Topics and Sentiments in COVID-19 Vaccine-related Discussion on Twitter

Joanne Chen Lyu, Eileen Le Han, Garving K Luli

Submitted to: Journal of Medical Internet Research
on: April 18, 2021

Table of Contents

Topics and Sentiments in COVID-19 Vaccine-related Discussion on Twitter

Joanne Chen Lyu[1] PhD; Eileen Le Han[2] PhD; Garving K Luli[3] PhD

[1]Center for Tobacco Control Research and Education, University of California, San Francisco San Francisco US
[2]School of Information, University of Michigan, Ann Arbor Ann Arbor US
[3]Department of Mathematics, University of California, Davis Davis US

Corresponding Author:
Joanne Chen Lyu PhD
Center for Tobacco Control Research and Education, University of California, San Francisco
530 Parnassus Avenue
Address 2 (optional)
San Francisco
US

Abstract

Background: Though vaccination is a cornerstone for the prevention of communicable infectious diseases, vaccination has traditionally faced public fears, hesitancy. COVID-19 vaccines are no exception. Social media use plays a role in low acceptance of vaccines.

Objective: This study will identify the topics and sentiments in the public COVID-19 vaccine-related discussion on social media and discern the salient changes in topics and sentiments over time in order to better understand the public perceptions, concerns, and emotions that may influence the achievement of herd immunity goals.

Methods: Tweets were downloaded from a large-scale COVID-19 Twitter chatter data set from March 11, 2020, when the World Health Organization declared COVID-19 a pandemic, to January 31, 2020. We used R (The R Foundation) to clean the tweets and retain tweets that contained the following keywords: “vaccination", "vaccinations", "vaccine", "vaccines", "immunization", "vaccinate", and "vaccinated”. The final data set included in the analysis consisted of 1,499,421 unique tweets from 583,499 different users. We used R to perform the latent Dirichlet allocation algorithm for topic modeling as well as the sentiment and emotion analysis using the NRC (National Research Council in Canada) Emotion Lexicon.

Results: Topic modeling COVID-19 vaccine tweets yielded 16 topics, which were grouped into 5 overarching themes. Opinions about vaccination (227,840 tweets, 15.2%) was the most tweeted topic and remained the hottest topic in majority of the time of our examination. Vaccine progress around the world once became the hottest topic around August 11, 2020 when Russia approved the world's first COVID vaccine. With the advancement of vaccine administration, the topic of instruction on getting vaccines gradually became more salient and leapt to the hottest topic after the first week of January, 2021. Weekly mean sentiment scores showed that despite fluctuations, in general, the sentiment was increasingly positive. Emotion analysis further showed that trust was the most predominant emotion, followed by anticipation, fear, and sadness etc. Trust reached its peak on November 19, 2020 when Pfizer announced that its vaccine is 90% effective.

Conclusions: Public COVID-19 vaccine-related discussion on Twitter was largely driven by major events about COVID-19 vaccines and mirrored the hot news in mainstream media. Discussion also demonstrated a global perspective. The increasingly positive sentiment around COVID-19 vaccines and the dominant emotion of trust shown in the social media discussion may imply a higher COVID-19 vaccine acceptance compared with previous vaccines. Clinical Trial: N/P

(JMIR Preprints 18/04/2021:24435)
DOI: https://doi.org/10.2196/preprints.24435

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?
Please make my preprint PDF available to anyone at any time (recommended).
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.

✓ **Only make the preprint title and abstract visible.**
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Topics and Sentiments in COVID-19 Vaccine-related Discussion on Twitter

Abstract

Background: Though vaccination is a cornerstone for the prevention of communicable infectious diseases, vaccination has traditionally faced public fears, hesitancy. COVID-19 vaccines are no exception. Social media use plays a role in low acceptance of vaccines.

Objective: This study will identify the topics and sentiments in the public COVID-19 vaccine-related discussion on social media and discern the salient changes in topics and sentiments over time in order to better understand the public perceptions, concerns, and emotions that may influence the achievement of herd immunity goals.

Methods: Tweets were downloaded from a large-scale COVID-19 Twitter chatter data set from March 11, 2020, when the World Health Organization declared COVID-19 a pandemic, to January 31, 2020. We used R (The R Foundation) to clean the tweets and retain tweets that contained the following keywords: “vaccination", "vaccinations", "vaccine", "vaccines", "immunization", "vaccinate", and "vaccinated”. The final data set included in the analysis consisted of 1,499,421 unique tweets from 583,499 different users. We used R to perform the latent Dirichlet allocation algorithm for topic modeling as well as the sentiment and emotion analysis using the NRC (National Research Council in Canada) Emotion Lexicon.

Results: Topic modeling COVID-19 vaccine tweets yielded 16 topics, which were grouped into 5 overarching themes. Opinions about vaccination (227,840 tweets, 15.2%) was the most tweeted topic and remained the most discussed topic in majority of the time of our examination. Vaccine progress around the world once became the most discussed topic around August 11, 2020 when Russia approved the world's first COVID vaccine. With the advancement of vaccine administration, the topic of instruction on getting vaccines gradually became more salient and leapt to the most discussed topic after the first week of January, 2021. Weekly mean sentiment scores showed that despite fluctuations, in general, the sentiment was increasingly positive. Emotion analysis further showed that trust was the most predominant emotion, followed by anticipation, fear, and sadness etc. Trust reached its peak on November 19, 2020 when Pfizer announced that its vaccine is 90% effective.

Conclusions: Public COVID-19 vaccine-related discussion on Twitter was largely driven by major events about COVID-19 vaccines and mirrored the hot news in mainstream media. Discussion also demonstrated a global perspective. The increasingly positive sentiment around COVID-19 vaccines and the dominant emotion of trust shown in the social media discussion may imply a higher COVID-19 vaccine acceptance compared with previous vaccines.

Introduction

As the COVID-19 pandemic spread globally, COVID-19 vaccine-related issues have caught more public attention. Multiple research teams in major pharmaceutical companies and research institutions across nations have been developing vaccines [1, 2]. Though vaccination is a cornerstone for the prevention of communicable infectious diseases [3], vaccination has traditionally faced public fears, hesitancy, and even opposition [4, 5]. In the COVID-19 pandemic, it is estimated that 55%-85% of the population, depending on the country and the infection rate, needs to receive the COVID-19 vaccine to provide herd immunity [6, 7]. However, a survey about COVID-19 vaccine intentions in September 2020 suggested that 21% of the public in the USA would definitely get vaccinated and 24 % would definitely not get vaccinated [8]. One factor leading to the low acceptance of vaccines is poor health literacy, which is significantly influenced by social media use [9]. Therefore, there is a pressing need to understand how COVID-19 vaccines have been discussed on social media in order to better understand the public perceptions, concerns, and sentiments that may influence the achievement of herd immunity goals.

