Number of tweets
China	Sinovac	1760
UK	AstraZeneca	1895
USA	Pfizer	1907
Total		5562



**Figure 2.** Clustering results of BERTopic and Latent Dirichlet Allocation (LDA). (a) Sinovac. (b) AstraZeneca. (c) Pfizer.

and significantly improve the ability of the language model to extract features.

**Encoder.** The embedding vector is also input into the encoder, which encodes it into an intermediate semantic vector. The encoder utilizes a two-layer bidirectional long short-term memory,<sup>46</sup> which can better capture the semantic dependence.

**Attention + decoder.** Since the decoder cannot use the following steps in advance when executing each step input, the decoder adopts a two-layer unidirectional long short-term memory structure. The attention mechanism<sup>47</sup> applies to the hidden states of the encoder and gets the context, which is the input. The target word is connected in series as the output of the long short-term memory on the decoder enters and loops to obtain hidden states. The hidden states are bound to perform softmax to calculate the output probability.

**CL.** CL learns the feature representation of the sample so that the feature representation is as close as possible.<sup>48</sup> Inspired by previous work,<sup>49</sup> CL is incorporated into the summarization model. Figure 3 shows the model of CL. The proxy job establishes the similarity between samples. For example, similar models are positive, and the samples that are dissimilar are negative. Data augmentation is a typical means of implementing agent tasks. This model places closer topics having ‘similar’ meaning without the specific knowledge about the topic-based distances.

Moreover, a negative sample is designed. The  $M$  quality measure candidate can be defined in many ways. The candidate summary score  $S_i$  and the reference summary  $S$  are also defined. The pre-trained abstractive model is employed to generate a summary of different candidates with varying degrees of quality, then the model is encouraged to assign higher estimation probabilities to better candidates by fine-tuning the model with contrastive loss.

$$\mathcal{L} = \sum_i \sum_{j>i} \max(0, f(S_j) - f(S_i) + \lambda_{ij})$$

where  $S_i$  and  $S_j$  are two different candidate summaries.  $\lambda_{ij}$  is the gap multiplied by the ranking difference between candidates. This loss gives the abstract model a dual purpose: as a reference-free evaluation model that can be used in a two-stage summarization pipeline to score the candidates generated by the pre-trained generation model and select their final output.

Table 2 gives examples of summary generation, and each instance contains the topic number and its corresponding summary. The summary-generation model of proposed CL generates richer, more comprehensive, and more accurate summaries of English-language documents than other models. This result indicates that the model understands the complete text, the meaning of sentences and words within the context of the sentences; it portrays the sentences and words more carefully, which aligns with human understanding.

## Result

The experimental results are divided into four parts: topic analysis, factors induction, visualization analysis in CL, and group difference analysis.

### Subject word analysis

Figure 4 shows the pre-processed dataset of the 20 most frequent words. The top 20 words include “effective, delta and variant” which differs from the vaccine brand across the dataset. This implies that the most frequent words reflect the dataset by impacting tweets and reflecting user sentiment.

### Factor induction

By comparing LDA with BERTopic, BERTopic shows the best effect in terms of topic consistency. This method

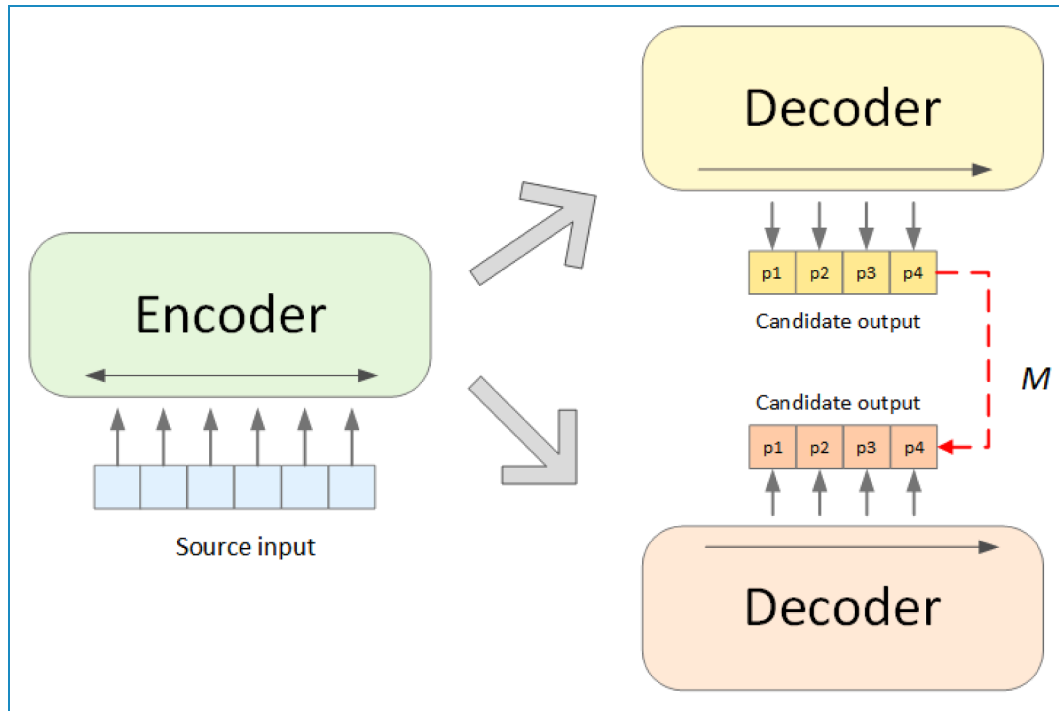


Figure 3. Summary generation based on contrastive learning (CL).<sup>49</sup>

directly outputs the most prominent keywords for each topic. Topic names for each brand are defined by first looking at the top 20 keywords for each topic. The tweets are analyzed and categorized based on three factors, as extracted from the top-five topics in each database. Since we were interested in the most commonly discussed topics in the three datasets, which also point to the reasons for the central vaccination attitudes, the obtained topics are ranked by extracting the importance of each dominant text topic. This topic ranking is shown in Tables 3(a), (b), and (c), along with the original number of topics. The order of each topic is analyzed based on these data: their optimal number of topics is 35 (Sinovac), 39 (AstraZeneca), and 30 (Pfizer). The name is then checked against the 20 most popular tweets assigned to that topic. The BERTopic model topic is a matrix of the highest frequency of each subject term from using the subject term  $H$ , which allows us to understand the content of the topic.

