Original Paper

Mining Trends of COVID-19 Vaccine Beliefs on Twitter With Lexical Embeddings: Longitudinal Observational Study

Harshita Chopra[1*], BTech; Aniket Vashishtha[1*], BTech; Ridam Pal[2], BTech; Ashima[2]; Ananya Tyagi[2], BTech, MTech; Tavpritesh Sethi[2,3], MBBS, PhD

[1]Guru Gobind Singh Indraprastha University, New Delhi, India

[2]Indraprastha Institute of Information Technology, New Delhi, India

[3]All India Institute of Medical Sciences, New Delhi, India

[*]these authors contributed equally

Corresponding Author:
Tavpritesh Sethi, MBBS, PhD
Indraprastha Institute of Information Technology
Okhla Industrial Estate
Phase III
New Delhi, 110020
India
Phone: 91 97799 08630
Email: tavpriteshsethi@iiitd.ac.in

Abstract

Background: Social media plays a pivotal role in disseminating news globally and acts as a platform for people to express their opinions on various topics. A wide variety of views accompany COVID-19 vaccination drives across the globe, often colored by emotions that change along with rising cases, approval of vaccines, and multiple factors discussed online.

Objective: This study aims to analyze the temporal evolution of different emotions and the related influencing factors in tweets belonging to 5 countries with vital vaccine rollout programs, namely India, the United States, Brazil, the United Kingdom, and Australia.

Methods: We extracted a corpus of nearly 1.8 million Twitter posts related to COVID-19 vaccination and created 2 classes of lexical categories—emotions and influencing factors. Using cosine distance from selected seed words' embeddings, we expanded the vocabulary of each category and tracked the longitudinal change in their strength from June 2020 to April 2021 in each country. Community detection algorithms were used to find modules in positive correlation networks.

Results: Our findings indicated the varying relationship among emotions and influencing factors across countries. Tweets expressing hesitancy toward vaccines represented the highest mentions of health-related effects in all countries, which reduced from 41% to 39% in India. We also observed a significant change (*P*<.001) in the linear trends of categories like hesitation and contentment before and after approval of vaccines. After the vaccine approval, 42% of tweets coming from India and 45% of tweets from the United States represented the "vaccine_rollout" category. Negative emotions like rage and sorrow gained the highest importance in the alluvial diagram and formed a significant module with all the influencing factors in April 2021, when India observed the second wave of COVID-19 cases.

Conclusions: By extracting and visualizing these tweets, we propose that such a framework may help guide the design of effective vaccine campaigns and be used by policy makers to model vaccine uptake and targeted interventions.

(JMIR Infodemiology 2023;3:e34315) doi: 10.2196/34315

KEYWORDS

COVID-19; COVID-19 vaccination; vaccine hesitancy; public health; unsupervised word embeddings; natural language preprocessing; social media; Twitter

Introduction

The unprecedented spread of COVID-19 has created massive turmoil in public health around the world [1]. The development of vaccines has played a pivotal role in eradicating and mitigating significant outbreaks of infectious diseases like smallpox, tuberculosis, measles, and similar contagious diseases [2]. Major pharmaceutical companies located across the globe are in the phase of developing vaccines, with only a handful of the vaccines authorized for clinical trials [3,4]. As the distribution of vaccines and associated campaigns expand, people continue to express their opinions and personal incidents on social media platforms.

Social media plays a decisive role in propagating information, leading to the emergence of varying perceptions related to the pandemic [5]. During the initial phase of national lockdown in several countries, Twitter had reported an increase of 24% in daily active users due to the increased usage of social media, the highest year-over-year growth rate reported by the company to date [6].

Mass media strongly influences vaccine uptake and vaccination rates, as shown previously for influenza [7,8]. Although some studies have also shown a positive impact of mass media on improving vaccine uptake and mitigating hesitancy [9], its role in the spread of vaccine misinformation and conspiracy theories has been widespread [10]. Recent studies such as "The 'Pandemic' of Disinformation in COVID-19" [11] reported several events for which mass media channels have misinformed the public by sharing incomplete or unverified updates on new treatments, myths about usage of masks, and errors of some hospital organizations that resulted in higher reluctance from patients to go to hospitals or medical centres. The surge in consumption of COVID-19 updates from mass media channels has impacted different age groups by inducing panic and anxiety [12].

The COVID-19 pandemic has been studied in multidisciplinary aspects, and the analysis of Twitter posts remains a widely explored area in public health research [13-15], primarily because of the rapidly evolving nature of the content. Over the last decade, researchers have used multiple methods such as sentiment classification [16], social network analysis [17], and topic identification [18] to study the presence of provaccine and antivaccine communities on social media. It has been observed that vaccine uptake is affected by multiple factors, including rising adverse effect reporting, socioeconomic inequities, and quantitative allocation [19]. In addition, the spread of misinformation online has been a concerning issue, and prior survey-based studies suggest that it is linked with vaccine hesitancy and effects on public health [20,21]. On the other hand, certain marginalized groups continue to face inaccessibility to vaccines [22].

This paper presents a temporal and demographic analysis of lexical categories mined from Twitter conversations around vaccines. We further subdivided these categories into 2 subtypes: emotions and their influencing factors. We examined the relationships between emotions such as hesitancy, rage, contentment, sorrow, faith, and anticipation with influencing factors such as conspiracy theories around vaccines, social inequities, and health effects using unsupervised word embeddings trained on the curated corpus of tweets during an 11-month period. Further, we created correlation-based networks of these categories and performed clustering using the Infomap algorithm. The alluvial diagrams generated by these networks demonstrate the flow of importance of each factor from one month to another. We performed a granular analysis of the temporal-based trends of various outlooks toward COVID-19 vaccine activities. We analyzed their correlation with prominent factors for 5 countries (India, the United States, Brazil, the United Kingdom, and Australia) located on 5 different continents to demonstrate the comparative results among them.

Recent research work has analyzed vaccine hesitancy or sentiment analysis to determine the overall general perception among people toward COVID-19 vaccines. Our work provides a more detailed insight into the variety of outlooks people had toward the emergence of continuous vaccine updates and possible correlations with reasons for these outlooks. Major analysis work on survey data in specific regions or a cohort of the population has helped understand people's opinions toward vaccine uptake or resistance. Still, we have worked on a large corpus of tweets (more than 1.8 million) from different countries. As the meteoric rise in the use of social media has become a substantial influencing source for formulating different perceptions in millions of users, working with such a data source helps gain a broader and better sense of various factors that might be associated with fueling vaccine resistance. We have also analyzed our findings with vaccine developments and news in each country during the specific time periods to support our results.

Methods

Design and Data Set

We performed an observational study by curating a longitudinal data set by scraping more than 1.8 million tweets using the Snscrape library [23] from June 2020 to April 2021. The query used to extract the tweets was created using an "OR" combination of hashtags and words related to vaccines and the names of the vaccines administered in the respective countries. Detailed queries for each country are mentioned in Table 1.

