Covid-19 Vaccination Hesitancy: Sentiment Analysis of Twitter Posts

Dadian Qu[1], Yuming Chen[2], Taoyu Mao[3], Xinyu Dai[4]

dadianqu@gmail.com[1], ttamymc@yeah.net[2], 1062609301@qq.com[3], lesleyd0624@gmail.com[4]

School of Business, George Washington University, Washington DC, 20052, USA[1]
N/A, BASIS International School, Nanjing, 210028, China[2]
Liangjiang International College, Chongqing University of Technology, Chongqing, 401135, China[3]
School of Economics and Management, Hubei University of Technology, Wuhan, 430068, China[4]

Abstract. Many people show concerns about Covid-19 vaccination and choose not to be vaccinated for different reasons. The understanding of those reasons is the key success factor in increasing the uptake rate and control the Covid-19 pandemic. To better understand the public opinions, we use Twitter posts as our dataset since it is highly used to express people's feelings in many countries. Our study is to use machine learning-based algorithms and human coding methods to perform sentiment analysis, topic modelling, and topic aggregation. We used Vader to filter out positive and neutral tweets. After that, we used LDA to group all the tweets into different clusters. Then we manually aggregated clusters into topics. Our findings show that people have vaccination hesitancy mainly because of the following topics: Safety Concerns; Belief Covid is Harmless; Distrust; Efficacy; Social Distribution.

Keywords: Covid-19; Sentiment Analysis; Twitter; opinion mining; Vaccination hesitancy

1. Introduction

COVID-19 has become a significant challenge in the world, affecting all humanity in many ways. Its prevalence worldwide has caused people many serious problems, including death, multiple organ failure, and pneumonia. Therefore, the pandemic continues to create challenges to our economies, societies, and the public health system. Thus, the COVID-19 vaccination has become a crucial step in combating COVID-19. And it also has been considered the most effective way of fighting the pandemic. However, people have shown mixed feelings about COVID-19 vaccination. Whereas some people have had positive and neutral emotions, some have displayed negative reactions. According to MacDonald et al, vaccine hesitancy was declared by WHO in 2019 as one the ten greatest global threats [1].

Experts plan to control the epidemic by establishing a colony immune barrier (herd immunity) for the whole population through novel coronavirus vaccination. Herd immunity is possible when a high percentage of the population either gets infected or vaccinated, this would decrease the likelihood of the rest of the population from getting infected. It is generally believed that herd immunity can be established only when the vaccination rate reaches 60% - 70%. This required vaccine coverage is very high and may be difficult to achieve for many reasons. This

ICBBEM 2023, May 19-21, Hangzhou, People's Republic of China

DOI 10.4108/eai.19-5-2023.2334252

is a huge challenge not only for pharmaceutical companies and finite healthcare resources, but also for Government agencies and regulatory authorities.

Vaccines offered a solution to control the outbreak by increasing people's immunity. But in social media, vaccine misinformation became a common problem that prevailed at the very beginning because of its wide usage and access [2]. That misinformation is often transmitted by different users, regardless of the source's origin. Even without misinformation, the uptake rate can be reduced by those negative sentiments. Positive sentiments on social media can potentially increase the uptake rate [3]. In light of the pandemic, Twitter has allowed people to express their emotions and opinions toward vaccination, either positive or negative. The negative consequences of social media on COVID-19 are apparent in society. People exposed to COVID-19 information on Twitter are less likely to have optimistic attitudes. Significant negative emotions and sentiments are strongly tied to society's tendency toward negative attitudes and perceptions toward vaccination, which could contribute to a low uptake rate.

According to popular tweets, the negative responses were influenced by the fact that people distrusted the vaccines for being effective in protecting people. According to Mishra et al., "people suspected whether sufficient safeguards were put in place to manufacture vaccines and raised doubts over the efficacy of vaccines." [4]. In this regard, the public has become fearful that vaccines are ineffective in protecting individuals' lives and cause many side effects. At the global level, many developed countries still lag in vaccination, despite the many vaccine supplies. Such a low uptake rate showed people's hesitancy, misperceptions, and negative emotions toward vaccines.

Accordingly, Bustos et al. claim that responses regarding public perception have changed for a good part over time [5]. Essential phrases such as vaccination and trials were recorded to attain a better vaccination and trials associated with trust established promotional and educational schemes for better vaccination coverage. In this manner, Cotfas et al. state that positive tweets are often made against the target entity, whereas negative tweets often express the target entity's favorable view [6]. Notably, the prevalence of positive sentiments in tweets indicates that tweets are definite signs of hope and enjoyment. Alam et al. claim that the difference between negative feelings and positive feelings is different, with "positive being dominant and planning stronger responses." [7]. With that in mind, COVID-19 vaccination presents highly satisfactory outcomes due to different sentiments. The conclusion Jun et al. draw is that efforts to address negative emotions and combat misinformation have proved essential in increasing global vaccination uptakes [8].

In this research, we have done the following tasks:

- Extracted COVID-19 vaccination-related text from Twitter.
- Investigated those texts to find out reasons why people have negative feelings.
- Categorized those negative sentiments and provided strategies to increase uptake rate.

