Public Perception of COVID-19 Vaccines through Analysis of Twitter Content and Users

Authors:

Sameh N. Saleh, MD*,1,2, Samuel A. McDonald, MD2,3, Mujeeb A. Basit, MD, MMSc1,2, Sanat Kumar2,4, Reuben J. Arasaratnam, MD, MPH1, Trish M. Perl, MD1, Christoph U. Lehmann, MD2,5, Richard J. Medford, MD1,2

1 Department of Internal Medicine, University of Texas Southwestern Medical Center, 5323 Harry Hines Boulevard, Dallas, TX 75390
2 Clinical Informatics Center, University of Texas Southwestern Medical Center, 5323 Harry Hines Boulevard, Dallas, TX 75390
3 Department of Emergency Medicine, University of Texas Southwestern Medical Center, 5323 Harry Hines Boulevard, Dallas, TX 75390
4 Lebanon Trail High School, 5151 Ohio Dr, Frisco, TX 75035
5 Departments of Pediatrics, Bioinformatics, Population & Data Sciences, University of Texas Southwestern Medical Center, 5323 Harry Hines Boulevard, Dallas, TX 75390

Corresponding Author:
Sameh N. Saleh, MD
5323 Harry Hines Blvd, Dallas, TX 75219
M: 571-338-3680
F: 214-648-9478
E: sameh.n.saleh@gmail.com

Keywords:
COVID-19, Vaccine, Twitter, Social Media, Public Opinion, COVID-19 Vaccines, SARS-CoV-2, Vaccination, Vaccination Refusal, Vaccine Hesitancy, Natural Language Processing, Sentiment Analysis, Topic Modeling, Demographic Inference

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.
Abstract

Twitter is a robust medium to understand wide-scale, organic public perception about the COVID-19 vaccine. In this cross-sectional observational study, we evaluated 2.4 million English tweets from nearly 1 million user accounts matching keywords (‘covid*’ OR ‘coronavirus’) AND ‘vaccine’) during vaccine development from Feb 1st through Dec 11th, 2020. We applied topic modeling, sentiment and emotion analysis, and demographic inference of users on the COVID-19 vaccine related tweets to provide insight into the evolution of public attitudes. Individuals generated 87.9% (n=834,224) of tweets. Of individuals, men (n=560,824) outnumbered women (n=273,400) by 2:1 and 39.5% (n=329,776) of individuals were ≥40 years old. Daily mean sentiment fluctuated congruent with news events, but overall trended positively. Trust, anticipation, and fear were the three most predominant emotions; while fear was the most predominant emotion early in the study period, trust outpaced fear from April 2020 onward. Fear was more prevalent in tweets by individuals (26.3% vs. organizations 19.4%; p<0.001), specifically among women (28.4% vs. males 25.4%; p <0.001). Multiple topics had a monthly trend towards more positive sentiment. Tweets comparing COVID-19 to the influenza vaccine had strongly negative early sentiment but improved over time. Our findings are concerning for COVID-19 vaccine hesitancy, but also identify targets for educational interventions.
Introduction

With the global continuation of the COVID-19 pandemic, the large-scale administration of a SARS-CoV-2 vaccine (referred from here on as the COVID-19 vaccine) is crucial to achieve herd immunity and curtail further spread of the virus\(^1\). As governments work to approve and distribute safe and effective vaccines,\(^2\) important questions regarding vaccination willingness persist: What are the attitudes and perceptions of the public\(^3\) to these vaccines and how can they affect vaccine uptake\(^4\)? These questions are important to develop an education and outreach approach to achieve the desired vaccine penetration to achieve herd immunity\(^5\). In 2019, prior to the COVID-19 pandemic, the World Health Organization (WHO) had identified vaccine hesitancy as one of the top 10 greatest global health threats\(^6\). While surveys on attitudes and perception of a COVID-19 vaccine show significant vaccine hesitancy among the general population\(^7-9\) and health care providers\(^10,11\), studies remain small in size, tend to focus on local participants, are prone to sampling error from non-probability sampling and reporting bias, and perhaps most poignantly, cannot capture real-time changes in vaccine willingness. Crowdfunding platforms may provide an indication of emerging community needs related to COVID-19 but fail to provide a continuous assessment of community sentiment\(^12\).

