Original Paper

Factors Driving the Popularity and Virality of COVID-19 Vaccine Discourse on Twitter: Text Mining and Data Visualization Study

Jueman Zhang[1], PhD; Yi Wang[2], PhD; Molu Shi[3], PhD; Xiuli Wang[4], PhD

[1]Polk School of Communications, Long Island University, Brooklyn, NY, United States

[2]Department of Communication, University of Louisville, Louisville, KY, United States

[3]Louisville, KY, United States

[4]School of New Media, Peking University, Beijing, China

Corresponding Author:
Xiuli Wang, PhD
School of New Media
Peking University
5 Yiheyuan Road
Haidian District
Beijing, 100871
China
Phone: 86 10 6276 6689
Email: xiuli.wang@pku.edu.cn

Abstract

Background: COVID-19 vaccination is considered a critical prevention measure to help end the pandemic. Social media platforms such as Twitter have played an important role in the public discussion about COVID-19 vaccines.

Objective: The aim of this study was to investigate message-level drivers of the popularity and virality of tweets about COVID-19 vaccines using machine-based text-mining techniques. We further aimed to examine the topic communities of the most liked and most retweeted tweets using network analysis and visualization.

Methods: We collected US-based English-language public tweets about COVID-19 vaccines from January 1, 2020, to April 30, 2021 (N=501,531). Topic modeling and sentiment analysis were used to identify latent topics and valence, which together with autoextracted information about media presence, linguistic features, and account verification were used in regression models to predict likes and retweets. Among the 2500 most liked tweets and 2500 most retweeted tweets, network analysis and visualization were used to detect topic communities and present the relationship between the topics and the tweets.

Results: Topic modeling yielded 12 topics. The regression analyses showed that 8 topics positively predicted likes and 7 topics positively predicted retweets, among which the topic of vaccine development and people's views and that of vaccine efficacy and rollout had relatively larger effects. Network analysis and visualization revealed that the 2500 most liked and most retweeted retweets clustered around the topics of vaccine access, vaccine efficacy and rollout, vaccine development and people's views, and vaccination status. The overall valence of the tweets was positive. Positive valence increased likes, but valence did not affect retweets. Media (photo, video, gif) presence and account verification increased likes and retweets. Linguistic features had mixed effects on likes and retweets.

Conclusions: This study suggests the public interest in and demand for information about vaccine development and people's views, and about vaccine efficacy and rollout. These topics, along with the use of media and verified accounts, have enhanced the popularity and virality of tweets. These topics could be addressed in vaccine campaigns to help the diffusion of content on Twitter.

(JMIR Public Health Surveill 2021;7(12):e32814) doi: 10.2196/32814

KEYWORDS

COVID-19; vaccine; topic modeling; LDA; valence; share; viral; Twitter; social media

Introduction

Background

Since the World Health Organization (WHO) declared the COVID-19 outbreak a pandemic in March 2020 [1], the United States has seen the highest number of confirmed cases and deaths [2]. Many health organizations, including the WHO [3] and the US Centers for Disease Control and Prevention (CDC) [4], consider vaccination as a critical prevention measure to help end the pandemic and restore society to its normal status. Owing to remarkable advances in vaccinology, scientists developed COVID-19 vaccines within an unprecedented short time. In December 2021, less than 1 year after the virus was identified, the first two vaccines were approved for emergency use in the United States: the Pfizer-BioNTech vaccine and the Moderna vaccine [5]. Both of these vaccines use messenger RNA (mRNA)-based technology, which had not been approved previously for general use in humans [5]. Johnson & Johnson's Janssen vaccine, which is based on a slightly more mature technology of a viral vector, became the third vaccine approved for emergency use in the United States in February 2020 [6]. Owing to their novelty, COVID-19 vaccines had potential to fuel the existing vaccine debate, including arguments over vaccine safety and effectiveness, which had received notable attention in recent years before the pandemic [7]. In addition, political polarization, reaffirmed in the 2020 presidential election, was manifested in a wide range of issues, including responses to the COVID-19 pandemic [8] and vaccines [9]. Generally, Democrats had more favorable attitudes toward COVID-19 vaccines than Republicans [9]. These political fissures further had potential to propel the vaccine debate. Amidst the heated discussion of COVID-19 vaccines, the United States has been rolling out the most massive vaccination campaign in its history to fight against the pandemic [10].

Investigating public discourse about COVID-19 vaccines will shed light on people's perception and attitudes. As a major social media platform and a vital source for text-based public discourse, Twitter has been studied to understand public discourse about vaccines in general [11-14] and about specific vaccines, including COVID-19 vaccines [15,16]. Text-mining techniques have been increasingly used in recent research to investigate tweets about the COVID-19 pandemic (eg, [17-21]) and about COVID-19 vaccines [15,16]. These studies have employed machine learning algorithms to automatically analyze massive amounts of tweets and capture latent textual information such as topics, sentiment, and trends.

Although text mining is clearly an effective way to identify underlying textual clusters and patterns from vast amounts of tweets, less is known about how such information can help to understand the diffusion of information and opinions on Twitter. The aim of this study was to investigate message-level drivers of the popularity and virality of tweets about COVID-19 vaccines using text-mining techniques. Specifically, the objective of the study was to investigate how text-mined topics and valence, together with social media message features affect likes and retweets. Another aim of the study was to examine the topic communities of the most liked and most retweeted tweets using network analysis and visualization. These findings have implications for the direction of vaccine campaigns.

Literature Review

The extent to which a message results in optimal diffusion on social media can be assessed by users' favorable responses such as clicking "like" and "share" buttons to overtly indicate individual interest and support [22,23]. On Twitter, users can click on the "Like" icon to show appreciation for a tweet or on the "Retweet" icon to share it publicly with their followers [24]. Prior research has considered the like count of a tweet as an indicator of its popularity and the retweet count of a tweet as an indicator of its virality [23,25]. Drawing on these studies [23,25], we assessed the popularity of a tweet by the number of likes and assessed the virality of a tweet by the number of retweets. Compared with liking, retweeting is a more social behavior [26]. For both responses, the bandwagon effect postulates that the adoption of trends increases more with respect to the number of people who have already done so [22].

