Sentiment Analysis and Topic Modeling on COVID-19 Vaccines using Twitter Data

Daniel Ojeda, Eric Landaverde, Ching-Yu Huang, and Daehan Kwak
Department of Computer Science and Technology, Kean University
Union, NJ USA
{ojedada, landaver, chuang, dkwak}@kean.edu

***Abstract*—Ever since the public began quarantining due to Covid-19 in 2020, people have been waiting for life to go back to normal. Until April 2021, "normal" activities were only allowed by being socially distant or wearing a mask. However, in early 2021, CDC announced that a vaccine would soon be released to the public. This announcement seemed to be good news for some, but for others, this was another obstacle on the way to normalcy. People have shown pushback against the Covid-19 vaccine due to the uncertainty of its effectiveness coupled with its potential side effects. Vaccine hesitancy has a negative impact on society and poses a real threat to public health. This is an important issue worldwide, and questions arise about how the general public feels about getting vaccinated. Therefore, the purpose of this study is to analyze the sentiment of Twitter users towards vaccination, specifically the Covid-19 vaccine. This study collects data through Twitter IDs to pick up on hashtags and keywords relating to the Covid-19 vaccine via the Twitter API and Tweepy. Tweets are put through a sentiment analysis tool to get a general idea of the sentiment. Furthermore, topic modeling is used to understand the topics discussed when mentioning the Covid-19 vaccine. By analyzing the sentiment towards the Covid-19 vaccine, we hope to provide the first step towards mitigating the risk associated with vaccine hesitancy.**

Keywords—Sentiment Analysis, Opinion Mining, Topic Modeling, Covid-19, Vaccine.

I. INTRODUCTION

In late 2019, patients in Wuhan, China, began coming into the hospital with flu-like symptoms. At the time, an official diagnosis had not yet been made or released to the public; however, these are thought to be the first few cases of Covid-19, according to the Centers for Disease Control and Prevention (CDC). After lots of research done by scientists across the globe, the CDC finally released to the public that this virus was thought to be the Coronavirus in January 2020. The first confirmed case of Covid-19 was on January 20, 2020. From that point on, the public began taking precautions such as wearing masks, social distancing, and quarantining [1].

In early 2021, vaccines to combat Covid-19 slowly began being distributed around the United States. Around this time, the public began discussing the vaccine online – sharing their opinions and concerns. Many people worried about the vaccine's potential side effects because it was released soon after the Coronavirus pandemic began. These concerns were taken to the internet, where others could weigh in, and soon there was a significant discourse around the vaccine and whether people should take it or not. Many took to Twitter to discuss the topic.

With so much discourse taking place and so many different opinions being shared, Covid-19 quickly became a gripping topic to be discussed, specifically the vaccine. It is essential to see what kinds of opinions people had and what they were explicitly discussing regarding the vaccine. Thus, this study utilized a dataset with Twitter IDs to better understand what the general public was saying about the Covid-19 vaccine.

Twitter is a platform where people can share short 280-character posts called tweets and post pictures. Twitter was first released in 2006 and is now a prevalent application and website. Twitter has over 229 million daily active users that generate nearly 500 million tweets per day and around 200 billion Tweets per year [2]. With the number of users on Twitter as a group expressing distinct opinions regarding the COVID-19 pandemic, it was considered for sentiment analysis and topic modeling of Covid-19 vaccines.

II. RELATED WORKS

Several studies have analyzed the sentiment surrounding Covid-19 as a whole, and many focus on the Covid-19 vaccine [3]. There have also been studies done that specifically utilize data from Twitter to analyze sentiment toward the Covid-19 vaccine [3-8].