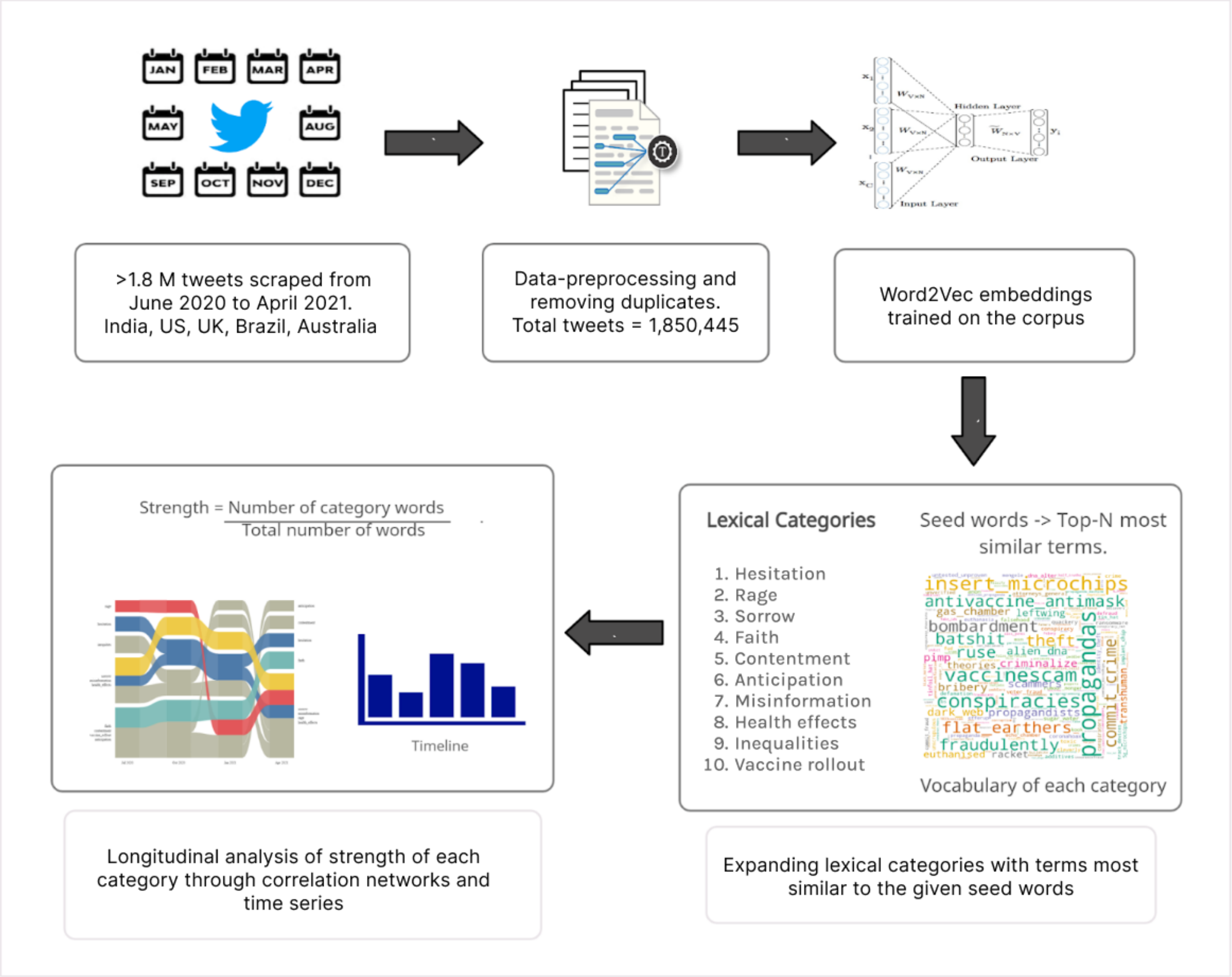
Preprocessing of tweets was carried out on lowercase-converted text by removing white spaces, punctuation, hashtags, mentions, digits, stop words, URLs, and HTML characters. The verbs present in the text were lemmatized using WordNet Lemmatizer from the Natural Language Toolkit (NLTK) package [24]. Duplicate tweets were removed based on identical username, time, and location. Figure 1 illustrates an abstract view of the study design. We list all the software and packages used in further analysis along with the corresponding versions and sources in Multimedia Appendix 1.

Table 1. Queries used for scraping tweets from each country and number of tweets used after preprocessing.

Country	Query[a]	Tweets, n
United States	(General keywords) OR (moderna OR pfizer OR biontech OR astrazeneca OR inovio OR novavax OR #pfizerbiontech)	1,121,216
United Kingdom	(General keywords) OR (pfizer OR biontech OR oxfordvaccine OR astrazeneca OR moderna OR #pfizerbiontech)	432,271
India	(General keywords) OR (covishield OR covaxin)	229,127
Australia	(General keywords) OR (pfizer OR biontech OR oxfordvaccine OR astrazeneca OR moderna OR novavax OR #pfizerbiontech)	50,224
Brazil	(General keywords) OR (coronavac OR Sinovac OR AstraZeneca OR Pfizer OR BioNTech OR #pfizerbiontech OR oxfordvaccine)	17,608

[a] General keywords: (vaccine OR vaccination OR vaccinate OR covax OR #covidvaccine OR #coronavaccine OR #covidvaccination).

Figure 1. Overview of the pipeline followed to create and analyze the strength of lexical categories.


JAN
FEB
MAR
APR
MAY
AUG
SEP
OCT
NOV
DEC
Hidden Layer
Output Layer
Input Layer
>1.8 M tweets scraped from June 2020 to April 2021. India, US, UK, Brazil, Australia
Data-preprocessing and removing duplicates. Total tweets = 1,850,445
Word2Vec embeddings trained on the corpus
Strength = Number of category words / Total number of words
Timeline
Lexical Categories
Seed words -> Top-N most similar terms.
1. Hesitation
2. Rage
3. Sorrow
4. Faith
5. Contentment
6. Anticipation
7. Misinformation
8. Health effects
9. Inequalities
10. Vaccine rollout
insert_microchips
antivaccine_antimask
bombardment
batshit
theft
ruse
vaccinescam
conspiracies
flat_earthers
fraudulently
propaganda
commit_crime
Vocabulary of each category
Longitudinal analysis of strength of each category through correlation networks and time series
Expanding lexical categories with terms most similar to the given seed words

Ethics Approval

Publicly available Twitter data were used, and an aggregated analysis was performed without any attempt to re-identify or link any personal information. The study received institutional review board approval (IIITD/IEC/08/2021-6) and was conducted under the oversight of the associated protocol.

Curating Categories Using Unsupervised Word Embeddings

We created 10 lexical categories for a psychometric evaluation of the tweet content in an approach similar to that by Empath [25]. The categories formed can be broken down into 2 classes: “emotions” and “influencing factors.” Emotions consist of the affective processes that help us understand how reactions, feelings, thoughts, and behavior of people evolve in a given situation. We selected 6 COVID-19–related emotions, namely hesitation, rage, sorrow, faith, contentment, and anticipation, along with their putative influencing factors such as misinformation, vaccine rollout, inequities, and health effects in contrast to the COVID-19 vaccines. We specified a set of seed words corresponding to these categories, as shown in Table 2.

We trained a low dimensional representation (d=100) as word embeddings for the unigrams and frequently occurring bigrams (co-occurring at least 5 times with the bigram scoring function [26] greater than a threshold of 50) present in our corpus using the skip-gram algorithm of the Word2Vec model [27] with a sliding window size of 5. We defined lexical categories as sets of words most similar to the assigned seed words. Each seed word, ensured to be present in the model's vocabulary, was mapped to a word vector. We used cosine similarity to measure proximity to find the top N(=50) words in the nearby vector space. Following this approach, k seed words were expanded to a list of maximum k×N words. A category was defined as the union set of seed words and their closest similar words (Table 2). Seed words used for the health effects category were taken from the adverse events mentioned in the Vaccine Adverse Event Reporting System (VAERS) database [28], which occurred in our data set's vocabulary. The resulting set of words in each lexical category was manually verified.

Table 2. Curated categories (emotions and influencing factors), their description, and seed words.