Even though social media data analysis has been widely used for both health-related issues and breaking public health crises [10-14], analysis of big data on social media discussion on COVID-19 vaccine was scant [15, 16]. To the best of our knowledge, the most recent published work about COVID-19 vaccine-related social media discussion ended in November 2020 [16]. However, many significant events around COVID-19 vaccines happened after that, such as the confirmation of more COVID-19 variant cases in North America by the Centers for Disease Control and Prevention (CDC), rollout of vaccines, and more vaccines showing high efficacy. Previous studies found that changes on social media discussion about vaccine-related topics correspond to the changing reality [15, 17]. Thus, research involving recent social media data is needed to fully understand the public discussion on COVID-19 vaccines during the pandemic. In addition, knowing the content of COVID-19 vaccine discussion on social media will provide possible explanation for their attitude toward COVID-19 vaccines and COVID-19 vaccine acceptance/hesitancy. However, previous research on COVID-19 vaccines lacked of efforts in line [16, 18-21]. To fill this gap, this study will examine the public discourses about COVID-19 vaccines on Twitter since World Health Organization (WHO) announced the global pandemic in March 11 2020, up until January 31, 2021 to identify the topics, overarching themes, and sentiments around COVID-19 vaccines and vaccination. This is the first study to cover almost a year of the pandemic. Such a long-time span would not only allow us to see a bigger picture of the public's discussion and concerns over COVID-19 vaccines, but also discern the salient changes in major topics and sentiments during the course of the pandemic, and further inform public health education and campaigns for increasing COVID-19 vaccine acceptance. Also, it may provide insight for other vaccine promotion.

Methods

Data Extraction and Preprocessing

The IDs from a total of 1,499,421 tweets, without retweets, from March 11 through January 31, 2021, were obtained using the data set maintained by Georgia State University’s Panacea Lab [22]. These tweets were collected by the Panacea Lab using the following 13 keywords: COVD19, CoronavirusPandemic, COVID-19, 2019nCoV, CoronaOutbreak, coronavirus, WuhanVirus, covid19, coronaviruspandemic, covid-19, 2019ncov, coronaoutbreak, and wuhanvirus. Since Twitter's Terms of Service do not allow the full JavaScript Object Notation (JSON) for data sets of tweets to be distributed to third parties, Georgia State University's Panacea Lab can only provide tweet IDs [23], which can be hydrated to obtain the JSON objects from these tweets.

During the tokenizing stage, we used the *gsub* function in R (The R Foundation) to extract the tweets whose language field in the tweets' metadata was specified as English. All text mining was done using RStudio Version 1.4.1103 on a Mac running 11.2.2 Big Sur. We converted all the tweets to lowercase. We further filtered the tweets by the following keys words: "vaccination", "vaccinations", "vaccine", "vaccines", "immunization", "vaccinate", and "vaccinated". We prepared two batches of tweets, one for text mining and the other for sentiment/emotion analysis. The data processing procedures for the two batches are almost the same except at the beginning: For sentiment analysis, we converted all the emojis to words whereas for text mining, we removed all the emojis. Next, we created a script to remove the URLs, mentioned names, non-ASCII (American Standard Code for Information Interchange) characters, and anything other than English letters or spaces (eg, "1," "?," etc). Using the R package *dplyr*, version 1.0.2, we cleaned the tweets by removing duplicates. In order to filter tweets created by faked or bots accounts, we used the document-term matrix (DTM) which contains rows corresponding to the tweets and columns corresponding to the terms. Each entry in the DTM denotes the number of times a term appears in a tweet. The similarity matrix $S=[S_{ij}]$, which measures how similar the *i*-th tweet is to the *j*-th tweet, is obtained by computing the dot product $S_{ij}=¿ R_i \cdot R_j \vee \frac{¿}{\|R_i\|\|R_j\|} ¿$ between the i-th row vector R_i and the j-th row vector R_j in the DTM (document-term matrix), which geometrically represents the cosine of the angle between the row vectors R_i, R_j. Therefore, if the i-th and the j-th tweets are identical, then $S_{ij}=1$; if they are completely different (that is, the angle between the corresponding row vectors is 90 degrees), then $S_{ij}=0$. For tweets that are 80% similar, we retain the most representative one (measured by the magnitude of the row vector in the document-term-matrix). Furthermore, we used the *tweetornot* Version 0.1.0 package to remove users that are identified as bots with a 95% probability or higher. "Overall, the default model is correct 93.8% of the time when classifying bots" [24].

The final cleaned data set consisted of 1,499,421 unique tweets from 583,499 different users. We further cleaned the tweets by removing words and characters that were of little or no analytical value (eg, "the," "very," "&," etc). We performed this task by creating our own list of stop words by appending the 13 keywords related to "COVID19" and the seven keywords related to "vaccine" to the English stop words list from the R package *tidytext*, version 0.2.6; this was done because we already knew that every tweet would contain one or more of those keywords and having them in the tweets does not contribute to further our understanding of the main content of the tweets. Lastly, we stemmed and lemmatized the words to their root forms using the R package *textstem*, version 0.1.4 (eg, *studying*, *studies*, and *studied* were converted to *study*). See Figure 1 for a summary of our data preprocessing procedure.

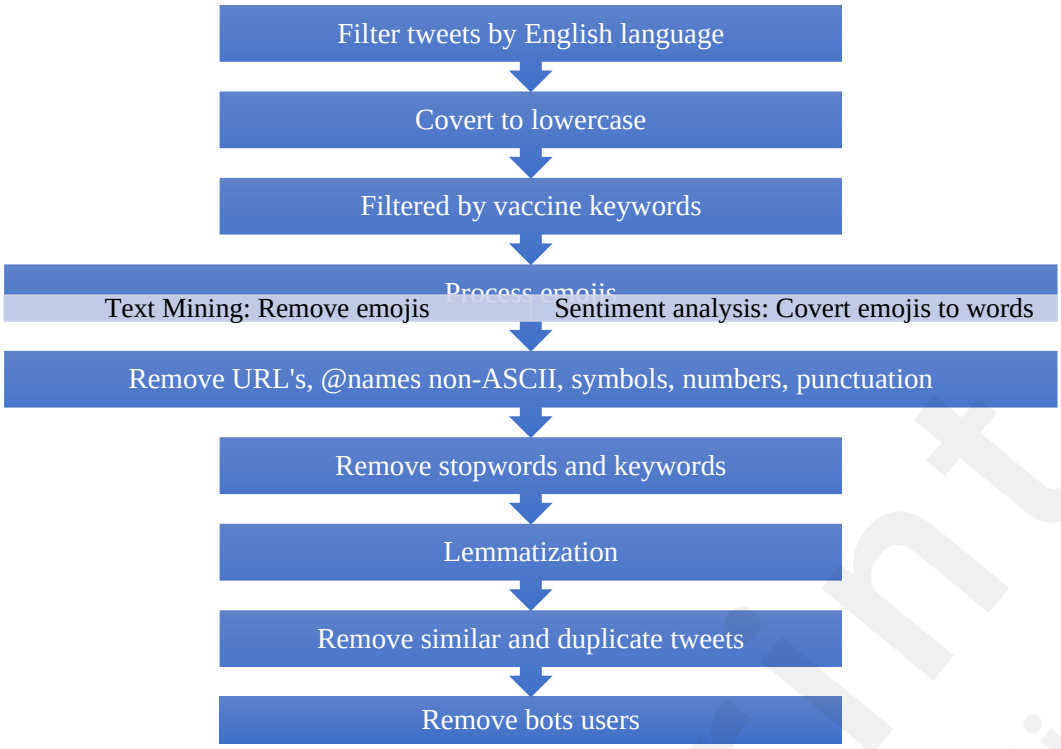

Figure 1. Data preprocessing procedure

Topic Modeling

Topic modeling provides an automatic, or unsupervised, way of summarizing a large collection of documents. It can help discover hidden themes in the collection, group documents into the discovered themes, and summarize the documents by topic. Topic modeling is often referred to as *soft* clustering, but it is more robust and provides better and more realistic results than typical clustering (eg, k-means clustering) or *hard* clustering [25]. A typical clustering algorithm assumes a distance measure between topics and assigns one topic to each document, whereas topic modeling assigns a document to a collection of topics with different weights or probabilities without any assumption on the distance measure between topics. There are many topic models available. “The most widely used model for topic modeling is the Latent Dirichlet Allocation (LDA) model” [26], developed by David Blei, Andrew Ng, and Michael I Jordan in 2002 [27].