These findings will help the CDC find compelling ways of convincing people to vaccinate.

2. Literature Review

The analysis drawn from Twitter has made several attempts to focus on entities, including detecting misleading information, topic analysis, and identifying emotions. These sentiments contribute to negative and positive reactions depending on government decisions and geographical location. For instance, Lanyi et al. argue that out of its 913 sampled tweets, 312 tweets defended the position of negative sentiments, mainly through political factors around vaccines [9]. Since emotions are strong vaccine risk perception predictors, people often tweet to show their fear emotions and other interrelated diffusions of depression, stress, and anxiety. This implies that negative sentiments are on the rise than positive sentiments from people.

For developed countries, the prevalence of negative sentiment is twice that of other countries. In fact, anger, sadness, and fear appear less frequently than positive sentiments such as joy. On the same note, a country's GDP per capita, human development index, institutional quality, and democracy index are strongly associated with adverse COVID-19 anger, sadness, and fear [10]. Within countries, differences in COVID-19 crisis exist regarding emotions. Similarly, the public distrust in the vaccine is influenced by cross-cultural variations among people with different health literacy, religions, ethnicity, and race.

Karami et al., Jun et al., and Hayawi et al. indicates that mixed reactions towards the vaccine have diverse topic priorities and focuses [8,11,12]. Thus, the CDC can know when to deploy the vaccines based on emotions evoked about the vaccine. On the other hand, the institution can easily read the tweets and identify the right population that has positively reacted to the vaccine. In this case, Hayawi et al. believe in the machine learning-based model, an effective tool to detect misinformation about the vaccine on social media [12]. Additionally, it allows the CDC to better understand public opinion by classifying Twitter content into pro-vaccine, vaccine-hesitant, and anti-vaccine views, which would support their choice-making endeavors.

Lazarus et al. found that the level of income is associated with the level of hesitancy in most Countries [13]. According to their research, people with lower incomes tend to be more vaccine-hesitant. The possible reason behind this is that people who lost their socioeconomic status and financial ability have stronger desires to return to normalcy. Thus, they are less likely to take vaccination when they believe it is risky.

The public responses before, after, and during the vaccination process would also help to understand different sentiments on vaccination. A study by Alam et al., Bustos et al., and Coftas et al. expounds that the public's experience of vaccines would matter a lot as a source of information [5-7]. The writers argue that social media users' sentiments and emotions toward the COVID-19 vaccine determine how the process will be conducted on other people. The changes in their perception and responses before and after the vaccine rollout will guide the CDC when administering.

Accordingly, Baj-Rogowska built up a 6A model that contains 6 categories including access, affordability, awareness, acceptance, activation, and assurance [14]. They found that the most important factor is awareness, which covers the availability of a wide range of actual and detailed information regarding vaccines in the population, such as safety, effectiveness, side-effects, immunization schedules, quantities of vaccination points and their localization. Thus, the necessary step to raise the update rate is to increase people’s awareness among others.

A study by Lyu et al. discovered that major events could be another factor that people change their emotions on social media [15]. They found that online sentiment was surprisingly positive after Russia approved the world's first COVID-19 vaccine. When Pfizer announced that their vaccine is 90% effective, the trust emotion reached its peak, and everyone online seemed very optimistic for the future.

The 3C (Confidence, Complacency, Convenience) Model of Vaccine Hesitancy developed by the SAGE Working Group is a useful tool to understand the main reasons that drive vaccine hesitancy [9]. It breaks them down into three main categories:

- Confidence: People have low confidence in the vaccine's effectiveness and safety and distrust scientists, policymakers, and health ministers.
- Complacency: People don't perceive themselves to be at risk and view the vaccine as not necessary.
- Convenience: There are a few barriers (physical, logistical, or economical) that hamper them from getting a vaccine.

3. Data Collection

We adopted the "Covid Vaccine Tweets" dataset from Kaggle to perform sentiment analysis and topic modeling. Tweets are short posts broadcasted by users of the online social media platform Twitter. The Twitter platform allows users to use "#hashtags" to present the topic of their post, so they can contribute to the discussion of the topic. The dataset contains 398,592 tweets which were created through filtering using the #CovidVaccine hashtags between 08/2020 and 08/2022. Figure 1 is the basic structure of the raw data. All tweets were recorded in a CSV file, including fourteen columns.

