## Letter

# Analysis of Evolutionary Social Media Activities: Pre-Vaccine and Post-Vaccine Emergency Use

Haoyue Liu, *Member, IEEE*, MengChu Zhou, *Fellow, IEEE*, Xiaoyu Lu, *Member, IEEE*, Abdullah Abusorrah, *Senior Member, IEEE*, and Yusuf Al-Turki, *Senior Member, IEEE*

Dear Editor,

In this letter, we analyze the public discourse sentiments over time and seek to understand the salient patterns around COVID-19 vaccines and vaccination from social media data. Globally, more than 373 million people have been diagnosed with COVID-19 and 5.66 million have died from this disease by 2022. It continues to have a negative impact on human daily life and the global economic development till now, due to the lack of effective treatment of COVID-19 induced issues and prevention of transmission methods.

Research related to COVID-19 is also proceeding at a rapid pace. Ferrag *et al.* [1] provide a discussion of potential solutions to the security and privacy challenges faced in using IoT applications to combat COVID-19. Ohata *et al.* [2] use a transfer learning method to extract features from x-ray images to automatically detect the COVID-19 infection. Developing and deploying an effective vaccine to slow down the spread of the outbreak becomes a principal task for researchers and pharmaceutical companies [3]. In 2020, multiple vaccines have received Emergency Use Authorization (EUA) from the US Foods and Drug Administration (FDA). On December 11, the FDA issued the first EUA for the Pfizer-BioNTech. A week later, Moderna COVID-19 vaccine received the second FDA-approved EDA. On February 27, 2021, the Johnson & Johnson received vaccine EUA as the third one. All the above vaccine-related information has made people eager to discuss vaccine-related issues. At the same time, massive amounts of information quickly spread on the Internet. Therefore, it is crucial for officials, policy makers and governments to correctly understand the attitudes and concerns of people such that proper policies can be proposed in a more targeted and efficient way and adopted to benefit human combat against this pandemic.

Nowadays, social media platforms, such as Twitter, Facebook and Instagram, provide a faster way that helps people to receive and disseminate information and express personal opinions in real-time. In this work, more than one year of COVID-19 vaccine related twitter data from selected areas are crawled and discussed in detail. Except analyzing the keywords or special terms in COVID-19 vaccine related tweets, obtaining human sentiment information from them is even more important. Automatic sentiment analysis is a way to study various aspects such as peoples' emotions, attitudes, and opinions from the target text without the help of humans. It is not enough to simply observe public attitudes towards vaccines through social networks. More hidden patterns among these tweets need to be explored in detail. Because some tweets discuss topics that are not related to their hashtags, and some tweets do not contain hashtags, topic modeling is one method to solve this problem, and to summarize relevant topics among millions of daily tweets.

This letter presents a sentiment analysis method to classify a more than one year public discourse about COVID-19 vaccines on Twitter and then reveals the trend of sentiment over time. It also presents a topic modeling method on tweets to investigate the salient topics for the COVID-19 vaccine in the USA and to discover the trends of different topics before and after vaccination periods. Ultimately based on the findings from the COVID-19 case study's trending topics and sentiment analysis around pre-vaccine and post-vaccine tweets, we can improve the readability of confusing messages about vaccines on social media and provide effective results to support government agencies and policymakers.

**Related work:** Some studies have used tweets to analyze different concerns about the COVID-19 vaccine. DeVerna *et al*. [4] present a collection of English-language Twitter. It focuses on discovering the prevalence of low-credibility vaccine-related information when it is disseminated over social media. A visualization of topic groups of hashtags is given in their experiments. Gallotti *et al*. [5] noticed that 'infodemics' spread rapidly and widely during the pandemic. This information may increase social panic and misleads the public. Jennings *et al*. [6] proved new evidence on how trust and information were linked with COVID-19 vaccine hesitancy and information policy by using social media posts from November 1, 2020 to December 31, 2020. Sharma *et al*. [7] investigate the characteristics of misinformation and conspiracy activities. They have identified some accounts on Twitter that promote conspiracy groups in vaccine-related discussions. Garcia and Berton [8] apply topic identification and sentiment analysis in both Brazil and the USA. Ten topics are investigated based on four-month tweets. A sentiment trends and their relation to announce news over this period are provided. Lyu *et al*. [9] analyze tweets posted from March 11, 2020 to January 31, 2021. They find that people have a higher acceptance of COVID-19 vaccines than some historical vaccines, e.g., influenza vaccine. Except the analysis of social media data from Twitter, Wu *et al*. [10] choose to study different population groups under Reddit. There have been many discussions on the trend of different topics over time, but most of discussion around the period of FDA issued the EUA, little attention has been paid to the change of topic trend over a period longer than one year from the early stage of vaccine development to the post-vaccine emergency use.