To extract common topics from this sheer number of tweets, we used the LDA algorithm for topic modeling. We performed the LDA algorithm on the data using the R *textmineR* package, version 3.0.4. The LDA algorithm requires manually inputting the number of expected topics. We ran the LDA algorithm on the data by varying the topic number from 2 through 40. For each topic number, we calculated the coherence score using the *textmineR* package; we ended up choosing 16 topics for the final model, as the topic number that was equal to 16 yielded the highest coherence score. (See Appendix 1)

The top eight terms from each of the 16 topics were generated. We also used the *geo_freqpoly* function in the R package *ggplot2*, version 3.3.2, to generate the frequency polygons (see Figure 2) to visualize the weekly frequency of the 16 topics from March 11, 2020 to January 31, 2021. For each tweet, the LDA assigned a probability to each of the 16 topics. We assigned the topic with the highest probability to a tweet and we grouped the tweets according to the most prevalent topics. To obtain representative tweets for each topic, we randomly sampled 100 tweets from each topic; the two authors then independently examined the sampled tweets, followed by a group discussion to

select the most representative ones. If one of the authors thought that there were no conspicuous topics that emerged from the first 100 sampled tweets, another 100 tweets would be sampled and further reviewed; the authors continued this process until the two judged that there was a clear common topic and they reached a consensus (see our previous paper [28] for more details). We used the *textmineR* package's topic label function to generate an initial labeling for the topics. After carefully reading through the sampled tweets from each topic, the two authors refined the machine-generated labeling to give each topic the most accurate, concise, and coherent description (see Table 1). Through discussions, the authors further grouped the topics into 5 overarching themes. Specifically, two of the authors first independently grouped the topics into any number of themes that made most sense to them and resolved conflicting views through discussion. The third author was involved for additional comments on both the agreement and disagreement between the two authors. The final decision of the grouping was made together by all three authors. For example, whether the topic of 'vaccination drive in India' should be grouped into the theme of 'vaccine administration' or 'vaccines as a global issue' was unresolved after the two-author discussion. By re-reading tweets and discussing among the three authors, we finally had it as a topic under the theme of 'vaccines as a global issue'.

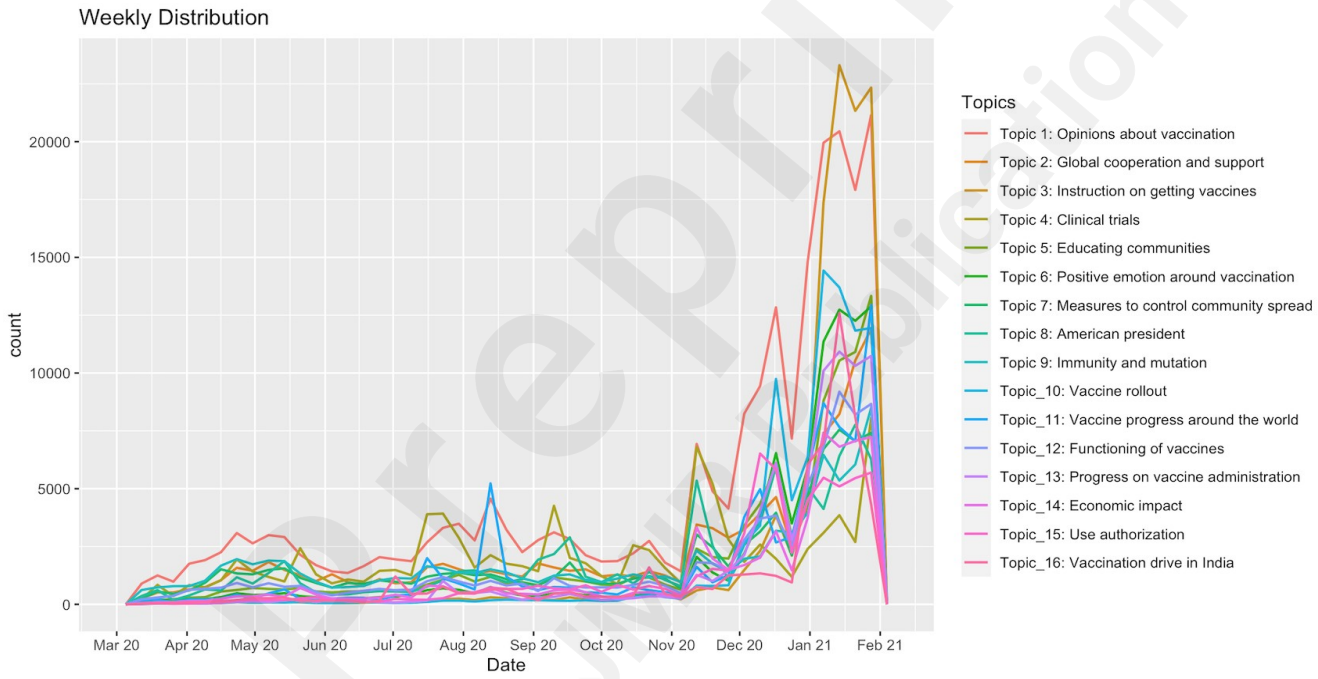

Figure 2. Weekly frequency of each topic on Twitter, from March 11, 2020 to January 31, 2021.

Sentiment and Emotion Analyses

Sentiment analysis can classify the polarity of a given document; it can give a score to a document to indicate whether the expressed opinion is positive, negative, or neutral. Emotion analysis goes beyond the simple polarity and can give a score to different emotions such as anger, fear, anticipation, trust, surprise, sadness, joy, and disgust (the so-called Plutchik's wheel of emotions) [29]. The *syuzhet* package (Jockers, 2017) is the most popular and efficient R package for sentiment/ emotion analysis [30]. To perform the emotion analysis, we used the NRC (National Research Council in Canada) Emotion Lexicon developed by Turney and Mohammad (2010) [31]. It is the most comprehensive dictionary for this task [32]. In Figure 3, we showed the weekly average polarity (sentiment) scores from March 11, 2020 to January 21, 2021; we fitted the data points with a best linear fit and obtained a slope of 0.003764 with p-value = 2.80e-05 and an intercept of 0.1653927 with p-value = 2.34e-09. In Figure 4, we plotted the weekly percentage emotions.

