Covid 19 Vaccine Public Opinion Analysis on Twitter Using Naive Bayes

Samar Ibrahim and Sherief Abdallah

May 18, 2023

Covid 19 vaccine public opinion analysis on Twitter using Naive Bayes

Samar Ibrahim and sheriff Abdallah

1The American University in Dubai, UAE, The British University in Dubai, UAE

sibrahim@adjunct.aud.edu sherief.abdallah@buid.ac.ae

Abstract. Twitter is a viable data source for studying public opinion. The study aims to identify public opinion and sentiments toward Covid-19 vaccine and examine conversations posted on Twitter. The study examined two Datasets; one of 7500 tweets collected using RapidMiner from June 7- 17, 2021, and 9865 tweets collected from Kaggle on the 3rd of January 2021. It used Naive Bayes model to classify, analyze and visualize tweets according to polarity, K-means clustering, and key tweet topics. The study showed that positive sentiments were dominant in both times; it also realized that positive polarity increased over time from January to June 2021. In addition, vaccine acceptance became more prevalent in the tweets' discussions and topics. Understanding sentiments and opinions toward Covid-19 vaccine using Twitter is critical to supporting public health organizations to execute promotions plans and encourage positive messages towards Covid-19 to improve vaccination mitigation and vaccine intake.

Keyword. Covid 19 Vaccine, Twitter, Public Opinion, Datasets, Naïve Bayes

1 Introduction

Corona Virus Disease (COVID-19) is a frightening pandemic discovered in early December 2019 in Wuhan, China, and spread rapidly across the globe. The World Health Organization (WHO) described this outbreak as a severe global threat [1]. The health crisis caused by this disastrous pandemic globally severely affected various sectors. In 2020, the global economy diminished by 3%, with a significant loss estimated at around 9 trillion USD [2].

Amid the unprecedented crisis, significant efforts have been embarked on in order to mitigate disease spread and find a cure or protection. Since December 2020, many COVID-19 vaccine candidates have been verified to be safe and effective in protecting against COVID-19 and were approved by regulators to be used [3]. Hence, many governments across the globe devised vaccination rollout plans according to priorities population and conducted a vaccination campaign where their residents were encouraged to be vaccinated and acquire immunity to tackle the pandemic [4]. In addition, Governments and health organizations have also used a variety of communications channels, traditional and nontraditional, such as social media, to share and provide learning knowledge related to the COVID-19 virus and its various vaccines [3].

Since the beginning of the COVID-19 pandemic, information circulating via social media networks such as Twitter, Facebook, YouTube, etc., has significantly affected public opinions. Furthermore, people's use and dependency on social media as a source of news and information have intensified because of the many implemented lockdown (Cinelli et al., 2020). These social media platforms also provided access to unprecedented information and knowledge through which users could express sentiments, views, and opinions. As a result, these platforms have become a viable source of data for researchers to analyze posts and predict trends for human sentiments and use these for better mitigating crises, especially pandemics (Cinelli et al., 2020). For example, Twitter, one of the world's leading micro-blogging social media platforms, has over 330 million active users, over 500 million tweets per day, and more than a billion visits per month, and people can genuinely express themselves in a timely mode [5].

The Emergence of mobile and ubiquitous computing and the advance in big data analytics and storage platform makes it more feasible for more social media data to be generated, collected, and analyzed [6]. There is a great advantage in using these platforms, especially Twitter, to analyze public sentiments and opinions instead of traditional face-to-face interviews. As a result, users can express themselves genuinely and freely share their locations, opinions, and feelings. Whereas in traditional ways, people's response is typically affected by the nervousness brought on by live communication situations [7].

Many researchers use many methods and techniques to analyze these tweets. They prove that an automated data mining system with real-time analysis of these posts is highly efficient for identifying people's opinions during times of disaster [8]. They mainly use sentiment analysis and topic clustering methods. By defining the topic used in these tweets using LDA (Latent Dirichlet allocation), they comprehend more the conversation subjects and sentiments about COVID-19. [9].

Moreover, many researchers have examined the effect of social media data on vaccine hesitancy. It is the role in spreading misinformation about Covid-19, such as the virus was the result of a 5G social network or the idea that this virus was created as a government bioweapon (Cinelli et al., 2020), to name a few examples. This discouraged immunization perception or decreased the vaccine's benefits perception, negatively affecting the vaccine uptake [10]. Therefore, detecting trends using social media data would provide public health authorities with essential opportunities to create policy decisions, provide new plans, set up arrangements or interventions, and make people aware of all public changes [5].

The objective of the research study is to compare public emotion toward Covid 19 vaccine using the microblog Twitter data and test its sentimental analysis across two different timelines.

1. The first research question compares the different polarities that change over time, using Twitter as a social media towards Covid -19 vaccine.
2. The second question compares different polarities using Twitter as social media data towards Covid -19 vaccine.
3. The third question is to examine the main topical issues in the tweet's conversations about the Covid 19 vaccine?

The rest of the paper is organized as follows: Section 2 looks at all previous research-related work analyzing public opinion and sentimental analysis towards Covid-19 and its vaccines using Twitter as a data source. Section 3 provides an overview of the methodology, methods, and algorithms used in the study. Section 4 presents the results and a detailed discussion of the results. Finally, section 5 provides concluding remarks and suggested future areas for research.