Category	Description	Seed words
Emotions		
1. Hesitation	Sceptic attitude and reluctance toward being vaccinated due to multiple negative factors affecting an individual's opinions	Anxious, nervous, fear, consequences, uncertain, hesitation, suspicion, harm
2. Sorrow	Dissatisfaction and disapproval toward the different phases of COVID-19 vaccine production and distribution	Sad, hopeless, worst, disappointment, setback
3. Faith	Signifies strong belief and confidence in vaccines along with optimistic behavior toward the success of vaccines	Faith, optimism, vaccines work, assurance, grateful
4. Contentment	Signifies a state of happiness, appreciation, and acceptance of the COVID-19 vaccines	Satisfy, glad, proud, gratitude, great, joy
5. Anticipation	State of urgent demand and necessity of vaccines	Anticipate, urgently, priority, quick, await
6. Rage	Anger or aggression is associated with conflict arising from a particular situation	Angry, annoyance, hate, mad, pathetic
Influencing factors		
7. Misinformation	Propagation of false information such as misinterpreted agendas and conceiving vaccines as conspiracy or scam	Propaganda, conspiracy, fraud, fake, poison
8. Vaccine rollout	Availability and distribution of vaccines through campaigns and mass vaccination drives	Vaccinate, distribution, supply, mass, dose, vaccination drive
9. Inequities	Socioeconomic disparities are based on societal norms such as caste, race, religion	Socioeconomic, deprive, racial injustice, racism, underrepresented
10. Health effects	Mentions of health-related adverse events caused by or affected by vaccines, including diseases, symptoms, and pre-existing conditions	From the VAERS[a] database (eg, headache, fatigue, inflammation)

[a]VAERS: Vaccine Adverse Event Reporting System.

Temporal Analysis of Lexical Categories

To measure each category's strength in a given text, we used the word count approach, similar to that by Empath [25] and other lexicon-based tools like Linguistic Inquiry and Word Count (LIWC) [29]. To obtain an unbiased value that is independent of the length of text, we divided the frequency by the total number of words using the following formula:

$$Strength\ of\ Category\ (S) = \frac{Number\ of\ occurrences\ of\ category\ words\ in\ text}{Total\ number\ of\ words\ in\ text}$$

We appended the preprocessed text of all tweets monthly to calculate the strength. The time series of the strength of emotion categories and influencing factors was helpful in analyzing the evolution of perceptions and opinions expressed by the public and how they vary with crucial time stamps like the news of the country's first vaccine approval.

Analysis of Change Before and After Approval

To understand the variation of emotions among social media users in the aftermath of the approval of vaccines, we conducted a before-after change analysis for each lexical category based on the date when the country's government approved the first COVID-19 vaccine.

We created a day-wise time series of the strength of each category from June 2020 to April 2021 and smoothened it using the Moving Average algorithm. The linear nature of the trend was captured using an ordinary linear regression model fit on the strength of a category in the 2 time periods preceding and succeeding the approval date. To calculate the significance of the change, we used the *z* test to compare the regression coefficients [30]:

$$z = \frac{b_1 - b_2}{\sqrt{SE_{b_1}^2 + SE_{b_2}^2}}$$

where b_1 and b_2 denote the slopes and $SE_{b_1}^2$ and $SE_{b_2}^2$ are the standard errors of the regression lines and before and after the approval, respectively.

Further, we used a change-point detection method based on dynamic programming using the Ruptures package [31] in Python3. The “Dynp” model was used with the “l1” cost function to detect one change point. This was done to verify if the date of approval was close to the change point.

To understand the Influencing factors co-occurring with hesitation, we resampled the tweets with a positive strength of hesitation (n=1000) and calculated the percentage of tweets that also had positive strength of anticipation, rage, misinformation, health effects, and inequities. The resampling was repeated for 100 iterations, and the mean and standard errors were plotted. The percentages of tweets from each of these categories that changed before and after the approval were recorded and tested for significance.

Longitudinal Correlation-Based Networks

The correlation between any 2 categories represents the degree to which they are linearly related. Daily strengths were calculated for each category followed by pairwise Pearson correlation [32]. Weighted networks of categories (nodes) and edge strengths (correlation coefficients) were constructed to evaluate the positive associations among classes (ρ≥0). Community detection on these networks was carried out using the Infomap algorithm [33], and the dynamic change in these associations was visualized as an alluvial diagram [34]. The use of the Pearson correlation typically requires the verification of some assumptions. We verified the assumption of outliers by plotting box plots of the samples and observed very few or no outliers. To check for a normal distribution, we used the Shapiro-Wilk test (used for n_samples<50), which was satisfied for most but not all months. Hence, we also present the analyses using Spearman correlation, a nonparametric measure, to construct the alluvial diagrams, as shown in Figure S1 in Multimedia Appendix 2.

Results

Analysis of Lexical Categories

Unsupervised word embeddings capture the context of words in the latent space based on their distribution and patterns of co-occurrence [35]. Given the noisy nature of social media data, it becomes difficult to implement a predefined lexicon-based approach with appropriate semantic inclusion. In this paper, we used unsupervised word embeddings trained on our corpus of tweets to find the words most similar to a given set of seed words, hence expanding the vocabulary of a lexical category. Table 3 shows the words belonging to the categories of hesitation and misinformation. The lexical category of hesitation represents words such as “skeptical,” “disillusionment,” “needle-phobic,” “dissonance,” and “consequence,” which demonstrate the uncertainty and doubt regarding vaccines and their effects. Some of the words most similar to “conspiracy” were found to be “implant_microchips” (cosθ=0.844), “qanon_conspiracy” (cosθ=0.820), “tinfoil_hat” (cosθ=0.808), and “echo_chamber” (cosθ=0.806). These terms denote how people link vaccines to unconventional concepts and propaganda.

Table 3. Words belonging to the lexical categories of hesitation and misinformation, representing the vocabulary expanded from the seed words of the respective categories.

Category	Category words
Hesitation	Confusions, trade_off, shortterm_longterm, frustrate, damage, popularize, apprehension, notions, tire, harmful
Misinformation	Frenzy, propaganda, lethal_injection, false_narratives, black_ market, insert_microchips, euthanised, unsafe_untested, non_believers, conspiracy_theory

Change in Trends Before and After Approval

The difference in slopes of the linear trends of the before and after periods for each category demonstrate 2 significant inferences: the magnitude of change and the direction of change. Figure 2A shows the trends for hesitation in India. A significant change in the direction of the slope is evident (z=10.37, P<.001), which depicts a decrease in its strength after the approval. There was a significant increase (z=–7.65, P<.001) in the magnitude of tweets expressing contentment during the vaccination phase in the United States as shown in Figure 2B. The detected change point was found to be lying within the ranges of 6 days (Figure 2A) and 10 days (Figure 2B) of the date of approval.

The percentage of tweets belonging to different categories was analyzed from the sample of tweets before and after the approval of vaccines in each country. Figure 3A shows that faith and contentment were both significantly higher (both P<.001) before the approval of the first vaccine in India on January 01, 2021 [36]. The factors co-occurring with hesitation were analyzed by calculating the percentage of tweets of 5 other categories (Figures 3C and 3D). Our findings suggest that mentions of health effects contributed the most in tweets with a positive hesitation score. Rage and discussions on misinformation became significantly higher (both P<.001) in the vaccination phase in India (Figure 3C), while an opposite trend was observed in the United States after approval on December 10, 2020 (Figure 3D) [37]. Similar analysis for the United Kingdom, Brazil, and Australia is shown in Figure S2 in Multimedia Appendix 2.

Figure 2. Linear variation in the strength of (A) hesitation in India and (B) contentment in the United States. The dotted line represents the date of approval, and the light blue line depicts the detected change point.