This study investigated three categories of message-level factors that, according to prior research, can drive the diffusion of media content online: information, emotion, and social media message features. As Twitter is a major source of text-based information, we drew on the literature related to the social transmission of online text information, including news articles and tweets. Past research on the virality of online news has suggested two categories of determinants: informational and emotional. From the informational perspective, information utility, as gauged by overall content usefulness, was found to prompt social media sharing of general news articles [27]. In the health context, a content attribute that taps into information utility is the presence of efficacy information [26], which provides ways to promote health or overcome a health risk [28]. Research has shown that overall content usefulness and presence of efficacy information both facilitate viewing and sharing of health news articles on social media [26]. In the situation of the COVID-19 pandemic, gaps in knowledge about the new coronavirus was evident in the United States early on [29] and demand for information of practical value was expected [25,30]. In addition, according to the uncertainty reduction theory, to alleviate risks in crises, people intend to engage in uncertainty reduction efforts by collecting credible information and sharing with others [25]. Nanath and Joy's [25] text mining study revealed that the optimism and solution topic as well as the mental health topic were positive predictors of retweet counts of COVID-19–related tweets. In addition to information utility, novel content in health news has been found to increase sharing [26]. COVID-19 vaccines were newly developed to help fight off the new coronavirus; thus, content related to aspects such as development and efficacy had the intrinsic feature of novelty and could potentially help to close the knowledge gaps.

Past research has generally shown that there were more positive than negative tweets on Twitter about vaccines in general [11-13] and about COVID-19 vaccines in particular [15,16]. Although positive content has been found to increase likes on social media [22,23], the findings are mixed regarding the impact of valence on the virality of online content. Berger and Milkman [27] found that positive sentiment increased social

media sharing of general news. A plausible explanation is that positive sharing reflects the positivity of the sender [26], which may enhance self-presentation [31] and identity communication [27]. However, Nanath and Joy [25] found that negative emotions increased the social transmission of COVID-19–related tweets. Moreover, Blankenship et al [11] revealed that antivaccine tweets were retweeted more than provaccine tweets. In comparison, Kim [26] revealed that content valence was unrelated to the virality of health news on social media.

In addition to content topic and valence, social media message features, including media presence, linguistic features, and account verification, could impact the popularity and virality of online content. Media presence and linguistic features can affect content processing fluency and further affect favorable online responses such as liking and retweeting. Content on social media may be of any mode such as text, photos, and videos. Past research has shown that a tweet with embedded media (ie, a photo or a video) stimulates likes and retweets [23]. It is postulated that the cognitive processing of photos is more fluent than that of words as it is faster to activate the semantic meaning of photos than that of words [32,33]. Therefore, tweets featuring embedded media are more likely to trigger favorable online responses.

In comparison, past research has revealed that linguistic features such as the number of hashtags, mentions, and external links decrease likes [23] and retweets [23,25]. It is suggested that these features increase content processing disfluency in two aspects. First, compared to the black color adopted by text, the blue color adopted by hashtags, mentions, and external links decreases the font-background contrast and causes visual perpetual disfluency [23,34]. Second, the nonalphanumeric symbols used by hashtags, mentions, and external links (ie, #, @, ://) create orthographical disfluency [23,35]. The content disfluency requires more cognitive effort to process the message and hence decreases favorable responses [23].

Finally, account features could potentially affect likes and retweets. In the face of information explosion in the digital age, account authenticity could be of particular importance in the diffusion of information. On Twitter, verified accounts have a blue badge next to the profile name to let users know that it is authentic. Twitter paused public submissions for account verification in 2017 and reopened the gate using a new application process in May 2021 [36]. The end date of our data retrieval was April 30, 2021, and therefore the data did not reflect the newly verified accounts. In addition, it is noteworthy that the tweets posted by verified accounts may not be verified.

Research Model and Questions

This study contributes to the literature by providing a conceptual model to understand the combined effects of the three above-mentioned categories of factors—content topics, content valance, and social media message features, including media presence, linguistic features, and account verification—on the popularity and virality of tweets about COVID-19 vaccines. We employed topic modeling to identify latent topics of tweets. We employed sentiment analysis to assess the valence of tweets. Automated extraction generated data about social media features. Therefore, we put forward the following research questions:

> *Research question 1 (RQ1): How do content topics, content valence, and social media message features affect the popularity of tweets about COVID-19 vaccines?*
>
> *Research question 2 (RQ2): How do content topics, content valence, and social media message features affect the virality of tweets about COVID-19 vaccines?*

In addition, among the 2500 most liked and most retweeted tweets, respectively, we used network analysis and visualization to detect topic communities and present the relationship between the topics and the tweets. We had the following research questions:

> *Research question 3 (RQ3): What are the salient topics of the most liked tweets?*
>
> *Research question 4 (RQ4): What are the salient topics of the most retweeted tweets?*

This study can help to advance knowledge on complex drivers of the popularity and virality of tweets about COVID-19 vaccines using machine-based text mining and network visualization in the context of a heated vaccine debate in the United States. These findings offer practical implications for health practitioners to employ more effective social media content.

Methods

Data Source

We collected publicly available original tweets about COVID-19 vaccines from January 1, 2020, to April 30, 2021, using snscrape [37], which were further filtered according to user profile data to include only English-language tweets and those from US-based users. This approach resulted in 501,531 tweets recorded in the final dataset.

Drawing on prior social media studies on vaccines [38,39], we developed keywords by balancing the general COVID-19 vaccine information with brand-specific information. As of April 30, 2021, which was our data retrieval end date, Pfizer-BioNTech, Moderna, and Johnson & Johnson/Janssen vaccines were authorized for emergency use in the United States [40]. At that time, the three vaccines, together with the AstraZeneca vaccine, had conditional marketing authorizations in European Union countries [41]. Although the AstraZeneca vaccine was not used in the United States, it garnered media and public attention in the United States, and therefore we also included this brand in the search. In addition, as COVID-19 vaccines varied in terms of the underlying technology, we considered technology-specific information. Pfizer-BioNTech and Moderna used mRNA technology, and Johnson & Johnson and AstraZeneca-Oxford used viral vector technology. Moreover, we checked government Twitter accounts such as the US CDC and Food and Drug Administration accounts to explore hashtags. Finally, the following strategy was used to scrape Twitter data. A tweet had to contain the keyword

(case-insensitive unless otherwise specified) "vaccine," together with one of the keywords "COVID," "COVID19," "COVID-19," "Pfizer," "Pfizer-BioNTech," "Moderna," "Johnson & Johnson," "Janssen," "AstraZeneca," and "Oxford-AstraZeneca"; or contain the keyword "vaccine" together with one of the following combinations: "mRNA" and "COVID," "viral vector" and "COVID," and "adenovirus" and "COVID"; or contain either of the two hashtags "#covid19vaccine" and "#covidvaccine."

Data Processing

The final dataset was preprocessed via genism [42] for topic modeling and sentiment analysis. We tokenized each tweet as a list of words [43], and removed high-frequency stop words such as "https" and "covid," in addition to the standard nltk stop words library [44], which were not expected to contribute to the uniqueness of each topic. The text corpus was then trained to recognize frequent bigrams such as "New York," using a gensim bigram model [42]. Next, all words were lemmatized to their dictionary form [43] to reduce redundancy in the bag of words (BOW) encoding. Finally, these lemmatized single words (ie, unigrams) and bigrams recognized by the bigram model were used to build the BOW representation for our latent Dirichlet allocation (LDA) model. That is, the corpus was encoded as a vector space, with each vector component representing a lemma.