	user_name	user_location	user_description	user_created	user_followers	user_friends	user_favourites	user_verified	date	text	hashtags	source	is_retweet
0	MyNewsNE	Assam	MyNewsNE a dedicated multi-lingual media house...	24-05-2020 10:18	64.0	11.0	110.0	False	18-08-2020 12:55	Australia to Manufacture Covid-19 Vaccine and ...	['CovidVaccine']	Twitter Web App	False
1	Shubham Gupta	NaN	I will tell about all experiences of my life f...	14-06-2020 16:42	1.0	17.0	0.0	False	18-08-2020 12:55	#CoronavirusVaccine #CoronaVaccine #CovidVacci...	['CoronavirusVaccine', 'CoronaVaccine', 'Covid...	Twitter for Android	False
2	Journal of Infectiology	NaN	Journal of Infectiology (ISSN 2689-9981) is ac...	14-12-2017 07:07	143.0	566.0	8.0	False	18-08-2020 12:46	Deaths due to COVID-19 in Affected Countries\n...	NaN	Twitter Web App	False
3	Zane	NaN	Fresher than you.	18-09-2019 11:01	29.0	25.0	620.0	False	18-08-2020 12:45	@Team_Subhashree @subhashreesotwe @iamrajchoco...	NaN	Twitter for Android	False
4	Ann-Maree O'Connor	Adelaide, South Australia	Retired university administrator. Melburnian b...	24-01-2013 14:53	83.0	497.0	10737.0	False	18-08-2020 12:45	@michellegrattan @ConversationEDU This is what...	NaN	Twitter Web App	False

Figure 1. Overview of the raw data

However, for purposes of this research we only use the text column. So, we excluded all unnecessary information and cleaned the text to have Figure 2 as our dataset.

	cleaned_text
0	is it the fate of rising covid cases in india ...
1	vaccine may be only partially effective warns ...
2	reveals how world's first will work
3	good news is all set to become the first count...
4	injected with covid coronavirus vaccine trial ...
5	what if we are just part of their pilot projec...
6	phase trials to begin in india
7	well that seems rather extreme vaccines are de...
8	plans to register st this week
9	how do we learn to live with if we're still wa...

Figure 2. Overview of the cleaned text

After exploring the data, we found that most people did not show their feelings about the vaccination, or they did not have a positive or negative tendency toward vaccination. Figure 3 shows the overall sentiment score distribution toward Covid-19 vaccines.

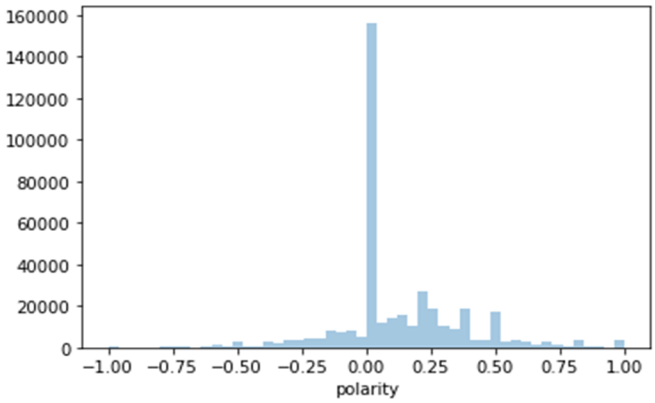


Figure 3. Sentiment scores towards Covid-19 vaccines

4. Research Methodology

Our methodology (Figure 4) consists of a preprocessing engine, analytical and modeling engine that incorporates unsupervised learning techniques to group similar topics together.

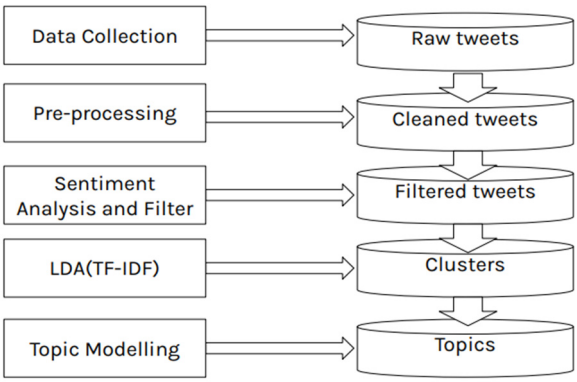


Figure 4. Research framework

Note: This is an illustration of the overall methodology in this research, which consists of 6 steps. We first extracted Twitter data from Kaggle. Then preprocessed to have a clean dataset. We used Vader to calculate the polarity score, and only used the ones less than -0.05 as they are our research targets. Then we used LDA to find out the best number of topics. In the end, we aggregated similar topics into one to have better categories.

4.1. Preprocessing

The original dataset contained 398,592 tweets. After removing 50 empty entries, we iterate through each tweet and remove URLs, mentions, hashtags, and non-ASCII characters. We used the python gensim library to remove stopwords and perform text stemming. Words such as "vaccine", "virus", and "covid" that are not meaningful for the research were also removed. We converted the time column into standard python datetime format using the DateTime library. After preprocessing we plotted the raw tweet count on a x-basis and illustrated when certain vaccines were introduced to show the volume of tweets during Covid-19 (Figure 5).