2 Related Work

Many researchers investigate public opinion towards Covid-19 using social media data. They use various methods and techniques to analyze social media data and study public attitudes toward Covid-19. One of the studies looked at public perception towards some non-pharmaceutical interventions (NPIs) used during the pandemic to mitigate it in different countries using the topic clustering technique applied to extracted tweets. They identified that some countries have more concerns and attention than other countries. People in New Zealand, for example, care more about using hand sanitizer and wearing masks than in the US [11]. Another study investigated different topics in the tweets related to Covid 19 and showed that analyzing the topic would allow a sense of trends of people's perceptions on Twitter. Tweets were investigated in Portuguese and English and compared to discuss the effectiveness of topic identification and sentiment analysis in these languages. Topics were ranked, and tweets' content was analyzed while providing an analytical assessment of the discourse evolution over time[12]. (2021) proved a high correlation between public attention level, analyzed on Twitter, and Covid-19 related to case number; this was done using topic analysis and sentiment analysis methods.

At the early stage of the pandemic, researchers explored COVID-19-related social media posts describing symptoms in order to identify Covid-19 different symptoms. As a result, a total of 36 symptoms were extracted and identified [13]. In addition, other researchers looked for frequent users. They analyzed their usage patterns and overall emotion during this period to understand the changing mood of the people in relation to the disease [14].

Other researchers designed and developed a novel learning machine based on classification, clustering, and topic extraction to analyze tweets and obtain significant sentiments. This model extracts main topics from the clusters using the K-means algorithm divides these topics into positive, neutral, and negative sentiments, and rapidly identifies commonly dominant characteristics of public opinions and attitudes related to COVID-19 [15]. Other studies examined tweets to explore sentiment analysis using a support vector machine (SVM), k-nearest neighbor (KNN), and Naïve Bayes [15]. Likewise, Samuel et al. (2020) in (Satu et al. 2021) analyze the tweets, provide non-textual variables using N-Gram, and analyze the sentiments using NB, Linear regression, LR, and KNN. Fake news during the pandemic was another focus for researchers to explore ways to detect and reasons for misleading the targeted population using a classification approach that implements natural language processing, machine learning, and deep learning [16]

Very few and limited research investigates the public opinion and perception of the vaccine on Twitter. One study assesses the opinion of the Indonesian people using social network analysis of the COVID-19 vaccine by implementing sentiment analysis using the Naïve Bayes Algorithm. Likewise, Yousefinaghani et al. (2021) study public sentiments and opinions toward COVID-19 vaccines based on the content of Twitter compared to their progression over time between January 2020 and January 2021, their geographical distribution, and provide discussion on vaccine rejection and hesitancy and show different patterns in various countries. This study also investigated public opinion and sentimental analysis but compared recent Twitter data collected in June 2021 to data collected in January 2021. In addition, it examined the topics of discussion included in these tweets.

3 Methodology and methods

The study was conducted following a 5-stage successive process, as depicted in Fugure1. First, data was collected, annotated, and pre-processed using Natural Language Processing techniques (NLP), then classified using sentimental analysis utilizing the Naïve Bayes classifiers algorithm in order to train a sentiment classification model. A Naive Bayes classifier is commonly used for solving classification problems and has been shown to be accurate at determining the true polarity of sentences as compared to other technique[2]. Next, data used performance classification to test the performance of the model. Furthermore, K-means clustering was implemented in the study to analyze the topics and apply them in visualization. Finally, visualization was used to present the result of the study.

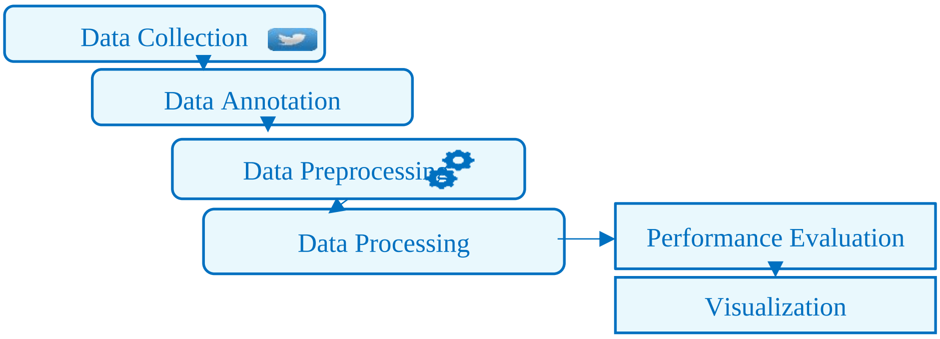

Fig. 1. Represents the different five stages of the study

3.1 Data collection

Two datasets were used in the study, one set of 9,862 data tweets collected from the Kaggle website dated from the third week of January of 2021. The second was collected directly from Twitter using RapidMiner searching twitter using different keywords, such as covid19-vaccine or #covid vaccine; English attribute was also used. Specific attributes were collected, including but not limited to creation date, username, user ID, user ID, language, source, text, geolocation, retweet count, and tweet ID. Only text and user ID attributes are used in the study. The tweets were collected between 10 and 17 June 2021, duplicate tweets were removed, and a total of 7500 tweets were collected. Tweets were collected at different times, the files were appended and tabulated in one data file, and Retweet and duplicate data were removed and prepared for annotation and preprocessing. Two extraction subsets of 900 tweets from each dataset were prepared for the second phase of the study.