The first theme has a more significant proportion, which indicates a relationship with vaccines, implying that vaccines are effective. The second refers to the health system, indicating that the health system is also influencing people to try vaccines, and the third is the property of vaccines, expressing the satisfaction associated with vaccination. Comparing the three regions, the vaccine factor dominates, but the European and American ones rank higher than the Chinese brand.

The proportion of Sinovac is relatively tiny in vaccine-related factors. Since users of Twitter are mainly

in the USA and the UK, people in China tend to use it less, resulting in fewer text samples. Users discussed the vaccine cost and whether they would spend money on vaccines. They are hesitant about which vaccine to get (Sinopharm or Sinovac). In particular, Sinovac has greater governmental support, and China is consistently in a state of popular trust in the government. As for users of AstraZeneca, discussions of its validity and efficacy predominate, such as, “Our young Australians deserve a safe effective vaccine not Morrison’s clearance stock AstraZeneca.” Americans are concerned about its side effects (regarding Pfizer), with tweets such as, “I got the Pfizer vaccine today and my arm hurts a little weirdly.”

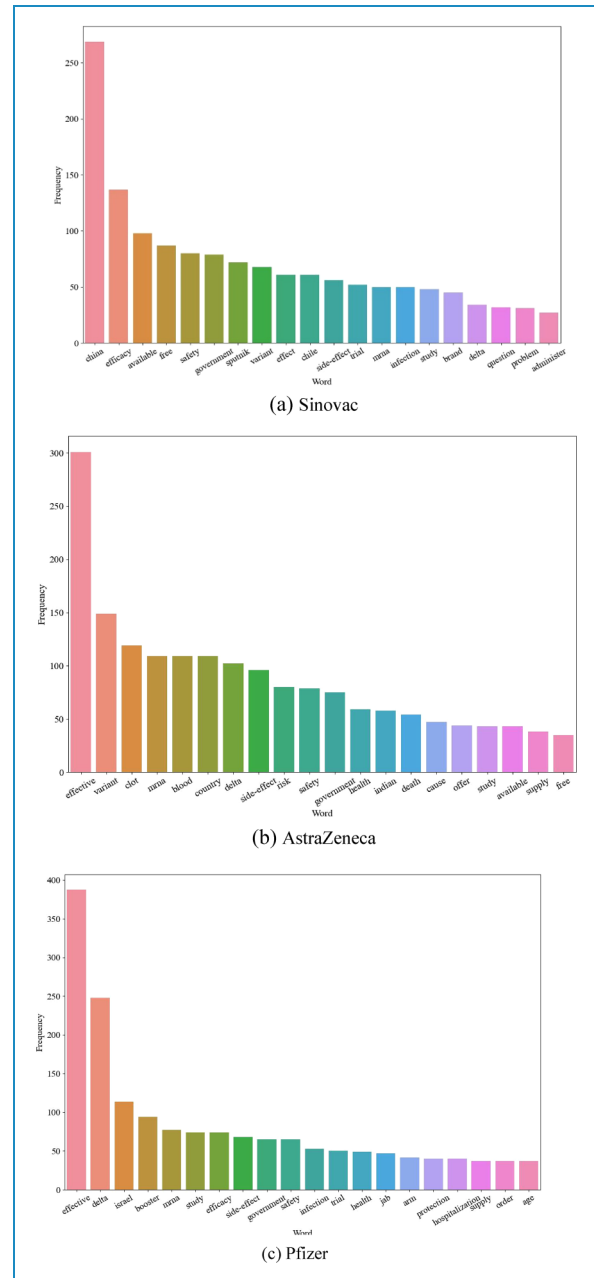
According to health system factors, Sinovac had the highest percentage of 4.9%, while the other brands were at 3.9% and 3.2%, mainly because China did not choose other vaccines, and the Chinese therefore have more reviews of Sinovac. However, Chinese vaccines are exported to South-east Asia. For instance, in Indonesia, “The Sinovac vaccine had 94% efficacy in an Indonesian study lmao.” AstraZeneca’s primary market is in the EU, which is WHO-approved, but it also sells to Thailand. For example, “SBS production capacity of AZ is 16 million per month and just 5-6 million doses will go to Thailand.” Pfizer mainly discusses the cost to the government, with important keywords such as “rich.” For example, the government paid for the research on the vaccines, and they’re paying for everyone to get the shot. There’s also the question of whether the government

**Table 2.** A typical summary generation (Pfizer).

Topic no.	Summary content
2	Pfizer says their vaccine is only 64% effective against the new Delta variant of the flu virus. The vaccine is 95% effective in the UK and 94% in the U.S. against the same strain. Israel says the vaccine is still 93% effective at preventing severe illness.
4	Pfizer vaccine is the only vaccine approved for 12 to 16 years olds in Australia. The vaccine is 95% effective against Delta and is available to everyone over the age of 12. Victoria's government is refusing to give the vaccine to the over 60s. Pfizer vaccine trials for under 12s are under way in NSW. A shortage of the vaccine in the UK has led to a delay.
5	The 23 <sup>rd</sup> Dr Pepper ingredient is the untested and pre-FDA approved Pfizer BioTech vaccine. The vaccine is a joke. Pfizer was the first successful vaccine and did not take a dime from the U.S. to create the vaccine.
6	The Taiwanese Government ordered 41 million doses of the Pfizer vaccine. The vaccine is not approved for use in the country. Pfizer spent over \$2 billion of their own money on the vaccine. CDC says the vaccine is only 64% effective for LaMDA. The company has agreed to sell the vaccine to the government.
12	Pfizer is seeking FDA approval of a booster vaccine for the flu vaccine. The company says the vaccine is losing efficacy after six months. Jackson says there is no evidence that a booster is needed. The vaccine will specifically target the variants of flu vaccine that are losing efficacy.

allows exports to the UK, for example, “EU unlike UK is arranging for the export of vaccines to covac countries.”