Lyu et al. [5] conducted a study that used Twitter to identify trends in the sentiment toward the Covid-19 vaccine. This study was done at a national level (specifically the United States) and world national level. They collected 2,678,372 COVID-19 vaccine-related tweets between November 1, 2020, and January 31, 2021, and used the Valence Aware Dictionary and sEntiment Reasoner tool to calculate the sentiment of each tweet. The scores indicated a positive, neutral, or negative sentiment for each tweet. They then applied the latent Dirichlet allocation analysis to obtain the topics discussed in both positive and negative sentiment tweets. After this, they performed a temporal analysis to explore the sentiment over time and a geographic analysis to observe the difference or similarities in sentiment in different locations. Their study concluded that overall, positive tweets had the most number compared to negative and neutral with the difference in the number of positive sentiments to negative tweets being 12.5%. They identified five overarching themes for positive sentiment tweets: trial results, administration, life, information, and efficacy. The five themes identified for tweets with negative sentiment were: trial results, conspiracy, trust, effectiveness, and administration. They also found that Brazil had the lowest sentiment score of −0.002, and the United Arab Emirates conversely had the highest sentiment

979-8-3503-2028-2/22/$31.00 ©2022 IEEE
DOI 10.1109/CSCI58124.2022.00218

score of 0.162. This study concluded that the sentiment towards the Covid-19 vaccine varied greatly over time and location.

Sattar et al. [7] conducted a study in which 1.2 million tweets, ranging from April–May 2021, were analyzed to draw conclusions about public sentiments toward vaccination. However, Sattar's study used two tools to gather sentiment analysis– Vader Lexicon and TextBlob. Sattar also compared the difference in the sentiment of raw tweets versus pre-processed. This was an insightful contribution, as many studies only observe tweets after being pre-processed. Their study found that there were far more positive sentiment tweets before being pre-processed than after. Conversely, there were many more neutral sentiment tweets after being pre-processed than before. Sattar's study observed sentiment per specific Covid-19 vaccine (Pfizer, Moderna, Johnson & Johnson, Oxford-AstraZeneca, Sputnik V, and Covaxin) with both Vader and TextBlob and decided to use observations of TextBlob as reference for the rest of the study. Similar to the previous studies mentioned, Sattar's study found that the overall number of tweets with positive sentiment was greater than that of negative sentiment.

Each study mentioned, along with others referenced, greatly contributed to the study of sentiment toward the Covid-19 vaccine. Although these studies are similar to the one conducted in this paper, few studies span the same time period as ours – March 2021 to November 2021, which is a contribution of this study. With this long span of time, we can analyze time trends throughout a more extended period to get an all-encompassing view of how the attitudes and opinions of the public fluctuated.

III. Methodology

A. Hydrating Tweets

The initial step of the research was data collection. In order to obtain the most data possible, a dataset of tweet IDs was used. This dataset is from IEEE and has been collecting tweet IDs in real-time since October 2019 and is still ongoing as of October 2022 [9]. However, for this study, the tweets analyzed ranged from March 2020 to November 2021. Around 1.717 billion tweet IDs were downloaded from the IEEE dataset, which could be used to reference the actual tweet. Because this dataset consists only of tweet IDs, the first step in this study was to obtain the complete tweet information from each ID. This was done by creating a python script and utilizing Tweepy, a package that provides a very convenient way to use the Twitter API. A total of eight months was needed to scrape the data set. The dataset downloaded was divided into 600 CSV files due to its size, and 10 CSV files were run with multiple Google Colab instances simultaneously to get through every single tweet ID. After hydrating all the tweet IDs, the dataset downsized to around 160 million tweets. This was because since some tweet IDs dated back to 2020, the actual tweet had already been deleted, or the account itself had been deleted. Furthermore, the original dataset consisted of tweet IDs that referenced retweets (this is essentially a repost of an existing tweet), and because we filtered out retweets when hydrating to avoid duplicates – many tweets were excluded.

B. Filtering Tweets

Because the tweets in the original dataset pertained to Covid-19 in general, the dataset needed to be filtered to contain only tweets related to the various Covid-19 vaccines. First, a list of keywords was generated to encompass the following vaccines: Pfizer, Moderna, AstraZeneca, Janssen, Covaxin, Novavax, Sputnik, Sinopharm, Sinovac, and Convidecia. Next, a python script was written to process the master CSV file containing all the hydrated tweets and keep only those that mentioned at least one of the vaccines listed above. This filtering resulted in about 22.03 million tweets in the dataset, which pertained to the Covid-19 vaccines.