Twitter, the microblogging platform, with over 187 million daily monetizable active users,\(^13\) serves as a robust medium to better understand wide-scale, organic public perception about the COVID-19 vaccine. With nearly 400 million mentions, \#COVID19 was the most used hashtag on Twitter in 2020\(^14\). Social media has become increasingly recognized for its rapid information dissemination (whether accurate or not) and dispersion of sentiment that quickly crosses geographic and social boundaries\(^15\). Analysis of social media text can inform real-time changes...
and evolution in population-level attitudes\textsuperscript{16}. As evident with the rise of the “infodemic” during the COVID-19 pandemic, Twitter has become a particularly useful data source in public health and healthcare-related research\textsuperscript{17} and has been repeatedly used to study public sentiment and understand trends throughout the COVID-19 pandemic\textsuperscript{18–21}. Earlier in the COVID-19 pandemic, we were able to demonstrate initial public sentiment regarding the virus, its origin and spread, and measures to limit its spread\textsuperscript{22} as well as early support for social distancing\textsuperscript{23} on Twitter.

Social media, and specifically Twitter, has been shown to be a major factor in vaccine uptake and should be monitored and potentially used for interventions to address vaccine hesitancy\textsuperscript{24}. Examining sentiments towards the influenza A H1N1 vaccine in 2009 showed that projected vaccination rates based on Twitter sentiment were similar to vaccination rates estimated by traditional phone surveys used by the Centers for Disease Control and Prevention (CDC)\textsuperscript{25}. A previous study noted information exposure on Twitter may account for differences in human papillomavirus (HPV) vaccine uptake that are not accounted for by socioeconomic factors like education, insurance, or poverty\textsuperscript{26}. Another study noted that there is a significant relationship between social media use by the public and organized action and public doubts of vaccine safety\textsuperscript{27}.

We aimed to apply content and sentiment analysis on COVID-19 vaccine related tweets as well as analysis of the responsible, originating user accounts to provide insight into the evolution of public attitudes about the COVID-19 vaccines over time. We hypothesized that content analysis from the start of the pandemic will identify important themes of discussion (especially those with negative sentiment or evidence of misinformation) throughout the vaccine development process.
that would inform health care officials, public health agencies, and policy makers and could be used to aid in the outreach and educational interventions for the COVID-19 vaccine to the general public.

Methods

Data Source

We performed a cross-sectional observational study of English-language tweets obtained by matching the keywords ((‘covid*’ OR ‘coronavirus’) AND ‘vaccine’) from February 1, 2020 to December 11th, 2020. December 11th was chosen as an end date to mark the United States Food and Drug Administration’s first emergency use authorization of a COVID-19 vaccine. We used the snscrape library to obtain (“scrape”) tweets identified through Twitter's advanced search tool, which returns a relevant sample of tweets. We manually reviewed a random subsample of 1,000 tweets and verified the tweets’ relevance to the topic of COVID-19 vaccination. We extracted 21 and 20 variables related to the tweets and to the posting user accounts, respectively (Supplemental Table S1 and S2).

Data Processing

We measured total daily tweets and completed descriptive statistics for collected variables. We applied natural language processing techniques to process, analyze, and visualize the text from tweets. To preprocess the tweet text for analysis (“cleaning”), we removed hyperlinks, user tags, and words of little analytical value. We also returned words to their root form and segmented text into one- and two-word terms. Further details are discussed in Supplemental Appendix A. We visualized the top 300 processed terms as a word cloud with larger font size representing greater term frequency. All analyses were conducted using Python, version 3.8.2 (Python
Software foundation). Institutional review board approval was not required because this study used only publicly available data.

Sentiment Analysis

Sentiment analysis describes the affect of a piece of text — the intrinsic attractiveness or aversiveness of a subject such as events, objects, or situations\textsuperscript{30}. We used the Valence Aware Dictionary and sEntiment Reasoner (VADER)\textsuperscript{31} to analyze the sentiment polarity of a tweet. VADER, a lexicon and rule-based tool, was particularly designed for sentiment analysis for social media text. In addition to regular words, VADER leverages punctuation, emoticons, emojis, sentiment-laden slang words and acronyms, as well as syntax and capitalization schemas to inform labeling of a positive, neutral, and negative score for each document. These three scores were combined to form a normalized, weighted composite score. Overall positive ($\geq 0.05$), neutral (-0.05 to 0.05), and negative ($\leq -0.05$) sentiments are defined at standardized composite score thresholds. When sentiment has been aggregated, we refer to an average sentiment of $\geq 0.05$ as positive and $\leq -0.05$ as negative. Trends in sentiment over time were determined using the Mann-Kendall trend test. We used the \texttt{TextBlob} library\textsuperscript{32} to label each tweet from a range of 0 (objective) to 1 (subjective) where objective tweets relay factual information and subjective tweets typically communicate an opinion or belief. Finally, we used the \texttt{NRCLex} library to label words within each tweet with corresponding emotional affects (i.e., Plutchik’s wheel of emotions which include anger, anticipation, fear, disgust, joy, sadness, surprise, and trust) based on the National Research Council Canada (NRC) affect lexicon\textsuperscript{33}. Based on these labels, we identified tweets with their primary emotion and visualized how the proportion of eligible tweets (i.e., those with an identified primary emotion) with a particular primary emotion changed over time.
Topic Modeling