**Data collection and pre-processing:** This work utilizes the Panacea Laboratory database [11]. Because this letter focuses on the period before and after FDA issued EUA for COVID-19 vaccine. The first EUA for COVID-19 vaccine was issued in December 2020. The third one was issued in Feb 2021. Therefore, tweets posted from March, 2020 to May, 2021 are selected in this study. It collects COVID-related tweets by using such keywords as COVID19, coronavirus pandemic, coronavirus, coronavirus pandemic, 2019ncov, and corona outbreak [12].

We focus on English tweets in the USA. Because the twitter user' specific privacy settings policy for geo-reference information, only a limited number of tweets containing geotag, which is not conductive to our analysis of public attitudes and easily causes the sampling bias issue. We assume that users' profile address is the address posting their tweets when their geo-location function is closed. Consequently, we obtain total 118 150 423 COVID-19 related tweets. Vaccine-related tweets are filtered from previous COVID-19 tweets by using the following selected keywords: vaccination, vaccine, immunization, vaccinate, pfizer, biotech, astrazeneca, moderna, J & J, and

Corresponding author: MengChu Zhou.

Citation: H. Y. Liu, M. C. Zhou, X. Y. Lu, A. Abusorrah, and Y. Al-Turki, "Analysis of evolutionary social media activities: Pre-vaccine and post-vaccine emergency use," *IEEE/CAA J. Autom. Sinica*, vol. 10, no. 4, pp. 1090–1092, Apr. 2023.

H. Y. Liu is with the ECE Department, Northern Illinois University, Dekalb 60115 USA (e-mail: hl394@njit.edu).

M. C. Zhou is with the ECE Department, New Jersey Institute of Technology, Newark 07102 USA (e-mail: zhou@njit.edu).

X. Y. Lu is with the School of Cyber Science and Engineering, Nanjing University of Science and Technology, Nanjing 210014, China (e-mail: xiaoyu.lu@njust.edu.cn).

A. Abusorrah and Y. Al-Turki are with the ECE Department, King Abdulaziz University, Jeddah 21589, Saudi Arabia (e-mail: aabusorrah@hotmail.com; yaturki@yahoo.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JAS.2023.123156

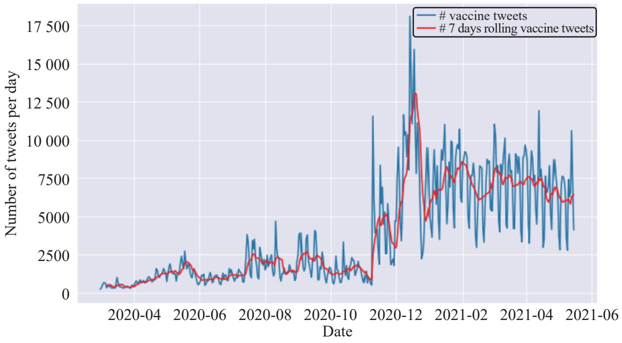

Fig. 1. Daily tweets count from March 2020 to May 2021.

Johnson & Johnson. Finally, the total number of 1 647 479 COVID-19 vaccines related tweets that contain at least one of the previous keywords are obtained. The graph of daily vaccine related count is shown in Fig. 1. To reduce the noise in the data, we select a 7-day rolling window to smooth the data. We observe that during the initial period of vaccine development, relatively low attention was paid to it. The first peak occurred in November 2020, when FDA issued the first EUA for Pfizer-BioNTech vaccine. Then as vaccination begin and other vaccines were successfully issued, there was an explosion of discussion about vaccine-related events. In 2021, the level of discussion about the vaccines became stabilized.

Some pre-processing need to be done to clean the data before proceeding to subsequent text analysis. First, we remove hashtags, URLs, mentions, and lowercase all text. Then NLTK toolkit [13] is used to tokenize the text and remove the stop words. Finally, we use Spacy toolkit to lemmatize the remaining words.