Measures

Like Count

The like count of each tweet, which is the number of likes a tweet gets, was captured in the data set. As a small number of tweets generated a great number of likes, the distribution was right-skewed. To reduce right skewness, we used the natural logarithm of like counts in statistical analyses, as in past research [23].

Retweet Count

The retweet count of each tweet, which is the number of retweets a tweet gets, was captured in the data set. Similar to like counts, retweet counts had a right-skewed distribution. To reduce right skewness, we used the natural logarithm of retweet counts in statistical analyses, as in past research [23,25].

Content Topic

The tweets were subjected to topic modeling using the LDA model [45]. Topic modeling is a commonly used unsupervised learning method that generates a probabilistic model for the corpus of text data [46]. As one of the two main topic models [46], LDA is increasingly being used to analyze textual data [47], including tweets (eg, [16-18,20,25]).

LDA depends on two matrices to define the latent topical structure: the word-topic matrix and the document-topic matrix [47]. In our study, a document was a tweet. The general idea is that a tweet is represented by a Dirichlet distribution of latent topics, where each latent topic is represented by a Dirichlet distribution of words [46].

The word-topic matrix reveals the conditional probability with which a word is likely to occur in a topic. The word-topic matrix is used to interpret the topics. A topic can be interpreted by examining a list of the most probable words ranked solely by their frequency to occur in that topic, using 3 to 30 words [48]. To aid topic interpretation, we also considered the ranking of the most probable topic-specific words by both frequency and relevance, as suggested by Sievert and Shirley [48]. The relevance for ranking words within a topic is indexed by a weight parameter, λ, with a value ranging from 0 to 1. A value closer to 0 highlights rare but exclusive words for the topic and a value closer to 1 highlights frequent but not necessarily exclusive words for the topic [48]. We adopted the recommended λ of 0.6 [48]. Lastly, we reviewed sample tweets with the highest topic-specific loadings to finalize topic interpretations.

The document-topic matrix reveals the conditional probability with which a topic is likely to occur in a tweet. In other words, it reveals the topic loadings for each tweet. The information was used in the regression models for prediction as well as in network analysis and visualization. The topic loading value ranges from 0 to 1, with a value closer to 1 indicating the higher topic loading of a tweet.

Content Valence

We used TextBlob [49], an open-source python library, to generate the valence score of each tweet. The range of the valence score is from –1 to 1, with the value of –1 indicating the most negative and the value of 1 indicating the most positive valence.

Media Presence

Data on whether a tweet had a photo, gif, or video were extracted, respectively.

Linguistic Features

The numbers of hashtags, mentions, and hyperlinks were extracted, respectively.

Account Verification

For each tweet, whether the account that posted it was verified or not was extracted.

Data Analysis

We performed linear regression analyses to examine the predictors of likes and retweets. Since the purpose of the study was to investigate the factors that affected the popularity and virality of tweets as indexed by like counts and retweet counts, we only considered the tweets that were liked and retweeted, as in past research [23,25]. In the models, the log-transformed like counts and retweet counts were respectively regressed on 12 topic loadings extracted from topic modeling, the valence score generated from sentiment analysis, three variables of media presence, three variables of linguistic features, and account verification.

Network Analysis and Visualization

We used two-mode visualization to present the relationship between topics and the 2500 most liked tweets and the 2500 most retweeted tweets, respectively. To prepare data for rendering each relationship network, we created a node list

consisting of topic and tweet nodes, and an edge list consisting of tweet IDs, the topics each tweet was connected to, and an edge weight representing the topic loading of each tweet. Each topic node with its name was sized in proportion to the sum of topic loadings of all tweets. To assist the viewer in discerning topics, we used a community detection algorithm built in Gephi [50], which is based on the Louvain modularity method that has been used in prior research [12]. Community detection algorithms [51] identify cohesive groups in the network [52,53]. In the network visualization, node color reflected topic community membership.

Results

Content Topics

We trained a topic model using LDA, with a search space on topic numbers from 3 to 21. Using a uniform search grid on Dirichlet concentration parameters, the model parameters were trained to optimize the coherence score C_v [54], which is a likelihood measure of word cooccurrence in the same topics. The best model was achieved at 12 topics with C_v=0.42. Table 1 summarizes the 12 topics. Interpretation of each topic was based on the top 10 probable words ranked solely by frequency and jointly by frequency and relevance, as well as review of sample tweets with high topic-specific loadings.

Table 1. Summary of topics and valence.

Topic number	Topic label	Top 10 words by frequency (λ=1)	Top 10 words by frequency and relevance (λ=0.6)	Valence
1	Vaccine access	vaccine, community, health, help, access, need, work, pandemic, country, support	vaccine, community, health, access, help, support, effort, global, distribution, ensure	0.137
2	Vaccine efficacy and rollout	vaccine, case, new, variant, show, death, test, risk, virus, report	case, vaccine, variant, show, new, test, death, study, pause, report	0.147
3	Vaccine development and people's views	vaccine, people, take, say, would, do, want, think, give, woman	vaccine, would, take, woman, people, think, enough, do, say, try	0.158
4	Vaccination status	get, vaccine, vaccinate, shot, people, shoot, vaccinated, first, fully, wait	get, vaccinate, shot, shoot, people, vaccinated, fully, family, wait, die	0.143
5	Feeling and side effect	get, vaccine, feel, go, good, day, side effect, make, work, arm	feel, get, side effect, good, go, arm, day, fact, science, normal	0.117
6	Vaccine appointment	vaccine, appointment, today, site, schedule, open, visit, call, clinic, vaccination	appointment, site, vaccine, open, schedule, visit, clinic, join, register, call	0.133
7	Vaccine availability	vaccine, available, week, say, year, question, old, last, next, come	available, question, old, year, week, say, last, next, answer, month	0.149
8	Vaccination eligibility and administration	dose, vaccine, receive, today, first, second, eligible, administer, day, start	dose, receive, second, eligible, today, first, administer, vaccine, day, begin	0.354
9	Age and issues	age, vaccine, offer, people, group, encourage, read, rollout, issue, concern	age, offer, group, encourage, rollout, reason, article, issue, explain, doctor	0.107
10	Preventive measures	safe, mask, keep, spread, stop, stay, wear, still, continue, passport	safe, mask, keep, spread, stop, stay, wear, passport, place, home	0.089
11	Student and county	retweet, check, student, event, walk, turn, county, staff, please, team	retweet, check, student, event, walk, turn, county, staff, please, team	0.093
12	Trust and communication	share, trust, watch, video, speak, play, minute, fall, head, availability	share, trust, video, speak, play, minute, watch, fall, head, availability	0.089

Content Valence

The overall valence was positive, with a score of 0.145. The range of the valence score is from –1 to 1, with –1 indicating the most negative and 1 indicating the most positive valence. As shown in Table 1, all 12 topics were associated with positive valence.