3.2 Data Annotation

After duplicate data was removed, the data was manually annotated, where tweets are classified into three polarities: Positive, neutral, and negative. Positive polarity, where tweets indicate a willingness to take the vaccine; Neutral polarity, \where tweets show neutral opinion towards the Covid19 vaccine; Finally, negative polarity, where tweets have a very clear opinion against the Covid-19 vaccine. Of the entire set, 60% of the data was annotated, and the other 40% was left for testing. Examples of the annotated tweets are depicted in Table 1.

Table 1. Examples of three annotated Tweets that correspond to three different kinds of polarity: Positive, Neutral, Negative

Polarity	Text
1-Positive	RT @CSi: No appointment required! Get vaccinated today! All individuals age 12 and older are currently eligible for a free vaccine.
2-Positive	Chi @ PA: Need another reason to get vaccinated? We have more data showing that getting a vaccine may keep you out of the hospital.
3-Positive	RT robinmonotti2: VACCINE ENHANCED DISEASE: "Almost one third of people in the UK who have so far died from the Indian variant Covid19.
1-Neutral	RT CT: Are you a youth aged 12-17 or do you work with youth who have questions about the #COVID19 vaccine. Toronto PFR &
2-Neutral	RT PU: The #COVID19 Dashboard has been updated On Thurs 17 June, 11,007 new cases & 19 deaths within 28 days
3-Neutral	RT thetruthin: #COVID19 Situation in Kamrup (M) seems to be under control as it reported only 158 new COVID-19 cases. According to official.
1-Negative	RT drjohnm: Another paper chronicling myocarditis after mRNA #covid19 vaccine in 8 young people. Some were admitted to ICU.
2-Negative	DrEricDing: Two patients positive for the #DeltaVariant in Calgary have died. One death was in patient with 2 doses of #COVID19 vaccine.
3-Negative	RT @abirballan: @afneil Many individuals don't benefit from the vaccine. It is unethical to ask an individual to take a risk for somebody else.

3.3 Data Preprocessing

Post annotation, research progressed to the data processing stage. Before performing any semantic data analysis, datasets were preprocessed to reduce all possible data noise. RapidMiner process Document operator was used to preprocess the data using NLP, depicted in Figure 2. This operator produced word vectors from string attributes and used the term frequency-inverse document frequency (TF–IDF) to preprocess the data. The first stage of preprocessing was tokenizing in order to create a list of single tokens that can include stand-alone words. The second stage was to filter many unnecessary items, URLs, special characters, emojis, non-ASCII codes, non-English letters, tabs, symbols, punctuations, and abbreviations. Then all tweets were converted to lower case, and all StopWords that occur frequently but don't carry distinct semantical meaning were removed. In addition, replace operator was used to replace all hashtags and special characters such as "@," "?", etc. These characters don't have significant meaning on NLP. Finally, the two extracted data sets were preprocessed. In addition, the complete dataset collected from Twitter was also preprocessed for this data to be used in the K-means clustering model and implemented in visualization techniques such as WordCloud.

Figure 2. Representation of the subprocess of the process document that performs preprocessing.

3.4 Data Processing

Sentiment analysis. Sentimental analysis, known as opinion mining, deals with text classification that divides text between positive, negative, and neutral. It is a computational study of opinions, where emotions are expressed in given words or sentences and are analyzed by NLP techniques [2]. Pang, Lee, and Vaithyanathan (2002) were the first to use a sentiment analysis approach to movie classification using machine learning approaches [8]. Vaccine sentiment analysis was implemented in order to explore the evolution of public opinion and classify sentiments of the extracted tweets related to Covid-19. This analysis utilized the Naïve Bayes classification algorithm, which was applied to classify the tweets according to their polarity. This algorithm is based on Bayes' Theorem:

$$\boldsymbol{P(A|B) = \big(P(B|A)P(A)\big)/P(B)}$$

Naïve Bayes is generally used in the classification technique, specifically on Twitter. Its main feature is to have a strong hypothesis of any condition. Using the annotated developed data, the Naïve Bayes operator in RapidMiner was implemented in the study to train and develop a model in order to classify tweets according to their polarity (positive, neutral. negative). The naïve Bayes has two-level, the training level, and the classification level; the training level will predict the input according to its polarity. We based on training 60% of the data and the remaining for testing the data for our model. Extraction data from the two datasets, the Dataset dated to January and the one collected from Twitter, were applied to the Naïve Bayes and tested for performance. For the purpose of the study, these extractions have the exact count of tweets for both subsets. Figure 3. depicts part of the process used with Naïve Bayes.