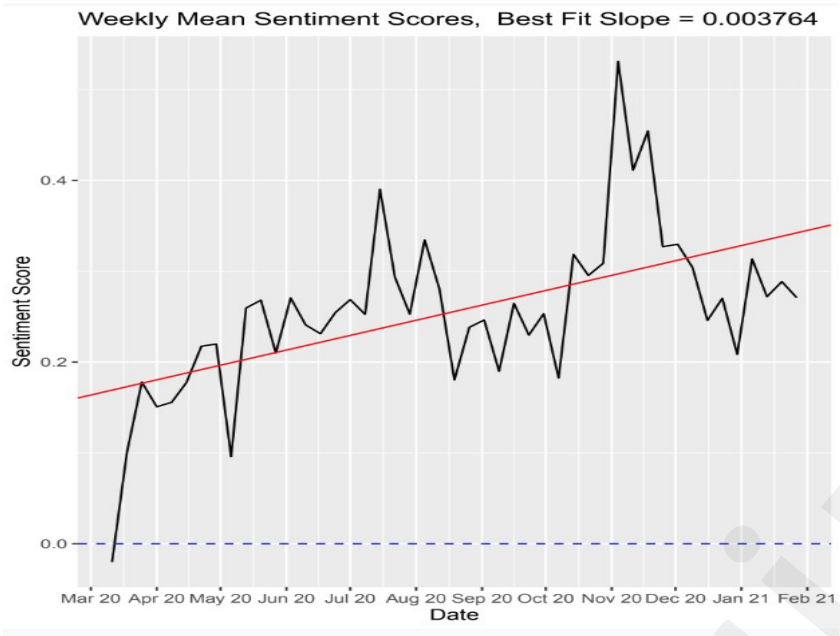

Figure 3. Weekly average polarity (sentiment) scores, from March 11, 2020 to January 21, 2021.

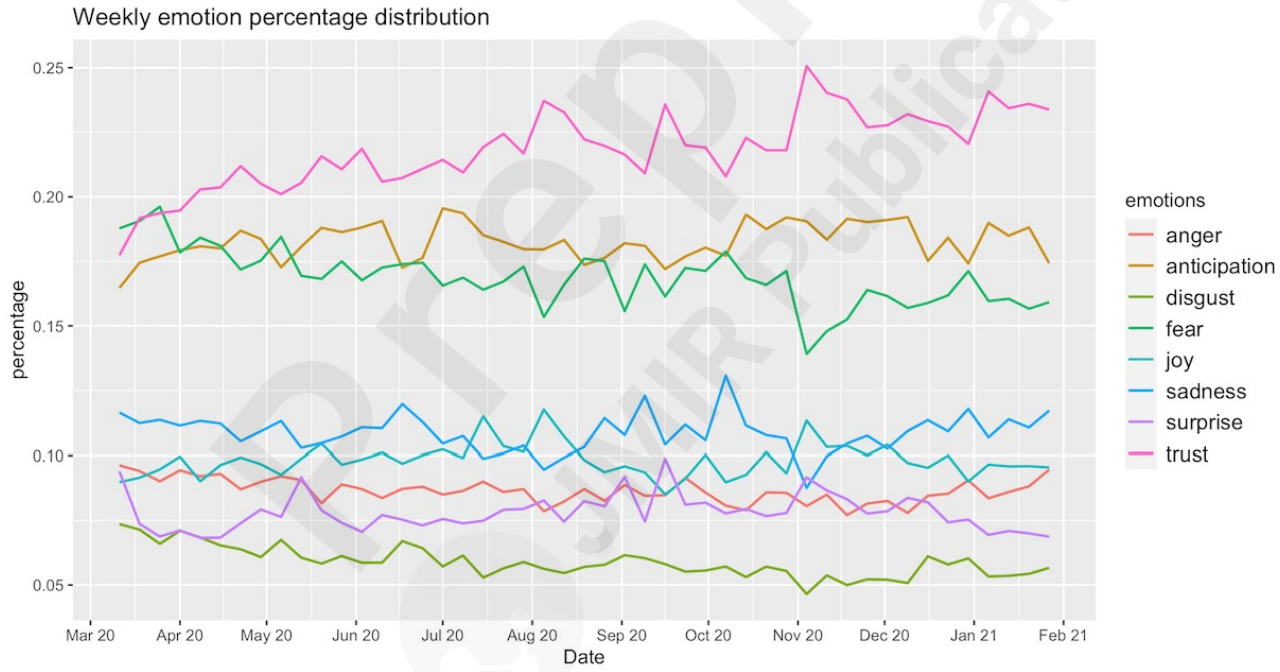

Figure 4. Weekly percentage emotions, from March 11, 2020 to January 21, 2021.

Results

We downloaded 144,332,894 tweets from March 11, 2020 through January 31, 2021 (for a total of 327 days) using the Panacea Lab database. After cleaning, a total of 1,499,421 tweets from 583,499 different users were included in the analysis. As shown in Figure 5, the number of daily tweets kept increasing; the daily average for the month of January 2021 is 22,202 tweets). Before November 9, 2020, the number of daily tweets is around and below 5,000 with only one exception on August 11, 2020 (n=7,486), when Russia approved the world's first COVID vaccine. The first wave of exponential increase in the number of daily tweets started from November 19, 2020 (n=12,720)

when Pfizer said its vaccine is 90% effective. The second wave of surge in the number of daily tweets started around Jan 3, 2021 when more COVID-19 variant cases were confirmed in North America by the Centers for Disease Control and Prevention (CDC) and the highest number of tweets in a single day (n=31,197) occurred on January 29, 2021 when Johnson & Johnson's and Novavax's vaccines showed 85% and 89.3% efficacy respectively.

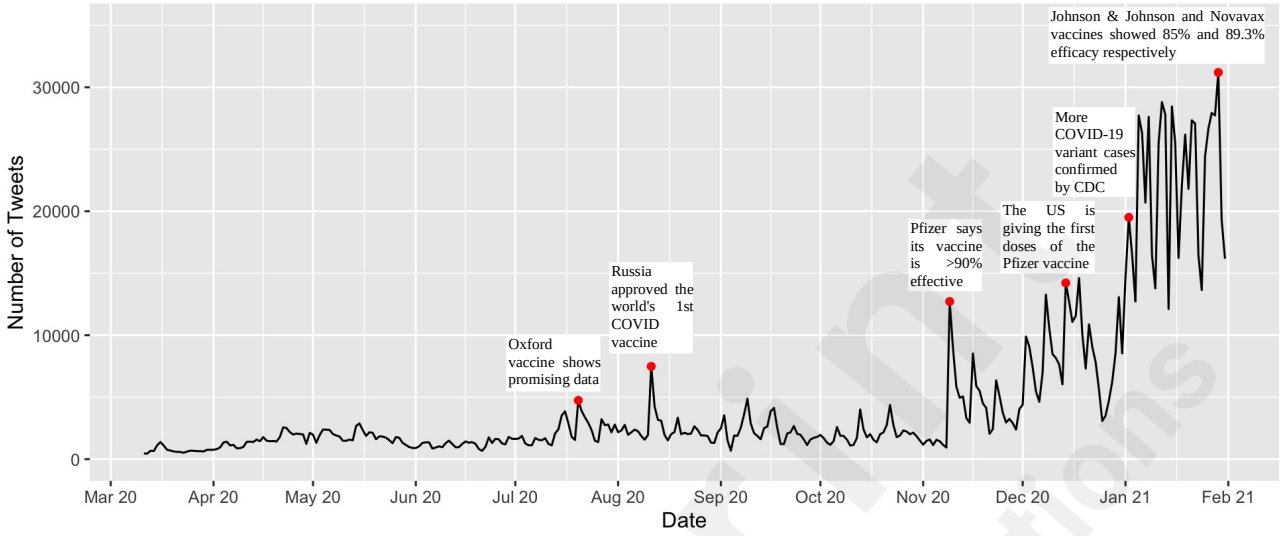

Figure 5. Daily number of COVID-19-related tweets, from March 11, 2020 to January 21, 2021.