Determinants of Like Counts

Table 2 reveals the effects of the four categories of independent variables on the log-transformed like counts. The regression model was significant at P<.001 (adjusted R^2=0.151). RQ1 was related to the determinants of likes. Out of the 12 latent topics identified by topic modeling, Topics 1 to 8 had weak but significant effects on likes. The valence also had a weak but significant effect on likes. Positive tweets increased likes. Media (photo, gif, or video) presence increased likes. Among linguistic features, the number of hashtags and that of external links decreased likes, whereas the number of mentions increased likes. Account verification increased likes.

Table 2. Linear regression models on predictors of popularity and virality of tweets.

Variables	Ln (like count)[a] (n=286,657)		Ln (retweet count)[a] (n=168,961)	
	β	*P* value	β	*P* value
Topics				
T1: Vaccine access	.029	.048	.062	<.001
T2: Vaccine efficacy and rollout	.049	<.001	.077	<.001
T3: Vaccine development and people's views	.055	<.001	.078	<.001
T4: Vaccination status	.048	<.001	.068	<.001
T5: Feeling and side effect	.040	<.001	.052	<.001
T6: Vaccine appointment	.027	<.001	.033	<.001
T7: Vaccine availability	.018	<.001	.019	<.001
T8: Vaccination eligibility	.011	<.001	.006	.08
T9: Age and issues	.009	.13	.009	.10
T10: Preventive measures	–.030	.26	–.037	.25
T11: Student and county	.076	.14	–.080	.14
T12: Trust and communication	–.079	.11	–.072	.21
Emotion (valence)	.059	<.001	.0003	.93
Media presence				
Has photo	.188	<.001	.088	<.001
Has gif	.019	<.001	.001	.64
Has video	.100	<.001	.084	<.001
Linguistic features				
Number of hashtags	–.072	<.001	–.059	<.001
Number of mentions	.007	.005	–.002	.45
Number of external links	–.126	<.001	.003	.18
Verified account	.452	<.001	.378	<.001

[a]To account for the right skewness of the data distribution, the natural log–transformed like counts and retweet counts were used in the analyses.

Determinants of Retweet Counts

Table 2 also reveals the effects of the four categories of independent variables on the log-transformed retweet counts. The regression model was significant at $P<.001$ (adjusted $R^2=0.130$). RQ2 focused on the determinants of retweets. Out of the 12 latent topics identified by topic modeling, Topics 1 to 7 had weak but significant effects on retweets. The valence had no effect on retweets. Media presence of a photo or video increased retweets. Among linguistic features, the number of hashtags decreased retweets. Account verification increased retweets.

Topic and Tweet Relationship Networks

RQ3 focused on salient topics among the most liked tweets. As shown in Figure 1, among the 2500 most liked tweets, Louvain clustering identified 4 out of the 12 topics. The tweets were clustered around vaccine access (Topic 1), followed closely by vaccine efficacy and rollout (Topic 2) and then vaccine development and people's views (Topic 3). The other topics were not salient and presented as one remaining cluster. Each topic community was represented by one color.

Table 3 summarizes the top 10 liked paraphrased tweets, like counts, dominant topics, and topic loadings. The first most liked tweet, which was posted in July 2020 and had 91,163 likes as of April 30, 2021, was clustered around vaccine access (Topic 1). It called for Medicare for All along with free COVID testing, treatment, and vaccines.

RQ4 focused on salient topics of the most retweeted tweets. As shown in Figure 2, among the top 2500 most retweeted tweets, Louvain clustering identified 5 out of the 12 topics the LDA identified in the total tweets. The top retweeted tweets mostly clustered around vaccine efficacy and rollout (Topic 2), closely followed by access to vaccine (Topic 1), and then vaccine development and people's views (Topic 3) and vaccination status (Topic 5). The other topics were not salient and presented as one remaining cluster. Each topic community was represented by one color.

Figure 1. Topic communities of the 2500 most liked tweets. Two-mode visualization was used to present the relationship between topics and the 2500 most liked tweets. The topics and the tweets are connected by edges weighted by topic loadings of each tweet. Each topic node with its name is sized in proportion to the sum of topic loadings of all tweets. Colors indicate topic communities as partitioned by the Louvain algorithm.

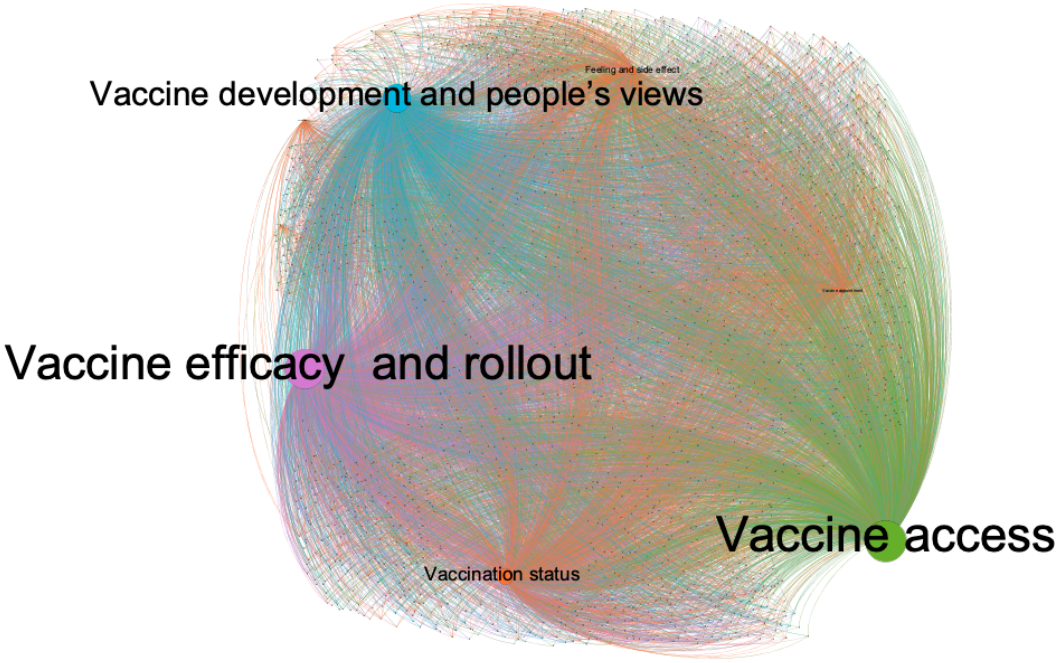

Table 3. Top 10 liked paraphrased tweets.