After cleaning the tweets to distill analyzable text as described in the methods, we applied a machine learning algorithm called Correlation Explanation (CorEx)\textsuperscript{34} to identify clusters of topics for all tweets. CorEx identifies the most informative topics based on a set of latent factors that best explain the correlations in the data in turn maximizing the total correlation or the multivariate mutual information\textsuperscript{35}. Each document (in our case, each tweet) may include multiple topics. We iterated through a range from 2 to 20 topics and trained a separate model for each number of topics with the goal of identifying the model with the maximum total correlation, ultimately choosing 15 topics for the topic model. We presented the top 20 words for each topic cluster to author CUL without prior access to individual tweets from the dataset to manually label a theme for each topic. The manually labeled topic labels were reviewed by two other authors SNS and RJM with unanimous agreement. We visualized the monthly distribution of topics over time and utilized a heat map to visualize how the mean sentiment of each topic has changed per month.

User Exploration and Demographic Inference

Given that each tweet has one authoring account, we identified all unique user accounts in our dataset and provided descriptive statistics with metadata available for the users, including the launch date of the account, followers (accounts following them), follows (accounts they follow), lifetime posts, likes, and media shared, as well as profile pictures, description information, and verified status (badge to indicate an account of public interest that has been verified to be authentic). To better understand demographic differences, we applied a previously validated deep learning system through the \textit{m3inference} library\textsuperscript{36} to infer the account user as an individual or an organization and if labeled as an individual, their gender (female or male) and age group.
(≤18 years old, 19-29 years old, 30-39 years old, and ≥40 years old) based on multimodal input that includes username, display name, description, and profile picture image. Each label using the algorithm has an accompanying probability. The automatic demographic detection was particularly designed for Twitter profiles for health-related cohort studies. We provided summary statistics for the demographics identified and stratified sentiment and subjectivity analyses by the different demographic groups to evaluate for differences. We used Mann-Whitney U and $\chi^2$ where appropriate to determine significance. Alpha level of significance was set a priori at 0.05 and all hypothesis testing was two-sided. We did not adjust for multiple comparisons as this was an exploratory study and should be interpreted as hypothesis-generating.

Results

A total of 2,356,285 tweets were extracted for the study period, of which 2,287,344 tweets were English-only and included for evaluation. The tweets were generated by 948,666 accounts which had been active for an average of 6.9 years (interquartile range [IQR], 2.6 - 10.0) with a median of 267 (IQR, 55 - 1,100) followers and 3,600 (519 - 15,572) lifetime likes. Only 2.9% (n=27,443) of accounts were verified (Table 1). Of the tweets analyzed, 54% (n=1,235,575) had a link, 40.1% (n=916,585) mentioned other twitter accounts, 18.1% (n=414,173) used hashtags, and 11.9% (n=273,278) contained media like an image or video. In terms of engagement, 41.3% (n=943,639), 24.0% (n=548,863), and 20.7% (n=473,204) of tweets received likes, replies, and retweets, respectively (Table 1). Individuals (vs. organizations) generated 87.9% (n=834,224) of tweets. Of individuals, men (n=560,824) outnumbered women (n=273,400) by 2:1 and 39.5% (n = 329,776) of individuals were ≥40 years old (Table 1).
Daily tweets abruptly spiked to 51,176 tweets on November 9th, the day Pfizer and BioNTech announced their vaccine’s effectiveness (up from 4,052 tweets on November 8th) and peaked on December 8th with 55,779 tweets. Tweets from November 1st to the end of the study period on December 11th accounted for 39.8% (n = 910,593) of all tweets (Figure S1). The corpus of tweets contained over 62.5 million words and 416 million characters. The ten most commonly tweeted terms and their frequencies were as follows: “people” (228,482), “trial” (206,310), “take” (181,598), “flu” (159,043), “trump” (149,042), “first” (147,103), “make” (142,242), “test” (131,719), “need” (126,846), and “one” (122,966). Figure 1 displays a word cloud of the top 300 words with larger font size concordant with frequency.