Topic modeling COVID-19 vaccine Tweets

Analysis of tweets yielded 16 topics, which were grouped into 5 overarching themes. In the descending order of quantity of tweets, they are: opinions and emotions around vaccines and vaccination (405,397 tweets, 27.04%), knowledge around vaccines and vaccination (355,305 tweets, 23.7%), vaccines as a global issue (311,251 tweets, 20.76%), vaccine administration (266,744 tweets, 17.79%), and progress on vaccine development and authorization (160,724 tweets, 10.72%). Table 1 summarizes the 16 topics, top terms in topics, the number and percentage of each topic, and provides a tweet example for each topic. More details about the themes and topics, including the salient temporal variance of topics (see Figure 2), are elaborated in the following subsections.

Table 1. Numbers and percentages of COVID-19 vaccine tweets related to each topic

Themes	Topics	Top Terms contributing to topic model	Total tweets (%)	Examples of paraphrased tweets* (date posted)
Theme 1: Opinions and emotions around vaccines and vaccination	Opinions about vaccination	People, get, take, go, want, make, like, think	227,840 (15.20)	It is pathetic to see the low trust in science and the government. People need to see leaders and politicians on TV to receive the vaccine to convince them that the vaccine is safe. (December 18, 2020)
	Positive emotion around vaccination	Get, good, first, day, today, work, one, feel	89,117 (5.94)	After getting the first dose of the covid-19 vaccine today, I can finally breath some fresh air and feel there is hope in life. (January 23, 2021)
	American president	Trump, Biden, American,	88,440 (5.90)	There are more lies than usual in today's press. Biden is being portraited

		president, gate, bill, plan, administration		as anti-vaccine. This is not correct. In fact, he said on Wednesday he trusted the vaccines and the scientists, while accusing President Trump of playing politics with a potential covid-19 vaccine. (September 16, 2020)
Theme 2: Knowledge about vaccines	Educating communities	Question, read, answer, expert, article, community, immunity, black	96,532 (6.44)	You can learn more about the covid-19 vaccines in our upcoming town hall meeting. We will address any questions and concerns on January 13, 2021, 2-3:30PM (EST). (January 7, 2021)
	Measures to control community spread	Mask, get, need, go, still, available, end, life	89,008 (5.94)	Even covid-19 vaccines are an important step in slowing down the spread of the virus, people still need to continue taking all precautions: wear a mask, maintain physical distance from others, and keep your hands clean. (January 19, 2021)
	Immunity and mutation	Flu, year, new, may, variant, work, effective, strain	88,327 (5.89)	They say that coronavirus resembles the flu virus. As we know, the flu virus mutates. Therefore, they need to create different flu shot every year into order to fight off the virus, but there hasn't been a flu vaccine that is 100% effective. So good luck with making the covid-19 vaccine. (May 11, 2020)
	Functioning of vaccines	People, effect, test, risk, side, side effect, immune, death	81,438 (5.43)	Both vaccines use mRNA technology which contains instructions that tell our cells how to make a protein that triggers an immune response inside our bodies. (December 21, 2020)
Theme 3: Vaccines as a global issue	Global cooperation and support	World, country, global, pandemic, access, need, develop, effort	108,366 (7.23)	On Thursday, WHO warns against "vaccine nationalism". No one country is safe if poor countries can't get access to the vaccines. (August, 9, 2020)
	Vaccine progress around the world	UK, china, EU, Russia, first, country, approve, news	83,156 (5.55)	On Tuesday, German health minister Jens Spahn joined EU to place restriction on exporting covid-19 vaccines. This came amid discontent over the delay in rolling out the AstraZeneca vaccine to member countries. (January 26, 2021)
	Economic impact	New, case, death, news, rollout, rate, high, report	61,360 (4.09)	A rise in new covid-19 cases outweigh positive vaccine news. Shutdown fears sink global markets. US coronavirus deaths are above 250,000. (November

				19, 2020)
	Vaccination drive in India	India, drive, minister, Indian, speed, warp, health, warp speed	58,369 (3.89)	India clears covid-19 vaccine makers Serum Institute of India (SII) and Bharat Biotech to start the world's largest immunization drive. (January 3, 2021)
Theme 4: Vaccine Administration	Instruction on getting vaccines	County, health, state, appointment, update, plan, site, distribution	106,544 (7.11)	Health departments will each announce their vaccine availability and locations. Eligible individuals at the federally qualified health center will be contacted regarding access to a covid-19 vaccine. (December 30, 2020)
	Vaccine rollout	Worker, care, health, first, receive, healthcare, hospital, staff	85,623 (5.71)	We have many sites delivering the covid-19 vaccines to the top priority group. Please be patient and if you are in the top priority group, you will be contacted. (January 17, 2021)
	Progress on vaccine administration	Dose, million, first, receive, week, Pfizer, people, first dose	74,577 (4.97)	Since Tuesday's update, Louisiana has administered 25,133 additional covid-19 vaccines. The second doses started being administered this week. 7,068 people have been fully vaccinated. In total, since mid-December, 81,585 doses have been administered. (January 7, 2021)
Theme 5: Progress on vaccine development and authorization	Clinical trials	Trial, clinical, Pfizer, Moderna, phase, effective, clinical trial, oxford	99,754 (6.65)	There are 70 covid-vaccines under development. Moderna is one of the first to test their covid-19 vaccine on humans. If the trial is successful, Moderna could reach the final stage of testing by Fall 2020. (April 26, 2020)
	Use authorization	Use, FDA, approve, emergency, approval, Pfizer, emergency use, health	60,970 (4.07)	The FDA's vaccine advisory committee has unanimously approved an emergency use of the covid-19 vaccine made by Pfizer and by Moderna. (January 19, 2021)

* The tweets were paraphrased for users' privacy.

Theme 1: Opinions and Emotions Around Vaccines and Vaccination

This theme contained three topics, among which the topic of opinions about vaccination (227,840 tweets, 15.2%) was the most tweeted topic among all the 16 topics and remained the most discussed topic in majority of the time of our examination (see Figure 2). It was about the mixed opinion of the tweeters about vaccination, including their doubt, hesitancy, trust, and advocacy. Some tweeters also asked for freedom about vaccination. The second topic, positive emotion around vaccination, featured happiness, hope, relief and other positive emotions that the public showed in their tweets. Many emotions were based on their direct and/or indirect experience with vaccination. The third

topic was named American president. The public's opinions and emotions were expressed through their comments on Trump and Biden's vaccine-related words and actions.