TABLE I. Keyword replacement using regex patterns. A list of all the possible valid words for each vaccine and correcting spelling errors.

Keyword	Replacements
Sinopharm	Sinopharm, BBIBP-CorV, BBIBP, CorV
Sinovac	Sinovac, CoronaVac
Pfizer	pfizer, BioNTech, fizer
Moderna	moderna, nnmoderna
AstraZeneca	AZ, Astra, Zeneca, Astrazeneca, Oxfordvacc*, Oxfordvax*
Covaxin	BBV152, Bharat, Bharat Biotech, Covaxin
Sputnik	Sputni*
Janssen	Johnson and Johnson, Johnson & Johnson, Jnj, jj, j&j
Novavax	Novava*
Convidecia	CanSinoBiologics, CanSino, CanSinoBio, Covidecia

C. Sentiment Analysis

This section aims to analyze the sentiment toward Covid-19 vaccines through tweets from Twitter. Sentiment analysis is used to get a numerical understanding of the feelings expressed through text. For this study's purposes, the text is a tweet. Sentiment analysis dates back to the 20th-century post World War II when it was used to analyze public opinions [10]. Today, it is used in numerous studies originating all over the world. It can be instrumental in getting a general overview of the general public's opinion and be a precursor for future research. Supervised or unsupervised machine learning can be used for sentiment analysis. The most challenging machine learning model for sentiment analysis is supervised learning due it requires a subset of the data to be labeled to train the model. The labeled data sets are used to feed the model, consisting of an input and the desired output. On the other hand, unsupervised learning has unlabeled data that the model tries to understand on its own. Two of the most popular sentiment analysis Python packages are VADER and TextBlob. VADER was chosen in this research because it focuses on spotting the sentiments in the text that frequently appears on social media, such as emoticons, recurring words, abbreviations, and punctuations. Tweets are put through Valence Aware Dictionary and sEntiment Reasoner (VADER) sentiment analysis to get a general idea of sentiment toward vaccines.

TABLE II. EXAMPLE TWEETS AND CORRESPONDING SENTIMENT SCORES.

Text	Tweet_id	Date	Vader Scores	Keyword
"Moderna, who are making the Corona vaccines, have suspended animal trials and suspended normal protocols to push this vaccine through" What could possibly go wrong? https://t.co/DDRgdcm7WF	1242508440561500000	2020-03-24 17:47:51+00:00	{'neg': 0.288, 'neu': 0.712, 'pos': 0.0, 'compound': -0.8519}	['Vaccine', 'Moderna']
@pfizer @CDCgov just a suggestion sending fake agents imitating a cell which corona virus grabs on.	1243414070604090000	2020-03-27 05:46:30+00:00	{'neg': 0.193, 'neu': 0.807, 'pos': 0.0, 'compound': -0.4767}	['Pfizer']

TABLE III. COVID VACCINE RELEVANT TOPIC FOR EACH MONTH.