Daily mean sentiment of tweets fluctuated congruent with news events, but overall trended positively throughout the study period (Mann-Kendall statistic=10,122; tau=0.218; p<0.001) (Figure 2a). Several days in early to mid-March and on October 13th saw particularly negative sentiments, coinciding with news of the declaration of a pandemic by the WHO and Johnson & Johnson’s halting of their vaccine trial on October 12th, respectively. Highest daily mean positive sentiment revolved around Moderna’s July 14th announcement of a safe vaccine with “robust immune response” in an early trial and Pfizer’s November 9th announcement of over 90% effectiveness of its vaccine. Twitter accounts representing organizations had more positive sentiments than tweets from individuals (median weekly difference, 0.118; IQR, 0.091 to 0.144), but there was no significant difference in polarity for age (median weekly difference, 0.006; IQR, -0.011 to 0.019) and only minimal positive difference for males (median weekly difference, 0.030; IQR, 0.012 to 0.044) (Figure 2b-d).
The sentiment trends were reflected by the primary emotions identified in the COVID-19 vaccine tweets by month (Figure 3a). Fear started as the most prevalent primary emotion in nearly 40% of eligible tweets early on but decreased to under 20% by the end of the study period. Conversely, trust increased from below 20% to around 40% and outpaced fear in April 2020, maintaining as the most prevalent primary emotion thereafter. Anticipation was the second most prevalent primary emotion for most of the study period, steadily ranging from 25% to 30%. All other emotions were consistently expressed as the predominant emotion in less than 10% of eligible tweets. Individuals had an increased predominance of fear (26.3% vs. 19.4%; p<0.001) and decreased predominance of anticipation (25.9% vs. 33.6%; p<0.001) and trust (32.5% vs. 35.2%; p<0.001). For individual accounts, women had more fear (28.4% vs. males 25.4%; p<0.001) with less anticipation (23.8% vs. 26.8%; p<0.001) than men, but no significant difference in trust (32.3% vs. 32.5%, p=0.11). Those less than 40 years old had more fear (26.6% vs. 26.0%; p<0.001) and less trust (32.0% vs. 33.0%; p<0.001) (Figure 3b-d). Tweets throughout the year tended to be more objective (where 0 is fully objective and 1 as fully subjective) with limited daily variation (overall mean 0.359; std 0.028) (Figure S2).

Table 2 shows each topic label with their key words and sample tweets. Figure S3 shows the 15 topics obtained from topic modeling with the proportion of tweets per month that contained each topic. The dominant topic (topic 15) focused on mask use and public reactions. Discussions about misinformation and conspiracy theories comprised the next most common topic, peaking in May and staying relatively consistent from July through December. Tweets related to the Indian and Russian governments’ decision on producing and using the Sputnik V vaccine (topic 2) spiked in August. Discussion of Emergency Use Authorizations (EUA) and vaccine approvals
(topic 12) did not spike until November 2020 with the approval of the Pfizer and Moderna vaccines. Several topics had strong mean positive sentiments throughout the study period, including discussions of biotechnology companies and the stock market (topic 3), vaccination firsts (topic 4), vaccine development (topic 6), and EUAs (topic 12). Other topics showed a progressive trend from positive to negative throughout the study period including discussion of US politics and the election (topic 1), the FDA and CDC (topic 14), and mask use and public reactions (topic 15). Tweets comparing COVID-19 to influenza (topic 5) and its vaccine had strongly negative early sentiment but improved over time (Figure 4). Compared to the rest of individual users (n=810,318), those exhibiting negative sentiment posting about topic 5 (n=51,686) were proportionally more likely to be ≥40 years old (45.1% vs. 39.6%; p <0.001) and female (34.0% vs. 32.7%; p <0.001). The only other topic with persistently negative sentiment was discussion of misinformation and conspiracy theories (topic 13). Those exhibiting negative sentiment posting about topic 13 (n=166,819) were more likely than other user accounts (n=741,388) to be individuals (90.9% vs. 87.3%; p <0.001) and of those individual accounts, more likely to be female (34.4% vs. 32.4%; p<0.001).