**Sentiment analysis:** Sentiment analysis is a way to investigate people's opinion (positive, neutral, and negative) toward text. This work proposes to adopt the Valence Aware Dictionary and sEntiment Reasoner (VADER) [14] to analyze the sentiment of the tweets. VADER is a lexicon and rule-based sentiment analysis tool, which is designed for social media. In this analysis tool, compound score values are used to classify the tweets into three different polarities. A score closer to 1 indicates a more positive opinion view of the text. The most negative opinion is indicated by −1. A tweet's sentiment is positive if its compound score is between 0.05 and 1, neutral between −0.05 and 0.05, and negative between −1 and −0.05.

**Topic modeling:** This work proposes to use latent Dirichlet allocation (LDA) [15], which is a generative probabilistic topic model, to identify the salient topics from tweets. The data-driven and computational natural of LDA makes it attractive for this study, as it can be used to quickly and efficiently derive the topic structure of a large number of text files. The basic idea is that each topic is characterized by a distribution over words, and each document is represented by a distribution over latent topics. In order to work with the LDA model in Fig. 2, we first use a one-hot encoder to vectorize the tokenized tweets, and then follow the process for generate $M$ tweets as follows:

1) For each topic $k$, sample a distribution over words: $\beta_k \sim \mathrm{Dirichlet}(\lambda)$;

2) For each document $d$ with $N$ words:

a) Sample a distribution over topics: $\theta_n \sim Dirichlet(\alpha)$;

b) For each word $w_{d,n}$:

Sample a topic $z_{d,n} \sim \mathrm{Multinomial}(\theta_d)$ and sample a word $w_{d,n} \sim \mathrm{Multinomial}(\varnothing_{z_{d,n}})$.

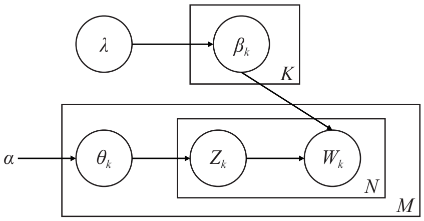

Fig. 2. LDA represented as a graphical model.

Each tweet contains $N$ words, $\beta$ depicts a probability matrix representing the probability of a word being in each topic, whereas $\alpha$ is a vector representing the probabilities of a given topic being represent in a given tweet. LDA is an unsupervised method, meaning that the number of topics in the corpus is unknown. A low number of topics produce a general classification result of topics, which not conductive to deeper exploration of the underlying topic. Many topics indicate that the classification results are too fine to be summarized, thus requiring further topic aggregation. This work proposes to apply the Python package Gensim [16] to evaluate the optimal number of topics in the LDA model. We test the number of topics ranging from 2 to 40. When it is 11, the highest coherence score is observed.

**Experiments:** Our experiments use an Intel Core i7-8700 CPU @ 3.20GHz, NVIDIA GeForce GTX1080 GPU, and Windows 10 64bit. We use Python 3.0, NLTK 3.3, Spacy 2.3, and Gensim 4.1 to realize the experiments. Fig. 3 shows the number of vaccine tweets per day for positive, neutral, and negative, respectively. In the few discussions about vaccines in March and April 2020, only a small percentage of people have positive attitudes towards vaccine development. Between April and November 2020, although the number of positive attitudes exceeded the number of negative ones, there is no clear distinction between them. Until the first vaccine is approved in the beginning of November 2020. The positive attitudes on the Internet substantially increased.

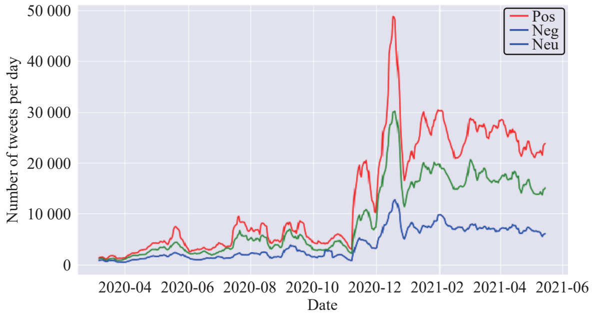

Fig. 3. Daily tweets count from March 2020 to May 2021.

Table 1 presents the topic modeling results. It shows the tweets percentage in each topic and their most 20 probable terms for the given 11 topics. Topic 6 contains the highest tweets percentage among all topics. According to the high-frequency terms in this topic, it focuses on the impact of vaccines on people's daily lives. For example, whether it is still necessary to wear a mask after vaccination and whether it is possible to go back to regular work. The second highest individuals concern is topic 10, which is related to American president election and government.