Like rank	Like count	Tweet	Dominant topic number and label	Dominant topic loading
1	91,163	Medicare for All along with free COVID testing, treatment, and vaccines are necessities of a decent society (July 2020).[a]	Topic 1: Vaccine access	0.518
2	90,177	Trump's attempt to deny vaccines to New York is playing politics with people's lives (November 2020).[a]	Topic 2: Vaccine efficacy and rollout	0.578
3	63,681	I participated in Moderna experiments to see if its vaccine and booster were safe and effective (April 2021)	Topic 3: Vaccine development and people's views	0.373
4	55,223	President Biden took credit for the vaccine from President Trump (March 2021)[a]	Topic 1: Vaccine access	0.964
5	48,631	The number of vaccine doses administered outnumbered that of new cases at a 10-to-1 ratio (February 2021)	Topic 2: Vaccine efficacy and rollout	0.514
6	46,997	I had ended my support for Trump and started taking COVID seriously. I got vaccinated, thanks to Biden and health workers (March 2021)	Topic 4: Vaccination status	0.578
7	36,753	Like with smallpox, vaccinations along with surveillance and contact tracing are essential to COVID's elimination (April 2020)[a]	Topic 2: Vaccine efficacy and rollout	0.547
8	36,250	Pfizer's mRNA vaccine candidate showed initial evidence of efficacy (November 2020)[a]	Topic 3: Vaccine development and people's views	0.844
9	35,604	President Trump delivered on his goal of having a safe and effective COVID vaccine by the end of the year (May 2020)	Topic 3: Vaccine development and people's views	0.533
10	35,514	The current vaccination pace will take 10 years to reach herd immunity. We need to speed this up (December 2020)[a]	Topic 2: Vaccine efficacy and rollout	0.385

[a]Tweet was among the top 10 liked and concurrently one of the top 10 retweeted tweets.

Figure 2. Topic communities of the 2500 most retweeted tweets. Two-mode visualization was used to present the relationship between topics and the 2500 most retweeted tweets. The topics and the tweets are connected by edges weighted by topic loadings of each tweet. Each topic node with its name is sized in proportion to the sum of topic loadings of all tweets. Colors indicate topic communities as partitioned by the Louvain algorithm.

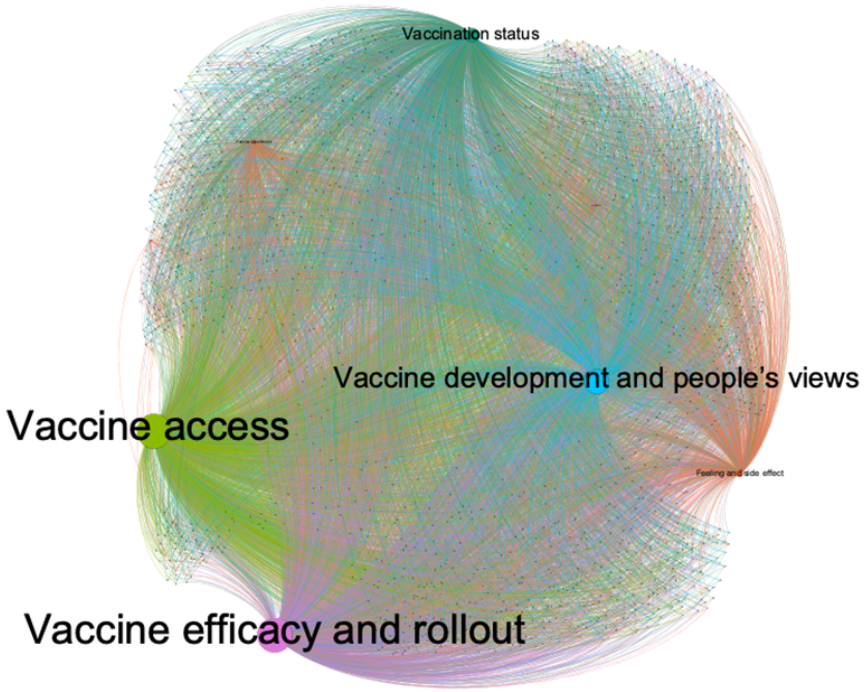

Table 4 summarizes the top 10 retweeted paraphrased tweets, their retweet counts, and dominant topics. The first most retweeted tweet, which was posted in December 2020 and garnered 17,427 retweets through April 2021, clustered around vaccine efficacy and rollout (Topic 2). This emphasized the long time needed to reach herd immunity based on the vaccination pace at that time.

Table 4. Top 10 retweeted paraphrased tweets.

Retweet rank	Retweet count	Tweet	Dominant topic number and label	Dominant topic loading
1	17,427	The current vaccination pace will take 10 years to reach herd immunity. We need to speed this up (December 2020)[a]	Topic 2: Vaccine efficacy and rollout	0.385
2	16,288	Medicare for All along with free COVID testing, treatment, and vaccines are necessities of a decent society (July 2020)[a]	Topic 1: Vaccine access	0.518
3	15,575	Trump's attempt to deny vaccines to New York is playing politics with people's lives (November 2020)[a]	Topic 2: Vaccine efficacy and rollout	0.578
4	14,536	The FDA[b] and CDC[c] recommend a pause in the use of the Johnson & Johnson COVID19 vaccine (April 2021)	Topic 1: Vaccine access	0.417
5	12,473	Pfizer's mRNA vaccine candidate showed initial evidence of efficacy (November 2020)[a]	Topic 3: Vaccine development and people's views	0.844
6	11,684	President Biden took credit for the vaccine from President Trump (March 2021)[a]	Topic 1: Vaccine access	0.964
7	11,046	Russian vaccine trial shows high efficacy (February 2021)	Topic 2: Vaccine efficacy and rollout	0.618
8	10,151	UK's vaccine is safe and induces an immune reaction (July 2020)	Topic 2: Vaccine efficacy and rollout	0.844
9	8586	Like with smallpox, vaccinations along with surveillance and contact tracing are essential to COVID's elimination (April 2020)[a]	Topic 2: Vaccine efficacy and rollout	0.547
10	8282	Why we need two doses of mRNA vaccines (April 2021)	Topic 1: Vaccine access	0.488

[a]Tweet was among the top 10 retweeted and concurrently one of the top 10 liked tweets.

[b]FDA: Food and Drug Administration.

[c]CDC: Centers for Disease Control and Prevention.

Discussion

Principal Results

This study investigated the combined effects of the three categories of message-level factors on the popularity and virality of tweets about COVID-19 vaccines using text-mining techniques. We also examined the topic communities of the most liked and most retweeted tweets using network analysis and visualization. In this section, we first discuss how text-mined topics and valence, together with autoextracted information about social media message features affected likes and retweets. We further discuss limitations and implications for the directions of vaccine campaigns.