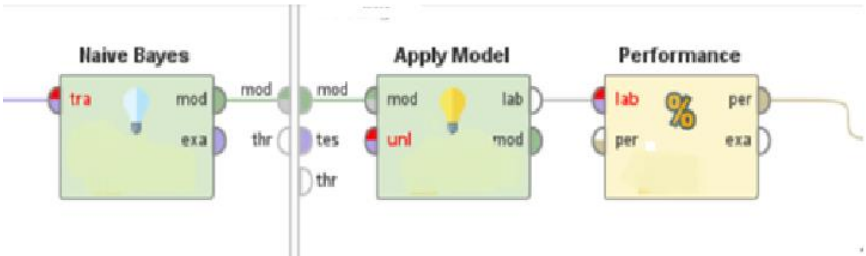

Figure 3. Representation of the Naïve Bayes classifier and the performance operator

Evaluation. Performance classification was implemented to evaluate the statistical performance of the model. K-fold Cross-validation Performance operator with a single K parameter from RapidMiner was used, where data was split between training and testing. The output of the performance operator is a confusion matrix.

Clustering is also a machine learning method that groups data according to common characteristics. K-means is a clustering technique implemented in the study to group data into K mutually exclusive clusters. With the least mean distance between the data points and the centroid. The objective is to apply this method in this study and find out the most popular topic and conversation in the tweets. It is a method of vector quantization. It has an objective to partition N observations into K clusters, with each term in a document assigned a score based on how prevalent the paper is. Any word with a high score implies more significance or representative than the other words. Therefore, on average, all top words in the different clusters are the most significant in the document. Cluster operator was also implemented using Rapid miner on the complete datasets from Twitter that counts for 7500 tweets. The number of clusters was achieved after consecutive iterations of trying for K number of clusters to reach the number 5.

4 Results and Discussions

In this study, the two extracted datasets that count for 900 tweets were implemented and applied to the Naïve Bayes classifier model. As a result, 400 out of 900 or 44.44% of the extracted Datasets from January showed a positive attitude towards the Covid-19 vaccine. Whereas 350 out of 900, 38.89%, showed a neutral attitude or opinion toward the datasets in June. The remaining 150 making 16.67%, were negative. Whereas for the datasets collected in June 2021, 550 out 900 or 61.12% adopted a positive attitude, 251 out 900 or 27.89% related to a neutral count of polarity, and 99 out 900 adopted a negative count of polarity or 11%. A summary is provided in Table 3.

These results showed an increase in positive polarity count, indicating an increase in the positive attitude towards the Covid-19 vaccine over time (from Jan to June 2021). Analyzing the data from both sets will show words such as “take the vaccine,” “appointment,” “doses,” and “received.” Also, there was a decrease in neutral polarity, which could mean that the neutral attitude towards the Covid19 vaccine was changing and decreasing. Looking at tweets with neutral polarity showed no implication for or against the vaccine. Finally, negative polarity counts decreased between data collected in January and June. In addition, from analyzing the tweets, these tweets included words such as “rejection” or mentioning “death cases.” Therefore, the decrease in negative polarity could be translated as fewer people rejecting the vaccine and more willing to take the covid19 vaccine. Two graphs that depict the result of the sentiment analysis of the tweets for January and June according to polarity are represented in Figure 4 and Figure 5 below.

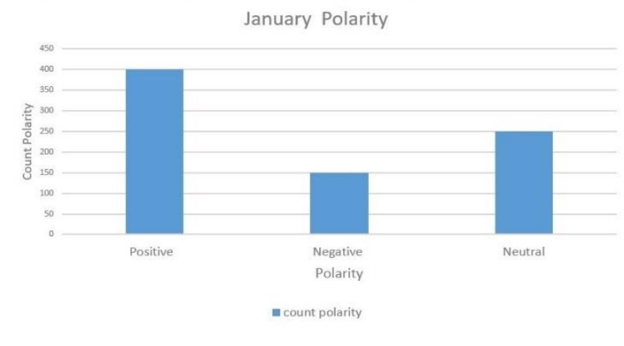

Figure 4. Visualization of the data of January according to polarity

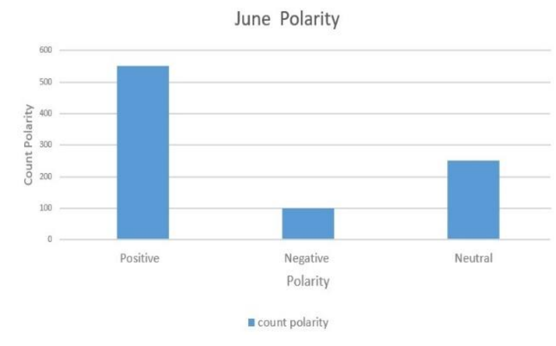

Figure 5. Visualization of the data of January according to polarity

Table 2. Percentage of polarity results for the two Datasets.

Percentage Polarity in January	Polarity percentage in June
Positive Polarity: 44.44%	Positive Polarity: 61.12%
Neutral Polarity: 38.89%	Neutral: Polarity 27.89%
Negative Polarity:16.67%	Negative Polarity: 11%

As mentioned earlier, the model is evaluated by K fold cross performance. The result was the Confusion matrix to estimate the classifier's performance by indicating the number of correct versus incorrect predictions. It identified true values such as True Positive, True Neutral, and True Negative. The confusion matrix's objective is to test the accuracy of the training and the testing data. We have two datasets for January and June that use two different Naïve Bayes classifiers, so we had two annotated Datasets used to train the classifiers; after performing training and testing, the accuracy of the confusion matrices is depicted in table 3 and 4, respectively. The Accuracy of the model for data in January is equal to 78.889%, and its Kappa=0.665, 330, 280, and 100 from the confusion matrix for data selected in January are correct True polarity for tweets, i.e., True positive, True neutral, and True negative. The rows presented the tweets for positive, neutral, and negative polarity. Similarly, for the data in June, the Accuracy = 80.556%, and the Kappa = 0.645489, 480, 170, 75 are the correct polarity for tweets, i.e., True positive, True neutral, and True negative, respectively.