Date	Topic	Top 10 keywords associated with the topic.
03-2020	Covid-19 vaccine tested on animals.	see, strain, first, mutate, recover, disease, available, new, covid, animal
04-2020	Accelerate the vaccine's development	prove, say, fairly, distribute, become, novel, widely, succeed, fast, readily
05-2020	Accelerate the vaccine's development	successful, shot, rapidly, insist, jab, deployment, decade, prove, urgency, mass
06-2020	Covid-19 testing	shot, jab, variety, successful, fairly, prove, quantity, decade, urgency, partially
07-2020	Vaccine development	shot, jab, say, version, currently, possibly, drug, potentially, worldwide, guarantee
08-2020	Potential vaccine distribution	possibly, currently, jab, shot, intend, globally, awhile, fully, vaccination, easily
09-2020	Potential vaccine distribution	possibly, shot, year, vaccination, prove, even, mass, fully, vaccinate, medication
10-2020	Potential vaccine distribution	shot, hastily, inoculation, currently, jab, consider, fully, vaccination, simultaneously, possibly
11-2020	Potential vaccine distribution	vaccination, jab, possibly, likely, ultimately, shot, especially, potentially, globally, currently
12-2020	Initial Covid-19 vaccine distribution	shot, say, currently, vaccinate, vaccination, jab, globally, suppose, initially, eventually
01-2021	Covid-19 vaccine distribution	vaccination, say, shot, currently, jab, dose, inoculation, initially, time, vaccinate
02-2021	Covid-19 vaccine distribution	shot, vaccination, jab, say, currently, vaccinate, time, initially, inoculation, covid
03-2021	Belief that initial vaccine doses are effective	vaccination, shot, jab, say, vaccinate, inoculation, currently, effectively, initially, dose
04-2021	Annual Covid boosters	vaccination, shot, jab, say, vaccinate, initially, currently, time, annually, covid
05-2021	Covid boosters	vaccination, jab, shot, vaccinate, currently, time, covid, say, dose, simply
06-2021	Covid-19 vaccines for all people	jab, vaccination, shot, say, vaccinate, covid, inoculation, time, currently, people
07-2021	Need vaccines for current Covid-19 variant	vaccination, shot, jab, vaccinate, currently, especially, time, people, covid, need
08-2021	Vaccines for current Covid-19 variant	vaccination, jab, shot, vaccinate, say, covid, especially, currently, time, people
09-2021	Vaccines for current Covid-19 variant	jab, vaccination, shot, covid, vaccinate, especially, say, currently, people, injection
10-2021	Belief vaccines are effective for current Covid-19 variant	jab, shot, vaccinate, covid, especially, effectively, people, essentially, say, currently
11-2021	Vaccines for current Covid-19 variant	shot, jab, vaccination, vaccinate, covid, say, injection, especially, currently, time

D. Topic Modeling

To understand the underlying topics present in the global discourse, we utilize a natural language processing technique called topic modeling. Topic modeling is an unsupervised machine learning technique used to extract abstract topics from a collection of texts called a corpus by scanning the documents within the corpus, detecting word and phrase patterns, and clustering similar word groups and expressions that best represent a set of documents. For topic modeling, gensim's Word2Vec model is utilized to identify relevant topics within the corpus. Word2Vec functions by detecting contextually similar words mathematically through its two-layer neural network. This is done by having the model learn word embeddings and then determine their similarity to each other by the distance between the embeddings, which is measured by cosine similarity. Our Word2Vec implementation scans the corpus for a keyword phrase such as "vaccine" and determines which words are contextually similar, thus providing us with a list of ten contextually similar words to the keyword phrase for each of the twenty-one months our data set spans over. The model runs on preprocessed tweet text and is optimized to a minimum count of fifty per word to improve the extraction of only semantically relevant words.

E. Visualization

Due to the size of the dataset created for this research, graphs were created to display the data in a way that makes it easier to analyze and develop a deeper understanding of COVID-19 vaccines. Thus, we created a database using MySQL Workbench to host and store VADER scores' data. Then, using markup and scripting languages such as HTML, CSS, JavaScript, and PHP, we were able to create graphs based on the data collected.

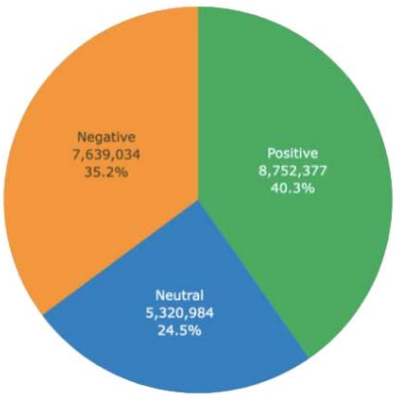

Fig. 1. Sentiment breakdown: 03-2020 to 11-2021.