Out of the 12 latent topics identified by topic modeling, Topics 1-8 increased likes and Topics 1-7 increased retweets. Vaccine development and people's views (Topic 3) had the largest positive impact on likes and retweets, as reflected by β coefficients. The intrinsic novelty feature of COVID-19 vaccines could provide plausible explanations. The vaccines were newly developed to help fight off the new coronavirus, and two out of the four brands examined in the study used mRNA, a technology that had not been approved previously for general use in humans [5]. Therefore, information about vaccine development and technology was more popular and viral. Relatedly, 3 out of the top 10 liked tweets reflected Topic 3, two of which were about mRNA vaccines. One out of the top 10 retweeted tweets reflected Topic 3, which was about mRNA vaccines. The findings were consistent with those in past research that suggested the impact of novel content in the social transmission of health news [26].

Vaccine efficacy and rollout (Topic 2) had the second largest positive impact on likes and retweets, as indicated by β coefficients. Prior research revealed the impact of efficacy information on the virality of online health news [26] and in tweets about the COVID-19 pandemic [25]. This study also underscores the importance of efficacy information on the virality of tweets about COVID-19 vaccines.

The findings suggest that tweets focusing on the topic of vaccine development and people's views, and the topic of vaccine efficacy and rollout highly meet the public's needs for information during the COVID-19 pandemic, and therefore tend to become popular and viral on Twitter. It is plausible that these tweets provide useful and novel information that help to reduce uncertainty in a health crisis. Vaccine campaigns could provide more information about these topics to help the diffusion of information on social media.

It is notable that polarized political information such as that supporting a political party could be intertwined with different topics. Polarized political information was contained in 5 out of the top 10 liked tweets and in 3 out of the top 10 retweeted tweets. As political stance may play a role in the vaccine debate in the United States [9], it would be interesting for future studies to investigate its impact in addition to other factors.

This study showed that the overall valence of the tweets was positive. This was consistent with findings in prior research on tweets about vaccines in general [11-13] and about COVID-19 vaccines in particular, regardless of country [15,16]. The results showed that positive valence increased likes. This is in alignment with findings in prior research [22,23]. In comparison, the results showed no impact of valence on retweets. Past research revealed mixed findings regarding the effects of valence on retweets [11,25-27]. The explanation may rest in the complex cognitive sources underlying retweeting behavior. Compared with liking, retweeting is a more social behavior that may involve expected reactions from recipients about the content and/or the sender [26].

Regarding social media message features, account verification had the largest positive impact on likes and retweets among all factors, as reflected by β coefficients. This finding underscores the importance of account authentication in the popularity and virality of tweets in the face of massive amounts of information. Credible information is vital to reduce uncertainty in a crisis according to the uncertainty reduction theory [25,55]. However, it is notable that account authentication does not always mean content authentication. Accordingly, misinformation spread by verified accounts could pose greater challenges to vaccine campaigns. Vaccine campaigns could try to use and motivate different verified accounts, including institutional and individual accounts, to share credible information for wider reach and to prevent the spread of misinformation.

Furthermore, in alignment with the literature [32,33], the presence of a photo or video enhanced likes and retweets. The presence of a gif increased likes but did not affect retweets. In addition, consistent with the literature [23,34,35], the number of hashtags decreased likes and retweets. The number of external links decreased likes, but did not affect retweets. Inconsistent with the literature [23,25], the number of mentions facilitated likes, but did not affect retweets.

The results revealed that among the examined factors, more could impact likes than retweets. Eight topics predicted likes, whereas seven predicted retweets. Valence predicted likes but did not predict retweets. The presence of a gif, the number of mentions, and the number of external links predicted likes but not retweets. A comparison between like counts of the top 10 liked tweets and retweet counts of the top 10 retweeted tweets also suggested that a tweet was much more likely to be liked than to be retweeted. The number of likes for the highest liked tweet was more than five times the number of retweets for the highest retweeted tweet. These findings indicate more challenges to make a tweet viral than popular.

Limitations

This study has several limitations. We used machine-based text mining to identify the underlying topics and valence in the vast amounts of tweets about COVID-19 vaccines. We then included the text-mined topics and valence, together with autoextracted information of social media message features in the regression models for prediction of the popularity and virality of tweets. Although this approach reduced manual coding, the results were mostly limited to autoidentified and autoextracted factors. Our manual reviews of sample tweets in each topic as well as the top 10 liked and retweeted tweets provided clues that politically polarized information could be intertwined with different topics. It would be interesting for future research to investigate how this may affect the popularity and virality of tweets. For instance, retweeting could derive from complex cognitive sources such as self-presentation [31] and identity communication [27]. A question arises whether consistency in the political stance between the sender and the recipients impact retweets.

Furthermore, the findings were limited to US-based public discourse about COVID-19 vaccines on Twitter. Social media platforms have played an important role in disseminating information and opinions during the COVID-19 pandemic [56]. It would be interesting for future research to compare Twitter with other social media platforms. For instance, the relative significance of examined factors in predicting popularity and virality may vary depending on the social media platform analyzed, as each has its own features.

Finally, the results revealed message-level drivers of the popularity and virality of tweets about COVID-19 vaccines. We included account verification as an independent variable in the regression models and the results showed that it had a positive impact on likes and retweets. However, we did not identify social bots in the massive amounts of tweets. It would be interesting for future studies to investigate the impact of social bots.

Conclusions

This study suggests the public interest in and demand for information about vaccine development and people's views, as well as vaccine efficacy and rollout during the COVID-19 pandemic. These topics, along with the use of media and verified accounts, enhance the popularity and virality of tweets. These issues could be addressed in vaccine campaigns to help the diffusion of content on Twitter.

Conflicts of Interest

None declared.