The factor of individual’s social attributes is mainly the impact of social events, such as the situation in the Middle East. “Please help and request to Saudi Arabia government to accept individual’s social attributes these suitable and safest vaccine in Saudi,” and “Chile used the Sinovac vaccine. It’s not that effective.” These sentences arise mainly from the Middle East and South American countries with relatively good diplomatic relations with China. The countries where AstraZeneca mainly plays are the USA and India. The main political factor is that both countries were formerly British colonies and had the same political system. However, Pfizer targets children and needs to improve its effectiveness in the USA. For instance, “my little sister bravely participated in a Pfizer vaccine trial for kids.”



**Figure 4.** Frequency of top-20 words by brand. (a) Sinovac. (b) AstraZeneca. (c) Pfizer.

### Visualization analysis in CL

To demonstrate the helpfulness of the framework of the CL model, the UMAP as a dimension-reduction technique is used. This algorithm is in three datasets by the embedding layer, with high-dimensional feature vectors for dimensionality reduction visualization. As shown in Figure 5, the distribution of topic models in the embedding space is reported. Before topic analysis, it is necessary to pre-train the text, specifically embedding the text and vectorizing

**Table 3.** (a) Top-five topics related to Sinovac by number of tweets.

Factor	Topic-ID	Proportion	Keywords	Example
Vaccine-related	2	7.4%	China, Sinovac, Sinopharm, people, country, covid, just, want, Sinovac vaccine China	Sinovac is not welcome here. No to China made vaccines.
	4	4.1%	good, effective, best, available, Sinovac vaccine good, vaccines, Sinopharm, infection, high, want	I think specially in older adults the Sinopharm Sinovac isn't seeming to be very effective.
	5	2.8%	did, Sinovac did, sure, vaccine administered, know, wonder, used, administered, using Sinovac, sarscov2	Which vaccine is being administered? Sinopharm or Sinovac?
	6	1.8%	free, cost, vaccine, doses, pay, effective priced, dose, letter, price, given free	I don't think there will be reimbursement since Sinovac is supposed to be free.
	12	1.8%	good, vaccine, Sinovac good, Sinovac good vaccine, aren, anti, worst, bad vaccine, bad, don	Just as worthless as the Sinovac vaccine.
Health system-related	1	3.6%	got, dose, 1st, Sinovac, today, 1st dose, got 1st, effects, just, got Sinovac vaccine	In my country we got the Sinovac vaccine, I'm getting my second dosage today.
	7	2.5%	Sinovac, sputnik Sinovac, vs, vaccine, SAHPRA, need, Sinovac sputnik vaccines, variants, Sinopharm, adenovirus	Just because WHO think SPUTNIK V vaccine and Sinovac vaccines are bad make them unsafe.
	8	2.4%	Philippines, vaccines, pinoys, fake, government, hate, serum, Sinopharm, Duterte, doesn	The Filipinos hate Sinovac because it's from China.
	10	2.0%	got, mom, Sinovac vaccine, said yes, like, dangerous, told, bc, friend, sure	Throw Sinovac and Sinopharm in the trash. Do not take these dangerous vaccines
	11	1.4%	Indonesia, effective, Sinovac, said, workers, delta, health, results, mass vaccination, healthcare workers inoculated	The Sinovac vaccine had 94% efficacy in an Indonesian study lmao.
Individual's social attributes	3	4.9%	people, think, want, just, Sinovac vaccine, make, like, govt, choose, fair	People just want a safe and effective covid vaccines.
	9	2.1%	India, request, Sinopharm, Saudi, kindly, govt, help, good, BJP, Sindh	Please help and request to Saudi Arabia government to accept these suitable and safest vaccine in Saudi.
	13	1.7%	Duque, welcome, arrival, president, vaccines, pres, doses Sinovac vaccines, says, roque says, million	Duterte should get vaccinated with Sinovac's CoronaVac from China, and Leni with a more effective vaccine.
	14	1.6%	passport, vaccine passport, able, list, approved, country, Europe, Sinovac vaccine, Sputnik, won, able to travel	You have obtained your vaccine passport; you are free to travel anywhere.
	15	1.5%	vaccine, used, effective, using, China, great, rollout, maybe, efficacy, using Sinovac	Chile used the Sinovac vaccine. It's not that effective.

(continued)



Table 3. Continued.