Table 3. Confusion Matrix resulted after implementing the performance operator: Accuracy =78.889%, & Kappa= 0.665.

	True Positive	True Neutral	True Negative	Class Precision
Prediction Positive/Actual	330	35	35	82.5%
Predictive Neutral/Actual	30	280	40	80%
Predictive Negative /Actual	20	30	100	66.667%
Class Recall	86.842%	81.15%	7.143%	

Table 4. Confusion Matrix resulted after implementing performance operator Accuracy = 80.556% & Kappa= 0.645489

	True Positive	True Neutral	True Negative	Class Precision
Prediction Positive/Actual	480	40	30	82.5%
Predictive Neutral/Actual	51	170	30	80%
Predictive Negative /Actual	9	15	75	66.667%
Class Recall	88.889%	75.556%	55.556%	

This study addressed people's opinions and attitudes towards Covid-19 by analyzing Twitter posts. Results showed an increase in positive polarity as time progressed and a decrease in neutral polarity and negative polarity on Twitter. It was noticeable that tweet conversations tend to be more toward positive or pro-vaccine attitudes in different times where positive polarity exceeded neutral and negative polarity in Datasets. This also shows that people’s attitudes were more pro-vaccine than anti-vaccine or neutral vaccine attitudes. And from the analysis of the result of the positive polarity, there was a more common general acceptance in the tweets for vaccination becoming a must. However, some of the analyses of negative tweets showed extreme side effects that were very rigid and refused the Covid-19 vaccine.

It was also important to look at tweet conversations and topics to investigate people's opinions and attitudes. Using K-means clustering to analyze the tweet, the most frequent 25 topics used in the clusters were selected from the tweets and ranked. Many graphical visualizations were used in this study using k means clustering methods.   Table 6 provides the top 25 popular frequent words with their scores in the five clusters related to the Covid-19 vaccine. Topics such as” third,” “dose,” “received,” “appointment,” dates,” “vaccination,” and “available” infer that people's tweets related to the pro-vaccine conversation. Also, words such as “total, “Pfizer,” “sputniK,” “Moderna,” “AstraZeneca” showed that the tweets’ conversations are more inline of pro-vaccine. In addition, a graph that depicts the ranking of the various topics is presented in figure 6.   The graph shows “dates” is the first in ranking and “availability “ranks last. This implies that people are looking to book appointments to register for the vaccine. The other topic in ranking stands for “coronavirus,” “Africa,” and “important,” indicating a sentiment that vaccine is important. “third” where people are considering a third vaccine, other topics such as “delta” variant where it seems that people are becoming more aware of the Indian variant and are more worried about it.

Table 5. The most common word used in the 5 clusters.

The 25 most Frequent Topics in the tweets	C0	C1	C2	C3	C4
Variant	0	0.305	0	0	0
Vaccine	0	0	0.09	0.231	0.245
Vaccination	0	0.268	0.073	0	0
Total	0	0.305	0	0	0
Third	0	0.159	0.237	0.045	0.213
Sputnik	0.002	0.258	0.112	0.001	0
Sorted	0	0	0.215	0.006	0
Sinopharm	0	0.305	0	0	0
Seceived	0	0	0.157	0	0
Pfizer	0	0	0	0.003	0.348
Pandemic	0.003	0	0	0.134	0
Mrna	0	0.305	0.031	0	0
Moderna	0	0.002	0.009	0.001	0.399
Limit	0.001	0	0.001	0	0.348
Important	0	0.158	0.235	0.055	0.216
Dose	0	0.312	0.001	0.001	0
Africa	0	0.001	0.103	0.324	0.248
Delta	0.001	0	0.002	0.003	0.347
Dates	0	0.31	0.455	0.001	0.21
Covishield	0	0	0	0.229	0
Coronavirus	0.001	0.131	0.114	0.457	0.161
Capacity	0	0.316	0.013	0	0
available	0	0	0.154	0	0
astraZeneca	0	0.314	0.013	0	0
Appointment	0.001	0	0	0	0.348