References

1. WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March. World Health Organization. 2020 Mar 11. URL: https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020 [accessed 2020-03-15]
2. COVID-19 Dashboard. Johns Hopkins University. URL: https://coronavirus.jhu.edu/map.html [accessed 2021-04-15]

3. COVID-19 vaccines. World Health Organization. 2021. URL: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/covid-19-vaccines [accessed 2021-07-20]
4. Science brief: COVID-19 vaccines and vaccination. Centers for Disease Control and Prevention. 2021 Jul 27. URL: https://www.cdc.gov/coronavirus/2019-ncov/science/science-briefs/fully-vaccinated-people.html [accessed 2021-07-28]
5. Ball P. The lightning-fast quest for COVID vaccines—and what it means for other diseases. 2020 Dec 18. URL: https://www.nature.com/articles/d41586-020-03626-1 [accessed 2020-12-20]
6. Zipkin M. New vaccine approaches present new possibilities, but new challenges. 2021 Jun 01. URL: https://www.nature.com/articles/d43747-021-00079-x [accessed 2021-06-05]
7. Iannelli V. An overview of the vaccine debate: looking at both sides of the argument. 2021 Jun 04. URL: https://www.verywellhealth.com/the-vaccine-debate-2633685 [accessed 2021-06-10]
8. Kerr J, Panagopoulos C, van der Linden S. Political polarization on COVID-19 pandemic response in the United States. Person Ind Diff 2021 Sep;179:110892. [doi: 10.1016/j.paid.2021.110892]
9. Fridman A, Gershon R, Gneezy A. COVID-19 and vaccine hesitancy: A longitudinal study. PLoS One 2021 Apr 16;16(4):e0250123 [FREE Full text] [doi: 10.1371/journal.pone.0250123] [Medline: 33861765]
10. Harris R. Why the COVID-19 vaccine distribution has gotten off to a slow start. 2021 Jan 01. URL: https://www.npr.org/2021/01/01/952652202/why-the-covid-19-vaccine-distribution-has-gotten-off-to-a-slow-start [accessed 2021-01-10]
11. Blankenship E, Goff ME, Yin J, Tse ZTH, Fu KW, Liang H, et al. Sentiment, contents, and retweets: a study of two vaccine-related Twitter datasets. Perm J 2018;22:17-138 [FREE Full text] [doi: 10.7812/TPP/17-138] [Medline: 29911966]
12. Gunaratne K, Coomes EA, Haghbayan H. Temporal trends in anti-vaccine discourse on Twitter. Vaccine 2019 Aug 14;37(35):4867-4871. [doi: 10.1016/j.vaccine.2019.06.086] [Medline: 31300292]
13. Love B, Himelboim I, Holton A, Stewart K. Twitter as a source of vaccination information: content drivers and what they are saying. Am J Infect Control 2013 Jun;41(6):568-570. [doi: 10.1016/j.ajic.2012.10.016] [Medline: 23726548]
14. Ortiz-Sánchez E, Velando-Soriano A, Pradas-Hernández L, Vargas-Román K, Gómez-Urquiza JL, Cañadas-De la Fuente GA, et al. Analysis of the anti-vaccine movement in social networks: a systematic review. Int J Environ Res Public Health 2020 Jul 27;17(15):5394 [FREE Full text] [doi: 10.3390/ijerph17155394] [Medline: 32727024]
15. Hussain A, Tahir A, Hussain Z, Sheikh Z, Gogate M, Dashtipour K, et al. Artificial intelligence-enabled analysis of public attitudes on Facebook and Twitter toward COVID-19 vaccines in the United Kingdom and the United States: observational study. J Med Internet Res 2021 Apr 05;23(4):e26627 [FREE Full text] [doi: 10.2196/26627] [Medline: 33724919]
16. Lyu JC, Han EL, Luli GK. COVID-19 vaccine-related discussion on Twitter: topic modeling and sentiment analysis. J Med Internet Res 2021 Jun 29;23(6):e24435 [FREE Full text] [doi: 10.2196/24435] [Medline: 34115608]
17. Abd-Alrazaq A, Alhuwail D, Househ M, Hamdi M, Shah Z. Top concerns of Tweeters during the COVID-19 pandemic: infoveillance study. J Med Internet Res 2020 Apr 21;22(4):e19016 [FREE Full text] [doi: 10.2196/19016] [Medline: 32287039]
18. Chandrasekaran R, Mehta V, Valkunde T, Moustakas E. Topics, trends, and sentiments of tweets about the COVID-19 pandemic: temporal infoveillance study. J Med Internet Res 2020 Oct 23;22(10):e22624 [FREE Full text] [doi: 10.2196/22624] [Medline: 33006937]
19. Doogan C, Buntine W, Linger H, Brunt S. Public perceptions and attitudes toward COVID-19 nonpharmaceutical interventions across six countries: a topic modeling analysis of Twitter data. J Med Internet Res 2020 Sep 03;22(9):e21419 [FREE Full text] [doi: 10.2196/21419] [Medline: 32784190]
20. Kwok SWH, Vadde SK, Wang G. Tweet topics and sentiments relating to COVID-19 vaccination among Australian Twitter users: machine learning analysis. J Med Internet Res 2021 May 19;23(5):e26953 [FREE Full text] [doi: 10.2196/26953] [Medline: 33886492]
21. Liu Q, Zheng Z, Zheng J, Chen Q, Liu G, Chen S, et al. Health communication through news media during the early stage of the COVID-19 outbreak in China: digital topic modeling approach. J Med Internet Res 2020 Apr 28;22(4):e19118 [FREE Full text] [doi: 10.2196/19118] [Medline: 32302966]
22. Lee J, Hong IB. Predicting positive user responses to social media advertising: The roles of emotional appeal, informativeness, and creativity. Int J Inf Manag 2016 Jun;36(3):360-373. [doi: 10.1016/j.ijinfomgt.2016.01.001]
23. Pancer E, Poole M. The popularity and virality of political social media: hashtags, mentions, and links predict likes and retweets of 2016 U.S. presidential nominees' tweets. Soc Influ 2016 Dec 12;11(4):259-270. [doi: 10.1080/15534510.2016.1265582]
24. Using Twitter. Twitter. URL: https://help.twitter.com/en/using-twitter/ [accessed 2021-01-05]
25. Nanath K, Joy G. Leveraging Twitter data to analyze the virality of Covid-19 tweets: a text mining approach. Behav Inf Technol 2021 Jun 17:1-19. [doi: 10.1080/0144929x.2021.1941259]
26. Kim HS. Attracting views and going viral: how message features and news-sharing channels affect health news diffusion. J Commun 2015 Jun 01;65(3):512-534 [FREE Full text] [doi: 10.1111/jcom.12160] [Medline: 26441472]
27. Berger J, Milkman KL. What makes online content viral? J Market Res 2012 Apr 01;49(2):192-205. [doi: 10.1509/jmr.10.0353]
28. Moriarty CM, Stryker JE. Prevention and screening efficacy messages in newspaper accounts of cancer. Health Educ Res 2008 Jun 01;23(3):487-498. [doi: 10.1093/her/cyl163] [Medline: 17289658]