Factor	Topic-ID	Proportion	Keywords	Example
(b) Top-five topics related to AstraZeneca by number of tweets.				
Vaccine-related	1	13.6%	Zeneca, astra, vaccine, vaccines, people, got, getting, FDA, im, good	I'm very happy to have the Astra Zeneca, which is a really good vaccine.
	3	3.9%	blood, clots, clotting, rare, clot, risk, AstraZeneca, vaccine, woman, linked	The reason AstraZeneca was not approved in the USA was the risk of Blood Clots.
	4	3.7%	delta, variant, effective, 92, variants, AstraZeneca, vaccines, alpha, vaccine, covid	AstraZeneca has been shown to be 92% effective against hospitalization with Delta alone.
	5	3.6%	Australia, Australians, NSW, Morrison, AstraZeneca, Melbourne, people, vaccine, Scott, CSL,	Our young Australians deserve a safe effective vaccine not Morrison's clearance stock AstraZeneca.
	7	3.0%	mRNA, effective, vector, vaccines, viral, jj, adenovirus, better, AstraZeneca, data	Astra Zeneca is not as effective as the MRNA vaccines.
Health system-related	2	3.9%	AstraZeneca, vaccine, good, corona98, effective, just, safe, vaccines, bs, come	The vaccine has been proven to be safe and effective by the WHO.
	9	2.6%	today, happy, just, got, im, AstraZeneca, great, feeling, vaccine, gonna,	Got my first shot for the AstraZeneca vaccine today. I am happy.
	20	1.1%	Covishield, Covaxin, government, EU, India, extend, travel, brand, vaccine, protection	The AstraZeneca vaccine that isn't recognized in the EU is called Covishield (made in India).
	25	0.7%	weeks, gap, 12, interval, longer, doses, apart, 412, minimum, week,	TGA has reviewed that AstraZeneca can be safely administered 4-12 weeks apart.
	27	0.6%	million, Thailand, 56, fully, majority, doses, 10, taken, month, jabs	SBS production capacity of AZ is 16 million per month and just 5-6 million doses will go to Thailand.
Individual's social attributes	6	3.0%	India, Indian, Bharat, Covaxin, AstraZeneca, vaccine, serum, biotech, Indias, vaccines	India might have something to do with worse outcomes.
	8	2.9%	40s, advice, abc, age, ATAGI, people, 40, AstraZeneca, allow, UK	The UK Government won't allow their under 40s to get the AstraZeneca vaccine
	10	2.2%	EU, Europe, court, UK, batch, AstraZeneca, travel, thing, India, obliged	Concerned not to be able to travel to Europe, because the EU recognized Covishield.
	11	2.1%	fever, effects, arm, pain, sore, bad, covid, got, headache, headaches	I got my first dose with AstraZeneca, and had a hell of a fever for 28 hours.
	12	2.0%	Az, Sinovac, org, win, USA, preventing, TUA, vaccine, symptomatic, efficacy	People vaccinated with AstraZeneca, will not be able to go to the USA.
(c) Top-five topics related to Pfizer by number of tweets.				
Vaccine-related	1	13.6%	delta, variant, effective, Israel, Pfizer, 64,	

(continued)

Table 3. Continued.

Factor	Topic-ID	Proportion	Keywords	Example
			effective delta variant, 64 effective, effective preventing, severe	Studies show vaccines like Pfizer are still effective against Delta.
	2	5.6%	arm, got, Pfizer vaccine, feel, hours, experienced, symptoms, effect, shot, just	I got the Pfizer vaccine today and my arm hurts a little weirdly.
	3	5.1%	Pfizer, 12, children, Pfizer vaccine, safe, year, group, olds, age group, people	For the Pfizer vaccines in kids it's a 100% effective.
	4	3.5%	know, just, make, listen, vaccine Pfizer, Johnson, thing, honest, hell	I got the wrong vaccine. It wasn't Pfizer.
	7	2.4%	approval, months, need, interval, said, just, need boosters, like, good, CDC	Pfizer just said their vaccine antibodies are "waning" and we need a booster. It's ridiculous!
Health system-related	5	3.2%	government, rich, don, BioNTech, free, companies, Pfizer vaccines, capitalism, make, pay	The government paid for the research on the vaccines, and they're paying for everyone to get the shot.
	6	2.5%	Pfizer vaccines, covid, Aussies, NSW health, end, good, gov, waiting, federal, waiting Pfizer	Ideally, but the Pfizer vaccine shortage is entirely the fault of the federal gov.
	11	2.1%	country, vaccines, shortages, 000, course, rollout, issue, government, want, mean	So unfair how we don't have the Pfizer vaccine here.
	15	1.7%	able, tomorrow, gp, able Pfizer, vaccinated, available, just, eligible, email, hair	I was finally able to book my vaccine appointments!
	16	1.4%	UK, vaccinated, Pfizer, Pfizer vaccine EU, allow, travel, think, EU rules, racist, Belgium	EU unlike UK is arranging for the export of vaccines to Covac countries.
Individual's social attributes	12	2.0%	got, dose Pfizer vaccine, dose, got dose Pfizer, Pfizer, got dose, happy, great, just got, proud	Got Pfizer vaccine a week ago lol so far so good.
	14	1.7%	India, Covaxin, vaccine, people, sputnik, Covishield, given, hate, cold, needs	And some Indians are mad for Pfizer vaccine.
	18	1.3%	masks, wear, anti, indoors, preventing, efficacy, effective, think, mask indoors, maskless	I am not anti-vaccine, I am merely anti-mask after being vaccinated.
	19	1.2%	Got jnj, shot, remember, able, booster, got jj vaccine, bonus, allergies	I know people who got the JJ vaccine and they were fine.
	22	1.1%	yeah, vax, mom, Pfizer vaccine unless, trial, covid, prove, posted, difficult, fight	My little sister bravely participated in a Pfizer vaccine trial for kids.

it. This research selected the top-five topics, each color representing a different number of topics, squares representing word embeddings, and stars representing embeddings joined by CL.

We also found that the points joined with CL are mainly closer to the center of the clusters or even overlap with the center, while only the word embeddings are far from the center. For example, the points of topic 1 pertaining to

Pfizer lie outside the topic clusters. The CL strategy corrects the parameters of the pre-trained model, which strengthens clustering ability of the topic model and shows that the CL strategy also has a certain anti-noise ability. Thus, the fine-tuned model distinguishes different topic clusters better, and constrains the influences of outlier points on the clustering effect to some extent.

### Group difference analysis

The cosine similarity method is used to explore the similarity between different topics and factors. The closer the obtained cosine similarity is to 1, the more similar the user group of

that category is to the topic: the results are summarized in Tables 4(a), (b), and (c). For Sinovac (Table 4(a)) and correlations between the different factors, there is a high correlation in the vaccine factor, with the highest being 15.25% and the lowest being 14.62%, which indicates the most prominent element related to vaccines. Significantly, the correlation between the two health systems is higher among the health system factors. Also, among the social attributes, the correlation is higher for topic 5, so it can be determined that the correlation is higher for factor overlap. The same outside from AstraZeneca and Pfizer can be seen that also the two factors are much the same, which both are highly correlated between two identical factors. However, the factors of AstraZeneca are also found to be smaller, and the correlation is not significant.