Discussion

Twitter is a rich medium that can serve as both thermometer and thermostat for the COVID-19 vaccine, which is a crucial public health strategy to combat the pandemic. It can provide insight into public perception of a COVID-19 vaccine, but can also be used to understand and combat knowledge deficits and vaccine hesitancy through information and education\textsuperscript{41,42}. A majority (59%) of US Twitter users regularly obtain news on Twitter, proportionally more than any other social media platform\textsuperscript{43}. We analyzed nearly 2.4 million COVID-19 vaccine related tweets in 2020, creating a dataset that exceeded the scope of related studies\textsuperscript{44,45}. Generally, we believe that
Twitter users favored the vaccine during its development phase. Tweets with positive sentiment were more prominent than tweets with negative sentiment and trust emerged as the predominant emotion. However, there were periods of time (usually linked to events in the public news cycle), demographic subgroups, and topic clusters that had more prominent negative sentiment and emotion.

Organizational accounts were significantly more positive, exhibiting more anticipation and trust and less fear. For individuals, the gender and age distribution in our dataset parallels the reported proportional share of Twitter’s global advertising audience\textsuperscript{46}. Women expressed more fear and less anticipation, but by the end of the study period, that gap had narrowed. Those less than 40 years old tended to express less trust and more fear, but the margin was small.

The topic most strongly associated with negative, albeit improving, sentiment was the discussion of the influenza vaccine in combination with the COVID-19 vaccine. These tweets often compared deaths and illness from both diseases or expressed general vaccine mistrust to both vaccines. Examples include: “@\texttt{user} Only time I’ve ever had the flu is the 2 times I got flu shots. It was not a minor case either it was the full blown flu. I refuse to get another flu shot and I also will refuse the covid vaccine” and “Flu Virus equals Flu Vaccine. Coronavirus Equals Covid-19 Vaccine...Now if the Flu shot gives you the flu, the Covid-19 Shot will give you Coronavirus....am I in the general area of Right??”. Notably, these users exhibiting negative sentiment about this topic were more likely to be $\geq 40$ years old and female. This focused topic-demographic cluster, for example, exposes a direct opportunity for intervention to correct
misinformation and mitigate vaccine hesitancy. Conversely, the emergency use authorizations of
the vaccine and reports of the first vaccine recipients, which arose later in the study period, were
celebrated with positive sentiment and mirror the overall increasing trend in positive sentiment
and trust.

While the percent population immunity needed to achieve herd immunity (either through innate
or acquired immunity) for COVID-19 is not yet known, estimates have increased from 60-70%
to possibly closer to 75-85%\(^47,48\). Achieving herd immunity through infection would come at an
untenable cost\(^49\), making the immunization effort critical to protect lives. Therefore, it was
concerning to us that fear was a common and persistent predominant emotion in COVID-19
vaccine tweets. While the proportion of ‘trust’ tweets outpaced ‘fear’ tweets relatively early in
the study period, approximately 20% of eligible tweets still expressed fear in association with the
vaccine. If this fear translates into refusal to become immunized, we are not only likely to see a
prolonged pandemic, but also further increases in COVID-19 related deaths as concerning virus
variants take hold. As more people receive the vaccine in the future, we anticipate that
sentiments will become more positive over time with increased trust and vaccine uptake, but this
will need to be consistently studied especially in the context of newly approved vaccines and
news events.

Limitations
Our study was limited by several factors. First, we recognize that our dataset is not all inclusive
of tweets discussing the COVID-19 vaccine. Our tweet search criterion was narrow to ensure
accuracy of captured tweets for this initial work and did not include terms such as “shot(s)
“immunization” and “inoculation.” Moreover, despite the volume of tweets analyzed, we are
limited to only a relevant sample of all tweets per Twitter’s advanced search tool. Second, we
used existing tools to analyze sentiments and emotion of tweets that are not specific to health
care topics, which could have skewed our analysis. Third, tweets related to COVID-19
vaccination could have been flagged or removed by Twitter for containing misinformation, but
we were not privy to that context to determine how that could have affected our sample. Finally,
since we targeted only tweets in English and are unable to determine geographic location for
users, we are limited in making conclusions about specific countries or countries where English
is the not the predominant language.

Conclusions
Leveraging 2.4 million COVID-19 vaccine related tweets in 2020, we were able to successfully
explore sentiment, emotion, topics, and user demographics to elucidate important trends in
public perception about the COVID-19 vaccine. Tweets were overall positive in sentiment and
with growing trust. However, fear maintained as a dominant emotion raising concern regarding
the willingness to receive the COVID-19 vaccine and subsets of negative sentiment emerged.
Comparison to influenza and the influenza vaccine as well as discussion about conspiracy
theories were important topics with negative sentiment and showed some demographic
differences that could allow for informed intervention. Future work will leverage these natural
language processing tools to engage in targeted messaging based on user interests and emotions.
Declarations

**Ethics approval and consent to participate:** The University of Texas Southwestern Human Research Protection Program Policies, Procedures, and Guidance did not require institutional review board approval as all data were publicly available.