29. McCormack LA, Squiers L, Frasier AM, Lynch M, Bann CM, MacDonald PDM. Gaps in knowledge about COVID-19 among US residents early in the outbreak. Public Health Rep 2021 Nov 11;136(1):107-116. [doi: 10.1177/0033354920970182] [Medline: 33176108]
30. Lee CH, Yu H. The impact of language on retweeting during acute natural disasters: uncertainty reduction and language expectancy perspectives. Ind Manag Data Syst 2020 Jun 29;120(8):1501-1519. [doi: 10.1108/imds-12-2019-0711]
31. Wojnicki AC, Godes D. Word-of-mouth as self-enhancement. SSRN J 2008 Apr 28:06-01. [doi: 10.2139/ssrn.908999]
32. Arieh Y, Algom D. Processing picture-word stimuli: the contingent nature of picture and of word superiority. J Exp Psychol Learn Mem Cogn 2002 Jan;28(1):221-232. [doi: 10.1037/0278-7393.28.1.221] [Medline: 11827082]
33. Shaki S, Algom D. The locus and nature of semantic congruity in symbolic comparison: evidence from the Stroop effect. Mem Cognit 2002 Jan;30(1):3-17. [doi: 10.3758/bf03195260] [Medline: 11958352]
34. Reber R, Schwarz N, Winkielman P. Processing fluency and aesthetic pleasure: is beauty in the perceiver's processing experience? Pers Soc Psychol Rev 2004 Dec 21;8(4):364-382. [doi: 10.1207/s15327957pspr0804_3] [Medline: 15582859]
35. Alter AL, Oppenheimer DM. Uniting the tribes of fluency to form a metacognitive nation. Pers Soc Psychol Rev 2009 Aug 28;13(3):219-235. [doi: 10.1177/1088868309341564] [Medline: 19638628]
36. Relaunching verification and what's next Twitter. Twitter. 2021 May 20. URL: https://blog.twitter.com/en_us/topics/company/2021/relaunching-verification-and-whats-next [accessed 2021-05-25]
37. snscrape. GitHub. URL: https://github.com/JustAnotherArchivist/snscrape [accessed 2021-04-05]
38. Massey PM, Leader A, Yom-Tov E, Budenz A, Fisher K, Klassen AC. Applying multiple data collection tools to quantify human papillomavirus vaccine communication on Twitter. J Med Internet Res 2016 Dec 05;18(12):e318 [FREE Full text] [doi: 10.2196/jmir.6670] [Medline: 27919863]
39. Massey PM, Kearney MD, Hauer MK, Selvan P, Koku E, Leader AE. Dimensions of misinformation about the HPV vaccine on Instagram: content and network analysis of social media characteristics. J Med Internet Res 2020 Dec 03;22(12):e21451 [FREE Full text] [doi: 10.2196/21451] [Medline: 33270038]
40. Different COVID-19 vaccines. Centers for Disease Control and Prevention. 2021 May 27. URL: https://www.cdc.gov/coronavirus/2019-ncov/vaccines/different-vaccines.html [accessed 2021-06-02]
41. Safe COVID-19 vaccines for Europeans. European Commission. 2021. URL: https://ec.europa.eu/info/live-work-travel-eu/coronavirus-response/safe-covid-19-vaccines-europeans_en [accessed 2021-04-15]
42. Rehurek R, Sojka P. Gensim—statistical semantics in Python. 2011 Presented at: European meeting on Python in Science; August 25-28, 2011; Paris, France.
43. Manning CD, Raghavan P, Schutze H. Introduction to Information Retrieval. Cambridge, UK: Cambridge University Press; 2008.
44. Loper E, Bird S. NLTK: the natural language toolkit. 2002 Presented at: Association for Computational Linguistics-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics; July 2002; Morristown, NJ p. 63-70. [doi: 10.3115/1118108.1118117]
45. Blei D, Ng A, Jordan M. Latent dirichlet allocation. J Mach Learn Res 2003 Mar 01;3(1):993-1022. [doi: 10.5555/944919.944937]
46. Allahyari M, Pouriyeh S, Assefi M, Safaei S, Trippe ED, Gutierrez JB, et al. Text summarization techniques: a brief survey. Int J Adv Comput Sci Appl 2017;8(10):397-405 [FREE Full text] [doi: 10.14569/ijacsa.2017.081052]
47. Maier D, Waldherr A, Miltner P, Wiedemann G, Niekler A, Keinert A, et al. Applying LDA topic modeling in communication research: toward a valid and reliable methodology. Commun Method Meas 2018 Feb 16;12(2-3):93-118. [doi: 10.1080/19312458.2018.1430754]
48. Sievert C, Shirley K. LDAvis: a method for visualizing and interpreting topics. 2014 Presented at: Proceedings of the Workshop on Interactive Language Learning, Visualization, Interfaces at the Association for Computational Linguistics; Jun 2014; Baltimore, MD. [doi: 10.13140/2.1.1394.3043]
49. TextBlob: simplified text processing. URL: https://textblob.readthedocs.io/ [accessed 2021-01-05]
50. Gephi. Version 0.9.2. URL: https://gephi.org/ [accessed 2021-04-15]
51. Newman MEJ, Girvan M. Finding and evaluating community structure in networks. Phys Rev E 2004 Feb 26;69(2):026113. [doi: 10.1103/physreve.69.026113]
52. Fortunato S. Community detection in graphs. Phys Rep 2010 Feb;486(3-5):75-174. [doi: 10.1016/j.physrep.2009.11.002]
53. Kang GJ, Ewing-Nelson SR, Mackey L, Schlitt JT, Marathe A, Abbas KM, et al. Semantic network analysis of vaccine sentiment in online social media. Vaccine 2017 Jun 22;35(29):3621-3638 [FREE Full text] [doi: 10.1016/j.vaccine.2017.05.052] [Medline: 28554500]
54. Röder M, Both A, Hinneburg A. Exploring the space of topic coherence measures. 2015 Presented at: Proceedings of the 8th ACM International Conference on Web Search and Data Mining; 2015; Shanghai, China p. 399-408. [doi: 10.1145/2684822.2685324]
55. Berger CR, Bradac JJ. Language and social knowledge: uncertainty in interpersonal relations (Vol. 2). London, UK: Hodder Education; 1982.
56. Merchant RM, Lurie N. Social media and emergency preparedness in response to novel coronavirus. JAMA 2020 May 26;323(20):2011-2012. [doi: 10.1001/jama.2020.4469] [Medline: 32202611]

Abbreviations

BOW: bag of words
CDC: Centers for Disease Control and Prevention
LDA: latent Dirichlet allocation
mRNA: messenger RNA
WHO: World Health Organization

Edited by G Eysenbach; submitted 10.08.21; peer-reviewed by W Xie, J Turner; comments to author 01.09.21; revised version received 12.10.21; accepted 13.10.21; published 03.12.21

Please cite as:
Zhang J, Wang Y, Shi M, Wang X
Factors Driving the Popularity and Virality of COVID-19 Vaccine Discourse on Twitter: Text Mining and Data Visualization Study
JMIR Public Health Surveill 2021;7(12):e32814
URL: https://publichealth.jmir.org/2021/12/e32814
doi: 10.2196/32814
PMID: 34665761