## Discussion

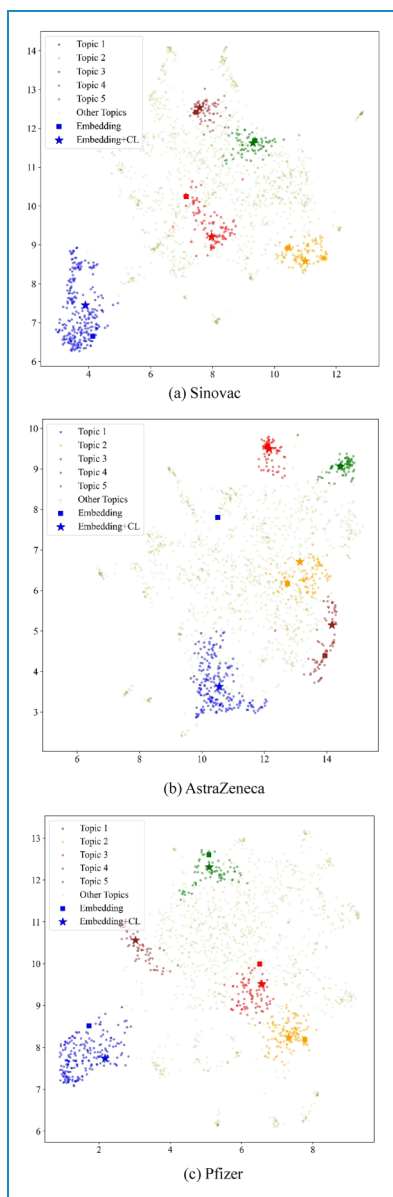
### Principal findings

This finding is higher than other previous on-line investigations from the UK,<sup>50</sup> the USA,<sup>51</sup> and France.<sup>52</sup> These countries arise probably because Twitter users are more prone to vent negative emotions on-line. The central theme of vaccination attitudes was identified by three groups in the discussion of the Tweets, including vaccine-related, health system-related, and individual societal attributes. Some themes demonstrate episodic changes and high degrees of co-occurrences in close association with vaccine developments. This study enriches our understanding of the public concerns related to vaccines. These shared concerns can inform public health organizations and professionals for more tailored health messages and vaccination policies.

Our results suggest that vaccine-related factors and effectiveness were the respondents' main reasons for the professed attitudes. People hoped to confirm their willingness to vaccinate by extending the test time and ensuring the safety and effectiveness of the vaccine. From the perspective of individual factors, older people should be a priority group for vaccination, but they did not show a greater desire to be vaccinated than younger people. Considering individual societal attributes, people who were currently vaccinated against seasonal influenza are more likely to receive the COVID-19 vaccine, which is related to the perception of, and trust in, the vaccine among this population. A Finnish scholar<sup>53</sup> proposed that people's intention towards vaccination can be predicted by their perception of the risks around vaccine safety. The group with high confidence in the effectiveness and safety of vaccines had a lower probability of negative vaccine attitudes than the group with quiet confidence therein.

### Summary generation and topic modeling

Although LDA introduces a way to attach topic content to a document, which treats each document as a hybrid of



**Figure 5.** Visualization of each dataset. (a) Sinovac. (b) AstraZeneca. (c) Pfizer.

**Table 4.** (a) Cosine similarity of different factors related to Sinovac (top-five topics).

Factors	Topic number	Vaccine-related	Health system-related	Individual's social attributes
Vaccine-related	2	15.25%	12.67%	12.43%
	4	14.88%	12.51%	13.72%
	5	12.45%	10.16%	10.39%
	6	13.03%	11.94%	11.99%
	12	14.62%	12.78%	12.86%
Health system-related	1	12.08%	11.39%	11.29%
	7	10.54%	10.01%	10.40%
	8	6.52%	8.50%	6.57%
	10	10.90%	11.73%	11.44%
	11	14.73%	12.58%	11.85%
Individual's social attributes	3	12.10%	10.71%	14.26%
	9	13.51%	11.82%	13.28%
	13	12.41%	11.91%	12.22%
	14	9.69%	8.45%	10.82%
	15	10.42%	8.49%	10.09%

**(b)** Cosine similarity of different factors related to AstraZeneca (top-five topics).

Vaccine-related	1	13.06%	8.53%	9.93%
	3	9.43%	8.47%	10.75%
	4	11.77%	10.04%	11.26%
	5	8.69%	10.96%	11.40%
	7	10.73%	9.57%	11.20%
Health system-related	2	8.17%	10.69%	11.15%
	9	7.77%	9.77%	10.94%
	20	8.16%	10.88%	10.96%
	25	8.47%	8.01%	11.38%
	27	5.46%	5.80%	5.97%
Individual's social attributes	6	10.17%	10.43%	13.24%

(continued)

Table 4. Continued.

Factors	Topic number	Vaccine-related	Health system-related	Individual's social attributes
	8	8.29%	8.82%	13.79%
	10	4.88%	6.39%	5.25%
	11	1.39%	0.82%	0.39%
	12	1.88%	2.18%	0.78%
(c) Cosine similarity of different factors related to Pfizer (top-five topics).				
Vaccine-related	1	16.80%	10.30%	12.50%
	2	11.40%	9.22%	11.49%
	3	10.65%	9.28%	9.95%
	4	11.06%	9.89%	10.44%
	7	11.42%	9.39%	10.48%
Health system-related	5	7.27%	8.30%	6.78%
	6	7.47%	8.46%	7.45%
	11	6.68%	7.97%	5.96%
	15	9.98%	9.71%	11.10%
	16	10.40%	10.82%	10.56%
Individual's social attributes	12	12.51%	11.05%	15.02%
	14	10.61%	10.22%	11.98%
	18	12.99%	10.48%	12.61%
	19	10.51%	9.27%	11.93%
	22	9.41%	8.42%	10.29%

several different topics, the tweets alluded to here usually need to meet this requirement, and most of them are short documents with one main topic. In addition, LDA suffers from a disorderly effect, which means that different topics cannot be generated in order, leading to differences in topic names when the words defining the topics or the order of importance differ.