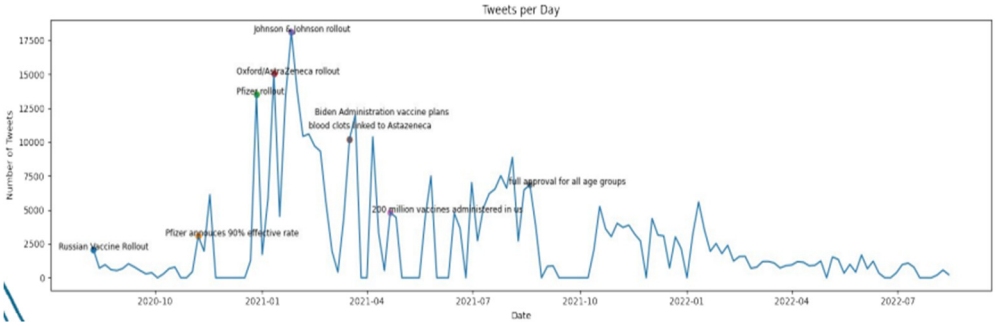


Figure 5. Count of tweets over time

4.2. Sentiment Analysis

Sentiment Analysis is a technique that identifies the emotions expressed in texts. It is widely adopted in natural language processing, text mining, and computational linguistics. In this

research, we performed sentiment analysis using the python library VaderSentiment and Textblob to investigate attitudes toward Covid-19 vaccination [16].

Vader is a lexicon and rule-based sentiment analysis tool. Compared with the traditional bag of words model, it reflects the true sentence semantics more accurately. In addition, Vader is attuned to sentiments expressed in social media, which perfectly suits our research purpose. In this paper, utilize Vader's Sentiment Intensity Analyzer to calculate a compound sentiment score. The score ranges from -1 to 1 with a higher score indicating a more positive sentiment. We label tweets with a score below -0.05 as negative, and since our research focuses on negative attitudes toward covid vaccines, we kept only negative tweets for subsequent topic modeling work.

4.3. Data Filtering

Besides filtering through sentiment, we further filtered our dataset based on 3 criteria to improve the performance and accuracy of topic modeling.

- Subjectivity

Since the research objective is to identify reasons for vaccine hesitancy, subjective tweets were preferred than objective tweets that state plain facts. A subjectivity score was evaluated for each tweet using the Textblob library. The score ranged from 0 to 1, with a higher score indicating a more subjective tone. The tweets with a subjectivity score lower than 0.2 were filtered out.

- Length

LDA performs poorly with short texts. Thus, we filtered out tweets with fewer than 6 words after removing stop words and stemming. This improves the probability that LDA generates meaningful clusters.

- Relevance to Vaccine Hesitancy

Although the dataset was created through the CovidVaccine Hashtag, a significant portion of tweets were irrelevant to vaccine hesitancy. We created a vaccine hesitancy word bank to solve this problem. The word bank was created by manually examining the top frequent words in the dataset and picking out ones relevant to vaccine hesitancy with the aid of the 3c model. We expand the word bank by identifying 25 on-topic tweets and appending the words contained in those tweets. A word bank of 341 words was finally derived, and we further filtered out tweets that did not contain words in the word bank.

After filtering, we obtained 10706 negative, subjective, long, and relevant tweets that would yield proper topic modeling results.

4.4. Topic Modeling

LDA is one of the most representative topic modeling methods, often used to classify and speculate the topic distribution of texts. That is, the algorithm would provide the topics in a corpus in the form of probability distribution.

To input texts into LDA, we convert preprocessed texts into bag-of-words structure. When using the LDA Algorithm to extract keywords, we will adopt the following scheme: TF-IDF is used to weight each word in the data set to obtain the weighted vector representation. Through word

space construction and vectorization, we can obtain the theme word distribution of the dataset. Eventually, we calculate the similarity between the word distribution and the document distribution and take the keyword_ num words with the highest similarity as the keywords.

In this paper, the TF-IDF algorithm is used to extract a keyword from Twitter data. Since twitter comments usually include some text irrelevant to the topic, such as special symbols, emotions or some text without practical significance, this algorithm can filter and exclude these non-important data and optimize the data.

4.5. Model Evaluation

As an unsupervised learning algorithm, LDA is a convenient tool for topic identification since tagging texts is not required. However, identifying the best number of topics for LDA remains a challenge. The perplexity index is a common way to determine the optimal number of topics. The perplexity score indicates the uncertainty of the subject to which the document belongs. The lower the degree of perplexity, the better the model clusters. The optimal number of tweets is identified at the point of inflection on the perplexity vs topic number graph.

$$P(Dtest) = exp\left\{\frac{-\Sigma_{d-1}^{M} log(p(w_d))}{\Sigma_{d-1}^{M} N_d}\right\} \quad (1)$$

Where, *M* is the number of the test corpus (number of documents), Nd is the size of the d-th text (number of words or tokens), *z* is the topic, *w* is the document, and *r* is the text topic distribution based on the training set.