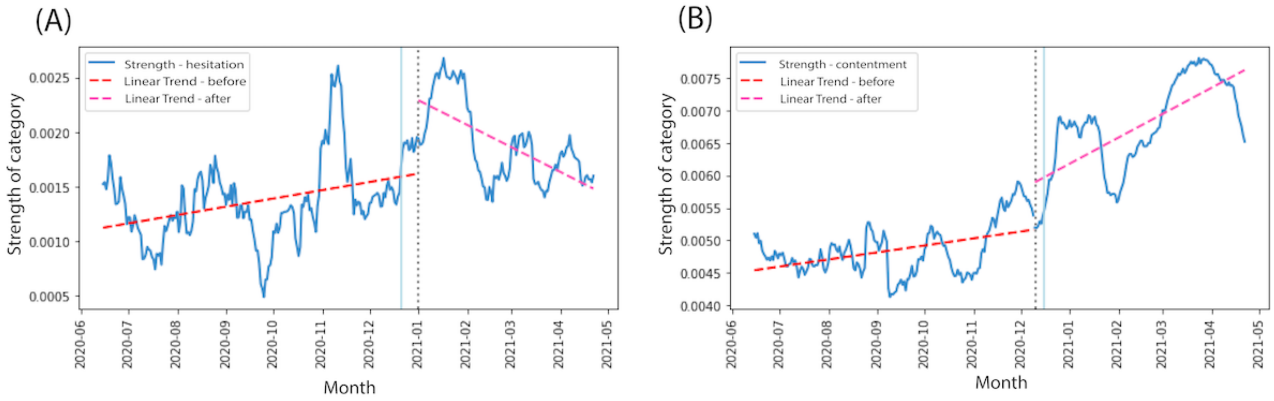
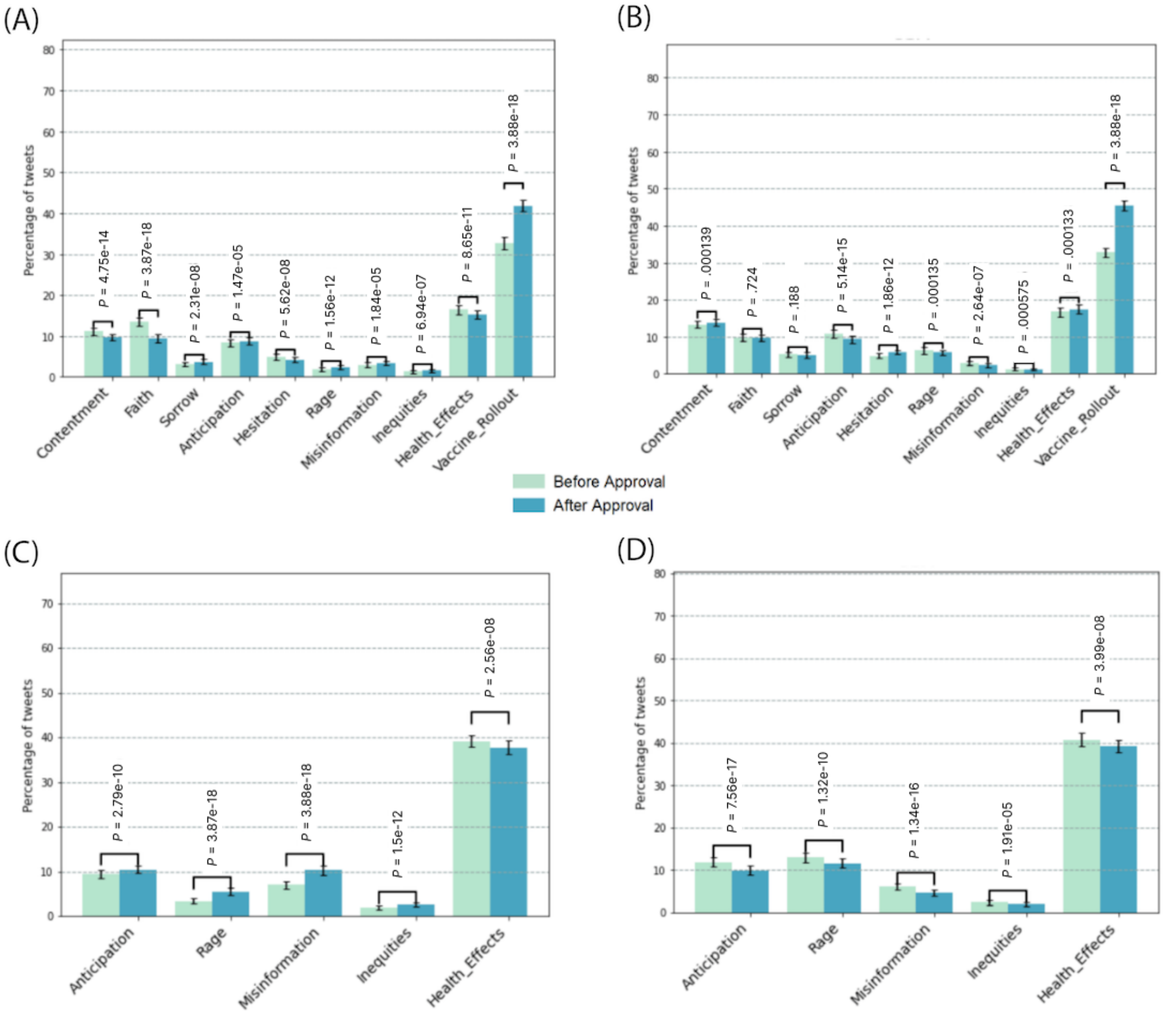


Figure 3. Percentage of tweets with a positive strength in each lexical category before and after approval of COVID-19 vaccine in (A) India (January 1, 2021) and (B) the United States (December 10, 2020) and the percentage of anticipation, rage, misinformation, inequities, and health effects in positive "hesitancy" tweets in (C) India and (D) the United States.



Longitudinal Analysis Using an Alluvial Diagram

Inferences from the alluvial diagrams (Figure 4A) based on Infomap clustering on Pearson correlation networks demonstrated that all the influencing factors (ie, misinformation, health effects, inequities, and vaccine rollout) formed a primary module with emotions of sorrow and rage, which gained the highest PageRank in April 2021, the time when India saw the second wave of COVID-19 cases while the vaccine rollout continued. This articulates the stern sentiment of disappointment due to rising issues and the nonavailability of vaccines for people under the age of 45 years. It also had a high correlation with tweets mentioning the spread of misinformation. Faith, contentment, and anticipation, which were found to be highly associated in the early months of July 2020 and October 2020, were found to be relatively less important and unrelated in April 2021.

On the contrary, lexical categories representing positive sentiment in the United States evolved to a significant module. Faith, contentment, and anticipation toward the vaccine were found to have a positive correlation with each other (Figure 4B). Hesitation was the emotion influenced by mentions of health effects and inequities, whereas rage, sorrow, and misinformation were seen as less central factors in the United States.

Analysis of the temporal trend of misinformation, hesitation, and rage in the 5 countries is depicted in Figure 5. Updates regarding vaccinations started increasing near the end of 2020, which led to changing trends for hesitation expressed on Twitter. A notable inference from the line plots is that hesitation started rising from the beginning of 2021 when primary vaccination drives were initiated. In addition to this, rage is highly expressed in the tweets from the United States, while mentions of misinformation-related terms represented more significant proportions in India and the United Kingdom. Lexical categories of hesitation and rage were found to have similar trends, suggesting a tentative association between the 2 categories.