During the model learning process, the sentence is learned in both directions to understand the contextual information of the word so that the same word can be better reflected in different contexts. They used a masked language model and sentence prediction multi-task training objective for pre-training, applying the model to other

specific tasks by fine-tuning to attain semantically richer word vector representations.

The present work is mainly applied to vaccine attitudes through summary generation, aiming to help users acquire information quickly and help the government promptly understand the factors affecting vaccine attitudes. The proposed algorithm is based on the characteristics of social media timeliness, which improves on previous scholars' research with the algorithm incorporating CL. Visualization results show that the summary-generation technique can improve the timeliness of the summary, however, the research has many algorithmic limitations that need to be investigated in depth, such as how to make the abstracts

more fluent and coherent when generating them.<sup>54</sup> Therefore, how to use semantic analysis to enhance the coherence of abstracts is a potential avenue of exploration for the future.

### Implications for management

The more significantly to promote the demand, use, and equitable distribution of COVID-19 vaccines, the community public health strategy should use a health belief model to minimize the possibility of vaccine hesitancy.<sup>55</sup> Strully et al.<sup>56</sup> suggested that governments should increase funding to promote vaccine popularization, and reduce vaccine inequity caused by racial discrimination and the marginalization of some traditional communities, which plays a role in the fair distribution and use of vaccines by establishing social organizations with different functions.

Unlike traditional media, social media allows individuals to create and share content rapidly across the globe without editorial oversight. Users may choose their content streams, resulting in a separation of meaning. For this reason, if anti-vaccination messages flood these platforms, considerable public health concerns can arise and may lead to vaccine hesitancy, undermining public confidence in future vaccine development for novel pathogens.<sup>6</sup> In response to false information about COVID-19, many foreign social media companies have issued a joint statement to combat “misinformation about the virus,” provide verified information, and use tools to flag and remove tweets containing misleading information and highly damaging content.<sup>57</sup>

Delivery of health knowledge of individual vaccination can reduce the incidence of negative vaccine attitudes and increase the rollout of any vaccination program. The public’s high expectations for vaccines and the rapid progress of vaccine research and development in the context of the COVID-19 epidemic have made society pay closer attention to safety, immunogenicity, protective efficacy, and response to virus variation of different types of COVID-19 vaccines. In Europe, it has been suggested that organizing groups or experts to engage with undecided people, encouraging them to ask questions, actively listening to their concerns, and providing clear, easily understood, and evidence-based information are effective ways in which to promote acceptance of COVID-19 vaccines.<sup>58</sup>

As an important disseminator of vaccine and vaccination knowledge and information, vaccinators are an important link to strengthening public confidence in vaccination.<sup>59</sup> Vaccinators should continue to learn and master the ability of vaccines to enhance their confidence in vaccines; in the face of patient consultation, they should be able to provide professional answers to alleviate the concerns of the patient.

### Limitations and future work

This study is based on only Tweet users, so a selection bias prevails compared to the broader target group for vaccination. This research will expand our range of social media platforms, such as Reedit or Weibo in future work; the range of vaccine brands analyzed (including, e.g. Janssen, Sinopharm, and Moderna) can also be broadened. Regarding methodology, this research only adopts CL for summary generation and adopts BERTopic to extract topic models: in future work, we will use deep learning for sentiment analysis or attitude detection. Although, this paper discusses the vaccine attitudes of the public in different regions, which does not consider human mobility and behavioral characteristics.<sup>60</sup> Future work can be improved by using geotagged mobility data.

Furthermore, most Tweets topics covered three themes: vaccine factors, individual factors, and cognitive factors. However, there remain some topics unable to be covered by these factors. There should also be longitudinal research conducted in the future.

### Conclusion

This paper presents a combination of NLP methods aimed at studying the reasons for vaccine attitudes in different regions, which focuses on information about users collected from Twitter and expressed by users. We first collect tweets from three brands with keywords: COVID-19 and vaccine. Moreover, this research establishes a text summary-generation model based on CL and then uses BERTopic to cluster topics automatically, which shows outstanding performance and provided high-quality results. The topic models are constructed for vaccine attitudes factors that can be best used in future opinion studies, especially for immunization programs, as the world remains uncertain about how pandemics evolve. For these reasons, our work can lead to a better understanding of the vaccination process. The findings will allow governments and pharmaceutical medical institutions to develop or redefine better solutions to the problem of vaccine attitudes. This work can improve the fight against the COVID-19 pandemic.

**Contributorship:** It is a single-author paper.


**Data availability statement:** The datasets generated for this article are not readily available because the raw data cannot be made public; if necessary, feature data can be provided. Requests to access the datasets should be directed to the corresponding author.

**Declaration of Conflicting Interests:** The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

**Ethical approval:** This retrospective study was approved by the Research Ethics Commission Wuhan University. The requirement for informed consent was waived due to its retrospective design.

**Funding:** The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the National Key Research and Development Program of China (No. 2021ZD0113304), the National Natural Science Foundation of China (No. 72204190), the Research Foundation of Ministry of Education of China (No. 22YJZH114), and the China Postdoctoral Science Foundation (No. 2022M722476).