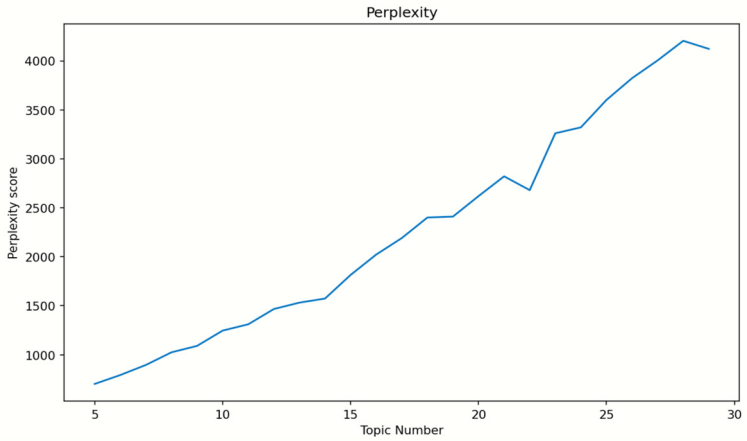


Figure 6. Perplexity score result

As shown in the graph above (Figure 6), no clear point of inflection can be identified. We speculate that the short length of tweets leads to unsatisfactory results of perplexity scores. To solve this problem, we propose a new research method based on input of new tweets to select the most appropriate number of topics.

Since the topic modeling part of this research aims to identify the public sentiment in each aspect of the 3C model, we proposed a new benchmark that directly evaluates a model's performance on classifying texts into predetermined topics [9].

We first manually generate 4 topics and 10 tweets in each topic. To evaluate an LDA model with *K* clusters, we obtain the cluster distribution of each tweet and record the index of the two clusters with highest probability. An array of cluster indexes for a generated topic was created by combining the recorded cluster indexes of each tweet within the generated topic. The score for the generated topic is calculated as the frequency of the mode divided by the length of the array, and the score for the overall model is calculated as the average score of the N-generated topics [17]. Since a proper model should maximize the number tweets of the same generated topic into the same LDA cluster, the model with the highest index performs the best.

We calculated the evaluation score of LDA models of 5 to 30 numbers of clusters and identified 15 clusters, scoring 0.35, as the best LDA model. Since we record the top two topics for each tweet, the maximum possible score is 0.5, and a score of 0.35 is at an acceptable level.

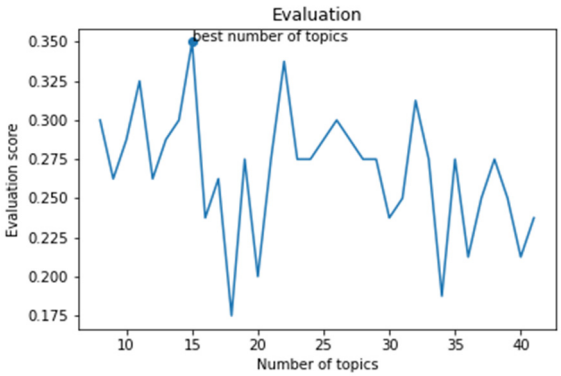


Figure 7. Evaluation score result

A topic description for each of the 15 clusters was written by examining the cluster keywords and tweets (Figure 7). Sort-cart method is used to aggregate those 15 clusters into different topics [18]. We first sampled 100 tweets for each cluster and assigned them to two independent authors. They would come up with a topic after reading through all the samples. If their topics agree with each other, then we have a topic for this specific cluster. If not, we would resample another 100 tweets and merge them with the first 100 ones, then repeat the whole process again until we have a topic. Through this so-called card sorting method, we subdivided the 15 topics into 5 themes extracted according to the 3C model of vaccine hesitancy: Safety Concerns, Belief that Covid is Harmless, Distrust, Efficacy, and Social Distribution (Figure 8).

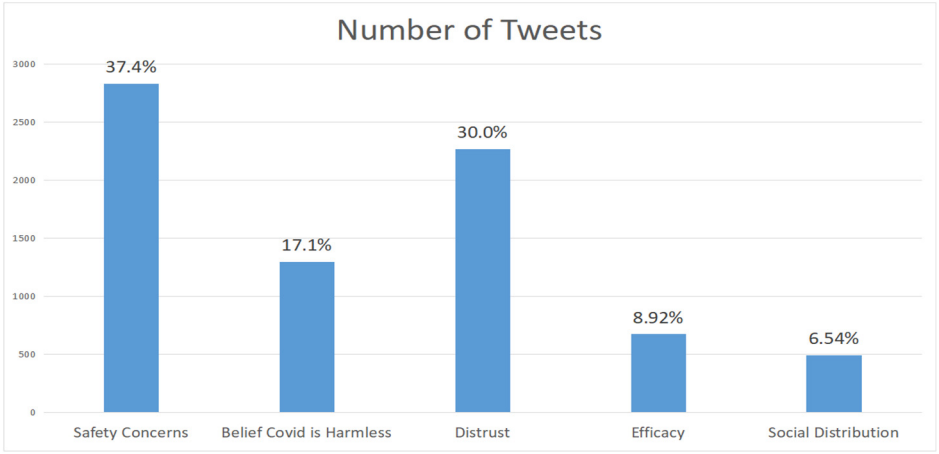
	Keywords	Topic discription	Number of Tweets	Mean Sentiment Intensity	Theme
Topic 1	effect, sore, booster, adverse	reports on adverse effects (fatigue, headache, fever)	1610	-0.511	safety concerns
Topic 2	walk, clinic, appoint	walk-in and appointments	1406	-0.412	Positive toward vaccine
Topic 3	get, risk, sick, severe	risk related to covid/vaccine	813	-0.499	safety concerns
Topic 4	risk, effect, long	side-effect and it's not necessary to take vaccine	807	-0.494	belief that covid is harmless
Topic 5	mask, sick, effect	People dont trust the vaccine and goveronment	718	-0.473	distrust
Topic 6	worker, refuse, health, risk	People feel vaccine is not effective so they dont take risk	674	-0.513	efficacy
Topic 7	reaction, booster	Anger towards people who don't take vaccine	668	-0.482	Positive toward vaccine
Topic 8	pregnant, women, death, effect	Discussion on whether pregnant women should receive vaccine	659	-0.486	Positive toward vaccine
Topic 9	health, effect, long public government	People don't trust the data released by the health department	601	-0.465	distrust
Topic 10	risk, minister, harsh, income	Different policies of different countries on the Coivd-19 vaccine and risks	572	-0.449	distrust
Topic 11	black, country, white, american	The vaccination rate of white people is much higher than that of black peopl	494	-0.487	social distribution
Topic 12	risk, children, need	children don't need vaccine	488	-0.453	belief that covid is harmless
Topic 13	need, effect, want	high risk health condition people need to take vaccine	419	-0.498	Positive toward vaccine
Topic 14	time, need, say, recieve	getting vaccine is painful	404	-0.485	safety concerns
Topic 15	adverse, effect, report	distrust of media	373	-0.467	distrust