**Data Availability, Materials, & Correspondence:** The data that support the findings of this study are available upon request. Please request from the corresponding author.

**Code Availability:** The code that support the findings of this study is available upon request. Please request from the corresponding author.

**Competing interests:** Dr. Lehmann reports stock ownership in Celanese Corporation and Colfax Corporation. There are no other competing interests.

**Funding:** None.

**Authors’ contributions:** Study concept and design: SNS, CUL, RJM; Data acquisition: SNS, SK; Analysis: SNS, SK, RJM; Interpretation of data: SNS, SK, CUL, RJM; Manuscript preparation: all authors. All authors read and approved the final manuscript.

**Acknowledgements:** Not applicable
Figures and Tables

Figure 1. Word cloud of top 300 words related to COVID-19 and vaccine. Larger fonts represent higher frequency in the corpus after preprocessing text.
**Figure 2a-d.**

**a)** Mean sentiment polarity shown by day (as points) and by week (as a dashed line). Each tweet was labeled as primarily negative (-1), neutral (0), or positive (1).

**b)** Mean weekly polarity stratified by individual versus organization.

**c)** Mean weekly polarity stratified by gender for individual accounts.

**d)** Mean weekly polarity stratified by age more or less than 40 years than for individual accounts.
Figure 3a-d. Percent of tweets with primary emotion per month a) overall, b) stratified by individual versus organization, c) stratified by gender for individual accounts, and d) stratified by age more or less than 40 years than for individual accounts. Only tweets with a predominant primary emotion (n = 1,489,027) are included.
Figure 4. Heat map showing mean sentiment by month for each topic. Note that a tweet can include multiple topics.
## Tables

**Table 1.** Tweet and user account characteristics are shown on top and inferred user demographics are shown on bottom.

<table>
<thead>
<tr>
<th>Tweets</th>
<th>User Accounts</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>n = 2,287,344</strong></td>
<td>N = 948,666</td>
</tr>
<tr>
<td>Has link(s)</td>
<td>Years account active</td>
</tr>
<tr>
<td>1,235,575 (54.0)</td>
<td>6.9 (2.6 – 10.0)</td>
</tr>
<tr>
<td>Mentions user(s)</td>
<td>Followers</td>
</tr>
<tr>
<td>916,585 (40.1)</td>
<td>267 (55 – 1,100)</td>
</tr>
<tr>
<td>Has hashtag(s)</td>
<td>Following</td>
</tr>
<tr>
<td>414,173 (18.1)</td>
<td>407 (137 – 1,069)</td>
</tr>
<tr>
<td>Has media</td>
<td>Lifetime statuses</td>
</tr>
<tr>
<td>273,278 (11.9)</td>
<td>4,605 (1,027 – 16,365)</td>
</tr>
<tr>
<td>Is quoted tweet</td>
<td>Lifetime likes</td>
</tr>
<tr>
<td>133,404 (5.8)</td>
<td>3,600 (519 – 15,572)</td>
</tr>
<tr>
<td>Has like</td>
<td>Media shared</td>
</tr>
<tr>
<td>943,639 (41.3)</td>
<td>205 (36 – 875)</td>
</tr>
<tr>
<td>Has reply</td>
<td>Public lists, member</td>
</tr>
<tr>
<td>548,863 (24.0)</td>
<td>2 (0 – 12)</td>
</tr>
<tr>
<td>Has retweet</td>
<td>Contains description</td>
</tr>
<tr>
<td>473,204 (20.7)</td>
<td>800,619 (84.4)</td>
</tr>
<tr>
<td>Has quoted tweet</td>
<td>Location listed</td>
</tr>
<tr>
<td>183,982 (8.0)</td>
<td>659,720 (69.5)</td>
</tr>
<tr>
<td><em>Twitter source</em></td>
<td>Contains profile picture</td>
</tr>
<tr>
<td>Web App</td>
<td>902,666 (95.2)</td>
</tr>
<tr>
<td>686,296 (30.0)</td>
<td>Contains banner picture</td>
</tr>
<tr>
<td>iPhone/iPad</td>
<td>721,542 (76.1)</td>
</tr>
<tr>
<td>660,382 (28.9)</td>
<td>Contains link in profile</td>
</tr>
<tr>
<td>Android</td>
<td>303,761 (32.0)</td>
</tr>
<tr>
<td>432,862 (18.9)</td>
<td>Verified account</td>
</tr>
<tr>
<td>TweetDeck</td>
<td>27,443 (2.9)</td>
</tr>
<tr>
<td>60,845 (2.7)</td>
<td></td>
</tr>
</tbody>
</table>