Theme 2: Knowledge Around Vaccines and Vaccination

This theme consisted of four topics focusing on understanding and facilitating understanding of COVID-19 vaccines and vaccination. The biggest topic under this theme was educating communities with a considerable number of tweets spreading information about the live webinars where health professionals would provide important information about COVID-19 vaccines and answer questions. The topic of measures to control community spread featured the recommendation of taking measures such as wearing mask, washing hands, and social distancing both before and after getting vaccinated. The topic of immunity and mutation were around the immunity of vaccines and mutation of coronavirus, which was frequently discussed in comparison with flu virus. Functioning of vaccines, as a topic, introduced how vaccines work and the subsequent symptoms and side effects after vaccination.

Theme 3: Vaccines as a Global Issue

The theme of vaccines as a global issue consisted of four topics, featuring the globality of the coronavirus and the COVID-19 vaccines; therefore, the topics under this theme were not centered on the US. It involves hot news about vaccine progress and updates around the world, such as international vaccine supplies, delivery, and purchases in many countries (i.e., the topic of vaccine progress around the world), vaccine impact on global economy (i.e., the topic of economic impact), and the world's largest inoculation drive (i.e., the topic of vaccination drive in India). The topic of vaccine progress around the world once became the most discussed topic around August 11, 2020 because Russia approved the world's first COVID vaccine (see Figure 2). The most salient topic under this theme was global cooperation and support, which called for global cooperation to accelerate the vaccine development and equitable access and advocated no vaccination nationalism. This was also the second most tweeted topic (108,366 tweets, 7.23%) among the 16 topics only after the topic of opinions about vaccination.

Theme 4: Vaccine Administration

This theme consisted of three topics covering several aspects of vaccine administration. The topic of vaccine rollout was mainly around vaccination to the top priority groups, including health care providers being the first in line. After the US gave the first doses of Pfizer vaccine on December 14, this topic remained one of the top three most discussed topic for about five weeks (see Figure 2). The second topic focused on the progress on the vaccine administration, including the shipment and supply of vaccines that directly relate to the progression of vaccine administration. The third topic under this theme was instruction on getting vaccines, which featured spreading the information from health authorities at various levels to guide the public to get vaccine shots. More detailed information was mentioned in the tweets in this topic, such as "if anyone is 75 and older, does not have internet access, and needs help to schedule an appointment, please call the central appointment desk for help at 1-866-960-0633 (a tweet on January 20, 2021)." With the advancement of vaccine administration, this topic gradually became more salient and leapt to the most discussed topic, surpassing the topic of opinions about vaccination after the first week of January and remained it till the end of January (see Figure 2).

Theme 5: Progress on Vaccine Development and Authorization

There are two topics under this theme, which focused on the development of COVID-19 vaccines and authorization by U.S. Food and Drug Administration (FDA). The topic of clinical trials was around the plan and process of clinical trials mainly from Pfizer and Moderna, and the updated results about clinical trials. This topic remained the most discussed topic for a week round July 20

when first human trial of Oxford coronavirus vaccine shows promise, and became the top two most discussed together with the topic of opinions about vaccination for three weeks after Pfizer said its vaccine is 90% effective on November 19, 2020 (see Figure 2). The topic of use authorization centered on FDA approved an emergency use authorization for COVID-19 vaccines.

Sentiment and emotion analyses on COVID-19 vaccine Tweets

Weekly mean sentiment scores showed that despite fluctuations, in general, the sentiment was increasingly positive from March 11, 2020 to January 31, 2021, with the linear best fit slope = 0.003764 (with p-value = 2.80e-05, which means that it is statistically significant) (see Figure 3). And the positive emotion reached its peak around November 19, 2020 when Pfizer announced its vaccine is 90% effective; on the same day the number of daily tweets became historically high before January as mentioned above. Emotion analysis further showed that trust was the most predominant emotion accounting for 22.78% of the eight emotions (anger, anticipation, disgust, fear, joy, sadness, surprise, and trust), followed by anticipation 18.34%, fear 16.29%, sadness 10.97%, joy 9.76%, anger 8.63%, surprise 7.60%, and disgust 5.63%. Noticeably, the most dominant emotion shown in COVID-19 vaccine tweets before April was fear but changed to trust since the week of April 1, 2020 (Figure 4). Trust remained the most dominant emotion since then and the number of tweets expressing trust continued to grow. In addition, it was observed that when the emotion of trust increased, the emotion of fear decreased. The trust emotion reached its peak on November 19, 2020 when Pfizer announced that its vaccine is 90% effective; on the same day, the fear emotion was expressed least throughout the time period of our examination. It was also noticed that besides the obvious change in the emotions of trust and fear over time, the other emotions were relatively stable in the period of March 11, 2020 to January 31, 2021 (see Figure 4).

Discussion

Principal Findings

This study has examined the sentiments and topics over a long-time span, covering the discussions about COVID-19 vaccine since it became a global pandemic (March 11 2020) to January 31, 2021, when multiple vaccines become available and mass vaccination has begun in the US and many other countries. This study adds to the latest research about the social impact of COVID-19 vaccine. For example, there are surveys addressing socio-demographic social media user characteristics and the social determinants of vaccine acceptance [18, 33]. Our research supplements studies like these by providing the discourse patterns on social media, that is, how people actually talked about vaccination intentions and other related issues. Researchers have taken a similar approach to studying very specific vaccine-related topics like side effects and the type of vaccines in the context of China. This study could provide a comparison in different cultural contexts, as the corpus was pulled from a global dataset and English tweets were analyzed. The identified specific topics and public sentiments could be used for further studies about specific vaccine-related topics among the English language social media users. Many of the current studies about COVID-19 vaccination focus on vaccine hesitancy and anti-vaccine messages [19] and some were using survey methods [34]. As the vaccines for COVID-19 are still very new, and the development took a very short period of time due to the urgent need, we would expect the expression of vaccine hesitancy in the public media channels. This study would help contextualize the major concerns about vaccine efficacy and safety.

This study has found some changing patterns of the discussion on Twitter along with the progress. Similar patterns of sentiments and topics found in another research using natural language processing and deep learning techniques on Facebook and Twitter posts [16]. The results of this research show that the patterns are valid across platforms. The number of tweets regarding COVID-19 vaccination

is largely driven by major events -- mainly the milestones in vaccine development and the new variants of the virus. Major spikes in the number of tweets are highly corresponding to these events. The analysis of each topic also shows such patterns. For example, from the weekly distribution of topics, what is noticeable is a sudden increase overall starting from early November, also right at the time of the Pfizer announcement. Scholars have studied how information flows from social media to mainstream news. In today's media ecology, the boundary between social and mainstream media is no longer clear but here we can see that social media discourse is largely mirroring what is happening in the news. Future studies could detect to what extent the social media discussion is shaped by mainstream media.