Figure 8. Description of the 15 topics

5. Result and Discussion

Throughout the analysis period (08/2020 to 08/2022), 398,592 Tweets referencing COVID-19 vaccines were posted by Twitter users. We classify all tweets into positive, neutral, or negative sentiments through sentiment analysis. Vaccine hesitancy is a complex issue. The 3C (Confidence, Complacency, Convenience) Model of Vaccine Hesitancy developed by the SAGE Working Group is a useful tool to understand the main reasons that drive vaccine hesitancy [9].

The results of this study show that concerns over vaccine Safety (37.4 %), and Distrust towards the government (30.0 %) are the two major themes relating to vaccine hesitancy. The topic of Social Distribution is the least representative, accounting for only 6.54% of the total tweets (Figure 9).



s

Figure 9. Number of tweets for each topic

In the present study, we calculated the average Vader sentiment score for each theme. The lower the score, the more negative the attitude, and the more difficult it is to convince. The results show that the score of Distrust is the largest, identifying that people with distrust views are the most easily persuaded. These people are the priority target audience (Figure 10).

Figure 10. Average sentiment score for each topic

The following subsections summarize the evidence for 5 themes. Selected example tweets that are presented in table 1 below:

Table 1. Sample tweet for each topic

Topic Name:	Example tweet:
Safety Concerns	The second dose is harsher, more myalgia, higher fevers, muscle pain, passing out - the second dose is quite toxic (looking at the raw data now in yet another COVID meeting) ...never had that with my annual (required) flu shot
Belief Covid is Harmless	Children will only be vaccinated if vulnerable @Telegraph Common sense. Without clinical trial data on the really young too big a risk.
Distrust	A covid-vaccine is coming to a place near you! Some people however are sceptical as the development went so fast. So what do the scientists say, how safe are they, should we be worried? Article produced and financed by @SINTEF
Efficacy	The fda has no method of determining if the vaccines produce immunity how can you approve eua and have no method of determining its

	effectiveness the inmates are running the asylum it's been one big giant con game lie,"AND the FDA has no method of determining if the vaccines produce immunity! https://t.co/Odg2xY0XXD
Social Distribution	I'm sure there are white Q-racist holdouts from the #CovidVaccine, I've seen 30 headlines about it today, yet the demo clearly under-vaccinated by % is black Americans. Media agenda and political correctness glossing over this is going to end up with more dead black Americans.

4.6. Tweets Mapped to the "Safety Concerns" Theme

It appeared the most significant aspect of safety concerns amongst Twitter users was the post-vaccine side-effects (e.g., about "urticaria," "Allergy," "dermatitis," and "menstrual period"), with a small minority encouraging others to refuse the vaccine as a result.

We see a lot of tweets about the side effects of COVID-19 vaccines on the Internet due to the widespread of the COVID-19, the large number of people affected, and the high attention paid to the virus. However, no matter what kind of vaccine is inoculated, there will be more or fewer side effects, which cannot be changed. The severity of adverse symptoms varies from individual to individual. Generally speaking, mild adverse reactions do not require treatment, and most people can recover on their own after two or three days. If the symptoms are serious or last for a long time, such as fever and body temperature higher than 39 ℃ for more than three days, the likelihood of secondary infection cannot be ruled out, and the patient should be diagnosed and treated in a hospital in time.

We should pay attention not to take antibiotic drugs before and after the vaccination. At the same time, when we are in the acute stage of diseases, such as physical fever, diarrhea, poor mental state, and other situations, we cannot be vaccinated. Furthermore, take care not to bathe after vaccination to prevent bacteria from entering the body and causing infection. As long as we take these precautions seriously, most side effects can be avoided.