### User Demographics

<table>
<thead>
<tr>
<th>Entity</th>
<th>N (%)</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>N = 948,666</strong></td>
<td></td>
<td>median (IQR)</td>
</tr>
<tr>
<td><strong>Individual</strong></td>
<td>834,224 (87.9)</td>
<td>0.999 (0.997 – 0.999)</td>
</tr>
<tr>
<td>Organization</td>
<td>114,443 (12.1)</td>
<td>0.867 (0.727 – 0.999)</td>
</tr>
<tr>
<td><strong>Sex (of individuals)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>273,400 (32.8)</td>
<td>0.992 (0.949 – 0.998)</td>
</tr>
<tr>
<td>Male</td>
<td>560,824 (67.2)</td>
<td>0.996 (0.980 – 0.999)</td>
</tr>
<tr>
<td><strong>Age (of individuals)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;40 years old</td>
<td>504,448 (60.5)</td>
<td>0.972 (0.896 – 0.994)</td>
</tr>
</tbody>
</table>
Table 2. Topic clusters identified by topic modeling. Words contributing to the model are shown in decreasing order of weighting. The topics are labeled manually based on these words.

<table>
<thead>
<tr>
<th>Possible Topic Label</th>
<th>Tweets/Topic</th>
<th>Words contributing to topic model (in ↓ order of weighting)</th>
<th>Representative Tweet</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mask use and general reactions</td>
<td>15, 811,844</td>
<td>people, mask, even, dont, would, take, know, die, death, need, one, still, many, kill, risk, never, work, way, yet, wear_mask</td>
<td>““Pretty much what it boils down to, at this point. Ignorance, arrogance, and stupidity will end up killing LOTS of people this year. I'm afraid! Be SMART. WEAR your mask. Wash your hands. Hold off on large gatherings until a safe, effective Covid-19 vaccine arrives. [link]”</td>
</tr>
<tr>
<td>Conspiracies and misinformation</td>
<td>13, 557,301</td>
<td>want, think, fake, make, try, believe, conspiracy, bill_gate, money, really, gonna, real, force, shit, god, anything, anyone, hoax, black, put</td>
<td>“@WhiteHouse Also, isnt this a RNA vaccine? Super experimental albeit dangerous, could mean with DNA as well. Human Guinea pigs. Wouldn't be surprised if the vaccine harms more then the COVID did.”</td>
</tr>
<tr>
<td>Impact on lockdowns on school, work, and economy</td>
<td>7, 406,227</td>
<td>wait, year, lockdown, open, month, next, life, time, long, end, last, come, back, next year, week, school, ago, day, economy, away</td>
<td>“@[tag] @[tag] @[tag] @[tag] And even with a vaccine they will continue with the lockdowns, the social distance and the fear mongering... If not for the Covid, they will find something...”</td>
</tr>
<tr>
<td>Vaccine mechanisms and immunity</td>
<td>10, 235,625</td>
<td>virus, mrna, immunity, antibody, prevent, infection, disease, spread, strain, protein, herd_immunity, mutate, symptom, immune_system, sarscov, prevent infection, mutation, cell, cause, infect</td>
<td>“If tests show one already had COVID-19 so one has antibodies and is now immune, CDC currently counts that as one infected and positive for COVID-19. After a vaccine, will every person vaccinated who therefore grows antibodies, be considered positive &amp; infected? @realDonaldTrump”</td>
</tr>
<tr>
<td>First vaccinations and recipients</td>
<td>4, 232,613</td>
<td>first, world, around, world first, become, first test, receive, country, first line, first person, world news, first country, person receive, world leader, government vote, make</td>
<td>“Thank the lord this is the beginning of the end: First patient receives Pfizer Covid-19 vaccine [link]”</td>
</tr>
<tr>
<td>Topic</td>
<td>Count</td>
<td>Total</td>
<td></td>
</tr>
<tr>
<td>-----------------------------------------</td>
<td>-------</td>
<td>--------</td>
<td></td>
</tr>
<tr>
<td>Emergency use authorizations and approvals</td>
<td>12</td>
<td>213,036</td>
<td></td>
</tr>
<tr>
<td>US politics and election</td>
<td>1</td>
<td>212,597</td>
<td></td>
</tr>
<tr>
<td>Stock market and pharma/biotech companies</td>
<td>3</td>
<td>205,557</td>
<td></td>
</tr>
<tr>
<td>Clinical trials and participants</td>
<td>8</td>
<td>205,468</td>
<td></td>
</tr>
<tr>
<td>Vaccine development and supporters</td>
<td>6</td>
<td>204,083</td>
<td></td>
</tr>
<tr>
<td>Russian response and global partners</td>
<td>2</td>
<td>187,369</td>
<td></td>
</tr>
<tr>
<td>FDA and CDC</td>
<td>14</td>
<td>179,396</td>
<td></td>
</tr>
</tbody>
</table>