Figure 4. Alluvial diagram for correlation-based networks showing the evolution of categories from July 2020 to April 2021 at an interval of 3 months in (A) India and (B) the United States.

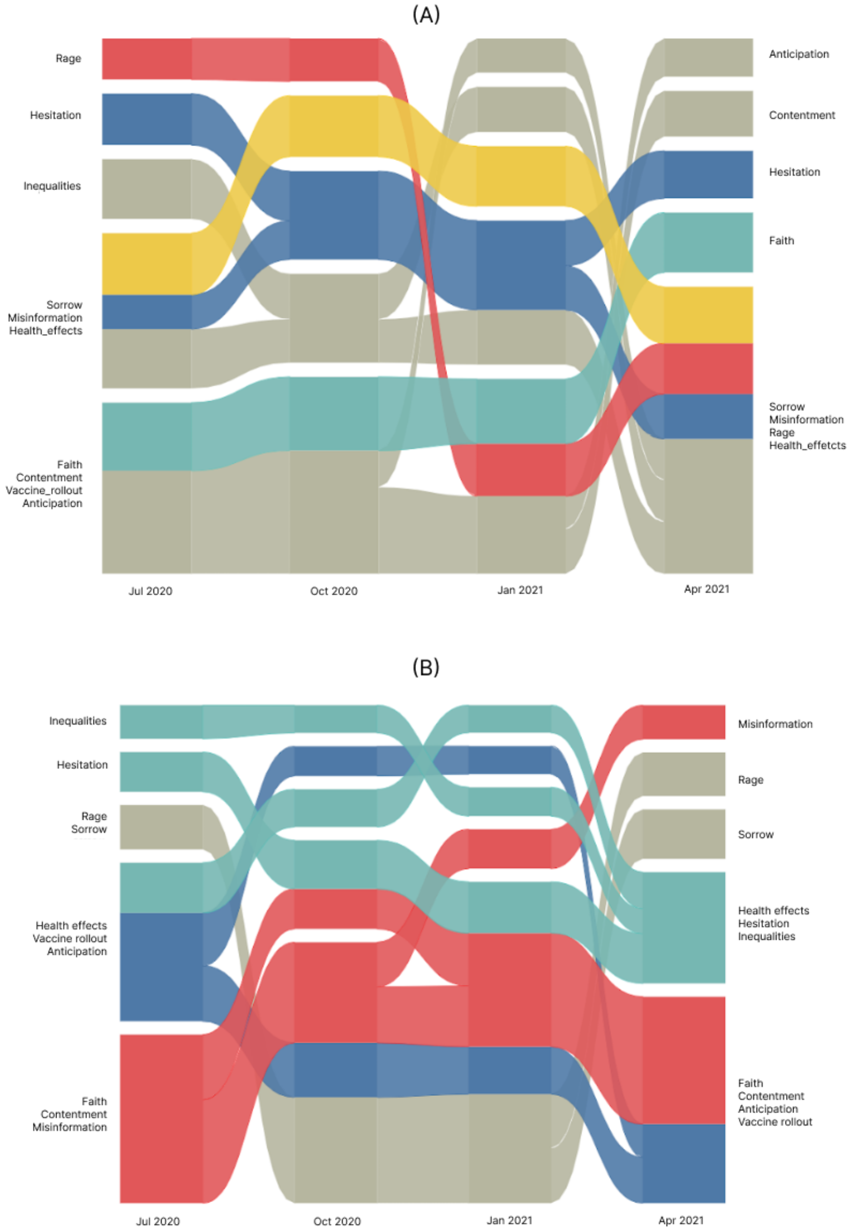
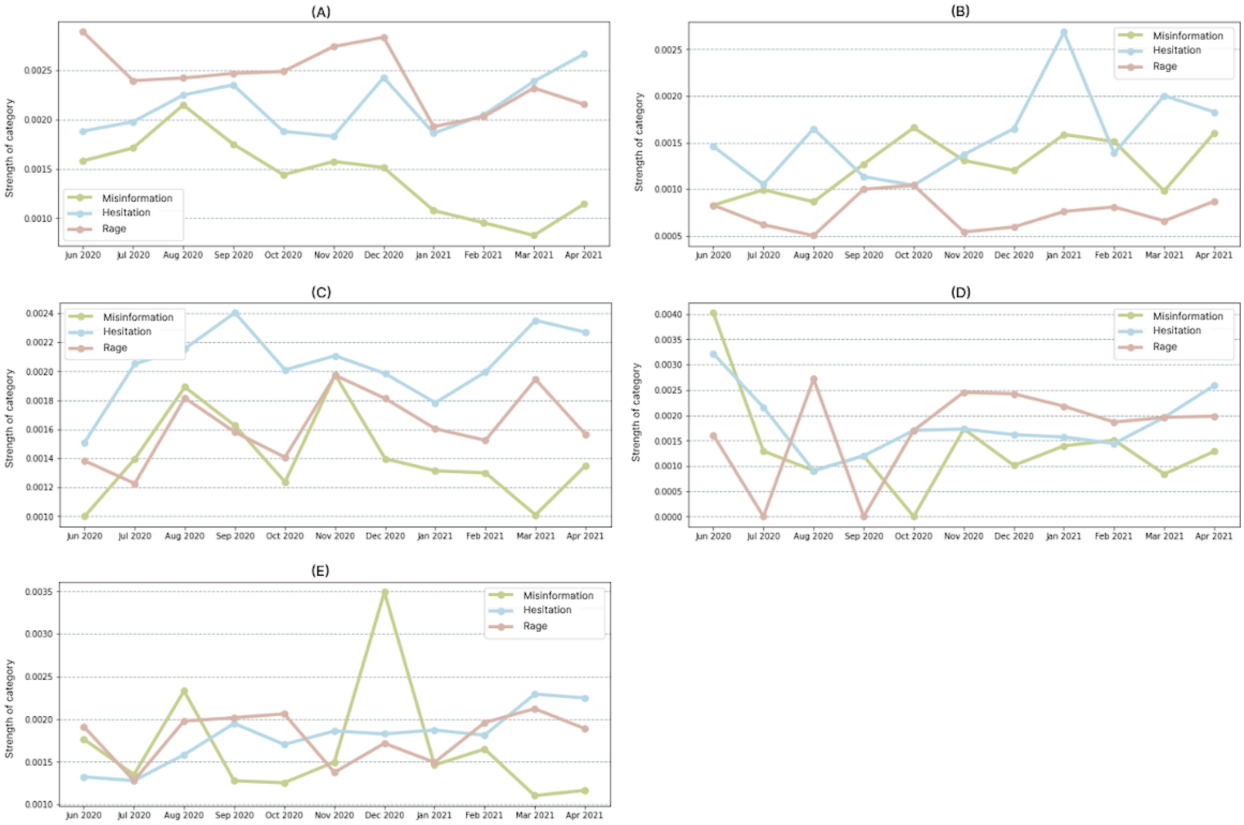


Figure 5. Comparing the temporal flow of strength of 3 categories (misinformation, hesitation, rage) for 5 countries: (A) United States, (B) India, (C) the United Kingdom, (D) Brazil, and (E) Australia.



Discussion

The rise in social media platforms, such as Twitter, has resulted in a valuable source to understand temporal variation in multiple affective and social categories. Influencing factors represented by word embedding–based lexical categories, namely misinformation, vaccine rollout, inequities, and health effects, significantly assisted in studying public perceptions toward emerging vaccine updates from initial approvals to rollout and administration.