Fig. 4 shows how the different topics flow in the timeframe of March 2020 to May 2021. The $y$-axis shows the percentage of 11 topics in the same day. By changing the percentage over time, we can observe the flow of these topics and gain insight into what topics people are paying more attention to at different stages. For example, in Topic 6, which has the largest overall weight. It can be observed that when the vaccination is not started. People assumed that their daily life would be like after enrolling in the vaccine. However, as the number of vaccinations increased, people could feel the change in real time and then got used to it, and this discussion gradually decreased. Topic 10 related to the US President and government, which received the most discussion during the election period. After the election day ends, the percentage of this topic begins to decrease. Topic 7 is related to the development of vaccines and their clinical results. As vaccines were approved one after another, people no longer had interest in this topic. On the contrary, people started to pay attention to the topic related to vaccination. As a result, Topic 8 is related to vaccine appointment, registration trails and location. Topic 9 is related to the eligible vaccination groups.

**Conclusions:** This letter identifies the individual opinions about

Table 1. Top 20 Frequencies Terms Per Topic

| Topic | Tweets percentage per topic (%) | Terms per topic |
|---|---|---|
| **Topic 1** | 10.06 | Update, state, information, live, distribution, plan, student, late, provide, learn, check, eligibility, require, join, local, watch, visit, community, today, pharmacy |
| **Topic 2** | 10.03 | Dose, week, receive, say, administer, shot, second, day, next, state, fully, least, official, expect, month, last, available, shoot, year, nearly |
| **Topic 3** | 4.75 | Say, may, question, variant, expert, need, ask, answer, know, concern, doctor, passport, medical, arm, warn, explain, many, require, could, shot |
| **Topic 4** | 9.64 | Risk, spread, may, protect, prevent, still, die, flu, even, infection, immunity, year, cause, could, disease, less, side_effect, protection, effective, know |
| **Topic 5** | 11.64 | Pandemic, help, access, effort, support, community, country, global, get vaccinate, need, work, world, develop, distribution, public, campaign, make, lead, hesitancy, ensure |
| **Topic 6** | 13.78 | Go, mask, good, wear, still, keep, work, let, need, safe, take, feel, make, back, thing, wait, know, family, time, kid |
| **Topic 7** | 7.41 | Use, test, show, trial, effective, datum, emergency, study, approve, say, positive, pfizer, safe, result, clinical, approval, efficacy, early, break, report |
| **Topic 8** | 9.25 | Appointment, today, clinic, worker, schedule, available, resident, county, call, care, receive, staff, offer, register, free, medicine, hospital, open, teacher, school |
| **Topic 9** | 5.38 | Eligible, covid vaccine, age, old, site, open, thank, child, sign, start, group, receive, adult, story, year, today, mass, phase, announce, begin |
| **Topic 10** | 12.67 | Take, would, say, want, trump, make, think, give, try, go, tell, refuse, right, biden, pay, know, lie, government, claim, trust |
| **Topic 11** | 5.38 | Case, death, report, number, rate, fully, state, high, continue, return, low, see, travel, rollout, rise, country, demand, increase, restriction, slow |

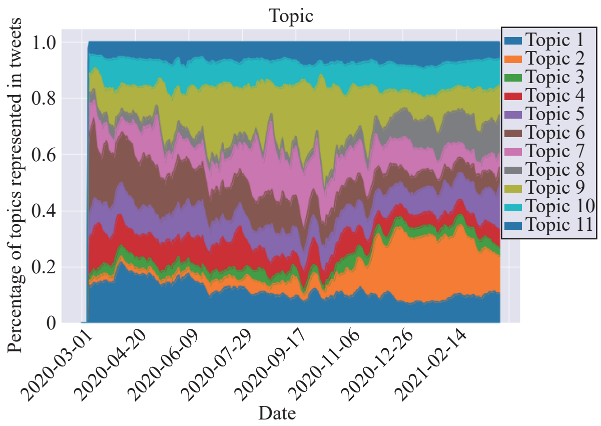

Fig. 4. Topic flow between the March, 2020 and May, 2021.