**Guarantor:** All authors are the guarantors of this manuscript.

**ORCID iD:** Yang Liu  <https://orcid.org/0000-0002-9410-1755>

## Reference

1. Eslami Jahromi M and Ayatollahi H. Utilization of telehealth to manage the Covid-19 pandemic in low- and middle-income countries: a scoping review. *J Am Med Inform Assoc* 2023; 30: 738–751.
2. Nicola M, Alsafi Z, Sohrabi C, et al. The socio-economic implications of the coronavirus pandemic (COVID-19): a review. *Int J Surg* 2020; 78: 185–193.
3. Fauci AS. The story behind COVID-19 vaccines. *Science* 2021; 372: 109–109.
4. Cordina M, Lauri MA and Lauri J. Attitudes towards COVID-19 vaccination, vaccine hesitancy and intention to take the vaccine. *Pharm Pract (Granada)* 2021; 19: 2317.
5. Al-Amer R, Maneze D, Everett B, et al. COVID-19 vaccination intention in the first year of the pandemic: a systematic review. *J Clin Nurs* 2022; 31: 62–86.
6. Puri N, Coomes EA, Haghbayan H, et al. Social media and vaccine hesitancy: new updates for the era of COVID-19 and globalized infectious diseases. *Hum Vaccin Immunother* 2020; 16: 2586–2593.
7. Mullard A. How COVID vaccines are being divvied up around the world. *Nature* 2020. DOI: 10.1038/d41586-020-03370-6
8. Oyeboode O, Ndulue C, Adib A, et al. Health, psychosocial, and social issues emanating from the COVID-19 pandemic based on social media comments: text mining and thematic analysis approach. *JMIR Med Inform* 2021; 9: e22734.
9. Baclic O, Tunis M, Young K, et al. Artificial intelligence in public health: challenges and opportunities for public health made possible by advances in natural language processing. *Can Commun Dis Rep* 2020; 46: 161.
10. El-Kassas WS, Salama CR, Rafea AA, et al. Automatic text summarization: a comprehensive survey. *Expert Syst Appl* 2021; 165: 113679.
11. Xiao T, Wang X, Efros AA, et al. What should not be contrastive in contrastive learning, <http://arxiv.org/abs/2008.05659> (2021, accessed 26 December 2022).
12. Du J, Luo C, Shegog R, et al. Use of deep learning to analyze social media discussions about the human papillomavirus vaccine. *JAMA Netw Open* 2020; 3: e2022025.
13. Qorib M, Oladunni T, Denis M, et al. Covid-19 vaccine hesitancy: text mining, sentiment analysis and machine learning on COVID-19 vaccination Twitter dataset. *Expert Syst Appl* 2023; 212: 118715.
14. Ljajić A, Prodanović N, Medvečki D, et al. Uncovering the reasons behind COVID-19 vaccine hesitancy in Serbia: sentiment-based topic modeling. *J Med Internet Res* 2022; 24: e42261.
15. Huangfu L, Mo Y, Zhang P, et al. COVID-19 vaccine tweets after vaccine rollout: sentiment-based topic modeling. *J Med Internet Res* 2022; 24: e31726.
16. Alamoodi AH, Zaidan BB, Al-Masawa M, et al. Multi-perspectives systematic review on the applications of sentiment analysis for vaccine hesitancy. *Comput Biol Med* 2021; 139: 104957.
17. Allahyari M, Pouriye S, Assefi M, et al. Text summarization techniques: a brief survey, <http://arxiv.org/abs/1707.02268> (2017, accessed 28 December 2022).
18. Comito C and Pizzuti C. Artificial intelligence for forecasting and diagnosing COVID-19 pandemic: a focused review. *Artif Intell Med* 2022; 128: 102286.
19. Liu Y, Fei H, Zeng Q, et al. Electronic word-of-mouth effects on studio performance leveraging attention-based model. *Neural Comput & Appl* 2020; 32: 17601–17622.
20. Yin H, Song X, Yang S, et al. Sentiment analysis and topic modeling for COVID-19 vaccine discussions. *World Wide Web* 2022; 25: 1067–1083.
21. Li X, Zhou M, Wu J, et al. Analyzing COVID-19 on online social media: trends, sentiments and emotions, <http://arxiv.org/abs/2005.14464> (2020, accessed 21 April 2023).
22. Zhou J, Zogan H, Yang S, et al. Detecting community depression dynamics due to COVID-19 pandemic in Australia. *IEEE Trans Comput Soc Syst* 2021; 8: 982–991.
23. Yin H, Yang S and Li J. Detecting topic and sentiment dynamics due to COVID-19 pandemic using social media. In: *Advanced Data Mining and Applications: 16th International Conference, ADMA 2020, Foshan, China, November 12–14, 2020, Proceedings 16*, pp.610–623. Springer International Publishing.
24. Kwok SWH, Vadde SK and Wang G. Tweet topics and sentiments relating to COVID-19 vaccination among Australian Twitter users: machine learning analysis. *J Med Internet Res* 2021; 23: e26953.
25. Lyu JC, Han EL and Luli GK. COVID-19 vaccine-related discussion on Twitter: topic modeling and sentiment analysis. *J Med Internet Res* 2021; 23: e24435.
26. Bonnevie E, Gallegos-Jeffrey A, Goldbarg J, et al. Quantifying the rise of vaccine opposition on Twitter during the COVID-19 pandemic. *J Commun Healthc* 2021; 14: 12–19.
27. Tavoschi L, Quattrone F, D'Andrea E, et al. Twitter as a sentinel tool to monitor public opinion on vaccination: an opinion mining analysis from September 2016 to August 2017 in Italy. *Hum Vaccin Immunother* 2020; 16: 1062–1069.
28. Bar-Lev S, Reichman S and Barnett-Itzhaki Z. Prediction of vaccine hesitancy based on social media traffic among Israeli parents using machine learning strategies. *Isr J Health Policy Res* 2021; 10: 49.
29. Piedrahita-Valdés H, Piedrahita-Castillo D, Bermejo-Higuera J, et al. Vaccine hesitancy on social media: sentiment analysis from June 2011 to April 2019. *Vaccines* 2021; 9: 28.