Principal Findings

Widespread misinformation being articulated through social media creates panic among users [38]. The misinformation category contains terms similar to “scam” and “conspiracy” from our data set that helped capture references of such words in the context of COVID-19 vaccines. High reporting of adverse effects and severe symptoms in rare cases leading to death [39] becomes a significant factor in increasing vaccination hesitation. The seed words given in the health effects category from the VAERS database led to the formation of its vocabulary containing “restless_sleep,” “skin_sensitivity,” “hot_flash,” “flulike_symptoms,” “complications,” and more. The semantic similarity-based approach allowed customization of categories according to our data set while ensuring the inclusion of rather noisy words like “feverish” and “achiness,” which cannot precisely be found in medical databases.

Inequalities based on socioeconomic status, religion, race, or demographics are standard in different countries, which can lead to inconsistencies while distributing vaccines. The inequities category encapsulated terms related to socioeconomic disparities and helped us identify the impact on other emotions. Based on inspection of our data set of tweets, we found words like “bigotry,” “underprivileged,” “financial_hardship,” and “institutional_racism” were occurring in a highly similar context toward vaccine distribution. Expression of inequities in April 2020 was found to be significantly anticorrelated with faith (*P*=.03) in India. Inaccessibility to vaccines in marginalized groups has led to lower gratification and higher anxiety among these groups [40].

We analyzed tweets from 5 countries belonging to different continents to get the generalized outlook toward vaccines and how they affect the global immunization process. Figure S3 in Multimedia Appendix 2 depicts sorrow, rage, and misinformation during April 2021 in the United Kingdom as the central module, with the highest PageRank. The Medicines and Healthcare products Regulatory Agency of the United Kingdom issued a new advisory during that period, concluding a possible link between AstraZeneca’s COVID-19 vaccine and extremely rare, unlikely occurrences of blood clots [41]. Upon a high-level investigation of the tweets from this period in the United Kingdom, we noticed that this press release had prompted multiple users to talk about blood clots due to the AstraZeneca vaccine. This could have been a potential contributing factor to the high strength of negative emotions expressed on social media platforms. Figure S4 in Multimedia Appendix 2 shows the alluvial diagram for Brazil. The category of rage, which was a relatively less important and independent module in the early months, had associations with sorrow and misinformation in April 2021 in Brazil. It aligned with a major peak in the numbers of cases and deaths during that period of the pandemic in Brazil [42]. In Figure S5 in Multimedia Appendix 2, we can see that faith, contentment, and vaccine rollout were relatively lower than other categories during July 2020, but later in April 2021, they formed a module with anticipation and gained the highest relative importance in the alluvial diagram. The announcement by the Australian government of securing an additional 20 million doses of the Pfizer-BioNTech COVID-19 vaccines overnight [43] happened in April 2021, and multiple tweets expressing optimism possibly contributed to the observed trend. Australia entered into 4 separate agreements with Pfizer, AstraZeneca, Novavax, and COVAX for the supply of COVID-19 vaccines, which resulted in a total number of approximately 170 million vaccine doses, as announced by the Prime Minister.

Related Work

Existing literature on understanding vaccine hesitancy primarily focuses on defined questions from a part of the population belonging to a specific country [44-46]. Although such studies using surveys can help understand the explicit reasoning provided by the individuals, they still pose a limitation on inculcating the variation in outlooks of a larger population over a long period of time. We aimed to fill these gaps by studying important events, such as vaccine trials, highest reported deaths, or import and export of new vaccines, that fueled different populations' emotions, as social media platforms are highly influential due to their comprehensive access and popularity. Our psychometric analysis considers important time stamps and a broader category of emotions to understand the before-after change and the factors with which they associate.

Identification of psychological processes that distinguish between vaccine-hesitant and receptive groups has been carried out in recent research [47]. This helps broadcast public health advisories on social media platforms by strategically taking into account the user's perspective. Effective public health interventions encouraging the uptake of COVID-19 vaccines have benefitted from psychologically oriented approaches [48,49].

Research around understanding the themes and general sentiments toward vaccination programs by analyzing social media posts has also been conducted [50,51]. Although their work provides an overview of positive, negative, or neutral sentiment around other important global developments affiliated with COVID-19 vaccine trials, our analysis provides intricate granularity in understanding the nature of emotions, temporal trends, and the influencing factors that have the highest correlations. Our pipeline effectively clusters the emotion categories and influencing factors around important time stamps based on vaccine approval with categories ranging from negative emotions like hesitation, rage, and sorrow to positive categories like contentment and faith. We further provide a framework to establish lexical categories for understanding the influencing factor correlation and its strength across crucial events. Identification of conspiracy theories related to COVID-19 vaccines has also been carried out [52], which can further be leveraged in addition to our work for improving the understanding of the underlying dynamics of social media posts and disrupting the spread of such content for improving vaccine uptake and tackling hesitancy.

Limitations

Our study has some limitations. We extracted the tweets based on an empirical search of keywords and hashtags relevant to our study in "OR" combination with names of vaccines in the respective countries. Although this approach casts a wide net to retrieve tweets representing discourse around these vaccines, it does not guarantee that all posts were related to COVID-19 vaccine conversations specifically. The chosen keywords for the queries also might not include all relevant terms for capturing tweets specific to our objective. Our framework scores the emotions and influencing factors based on a normalized word count criteria and may miss nuanced language such as sarcasm. However, we interpreted our scores as the amount of discussion happening related to that category, such as hesitancy. Further, the selected categories for our framework are commonly identified emotions that indicate people's perception toward vaccines. Our framework is designed to capture new categories and can be easily expanded and updated periodically to include relevant factors and emotion categories guided by contemporary patterns. Finally, a limitation of our study pertains to the representation bias inherent to social media–based analytics. However, considering that misinformation spreads the fastest through social media and we are considering trends, instead of absolute values, the results are expected to be fairly reliable. Future work may include segmentation of the trends by user demographics, and this information can help in developing tailored solutions for promoting inclusion of minority communities in campaigns. Vaccination drives and policies are targeted heavily toward older populations and minority groups that might not be an active part of such social media platforms. Therefore, for a better understanding of people's opinions toward vaccines, further exploration via other mediums targeting various communities is essential.

Conclusion

Our study provides research and practical implications for public policy making and research on vaccine hesitancy. Our findings offer insights into how the different stages of a pandemic and vaccination process influence emotions and crucial factors like misinformation, health discussions, and socioeconomic disparities on Twitter. This can help decision makers to navigate better solutions in future waves of COVID-19 or similar outbreaks and design appropriate interventions. Our approach can also be utilized to understand the general perception of people during such situations and what preventive measures should be implemented, taking the various influencing factors into account.

Future work can take the direction of local region-level analysis for a specific country to understand the granular emotions within different sections of people and the contributing factors behind them. Providing some weight to the number of reshares and likes the social media post gets can also play an essential role in including the influence the post had in calculating overall strength. Our approach has high adaptability and can be utilized for any online forum, news, or survey data to extract various insights. Designing categories and performing temporal analysis on social media data can also be used to identify multiple ongoing issues like the unavailability of medical resources like oxygen concentrators, intensive care unit beds, and drugs during the second wave of COVID-19. Such analysis can be taken into account while formulating quality allocation of scarce resources based on various factors and their strength. Better information extraction and understanding of such data can be facilitated through our work.

Acknowledgments

TS acknowledges support from the Department of Biotechnology (DBT) Project BT/PR/34245/AI/133/9/2019 and Delhi Knowledge Development Foundation (DKDF) Center of Excellence in Healthcare Project SRP-206 and Delhi Cluster - Delhi Research Implementation and Innovation (DRIIV) Project supported by the Principal Scientific Advisor Office.