COVID-19 vaccine and investigates the salient topic by using sentiment analysis [12], [17] and LDA on tweets. The positive and negative polarity change curves demonstrate that initially only a minority of people had positive attitudes toward vaccines. There was a significant increase in overall positive attitudes as vaccine research progressed. We have identified 11 major topics and analyzed their flow in the time range of some specific events. Overall, with the release of multiple vaccines, the focus shifted more to topics related to vaccination. Our next work is to identify the sentiment polarity change under these 11 topics. We should address the challenging issues about how the vaccine related tweets can be selected more accurately with less manual addition or deletion of keywords, and how the major topic can be automatically updated in real-time as timeline extends.

**Acknowlegement:** This work was funded by the Deanship of Scientific Research (DSR) at King Abdulaziz University, Jeddah, Saudi Arabia (GCV19-37-1441).

## References

[1] M. A. Ferrag, *et al*., "Fighting COVID-19 and future pandemics with the internet of things: Security and privacy perspectives," *IEEE/CAA J. Autom. Sinica*, vol. 8, no. 9, pp. 1477–1499, Sept. 2021.

[2] E. F. Ohata, *et al*., "Automatic detection of COVID-19 infection using chest X-ray images through transfer learning," *IEEE/CAA J. Autom. Sinica*, vol. 8, no. 1, pp. 239–248, Jan. 2021.

[3] S. Gottlieb, "America needs to win the coronavirus vaccine race," *Wall Street J.*. Apr. 26, 2020. [Online], Available: https://www.wsj.com/articles/america-needs-to-win-the-coronavirus-vaccine-race-11587924258, Accessed Feb. 4, 2021.

[4] M. DeVerna, *et al*., "CoVaxxy: A global collection of English Twitter posts about COVID-19 vaccines," arXiv preprints: arXiv-2101, 2021.

[5] R. Gallotti, F. Valle, N., P. Sacco, and D. Manlio, "Assessing the risks of 'infodemics' in response to COVID-19 epidemics," *Nature Human Behaviour*, vol. 4, no. 12, pp. 1285–1293, 2020.

[6] W. Jennings, G. Stoker, H. Bunting, V. Valgarðsson, J. Gaskell, D. Devine, L. McKay, and M. C. Mills, "Lack of trust, conspiracy beliefs, and social media use predict COVID-19 vaccine hesitancy," *Vaccines*, vol. 9, no. 6, p. 593, 2021.

[7] K. Sharma, *et al*., "COVID-19 vaccines: Characterizing misinformation campaigns and vaccine hesitancy on Twitter," arXiv preprint arXiv: 2106.08423, 2021.

[8] K. Garcia and L. Berton, "Topic detection and sentiment analysis in Twitter content related to COVID-19 from Brazil and the USA." *Applied Soft Computing*, vol. 101, p. 107057, 2021.

[9] J. Lyu, *et al*., "COVID-19 vaccine–related discussion on Twitter: Topic modeling and sentiment analysis," *J. Medical Internet Research*, vol. 23, no. 6, 2021.

[10] W. Wu, *et al*., "Characterizing discourse about COVID-19 vaccines: A reddit version of the pandemic story," arXiv preprint arXiv: 2101.06321, 2021.

[11] J. M. Banda, *et al*., "A large-scale COVID-19 Twitter chatter dataset for open scientific research–An international collaboration," *Epidemiologia*, vol. 2, no. 3, pp. 315–324, 2021.

[12] H. Liu, "Analyzing fluctuation of topics and public sentiment through social media data," *Thesis*, Newark College of Engineering, New Jersey Institute of Technology, USA, 2022.

[13] S. G. Bird and L. Edward. "NLTK: The natural language toolkit." in *Proc. COLING/ACL Interactive Presentation Sessions: Association Comput. Linguistics*, 2006, pp. 69–72.

[14] C. J Hutto and E. E. Gilbert, "VADER: A parsimonious rule-based model for sentiment analysis of social media text," in *Proc. Eighth Int. Conf. Weblogs and Social Media*, Ann Arbor, USA, June 2014.

[15] H. Jelodar, *et al*., "Latent dirichlet allocation (LDA) and topic Modeling: Models, applications, A survey," *Multimedia Tools and Applications 78*, vol. 78, no. 11, pp. 15169–15211, 2019.

[16] R. Rehurek and P. Sojka, "Gensim–statistical semantics in python," 2011. [Online], Available: https://www.fi.muni.cz/usr/sojka/posters/rehurek-sojka-scipy2011.pdf.

[17] H. Liu, *et al*., "Aspect-based sentiment analysis: A survey of deep learning methods," *IEEE Trans. Computational Social Syst*,, vol. 7, no. 6, pp. 1358–1375, Dec. 2020.