The sentiment analysis shows that the general sentiment towards COVID-19 vaccination is getting more positive over time. The overall sentiment score reached the highest in early November 2020, which also corresponded to the news about the high efficacy of Pfizer vaccine. As for the emotions, trust has always dominated the discussion, and reaches the highest point around early November, following Pfizer's announcement about the efficacy of its vaccine. The percentage of tweets is overall going up, showing that more people express their trust in the discussion of COVID-19 vaccine. The change of the percentage of tweets expressing fear mirrors that of trust. The overall percentage is going down, showing that as the vaccine development progresses, people are having less fear about the pandemic. The highest point was in mid-March 2020, after the declaration of global pandemic, and the lowest point was in early November at Pfizer's announcement. As the vaccine research and testing are getting closer to a promising result, the expression of fear declines. Other emotions, in terms of their percentages in overall tweets, remain more or less stable over time. Trust is the dominant emotion could be understood as a reflection of the vaccination as the only option. Unlike other kinds of vaccines that people can choose to take or not, given the prevalence of COVID-19, its speed of community spread, the disruption of normal life, and no other options proved to be efficient, vaccination has been increasingly taken as the only promising way out.

The topics related to opinions and emotions are most common, and among the three topics in this theme, the topic on opinions about vaccination takes the largest proportion. The COVID-19 vaccine development has been going along with the spread and variants of the virus, together with our increasing knowledge about the disease, all of which would become hot topics in public discourse. With still many uncertainties about the vaccine and options, we would expect mixed opinions surfacing on the platform. Such a mixed opinion about vaccination should also be situated in the larger context of the anti-vaccine movement in the United States and other countries as well. Our results show that doubts, vaccine hesitancy, conspiracy theories, and the argument of vaccination as individual freedom are all common themes in the anti-vaccine discourse. However, as mentioned earlier, as COVID-19 has affected people's day-to-day life, a vaccine is crucial for returning to normal life, people still have a lot of hope in it, leading to the sharing of positive emotions. Finally, with the political climate and the ongoing presidential election, both candidates were having vaccination on their agenda, which makes the discussion about the vaccine very politicized.

As COVID-19 is an ongoing crisis with a global scale, the discussion about the vaccine is also global. The pandemic reveals how much the world is now connected, and thus vaccination becomes a global issue – if a country cannot reach a certain level of vaccination of its population, it has high risk of contagion and virus mutation, making it hard for the country to get back to its role in the global economy, and global cooperation is needed in order to defeat the disease. Therefore, the economic impact of the pandemic and the vaccine development are salient issues.
The current available vaccines are also results of cross-national collaboration, which is why "vaccine nationalism" is frequently mentioned in the tweets, as something that would harm the progress of fighting the pandemic. In a global crisis, there are increasing needs of information seeking,

particularly when vaccines are becoming available, and people need instructive information for the vaccine rollout, getting vaccines, and the vaccine administration. As the COVID-19 is new and still being studied, and the vaccines also became available after a short period of research and development, there are a lot of uncertainties for the public to accept it. For lay people, the science behind the vaccine is still not well understood, particularly as COVID-19 is so new and many things are still unknown. The uncertainties provoked people to seek and share information about the vaccines, which becomes the topic of "knowledge around vaccines", such as the sciences behind the vaccines (immunity and mutation), the techniques used, and results from clinical trials. The uncertainty also provokes people to seek information about the vaccine rollout and administration, because they are closely connected to when and how they would be getting the vaccine.

The results of this study show that the discussions about COVID-19 vaccines are multi-faceted, and the public are actively seeking and sharing information about them. It is important for public health agencies to understand the major public interests and concerns regarding vaccination, that is, the major factors that would affect vaccine acceptance and hesitancy. In this way they could make appropriate strategies to facilitate the public communication. As our study shows that the major spikes in vaccine discussions corresponded to the breakthroughs of vaccine development, as they appeared in the news. Public health agencies need to pay attention to this pattern and monitor the online discussions on those days with major news about vaccine development or other significant events like the report of severe adverse events of a particular vaccination. They could develop immediate responses based on identifying the instant reactions – the dominant emotions and topics – on social media. As this study finds that trust is the predominant emotion regarding vaccines, it is promising and assuring for the public health agencies promoting vaccination. For individual themes and topics, public health agencies should pay close attention to the discussions about vaccine knowledge and administration, to close the information gap between the public needs and what they have provided.

Limitations

By looking at how topics and sentiments evolve along the timeline of the pandemic and vaccine development, one can see the correspondence of the volume of tweets and major research breakthroughs in the news. While it is clear that the number of tweets has a major spike at the news of Pfizer reporting its high vaccine efficacy in early November, a statistical analysis may provide more details about whether there are significant differences. Though not the focus of this study, close examination of the Twitter users may provide more meaningful information, such as how the contents shared by different kinds of users vary and whether certain types of users are more likely to post or comment on certain topics. In addition, this study may have geographic bias in tweets as that in many other unstructured textual data [35]. That means a given dataset could overrepresent some geographic areas. Last but not least, since Twitter users are not representative of the US population [36]. Our dataset could be overly representing a subset of the population with specific characteristics. Therefore, generalizing findings of this study should be very cautious.

Conclusion

This study identifies the major topics and sentiments about the COVID-19 vaccine-related issues discussed on social media. It also examines the changes over time to better understand a larger trend. Among the 16 distinct topics, opinions about vaccination was the most common topic and remains so over time. As the vaccine development were making progresses around the world, the dominant topics also shift. Instruction on getting the vaccine has become the most discussed topic around early January 2021. The discussion of COVID-19 vaccination on social media was largely driven by major news events about COVID-19 vaccines and mirrored the hot news in mainstream media. Also, the discussion has a global perspective. The overall sentiment was increasingly positive over time, and

trust was the predominant emotion, which shows that social media discussions may imply a higher COVID-19 vaccine acceptance compared with previous vaccines. Due to the timeline of our dataset, in this study we did not look further into the sentiments about specific brands of vaccines. We would expect that the discussions to be different when using different brand names to search for tweets and conduct sentiment analysis. Particularly, after the Johnson & Johnson vaccine being halted by CDC, there could be a surge of related discussions and a topic about the side effects of the vaccines may emerge. Therefore, a further study in this line is highly recommended.

Acknowledgments

This research was supported by National Cancer Institute Grant T32 CA 113710 (JL).

Conflicts of Interest

None declared.