All authors declared that they had insufficient or no funding to support open access publication of this manuscript, including from affiliated organizations or institutions, funding agencies, or other organizations. *JMIR Publications* provided APF support for the publication of this article.

Authors' Contributions

HC and AV designed and implemented the computational framework, interpreted the results, and wrote the paper. RP contributed to writing and provided feedback on statistical methods. Authors A and AT scraped the tweets and curated the data set. TS designed the study, analyzed the results, and contributed to writing. All authors read and approved the final paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1

List of software and packages used for our study with their sources and identifiers for the reproducibility of this study.
[DOCX File , 18 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Supplementary figures.
[DOCX File , 1044 KB-Multimedia Appendix 2]

References

1. Timeline: WHO's COVID-19 response. World Health Organization. URL: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/interactive-timeline [accessed 2023-03-18]
2. Greenwood B. The contribution of vaccination to global health: past, present and future. Philos Trans R Soc Lond B Biol Sci 2014 Jun 19;369(1645):20130433 [FREE Full text] [doi: 10.1098/rstb.2013.0433] [Medline: 24821919]
3. Baden LR, El Sahly HM, Essink B, Kotloff K, Frey S, Novak R, et al. Efficacy and safety of the mRNA-1273 SARS-CoV-2 vaccine. N Engl J Med 2021 Feb 04;384(5):403-416. [doi: 10.1056/nejmoa2035389]
4. Pfizer and BioNTech Conclude Phase 3 Study of COVID-19 Vaccine Candidate, Meeting All Primary Efficacy Endpoints. Pfizer. 2020 Nov 18. URL: https://www.pfizer.com/news/press-release/press-release-detail/pfizer-and-biontech-conclude-phase-3-study-covid-19-vaccine [accessed 2023-03-18]
5. Cinelli M, Quattrociocchi W, Galeazzi A, Valensise CM, Brugnoli E, Schmidt AL, et al. The COVID-19 social media infodemic. Sci Rep 2020 Oct 06;10(1):16598 [FREE Full text] [doi: 10.1038/s41598-020-73510-5] [Medline: 33024152]
6. Q1 2020 Letter to Shareholders. Twitter. 2020 Apr 30. URL: http://q4live.s22.clientfiles.s3-website-us-east-1.amazonaws.com/826641620/files/doc_financials/2020/q1/Q1-2020-Shareholder-Letter.pdf [accessed 2023-03-18]
7. Chen W, Stoecker C. Mass media coverage and influenza vaccine uptake. Vaccine 2020 Jan 10;38(2):271-277. [doi: 10.1016/j.vaccine.2019.10.019] [Medline: 31699506]
8. Yoo B, Holland M, Bhattacharya J, Phelps C, Szilagyi P. Effects of mass media coverage on timing and annual receipt of influenza vaccination among Medicare elderly. Health Serv Res 2010 Oct;45(5 Pt 1):1287-1309 [FREE Full text] [doi: 10.1111/j.1475-6773.2010.01127.x] [Medline: 20579128]
9. Piltch-Loeb R, Savoia E, Goldberg B, Hughes B, Verhey T, Kayyem J, et al. Examining the effect of information channel on COVID-19 vaccine acceptance. PLoS One 2021 May 12;16(5):e0251095 [FREE Full text] [doi: 10.1371/journal.pone.0251095] [Medline: 33979370]
10. Rovetta A. The impact of COVID-19 on conspiracy hypotheses and risk perception in Italy: infodemiological survey study using Google Trends. JMIR Infodemiology 2021 Aug 6;1(1):e29929 [FREE Full text] [doi: 10.2196/29929] [Medline: 34447925]
11. Tagliabue F, Galassi L, Mariani P. The "Pandemic" of disinformation in COVID-19. SN Compr Clin Med 2020;2(9):1287-1289 [FREE Full text] [doi: 10.1007/s42399-020-00439-1] [Medline: 32838179]

12. Dhanashree, Garg H, Chauhan A, Bhatia M, Sethi G, Chauhan G. Role of mass media and it's impact on general public during coronavirus disease 2019 pandemic in North India: An online assessment. Indian Journal of Medical Sciences 2021 May 29;73(1):21-25. [doi: 10.25259/IJMS_312_2020]
13. Tsao S, Chen H, Tisseverasinghe T, Yang Y, Li L, Butt ZA. What social media told us in the time of COVID-19: a scoping review. The Lancet Digital Health 2021 Mar;3(3):e175-e194. [doi: 10.1016/s2589-7500(20)30315-0]
14. Arora A, Chakraborty P, Bhatia MPS, Mittal P. Role of emotion in excessive use of Twitter during COVID-19 imposed lockdown in India. J Technol Behav Sci 2021 Oct 20;6(2):370-377 [FREE Full text] [doi: 10.1007/s41347-020-00174-3] [Medline: 33102690]
15. Aggrawal P, Jolly BLK, Gulati A, Sethi A, Kumaraguru P, Sethi T. Psychometric analysis and coupling of emotions between state bulletins and Twitter in India during COVID-19 infodemic. Front. Commun 2021 Sep 24;6:2005 [FREE Full text] [doi: 10.3389/fcomm.2021.695913]
16. Kim EH, Jeong YK, Kim Y, Kang KY, Song M. Topic-based content and sentiment analysis of Ebola virus on Twitter and in the news. Journal of Information Science 2016 Jul 11;42(6):763-781. [doi: 10.1177/0165551515608733]
17. Nagarajan K, Muniyandi M, Palani B, Sellappan S. Social network analysis methods for exploring SARS-CoV-2 contact tracing data. BMC Med Res Methodol 2020 Sep 17;20(1):233 [FREE Full text] [doi: 10.1186/s12874-020-01119-3] [Medline: 32942988]
18. Surian D, Nguyen DQ, Kennedy G, Johnson M, Coiera E, Dunn AG. Characterizing Twitter discussions about HPV vaccines using topic modeling and community detection. J Med Internet Res 2016 Aug 29;18(8):e232 [FREE Full text] [doi: 10.2196/jmir.6045] [Medline: 27573910]
19. Rodrigues CMC, Plotkin SA. Impact of vaccines; health, economic and social perspectives. Front Microbiol 2020 Jul 14;11:1526 [FREE Full text] [doi: 10.3389/fmicb.2020.01526] [Medline: 32760367]
20. Loomba S, de Figueiredo A, Piatek SJ, de Graaf K, Larson HJ. Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA. Nat Hum Behav 2021 Mar 05;5(3):337-348. [doi: 10.1038/s41562-021-01056-1] [Medline: 33547453]
21. Roozenbeek J, Schneider CR, Dryhurst S, Kerr J, Freeman ALJ, Recchia G, et al. Susceptibility to misinformation about COVID-19 around the world. R Soc Open Sci 2020 Oct 14;7(10):201199 [FREE Full text] [doi: 10.1098/rsos.201199] [Medline: 33204475]
22. Kantamneni N. The impact of the COVID-19 pandemic on marginalized populations in the United States: A research agenda. J Vocat Behav 2020 Jun;119:103439 [FREE Full text] [doi: 10.1016/j.jvb.2020.103439] [Medline: 32390658]
23. JustAnotherArchivist / snscrape. GitHub. URL: https://github.com/JustAnotherArchivist/snscrape [accessed 2023-03-18]
24. Bird S, Loper E. NLTK: The Natural Language Toolkit. Proceedings of the ACL Interactive Poster and Demonstration Sessions 2004:214-217 [FREE Full text] [doi: 10.3115/1219044.1219075]
25. Fast E, Chen B, Bernstein MS. Empath: Understanding Topic Signals in Large-Scale Text. 2016 Presented at: CHI Conference on Human Factors in Computing Systems; May 7, 2016; San Jose, CA. [doi: 10.1145/2858036.2858535]
26. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed Representations of Words and Phrases and their Compositionality. 2013 Presented at: Advances in Neural Information Processing Systems 26 (NIPS 2013); December 5-10, 2013; Lake Tahoe, NV URL: https://arxiv.org/abs/1310.4546
27. Ma L, Zhang Y. Using Word2Vec to process big text data. 2015 Presented at: IEEE International Conference on Big Data (Big Data); October 29-November 1, 2015; Santa Clara, CA. [doi: 10.1109/bigdata.2015.7364114]
28. VAERS Data Sets. Vaccine Adverse Event Reporting System. URL: https://vaers.hhs.gov/data/datasets.html [accessed 2023-03-18]
29. Tausczik YR, Pennebaker JW. The psychological meaning of words: LIWC and computerized text analysis methods. Journal of Language and Social Psychology 2009 Dec 08;29(1):24-54. [doi: 10.1177/0261927X09351676]
30. Paternoster R, Brame R, Mazerolle P, Piquero A. Using the correct statistical test for the equality of regression coefficients. Criminology 1998 Nov;36(4):859-866. [doi: 10.1111/j.1745-9125.1998.tb01268.x]
31. Truong C, Oudre L, Vayatis N. Selective review of offline change point detection methods. Signal Processing 2020 Feb;167:107299. [doi: 10.1016/j.sigpro.2019.107299]
32. Benesty J, Chen J, Huang Y, Cohen I. Pearson Correlation Coefficient. In: Noise Reduction in Speech Processing. Springer Topics in Signal Processing, vol 2. Berlin, Heidelberg: Springer; 2009:1-4.
33. Bohlin L, Edler D, Lancichinetti A, Rosvall M. Community Detection and Visualization of Networks with the Map Equation Framework. In: Ding Y, Rousseau R, Wolfram D, editors. Measuring Scholarly Impact. Cham, Switzerland: Springer; 2014:3-34.
34. Rosvall M, Bergstrom CT. Mapping change in large networks. PLoS One 2010 Jan 27;5(1):e8694 [FREE Full text] [doi: 10.1371/journal.pone.0008694] [Medline: 20111700]
35. Tshitoyan V, Dagdelen J, Weston L, Dunn A, Rong Z, Kononova O, et al. Unsupervised word embeddings capture latent knowledge from materials science literature. Nature 2019 Jul 3;571(7763):95-98 [FREE Full text] [doi: 10.1038/s41586-019-1335-8] [Medline: 31270483]
36. Coronavirus: India approves vaccines from Bharat Biotech and Oxford/AstraZeneca. BBC News. 2021 Jan 03. URL: https://www.bbc.com/news/world-asia-india-55520658 [accessed 2023-03-18]