4.7. Tweets Mapped to the "Distrust" Theme

Normally, a vaccine will go through a long research and development process of at least 8 years or even more than 20 years from R&D to marketing. The previous fastest R&D record was the mumps vaccine, which took 4 years. So, a large number of people thought a vaccine developed within a year was impossible. A sense of fear that a fake vaccine was being rolled out to the public prompted some people to take a "wait and see" attitude towards vaccination.

There are several reasons for the rapid development of new crown vaccines. First, the virus was identified at the first time. Although COVID-19 is a new disease, the research on coronavirus has been started since Sara in 2002. Scientists have accumulated a lot of research and development experience. Second, the great development of global biotechnology and unprecedented international cooperation. Third, due to government support, R&D funds are abundant. Fourth, it is easy to recruit volunteers due to the outbreak of the epidemic.

In the United States, when approving the first new vaccine, it was the first time to fully disclose the whole day's review, query and reply to the public in an open and live broadcast manner. Through this open, transparent, scientific, and rigorous way, people can trust the government and the vaccine.

4.8. Tweets Mapped to the "Belief that Covid is Harmless" Theme

This theme is extracted from "complacency" in the 3C model. People holding the opinion that covid is harmless were often believers in Some people believe that covid symptoms are not serious enough. Thus, there is no need to take measures to prevent the disease. Specifically, some believe that since children don't often experience severe symptoms, they need not take the vaccine.

As suggested by Finney Rutten et al in a paper published on NIH, this problem can be solved by organizational-level intervention, requiring vaccination for childcare, school and college attendance [19]. Furthermore, interpersonal-level interventions, such as suggestions from clinicians also should be carried out.

4.9. Tweets Mapped to the "Efficacy" Theme

Some people chose not to believe that vaccination is an effective way to fight Covid-19. Some indicate that people who are vaccinated can still get Covid-19, so there is no need to take that risk when there is no clear sign that vaccination is effective enough.

A proper approach for those people is that we can provide them with statistics on the percentage of people who get Covid-19 for vaccinated people Vs. the percentage of people who get Covid-19 for non-vaccinated people. In that way, it will be easier to convince them that vaccination is effective enough to protect them.

4.10. Tweets Mapped to the "Social Distribution" Theme

Because of the influence of the constitution, the legal system and social system allocation, social exchange, and retribution, the component of the social system is divided into different classes and levels, such as age, income, race, nationality, level of education, living environment, immigrant and non-immigrant, working conditions, all of these may be considered to be the basic fair. It may also be criticized as unjust, and economies and tax systems may be criticized for allowing large wealth inequalities to develop or for undermining individual productivity; Education systems can be blamed for failing to provide equal opportunities for socially disadvantaged groups or for failing to meet the needs of gifted students. Through research, we found that white people's vaccination rates are much higher than blacks, perhaps with a level of education and family income connection, but no matter what language are you saying, how much is your annual household income, in dealing with the new champions this strong global infectious virus, we treat the seriousness of cautious attitude should not have any change, this is for the sake of our own, For our families, and for our country.

6. Conclusion

This study contributes to a growing body of work investigating Twitter as a source of soft intelligence, which can be used to capture real-time public insights, attitudes, and emerging trends concerning a particular health issue. Knowledge hidden in these tweets can help the CDC find compelling ways of convincing people to vaccinate. Policymakers need to design a well-researched immunization campaigns to remove vaccination barriers, safety concerns, and misconceptions about the Covid-19 vaccines.

However, our current research has many limitations. Firstly, in terms of topic modeling, the LDA topic model is more suitable for processing long text. Therefore, the results of using LDA to process short text-like tweets are not good enough. Secondly, as for data filtering, we found that although the Vader sentiment score of some tweets was less than -0.05, they expressed a positive attitude toward vaccines. We tried to overcome this limitation by manual examination to ensure the validity of the results.

We believe that our research proves the practicability of existing NLP tools. In the future, we can employ comments from social media data to support public health research.

Reference

[1] MacDonald, N. E. (2015). Vaccine hesitancy: Definition, scope and determinants. Vaccine, 33(34), 4161–4164. https://doi.org/10.1016/j.vaccine.2015.04.036

[2] Muric, G., Wu, Y., & Ferrara, E. (2021). COVID-19 vaccine hesitancy on social media: Building a public Twitter data set of antivaccine content, vaccine misinformation, and conspiracies. *JMIR Public Health and Surveillance*, *7*(11), e30642.