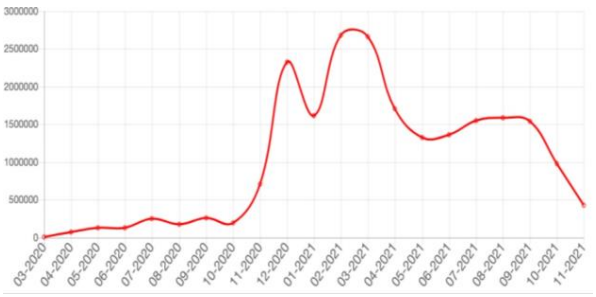

Fig. 2. Frequency of word vaccine per month.

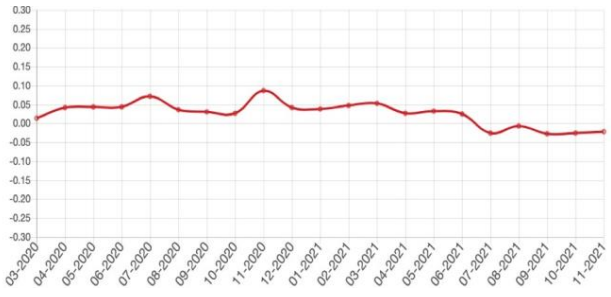

Fig. 3. Compound sentiment distribution for the word vacccine.

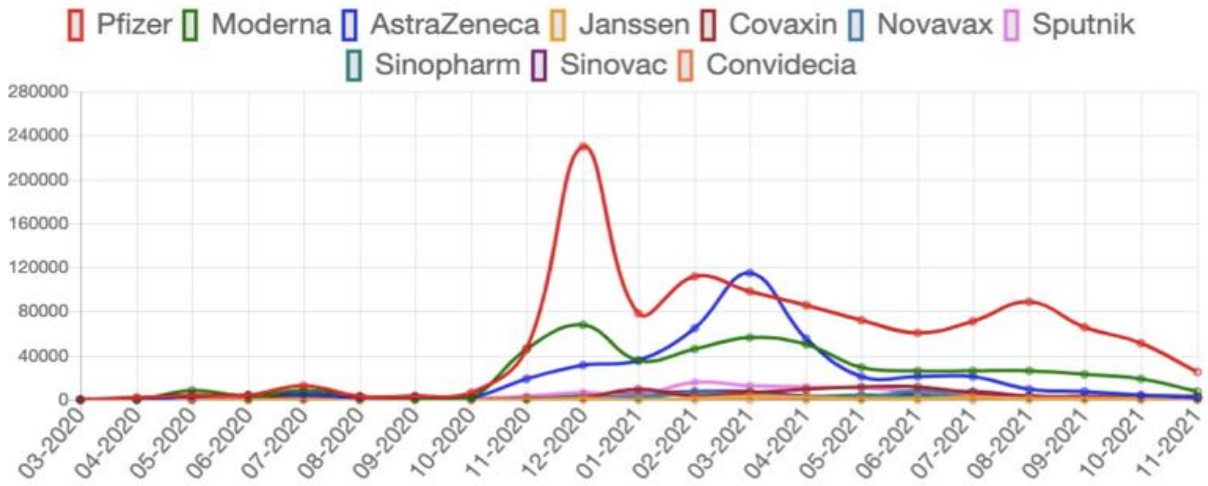

Fig. 4. Frequency of each vaccine per month.