37. Thomas K, Weiland N, LaFraniere S. F.D.A. Advisory Panel Gives Green Light to Pfizer Vaccine. The New York Times. 2020 Dec 17. URL: https://www.nytimes.com/2020/12/10/health/covid-vaccine-pfizer-fda.html [accessed 2023-03-18]
38. Burki T. Vaccine misinformation and social media. The Lancet Digital Health 2019 Oct;1(6):e258-e259. [doi: 10.1016/S2589-7500(19)30136-0]
39. Torjesen I. Covid-19: Norway investigates 23 deaths in frail elderly patients after vaccination. BMJ 2021 Jan 15;372:n149. [doi: 10.1136/bmj.n149] [Medline: 33451975]
40. Larson HJ, Cooper LZ, Eskola J, Katz SL, Ratzan S. Addressing the vaccine confidence gap. The Lancet 2011 Aug;378(9790):526-535. [doi: 10.1016/s0140-6736(11)60678-8]
41. Medicines and Healthcare products Regulatory Agency. MHRA issues new advice, concluding a possible link between COVID-19 Vaccine AstraZeneca and extremely rare, unlikely to occur blood clots. GOV.UK. 2021 Apr 07. URL: https://tinyurl.com/4d2ykmyv [accessed 2023-03-18]
42. Roser M, Ritchie H. Coronavirus Pandemic (COVID-19). Our World in Data. URL: https://ourworldindata.org/coronavirus [accessed 2023-03-18]
43. Press Conference - Australian Parliament House. Prime Minister of Australia [Internet]. URL: https://www.pm.gov.au/media/press-conference-australian-parliament-house-act-09april21 [accessed 2021-08-03]
44. Bendau A, Plag J, Petzold MB, Ströhle A. COVID-19 vaccine hesitancy and related fears and anxiety. Int Immunopharmacol 2021 Aug;97:107724 [FREE Full text] [doi: 10.1016/j.intimp.2021.107724] [Medline: 33951558]
45. Larson HJ, Jarrett C, Eckersberger E, Smith DM, Paterson P. Understanding vaccine hesitancy around vaccines and vaccination from a global perspective: a systematic review of published literature, 2007-2012. Vaccine 2014 Apr 17;32(19):2150-2159 [FREE Full text] [doi: 10.1016/j.vaccine.2014.01.081] [Medline: 24598724]
46. Marti M, de Cola M, MacDonald NE, Dumolard L, Duclos P. Assessments of global drivers of vaccine hesitancy in 2014-Looking beyond safety concerns. PLoS One 2017;12(3):e0172310 [FREE Full text] [doi: 10.1371/journal.pone.0172310] [Medline: 28249006]
47. Murphy J, Vallières F, Bentall RP, Shevlin M, McBride O, Hartman TK, et al. Psychological characteristics associated with COVID-19 vaccine hesitancy and resistance in Ireland and the United Kingdom. Nat Commun 2021 Jan 04;12(1):29 [FREE Full text] [doi: 10.1038/s41467-020-20226-9] [Medline: 33397962]
48. Cameron L, Leventhal H. The self-regulation of health and illness behaviour. New York, NY: Routledge; 2003.
49. Salovey P, Williams-Piehota P. Field experiments in social psychology. American Behavioral Scientist 2016 Jul 27;47(5):488-505. [doi: 10.1177/0002764203259293]
50. 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]
51. Muric G, Wu Y, Ferrara E. COVID-19 vaccine hesitancy on social media: building a public Twitter data set of antivaccine content, vaccine misinformation, and conspiracies. JMIR Public Health Surveill 2021 Nov 17;7(11):e30642 [FREE Full text] [doi: 10.2196/30642] [Medline: 34653016]
52. Shahsavari S, Holur P, Wang T, Tangherlini TR, Roychowdhury V. Conspiracy in the time of corona: automatic detection of emerging COVID-19 conspiracy theories in social media and the news. J Comput Soc Sci 2020;3(2):279-317 [FREE Full text] [doi: 10.1007/s42001-020-00086-5] [Medline: 33134595]

Abbreviations

LIWC: Linguistic Inquiry and Word Count
NLTK: Natural Language Toolkit
VAERS: Vaccine Adverse Event Reporting System

Edited by T Mackey; submitted 17.10.21; peer-reviewed by A Rovetta, K Dashtipour; comments to author 11.01.22; revised version received 09.03.22; accepted 31.03.22; published 02.05.23

Please cite as:
Chopra H, Vashishtha A, Pal R, Ashima, Tyagi A, Sethi T
Mining Trends of COVID-19 Vaccine Beliefs on Twitter With Lexical Embeddings: Longitudinal Observational Study
JMIR Infodemiology 2023;3:e34315
URL: https://infodemiology.jmir.org/2023/1/e34315
doi: 10.2196/34315
PMID: 37192952