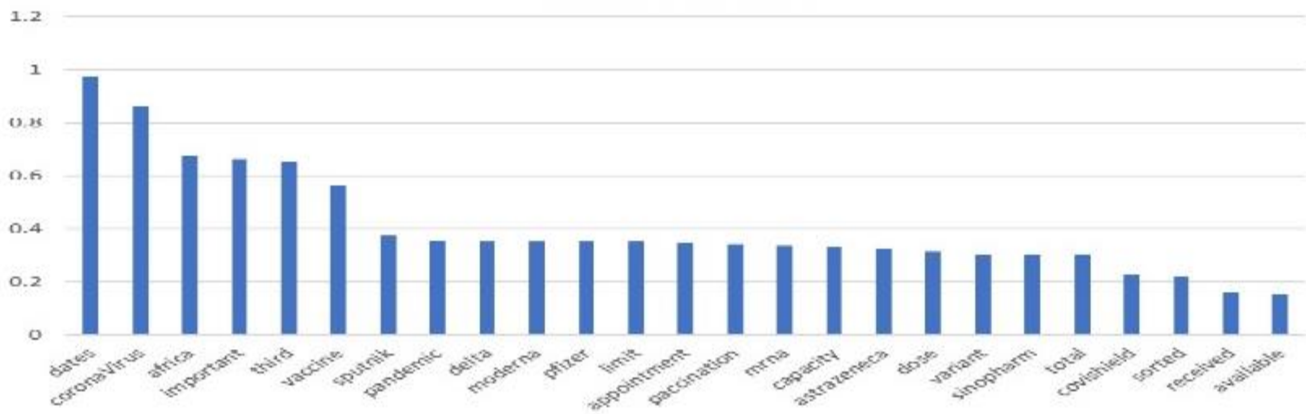

Figure 6. Represents the top 25 topics according to io its rankings with their frequency

The visual graph in Figure 7 depicts how each topic is represented according to its usage percentage, and different colors indicate each cluster. For example, it is clear from the graph that " dates " are very common and frequently used in clusters 1,2,3 and 4. Similarly, “coronavirus” and “third” are used in clusters 1,2,3 and 4. While some other topics, such as covishield, are the least used in all clusters, “capacity” has minimal use in clusters 4,3,2, etc. Furthermore, if you examine cluster 0 is barely presented.

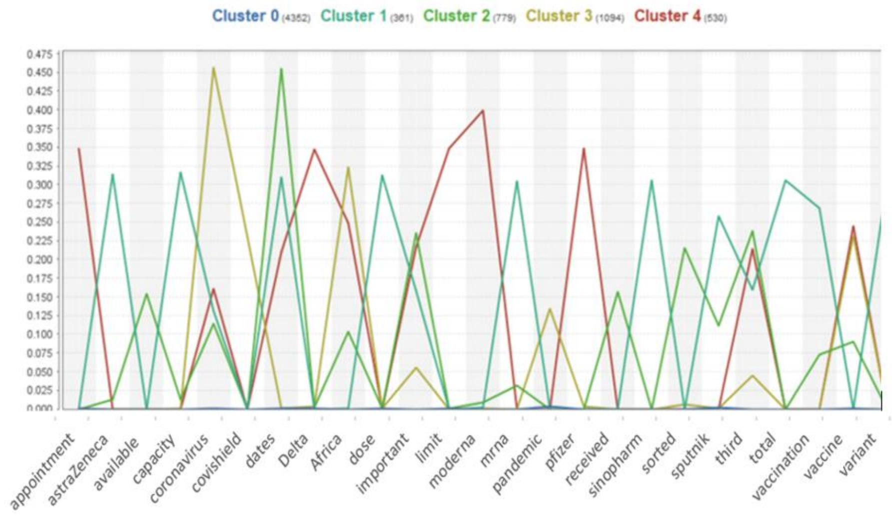

Figure 7. Prevalent of the most 25 frequent topics and their prevalence in each cluster

Similarly, Figure 8 visualizes K- means clustering. Still, this graph represents the frequent use of the topic using the intensity of the shade of color and according to each cluster where each cluster is presented in a single row. It is so clear that cluster 0 is unshaded or has white color and cluster one has the most frequent topic such as “dates’, “vaccine,” “variant,” and “vaccination.” This clearly shows that more conversation leads to more positive polarity and a positive attitude towards the vaccine.

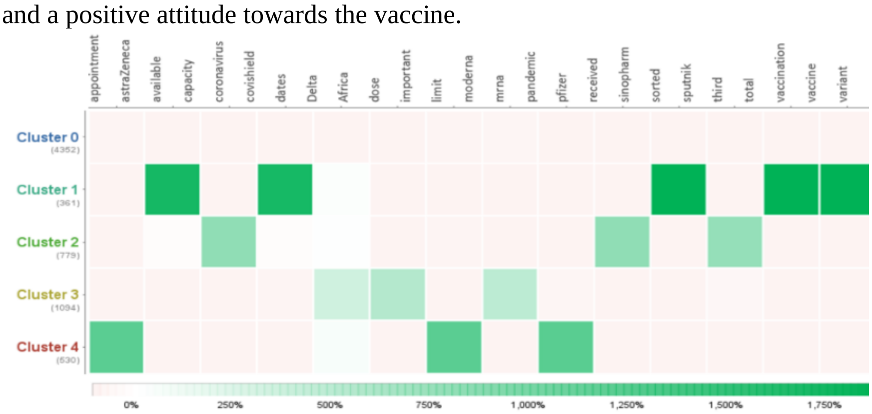

Figure 8. Heat map Visualizing the different topics represented in different clusters described as a shade of color.