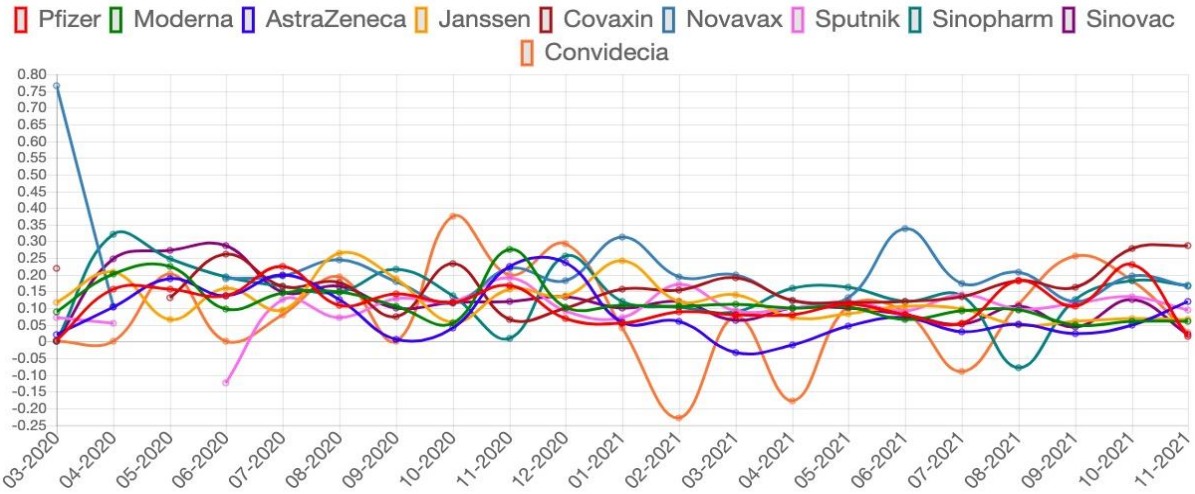

Fig. 5. Compound sentiment distribution for each vaccine per month.

IV. Results

Topic modeling summarizes the most popular topic discussed each month using the gensim's Word2Vec model. A total of 21 topics are summarized and presented in Table 3. According to the polarity analysis, as illustrated in Fig. 1, the public sentiment toward the COVID-19 vaccine is more positive, forming 40.3 % of the dataset. Negative tweets make up 35.2 % of the total tweet, and neutral tweets have the lowest proportion at 24.5%. During March 2020 (the first month we started to collect data), relevant keywords were: "strain," "mutate", "disease", "new", "covid", "animal" as can be seen in Table 3, when different news outlets reported about animals used in Covid-19 vaccine trials. Other notable keywords during the following months (April, May, and June 2020) were "jab," meaning taking the vaccine shot, the keywords "fairly", "distribute", "widely", "mass", pointing to concerns on guaranteeing fair and equitable access of vaccines for everyone. During July and August 2020, relevant keywords were: "potentially", "worldwide", "guarantee", "globally," "shot," "vaccination, "intend" as can be seen in Table 3, when Russia became the first country in the world to approve a possible vaccine against COVID-19 virus around August 2020. Also, during those months, the public sentiment toward COVID-19 was positive and neutral, as shown in Fig. 3. Following Pfizer's announcement that the Pfizer COVID-19 vaccine had achieved 90% efficacy, the number of tweets increased dramatically in November 2020 as shown in Fig. 4 and Topic modeling suggests that keywords related to "globally", "potentially", "jab", and "vaccination" were widely discussed during the same month as can be seen in Table 3. Convidecia was not approved by the WHO for the prevention of COVID-19 induced by SARS-CoV-2 during the study period, which may explain the negative sentiment spikes towards the CanSino CONVIDECIA vaccine as reflected in Fig. 5. Finally, Topic modeling suggests that keywords related to "currently", "especially", "people", "need", "inoculation" were widely discussed during the summer of 2021, when the Delta variant, which was first identified in India, takes over as the predominant variant in the United States.

V. Limitations

There was a limited amount of data we were able to process due to data collection being extremely difficult, which took eight months to collect the data. Furthermore, we were only able to hydrate tweets up until November 2021. Processing was also a challenge because of the dataset size we were able to hydrate (16 gigabytes worth of hydrated tweets data).

VI. Conclusion and Future Work

We conducted a sentiment analysis and topic modeling of 22,027,775 COVID-19 vaccine-related tweets dated from March 2020 to November 2021. Understanding the public sentiment toward COVID-19 vaccines can help public health agencies create appropriate policies. This research only accounts for the reactions from Twitter users to vaccines. Future work needs to consider other social media platforms like Reddit and Facebook to encompass a broad range of demographics. Also, future work using other techniques [11-15] to analyze the Covid-19 data will be explored.