References

1. Le TT, Cramer JP, Chen R, Mayhew S Evolution of the COVID-19 vaccine development landscape. Nature Reviews. Drug Discovery. 2020; 19 (10): 667–668. doi:10.1038/d41573-020-00151-8
2. Gottlieb S. America Needs to Win the Coronavirus Vaccine Race. The Wall Street Journal. https://www.wsj.com/articles/america-needs-to-win-the-coronavirus-vaccine-race-11587924258. published April 26, 2020 [accessed 2021-04-09].
3. Andre F, Booy R, Bock H, Clemens J, Datta SK, John TJ, Lee BW, Lolekha S, Peltola H, Ruff TA, et al. Vaccination greatly reduces disease, disability, death and inequity worldwide. Bull World Health Organ. 2008;86(2):140–46. doi:10.2471/BLT.07.040089
4. Abbasi, J. COVID-19 Conspiracies and Beyond: How Physicians Can Deal with Patients' Misinformation. JAMA 2021, 325, 208–210.
5. Ball P. Anti-vaccine movement could undermine efforts to end coronavirus pandemic, researchers warn. Nature 2020, 581, 251.
6. Kwok KO, Lai F, Wei WI, Wong SYS, Tang JWT. Herd immunity - estimating the level required to halt the COVID-19 epidemics in affected countries. J Infect. 2020;80(6):e32-e33.
7. Sanche S, Lin YT, Xu C, Romero-Severson E, Hengartner N, Ke R. High contagiousness and rapid spread of severe acute respiratory syndrome Coronavirus 2. Emerg Infect Dis. 2020;26(7):1470-1477.
8. Tyson A, Johnson C, Funk C. U.S. public now divided over whether to get COVID-19 vaccine. Pew Research Center. 2020. https://www.pewresearch.org/science/2020/09/17/u-s-public-now-divided-over-whether-to-get-covid-19-vaccine/ [accessed 2021-04-15]
9. Ashkenazi S, Livni G, Klein A, Kremer N, Havlin A, Berkowitz O. The relationship between parental source of information and knowledge about measles / measles vaccine and vaccine hesitancy. Vaccine. 2020;38(46):7292-7298.
10. Cole-Lewis H, Pugatch J, Sanders A, Varghese A, Posada S, Yun C, et al. Social listening: a content analysis of e-cigarette discussions on Twitter. J Med Internet Res 2015 Oct 27;17(10):e243
11. Scanfeld D, Scanfeld V, Larson EL. Dissemination of health information through social networks: twitter and antibiotics. Am J Infect Control 2010 Apr;38(3):182-188.
12. Lazard AJ, Scheinfeld E, Bernhardt JM, Wilcox GB, Suran M. Detecting themes of public concern: a text mining analysis of the Centers for Disease Control and Prevention's Ebola live

Twitter chat. Am J Infect Control 2015 Oct 1;43(10):1109-1111
13. Masri S, Jia J, Li C, et al. Use of Twitter data to improve Zika virus surveillance in the United States during the 2016 epidemic. BMC Public Health. 2019;19(1):761.
14. Signorini A, Segre AM, Polgreen PM (2011) The Use of Twitter to Track Levels of Disease Activity and Public Concern in the U.S. during the Influenza A H1N1 Pandemic. PLoS ONE 6(5): e19467. https://doi.org/10.1371/journal.pone.0019467
15. Bonnevie E, Gallegos-Jeffrey A, Goldbarg J, Byrd B, Smyser J. Quantifying the rise of vaccine opposition on Twitter during the COVID-19 pandemic. J Commun Healthc. 2021;14(1):12-19. DOI: 10.1080/17538068.2020.1858222
16. Hussain A, Tahir A, Hussain Z, et al. Artificial intelligence-enabled analysis of UK and US public attitudes on Facebook and Twitter towards COVID-19 vaccinations. J Med Internet Res. 2021;23(4):e26627
17. Deiner MS, Fathy C, Kim J, et al. Facebook and Twitter vaccine sentiment in response to measles outbreaks. Health Informatics Journal. September 2019:1116-1132. doi:10.1177/1460458217740723
18. Benis A, Seidmann A, Ashkenazi S. Reasons for taking the COVID-19 vaccine by US social media users. Vaccines. 2021;9(4):315; https://doi.org/10.3390/vaccines9040315
19. Puri N, Coomes EA, Haghbayan H, Gunaratne K. Social media and vaccine hesitancy: new updates for the era of COVID-19 and globalized infectious diseases. Hum Vaccin Immunother. 2020;16(11):2586-2593.
20. Malik, AA., McFadden, SM., Elharake, J., Omer, SB. Determinants of COVID-19 vaccine acceptance in the US, EClinicalMedicine, 2020; 26 (100495):100495.
21. Loomba S, de Figueiredo A, Piatek SJ, de Graaf K, Larson HJ. Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA. Nat Hum Behav. 2021;5(3):337-348.
22. Banda JM, Tekumalla R, Wang G, Yu J, Liu T, Ding Y, et al. A large-scale COVID-19 Twitter chatter dataset for open scientific research - An international collaboration (Version 32) [Data set]. Zenodo. 2020 Oct 18. URL: http://doi.org/10.5281/zenodo.4104587 [accessed 2021-01-31]
23. The Panacea Lab: COVID-19 Twitter. GitHub. 2020. [2021-02-02]. https://github.com/thepanacealab/covid19_twitter.
24. https://github.com/mkearney/tweetbotornot
25. Blum A, Hopcroft J, Kannan R. Foundations of Data Science. Cambridge, UK: Cambridge University Press; 2020.
26. Blum A, Hopcroft J, Kannan R. Foundations of Data Science. Cambridge, UK: Cambridge University Press; 2020.
27. Blei DM, Ng A, Jordan MI. Latent Dirichlet allocation. J Mach Learn Res 2003 Mar;3:993-1022.
28. Lyu J, Luli G. Understanding the Public Discussion About the Centers for Disease Control and Prevention During the COVID-19 Pandemic Using Twitter Data: Text Mining Analysis Study. J Med Internet Res 2021;23(2):e25108. DOI: 10.2196/25108
29. Plutchik R. A General Psychoevolutionary Theory of Emotion In: Robert P, Henry K, editors. Theories of Emotion. USA: Academic Press; 1980. pp. 3–33.
30. Misuraca, Michelangelo, Alessia Forciniti, Germana Scepi, and Maria Spano. "Sentiment Analysis for Education with R: packages, methods and practical applications." arXiv preprint arXiv:2005.12840 (2020).
31. Turney, P.D., & Mohammad, S. (2010). Emotions Evoked by Common Words and Phrases: Using Mechanical Turk to Create an Emotion Lexicon. In Proceedings of the NAACL-HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text (pp. 26–34), Los Angeles, CA: ACL. Retrieved from https://www.aclweb.org/anthology/W10-0204.
32. Naldi, M. (2019). A review of sentiment computation methods with R packages. Available at: https://arxiv.org/abs/1901.08319

33. Malik, AA., McFadden, SM., Elharake, J., Omer, SB. Determinants of COVID-19 vaccine acceptance in the US, EClinicalMedicine, 2020; 26 e100495https://doi.org/10.1016/j.eclinm.2020.100495
34. Alley SJ, Stanton R, Browne M, To QG, Khalesi S, Williams SL, Thwaite TL, Fenning AS, Vandelanotte C. As the Pandemic Progresses, How Does Willingness to Vaccinate against COVID-19 Evolve? International Journal of Environmental Research and Public Health. 2021; 18(2):797. https://doi.org/10.3390/ijerph18020797
35. Gore RJ, Diallo S, Padilla J. You Are What You Tweet: Connecting the Geographic Variation in America's Obesity Rate to Twitter Content. PLoS One. 2015 Sep 2;10(9):e0133505. doi: 10.1371/journal.pone.0133505
36. Wojcik, S, Hughes, A. Sizing Up Twitter Users. Pew Research Center. 2019. Available at: https://www.pewresearch.org/internet/2019/04/24/sizing-up-twitter-users/

Appendix. Coherence scores of different numbers of topics