WordCloud was also implemented in the study. The tool is a standard text analysis tool that visualizes the word of the most frequent term used in the tweet. It was created using the K-means clustering results using the most popular topic in these 5 clusters. It is used according to the frequency of occurrence of these topics. If you examine the word cloud, words such as “appointment,” “vaccine,” “take,” “third,” “vaccinated,” “Sinopharm,” Pfizer, and Moderna stand out. This word cloud can create a story to tell. Examining this word cloud tells a clear story to understand and analyze the situation. Here again, the story will continue towards more acceptance of the vaccine words such as “vaccinated,” “dose,” etc.

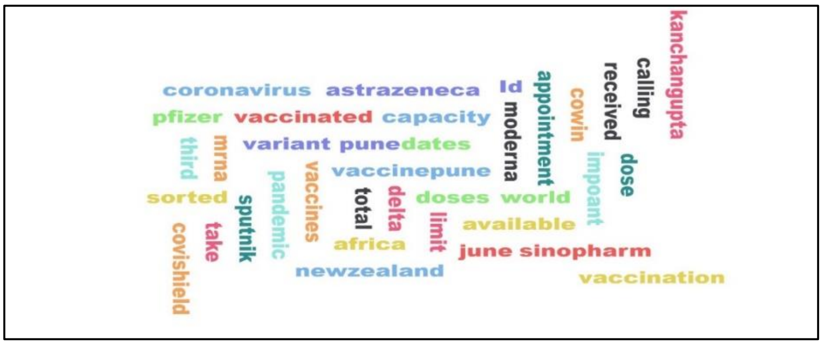

Figure 9. WordCloud for the most 25 frequent words

The words at the center and most in each cluster are depicted in Figure 9 below. These five graphs show each cluster and represent the two center words prevalent in each cluster. For example, cluster 0 has "sputnik" and "pandemic" forming its center. Whereas cluster 1 has "doses" and "capacity," similarly, cluster 2 has "dates" and "third," cluster 3 has "coronavirus" and "Africa," and cluster 4 has "Moderna" and "Pfizer."

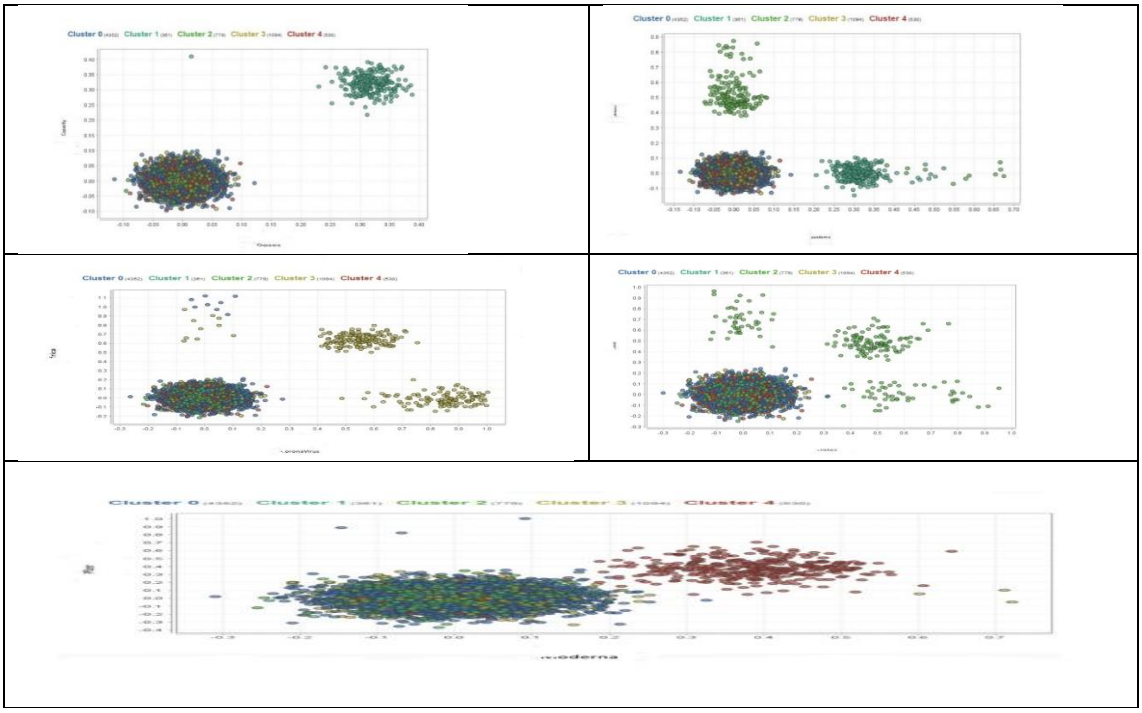

Figure 10. Scatter plots that visualize the center of the cluster using two words

5 Conclusion

Social media data such as tweets on Twitter is becoming a platform for disseminating ideas and individual opinions, enabling this platform to forecast and shape behavior. With the current pandemic, understanding and reacting to people's emotions, attitudes and opinions are a key component of public health plans, and it is very critical for all organizations and governments to encourage mitigation and strategic public health promotion plan in order to encourage vaccine acceptance and vaccine intake and decrease vaccine hesitancy and opposition.

Twitter is an excellent viable source for testing and examining issues related to people's attitudes and emotions. Insights gained from analyzing real-time conversations create more opportunities and measures for public health organizations to promote and affect possible change in a more dynamic and timely manner.