30. Yuan X, Schuchard RJ and Crooks AT. Examining emergent communities and social bots within the polarized online vaccination debate in Twitter. *Social Media + Society* 2019; 5: 205630511986546.
31. Liu Y, Zeng Q, Li B, et al. Anticipating financial distress of high-tech startups in the European Union: a machine learning approach for imbalanced samples. *J Forecast* 2022; 41: 1131–1155.
32. LeCun Y, Bengio Y and Hinton G. Deep learning. *Nature* 2015; 521: 436–444.
33. Hussain A, Tahir A, Hussain Z, et al. Artificial intelligence-enabled analysis of public attitudes on Facebook and Twitter toward COVID-19 vaccines in the United Kingdom and the United States: observational study. *J Med Internet Res* 2021; 23: e26627.
34. Shahid F, Zameer A and Muneeb M. Predictions for COVID-19 with deep learning models of LSTM, GRU and Bi-LSTM. *Chaos, Solitons Fractals* 2020; 140: 110212.
35. Devaraj J, Madurai Elavarasan R, Pugazhendhi R, et al. Forecasting of COVID-19 cases using deep learning models: is it reliable and practically significant? *Results Phys* 2021; 21: 103817.
36. Shastri S, Singh K, Kumar S, et al. Time series forecasting of Covid-19 using deep learning models: India-USA comparative case study. *Chaos, Solitons Fractals* 2020; 140: 110227.
37. Poria S, Cambria E, Bajpai R, et al. A review of affective computing: from unimodal analysis to multimodal fusion. *Inf Fusion* 2017; 37: 98–125.
38. Canini KR, Shi L and Grieths TL. Online inference of topics with latent Dirichlet allocation. 2009: 65–72.
39. Devlin J, Chang M-W, Lee K, et al. BERT: pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805 [cs]*, <http://arxiv.org/abs/1810.04805> (2019, accessed 16 April 2022).
40. Goldberg Y and Levy O. word2vec explained: deriving Mikolov, et al.'s negative-sampling word-embedding method, <http://arxiv.org/abs/1402.3722> (2014, accessed 28 December 2022).
41. Zhang T, Wu F, Katiyar A, et al. Revisiting few-sample BERT fine-tuning, <http://arxiv.org/abs/2006.05987> (2021, accessed 28 December 2022).
42. McInnes L, Healy J and Melville J. UMAP: uniform manifold approximation and projection for dimension reduction, <http://arxiv.org/abs/1802.03426> (2020, accessed 4 January 2023).
43. McInnes L, Healy J and Astels S. HdbSCAN: hierarchical density based clustering. *JOSS* 2017; 2: 205.
44. Röder M, Both A and Hinneburg A. Exploring the space of topic coherence measures. In: Proceedings of the eighth ACM international conference on web search and data mining, pp.399–408. Shanghai, China: ACM.
45. Mimno D, Wallach H, Talley E, et al. Optimizing semantic coherence in topic models. 2011, pp.262–272.
46. Graves A, Jaitly N and Mohamed A. Hybrid speech recognition with deep bidirectional LSTM. In: 2013 IEEE workshop on automatic speech recognition and understanding, pp.273–278. Olomouc, Czech Republic: IEEE.
47. Niu Z, Zhong G and Yu H. A review on the attention mechanism of deep learning. *Neurocomputing* 2021; 452: 48–62.
48. Le-Khac PH, Healy G and Smeaton AF. Contrastive representation learning: a framework and review. *IEEE Access* 2020; 8: 193907–193934.
49. Liu Y, Liu P, Radev D, et al. BRIO: bringing order to abstractive summarization, <http://arxiv.org/abs/2203.16804> (2022, accessed 19 December 2022).
50. Hussain A, Tahir A, Hussain Z, et al. Artificial intelligence-enabled analysis of public attitudes on Facebook and Twitter toward COVID-19 vaccines in the United Kingdom and the United States: observational study. *J Med Internet Res* 2021; 23: e26627.
51. Pogue K, Jensen JL, Stancil CK, et al. Influences on attitudes regarding potential COVID-19 vaccination in the United States. *Vaccines* 2020; 8: 582.
52. Detoc M, Bruel S, Frappe P, et al. Intention to participate in a COVID-19 vaccine clinical trial and to get vaccinated against COVID-19 in France during the pandemic. *Vaccine* 2020; 38: 7002–7006.
53. Karlsson LC, Soveri A, Lewandowsky S, et al. Fearing the disease or the vaccine: the case of COVID-19. *Pers Individ Dif* 2021; 172: 110590.
54. Gupta S and Gupta SK. Abstractive summarization: an overview of the state of the art. *Expert Syst Appl* 2019; 121: 49–65.
55. Abila DB, Dei-Tumi SD, Humura F, et al. We need to start thinking about promoting the demand, uptake, and equitable distribution of COVID-19 vaccines NOW! *Public Health in Practice* 2020; 1: 100063.
56. Strully KW, Harrison TM, Pardo TA, et al. Strategies to address COVID-19 vaccine hesitancy and mitigate health disparities in minority populations. *Front Public Health* 2021; 9: 645268.
57. Pennycook G, McPhetres J, Zhang Y, et al. Fighting COVID-19 misinformation on social media: experimental evidence for a scalable accuracy-nudge intervention. *Psychol Sci* 2020; 31: 770–780.
58. Soares P, Rocha JV, Moniz M, et al. Factors associated with COVID-19 vaccine hesitancy. *Vaccines* 2021; 9: 300.
59. Chung Y, Schamel J, Fisher A, et al. Influences on immunization decision-making among US parents of young children. *Matern Child Health J* 2017; 21: 2178–2187.
60. Cesario E, Comito C and Talia D. An approach for the discovery and validation of urban mobility patterns. *Pervasive Mob Comput* 2017; 42: 77–92.