**Emergency use authorizations and approvals**

- Pfizer, Moderna, approval, effective, Pfizer_Biontech, emergency use, data, biotech, authorization, receive, regulator, next week, approve Pfizer, effective prevent, Pfizer CEO, show effective, Moderna, effective, approve, data show, early data

- "Pfizer's Covid vaccine is days away from approval after data reveals it is 95% effective [link]"

**US politics and election**

- Trump, Biden, Realdonaldtrump, president, election, operation_warp, American, credit, speed, lie, take credit, Joe Biden, GOP, Democrat, Potus, WIN, vote, Joe Biden, Admin, America

- "@realDonaldTrump If you want to take partial credit for the Covid-19 vaccine fine. You still LOST the election. In Georgia for example you are behind there by 12k votes. The recount won't change the outcome. I look forward to your predictable reply and the end of your regime."

**Stock market and pharma/biotech companies**

- Market, stock, news, company, good news, biotech, drug company, pharma, price, billion, drug, surge, late, update, rise, break news, positive news, investor, pharma company, announce

- "Markets are supported by both the cumulative upside surprises to the economy since the end of the recession and the apparently faster-than-expected progress toward a COVID-19 vaccine. [link]"

**Clinical trials and participants**

- Trial, clinical, human trial, phase, human, volunteer, participant, Oxford, begin, phase clinical, show, trial participant, result, volunteer trial, ahead large trial, ahead large, show promise, immune response, test, number

- "Coronavirus Vaccine Update | Oxford’s COVID-19 vaccine trial in Brazil begins: Scientists say coronavirus jab may not work for older adults [link]"

**Vaccine development and supporters**

- Research, development, global, effort, develop, fund, researcher, global effort, join, effort develop, access, help, target, hacker, treatment, support, Dolly Parton, research development, accelerate, collaboration

- "As the world continues to feel the impact of COVID-19, the biopharmaceutical industry is working around the clock to identify and develop safe and effective vaccines to prevent infection, while also researching and developing new therapies to treat those infected with the virus."

**Russian response and global partners**

- Russia, India, VIA, Sputnik, Russian, China, Putin, Serum Institute, Indian, Covaxin, Chinese, Bharat Biotech, ICMR, Hacker Target, TIME India, via NBC News, Russia Sputnik, Indian Serum, Narendra Modi, possible

- "A Sputnik moment, president #Putin has announced that #Russia is the first country in the world to register a #Covid_19 vaccine. 10s of countries already requested it [link]"

**FDA and CDC**

- Trump, FDA, failure, administration, Trump administration, FDA approval, food_drug, food_drug administration, FDA approve, CDC

- "@CDCgov if you try and push through an unproven vaccine because of Trump’s desperation to recover from his abysmal handling of Covid-
<table>
<thead>
<tr>
<th>Section</th>
<th>Value</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Philanthropy and public health</td>
<td>9</td>
<td>156,183</td>
</tr>
<tr>
<td>Comparison to influenza</td>
<td>5</td>
<td>151,177</td>
</tr>
<tr>
<td>Safety and side effects from trials</td>
<td>11</td>
<td>137,445</td>
</tr>
</tbody>
</table>

19...good luck. No one I know myself included will be getting vaccinated.”

“[@tag] This presents a problem and crashes into the argument, should covid vaccines be mandated. I initially thought that it will need more then encouragement and common sense from the public but these vaccine deniers are going to deprive people of protection through fear. Arrest them.”

“Ok so I’m usually not super crunchy about everything but I’ve been hospitalized 2x after getting the flu shot bc of how badly I got the flu within months so I was told not to get the shot by my drs. what does that mean for COVID’s vaccine? Like what if I react the same?”

'AstraZeneca COVID-19 vaccine study put on hold due to suspected adverse reaction in UK participant [link]'
References


29. *Snscrape*. (Github).


32. Loria, S. *Textblob: simplified text processing*. (Textblob).


doi:10.1101/2021.01.06.21249350.