[3] Bari, A., Heymann, M., Cohen, R. J., Zhao, R., Szabo, L., Apas Vasandani, S., Khubchandani, A., DiLorenzo, M., & Coffee, M. (2022). Exploring Coronavirus Disease 2019 Vaccine Hesitancy on Twitter Using Sentiment Analysis and Natural Language Processing Algorithms. *Clinical Infectious Diseases*, *74*(Supplement_3), e4–e9. https://doi.org/10.1093/cid/ciac141

[4] Mishra, S., Verma, A., Meena, K., & Kaushal, R. (2022). Public reactions towards Covid-19 vaccination through Twitter before and after the second wave in India. *Social Network Analysis and Mining*, *12*(1), 57. https://doi.org/10.1007/s13278-022-00885-w

[5] Bustos, V. P., Comer, C. D., Manstein, S. M., Laikhter, E., Shiah, E., Xun, H., Lee, B. T., & Lin, S. J. (2022). Twitter Voices: Twitter Users' Sentiments and Emotions About COVID-19 Vaccination within the United States. *European Journal of Environment and Public Health*, *6*(1), em0096. https://doi.org/10.21601/ejeph/11499

[6] Cotfas, L.-A., Delcea, C., Roxin, I., Ioanăş, C., Gherai, D. S., & Tajariol, F. (2021). The Longest Month: Analyzing COVID-19 Vaccination Opinions Dynamics From Tweets in the Month Following the First Vaccine Announcement. *IEEE Access*, *9*, 33203–33223. https://doi.org/10.1109/ACCESS.2021.3059821

[7] Alam, K. N., Khan, M. S., Dhruba, A. R., Khan, M. M., Al-Amri, J. F., Masud, M., & Rawashdeh, M. (2021). Deep Learning-Based Sentiment Analysis of COVID-19 Vaccination Responses from Twitter Data. *Computational and Mathematical Methods in Medicine*, *2021*, e4321131. https://doi.org/10.1155/2021/4321131

[8] Jun, J., Zain, A., Chen, Y., & Kim, S.-H. (2022). Adverse Mentions, Negative Sentiment, and Emotions in COVID-19 Vaccine Tweets and Their Association with Vaccination Uptake: Global Comparison of 192 Countries. *Vaccines*, *10*(5), 735. https://doi.org/10.3390/vaccines10050735
[9] Lanyi, K., Green, R., Craig, D., & Marshall, C. (2022). COVID-19 Vaccine Hesitancy: Analysing Twitter to Identify Barriers to Vaccination in a Low Uptake Region of the U.K. *Frontiers in Digital Health*, *3*. https://www.frontiersin.org/articles/10.3389/fdgth.2021.804855
[10] Huang, P. H. (2020). Pandemic Emotions: The Good, the Bad, and the Unconscious-Implications for Public Health, Financial Economics, Law, and Leadership. *Nw. J.L. & Soc. Pol'y*, *16*, 81.
[11] Karami, A., Zhu, M., Goldschmidt, B., Boyajieff, H. R., & Najafabadi, M. M. (2021). COVID-19 Vaccine and Social Media in the U.S.: Exploring Emotions and Discussions on Twitter. *Vaccines*, *9*(10), 1059. https://doi.org/10.3390/vaccines9101059
[12] Hayawi, K., Shahriar, S., Serhani, M. A., Taleb, I., & Mathew, S. S. (2022). ANTi-Vax: A novel Twitter dataset for COVID-19 vaccine misinformation detection. *Public Health*, *203*, 23–30. https://doi.org/10.1016/j.puhe.2021.11.022
[13] Lazarus, J. V., Wyka, K., White, T. M., Picchio, C. A., Rabin, K., Ratzan, S. C., Parsons Leigh, J., Hu, J., & El-Mohandes, A. (2022). Revisiting COVID-19 vaccine hesitancy around the world using data from 23 countries in 2021. Nature Communications, 13(1), 3801. https://doi.org/10.1038/s41467-022-31441-x
[14] Baj-Rogowska, A. (2021). Mapping of the Covid-19 Vaccine Uptake Determinants From Mining Twitter Data. *IEEE Access*, *9*, 134929–134944. https://doi.org/10.1109/ACCESS.2021.3115554
[15] Lyu, J. C., Han, E. L., & Luli, G. K. (2021). COVID-19 Vaccine–Related Discussion on Twitter: Topic Modeling and Sentiment Analysis. *Journal of Medical Internet Research*, *23*(6), e24435. https://doi.org/10.2196/24435
[16] Hutto, C., & Gilbert, E. (2014). VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. Proceedings of the International AAAI Conference on Web and Social Media, 8(1), 216-225. Retrieved from https://ojs.aaai.org/index.php/ICWSM/article/view/14550
[17] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.
[18] J. Tomaszczyk, ‘‘Taksonomia jako narze˛dzie organizacji informacji,’’ Zagadnienia Informacji Naukowej, Wydawnictwo Uniwersytetu Śląskiego, Katowice, Poland, Tech. Rep., 2007, pp. 40–49.
[19] Finney Rutten, L. J., Zhu, X., Leppin, A. L., Ridgeway, J. L., Swift, M. D., Griffin, J. M., StSauver, J. L.,Virk, A., & Jacobson, R. M. (2021b). Evidence-Based Strategies for Clinical Organizations to Address COVID-19 Vaccine Hesitancy. *Mayo Clinic Proceedings*, *96*(3), 699–707. https://doi.org/10.1016/j.mayocp.2020.12.024