Acknowledgment

This material is based upon work supported in part by the National Science Foundation under Grant No. 1909824 and the Office of Research and Sponsored Programs, Kean University. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

[1] D. J. Sencer, "CDC Museum Covid-19 Timeline," Centers for Disease Control and Prevention, 16-Aug-2022. Available: https://www.cdc.gov/museum/timeline/covid19.html.

[2] C. Beveridge, "33 twitter statistics that matter to marketers in 2022," Social Media Marketing; Management Dashboard, 16-Mar-2022. Available: https://blog.hootsuite.com/twitter-statistics/.

[3] T. Hu, S. Wang, W. Luo, M. Zhang, X. Huang, Y. Yan, R. Liu, K. Ly, V. Kacker, B. She, and Z. Li, "Revealing public opinion towards covid-19 vaccines with Twitter data in the United States: A Spatiotemporal perspective," *Journal of Medical Internet Research,* 23(9), e30854, 2021.

[4] E. Bonnevie, A. Gallegos-Jeffrey, J. Goldbarg, B. Byrd, and J. Smyser, "Quantifying the rise of vaccine opposition on Twitter during the COVID-19 pandemic," *Journal of Communication in Healthcare*, vol. 14, no. 1, pp. 12–19, 2020.

[5] J. C. Lyu, E. L. Han, and G. K. Luli, "Covid-19 vaccine–related discussion on Twitter: Topic Modeling and sentiment analysis," *Journal of Medical Internet Research*, vol. 23, no. 6, 2021.

[6] M. T. Ansari and N. A. Khan, "Worldwide covid-19 vaccines sentiment analysis through Twitter content," *Electronic Journal of General Medicine*, vol. 18, no. 6, 2021.

[7] N. S. Sattar and S. Arifuzzaman, "Covid-19 vaccination awareness and aftermath: Public sentiment analysis on Twitter data and vaccinated population prediction in the USA," *Applied Sciences*, vol. 11, no. 13, p. 6128, 2021.

[8] S. Liu and J. Liu, "Public attitudes toward covid-19 vaccines on English-language twitter: A sentiment analysis," *Vaccine*, vol. 39, no. 39, pp. 5499–5505, 2021.

[9] R. Lamsal, "Design and analysis of a large-scale COVID-19 tweets dataset," *Applied Intelligence*, vol. 51, no. 5, pp. 2790–2804, 2020.

[10] M. V. Mäntylä, D. Graziotin, and M. Kuutila, "The evolution of sentiment analysis—a review of research topics, venues, and top cited papers," *Computer Science Review*, vol. 27, pp. 16–32, 2018.

[11] D. Kwak, D. Kim, R. Liu, L. Iftode, and B. Nath , "Tweeting traffic image reports on the road," In *IEEE 6th International Conference on Mobile Computing, Applications and Services*, pp. 40-48, 2014.

[12] F.Ali, D. Kwak, P.Khan, S.H.A. Ei-Sappagh, S.M.R. Islam, D. Park, and K.S. Kwak, "Merged ontology and SVM-based information extraction and recommendation system for social robots," *IEEE Access*, 5, 12364-12379, 2017.

[13] F. Ali, D. Kwak, P. Khan, S.M.R. Islam, K. H. Kim, K.S. Kwak, "Fuzzy ontology-based sentiment analysis of transportation and city feature reviews for safe traveling," *Transportation Research Part C: Emerging Technologies* 77, 33-48, 2017.

[14] F. Ali, P. Khan, K. Riaz, D. Kwak, T. Abuhmed, D. Park, and K. S. Kwak, "A fuzzy ontology and SVM–based Web content classification system," *IEEE Access*, 5, 25781-25797, 2017.

[15] F. Ali, D. Kwak, P. Khan, S. El-Sappagh, A. Ali, S. Ullah, K.H. Kim, and K. S. Kwak, "Transportation sentiment analysis using word embedding and ontology-based topic modeling," *Knowledge-Based Systems*, 174, 27-42, 2019.