This study used Naïve Bayes to investigate public opinion and attitudes related to the Covid-19 vaccine. In addition, our study examined the clustering technique using the K-means algorithm to look at the diverse topical issues on people's minds and expressed in tweets. Our research examines the progression of people's attitudes and opinions by comparing two different datasets from different times. Our findings demonstrated that people's opinions and attitudes show that tweets' positive polarity towards the Covid-19 vaccine increases over time. On the other hand, negative and neutral opinions are decreasing, and people on Twitter express more willingness to take the vaccine. This is aligned with findings in other recent studies on the vaccine [17]. The visualization technique using k-means clustering methods provides a way to extract popular topics and better understand the tweets, opinions, conversations, and topics relating to the Covid-19 vaccine.

Limitations: Few limitations are affecting this study. The data size that can be applied on a Naive Bayes classifier on a desktop is minimal, increasing the chance that the system will not accomplish the objective with a moderate data size. Second, English is not the only language used in the tweets.

These limitations will provide opportunities for further research to analyze tweets from different languages, or in other geographical locations, and with different cultures that might affect public opinion and attitude

References

1. M. Cinelli *et al.*, "The COVID - 19 social media infodemic," *Sci. Rep.*, pp. 1–11,(2020).
2. C. Villavicencio, J. J. Macrohon, X. A. Inbaraj, J. Jeng, and J. Hsieh, "Twitter Sentiment Analysis towards COVID-19 Vaccines in the Philippines Using Naïve Bayes," *Informatiom*, vol. 12, (2021).
3. F. Eibensteiner, V. Ritschl, F. A. Nawaz, and S. S. Fazel, "People ' s Willingness to Vaccinate Against COVID-19 Despite Their Safety Concerns : Twitter Poll Analysis," *J. Med. Internet Res.*, vol. 23, February, pp. 1–10, (2021),
4. Pristiyono, M. Ritonga, M. A. Al Ihsan, A. Anjar, and F. H. Rambe, "Sentiment analysis of COVID-19 vaccine in Indonesia using Naïve Bayes Algorithm," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 1088, no. 1,, (2021),
5. J. Sooknanan and N. Mays, "Harnessing Social Media in the Modelling of Pandemics — Challenges and Opportunities," *Bull. Math. Biol.*, vol. 83, no. 5, pp. 1–11, (2021),
6. M. B. Mutanga and A. Abayomi, "Tweeting on COVID-19 pandemic in South Africa : LDA-based topic modelling approach," *African J. Sci. Technol. Innov. Dev.*, pp. 1–10, (2020),
7. K. Hou, T. Hou, and L. Cai, "Public attention about COVID-19 on social media : An investigation based on data mining and text analysis," *Pers. Individ. Dif.*, vol. 175, no. January, p. 110701, (2021)

[8] J. R. Ragini, P. M. Rubesh Anand, and V. Bhaskar, "Big data analytics for disaster response and recovery through sentiment analysis," (2018,)

[9] R. Alshalan, H. Al-khalifa, D. Alsaeed, H. Al-baity, and S. Alshalan, "Detection of Hate Speech in COVID-19 – Related Tweets in the Arab Region : Deep Learning and Topic Modeling Approach," *J. Med. Internet Res.*, vol. 22, no. 12, 2020, doi: 10.2196/22609.

[10] R. P. Id *et al.*, "Examining the effect of information channel on COVID-19 vaccine acceptance," *PLoS One*, pp. 1–14, (2021).

[11] J. C. Lyu, G. K. Luli, and J. C. Lyu, "Understanding the Public Discussion About the Centers for Disease Control and Prevention During the COVID-19 Pandemic Using Twitter Data : Text Mining Analysis Study," *J. Med. Internet Res.*, vol. 23, (2021)

[12] K. Garcia and L. Berton, "Topic detection and sentiment analysis in Twitter content related to COVID-19 from Brazil and the USA," *Appl. Soft Comput. J.*, vol. 101, p. 107057,(2021)

[13] J. Guo, A. Christina, L. R. Ms, S. E. W. Bsn, and K. G. Cloyes, "Mining twitter to explore the emergence of COVID-19 symptoms," *Willey*, no. July, pp. 934–940, (2020),

[14] A. Mathur, P. Kubde, and S. Vaidya, "Emotional analysis using twitter data during pandemic situation: Covid-19," *Proc. 5th Int. Conf. Commun. Electron. Syst. ICCES 2020*, no. ICCES, pp. 845–848, (2020)

[15] M. S. Satu *et al.*, "TClustVID: A novel machine learning classification model to investigate topics and sentiment in COVID-19 tweets," *Knowledge-Based Syst.*, vol. 226, p. 107126, (2021),

[16] Y. Madani, M. Erritali, and B. Bouikhalene, "Using artificial intelligence techniques for detecting Covid-19 epidemic fake news in Moroccan tweets," *Results Phys.*, vol. 25, p. 104266, (2021),

[17] S. Yousefinaghani, R. Dara, S. Mubareka, A. Papadopoulos, and S. Sharif, "An analysis of COVID-19 vaccine sentiments and opinions on Twitter," *Int. J. Infect. Dis.*, vol. 108, pp. 256–262, (2021)