Prediction of the Impact of Covid-19 Vaccine on Public Health Using Twitter

Ayesha Ayub Syed, Ford Lumban Gaol, Wayan Suparta, Edi Abdurachman, Agung Trisetyarso, Tukoro Matsuo

Abstract--**Social media is a source of big data. Media like Twitter and Facebook has been used for collecting and analyzing user data for different purposes. The data can be used to analyze people opinions towards certain topics and incidents by applying sentiment analysis and then certain useful insights can be drawn from the analyzed data. During the current time of Covid-19, people have been sharing information regarding Covid-19 statistics, vaccines, and discussing the effects of the vaccine concerning public health. The purpose of this study is to analyze tweet data regarding the Covid-19 vaccine by applying sentiment analysis and predicting the impact of the vaccine on public health. Also, the tweets are analyzed for hidden topics by applying Topic Modelling using Latent Dirichlet Allocation (LDA) and Latent Semantic Analysis (LSA). The source of data for this study is Twitter API. The coding and data analysis is done using Python programming language in the Spyder (Scientific Python Development Environment) that is an integrated development environment for scientific programming, testing, and data analysis. The results of the study indicate a greater positive sentiment reflecting a healthy public discussion about the Covid-19 vaccine, information, awareness, and public acceptance. With these results, a positive impact of the Covid-19 vaccine on public health is predicted. The results of topic modeling discovered 10 hidden topics from the tweet dataset.**

Index Terms—**Covid-19, public health, positive, sentiment analysis, vaccine, topic modeling**

I. INTRODUCTION

THE global Covid-19 pandemic emerged at the end of the year 2019 in Wuhan, a city in China that spread across the globe within few months and was declared by WHO as an emergency for public health [1]. Along with some common strategies, nations of the world have implemented various public health strategies to reduce the transmission and control the pandemic spread [2]. The scientific research community has been engaged in finding better treatment options for the disease as well as developing a safe and effective vaccine to control the spread of the disease [3].

Together with a direct influence of the pandemic on public health, there is a collective influence on peoples' emotions as well due to change of norms, social distancing and, style of living [4]. People with heightened or charged emotions are less likely to support and get involved in preventive behaviors. Many are concerned over vaccine safety and side effects, thus exhibiting fear, a negative emotion. The intentions of the people towards the Covid-19 vaccine have also been affected by the prevailing anxiety and loss factors [5]. Still, a part of the population is hopeful and shows a good acceptance of the vaccine and other preventive measures.

The motivation for this work is to make a scientific contribution to the efforts for disease reduction or mitigation strategies in this time of crisis by conveying useful information regarding public opinions about the Covid-19 vaccine to healthcare practitioners and policymakers.

This paper presents an analysis of tweets related to the Covid-19 vaccine and public health using Twitter. Sentiment analysis is applied to examine and understand people's thoughts and emotions towards the Covid-19 vaccine concerning public health. Topic modeling is done to uncover topics discussed in the tweets. Theoretically, it is known that vaccination is a core factor of public health [6]. In a real scenario, it surely depends on the vaccine communication, education, awareness, and acceptance by the public. The misinformation and hoax about the Covid-19 vaccine found on the internet is also threatening to the health of people because it makes people reluctant towards preventative health behaviors and unwilling to get vaccinated [1] and this attitude is also passed onto their friends and family. Therefore, analyzing emotion in the case of the Covid-19 vaccine is very useful to inform the public healthcare community. The results of the sentiment analysis are used to predict the impact of the Covid-19 vaccine on public health.

Section II of the paper provides a background for the research and discusses some of the existing related work. Section III discusses the research methodology and experiment details. Section IV presents the results of the analysis extensively. Section V is the discussion section that provides some implications from the research. Section VI concludes the paper.

Manuscript received March 07, 2021; revised October 19, 2021.

Ayesha Ayub Syed is a Ph.D. student at the Department of Doctor of Computer Science, Bina Nusantara University, Jakarta, Indonesia; (e-mail: ayeshaayubsyed@yahoo.com).

Ford Lumban Gaol is an Associate Professor at Informatics Engineering and Information System at Bina Nusantara University, Jakarta, Indonesia; (e-mail: fgaol@binus.edu).

Wayan Suparta is an Associate Professor at Space Science Centre (ANGKASA), Institute of Climate Change, Universiti Kebangsaan Malaysia (UKM), Malaysia; (email: drwaynesparta@gmail.com)

Edi Abdurachman is an Associate Professor at the Department of Doctor of Research in Management at Bina Nusantara University, Jakarta, Indonesia; (e-mail: edia@binus.ac.id)

Agung Trisetyarso is an Assistant Professor at the Department of Doctor of Computer Science, Bina Nusantara University, Jakarta, Indonesia; (e-mail: atrisetyarso@binus.edu)

Tukoro Matsuo is a Professor at the Graduate School of Industrial Technology, Advanced Institute of Industrial Technology, Japan and Department of M-Commerce and Multimedia Applications, Asia University, Taiwan; (e-mail: matsuo@aiit.ac.jp)

II. Background and Related Work

In times of crisis especially that of the Covid-19 pandemic, people turn towards social media to voice their opinions and share their thoughts. Social media is a digital tool that allows users to connect and engage with others and share their content. There are several social media tools and applications available currently. A huge volume of data is generated by these media every day. Twitter is a very popular social media that generates a huge number of tweets posted by its users every day about different topics, events, incidents, and issues. Sentiment Analysis is a technique that can be applied to analyze the attitudes, opinions, and emotions of people towards certain topics, issues, events, incidents, products, or services [7]. Sentiment Analysis using Twitter provides useful insights into a topic, product, or incident by analyzing people's thoughts, opinions, and emotions in a certain context [8].

Manual classification and analysis of social media data is not a feasible approach. However, several automatic techniques have been developed and available for automatic text polarity classification. These are either lexicon-based methods or learning-based techniques [9]. When the sentiment is identified by breaking down the text into words whose polarity is already specified in the dictionary, the approach is known as the lexicon-based approach and when the sentiment classification is done by training classifier with pre-annotated text in terms of sentiment polarity, it is known as machine learning-based approach [10]. Examples of lexicon-based methods are bag-of-words, n-grams and, n-grams graphs while learning-based methods include Support Vector Machines (SVM), Naïve Bayesian, Logistic Regression, etc.

[11] applied sentiment analysis on Twitter data regarding Covid-19 during April-May 2020. The data was analyzed for the polarity of the sentiment as well as its subjectivity. The results indicated the largest number of neutral sentiments, medium number of positive sentiments, and lowest negative sentiments. Regarding subjectivity, 64% of data was found to be objective, 22% was indicated as subjective and the rest 14% was neutral.

[12] utilized Twitter content to observe the opinions and attitudes of Saudi nationals towards the Covid-19 disease prevention measures and initiatives taken by the government. The model for sentiment analysis was developed using a machine learning approach and the Naïve Bayes classifier was used. Out of seven measures, the results of six measures contained more positive sentiment as compared to the negative. Negative polarity was dominant only in a single case. The results indicated an overall positive attitude of the people towards prevention policies implemented by the government to control the spread of the virus.

[12] conducted an Infodemiology study using a machine learning approach. Infodemiology is a research domain that utilizes unstructured textual data from the public available across internet media (social media, blogs, websites) for collection and analysis to inform and improve public health and policies related to public health [13]. The study was proposed to observe public discussions and sentiments regarding Covid-19. In addition to sentiment polarity, the researchers also identified discussion topics under five themes. The results revealed an emotion of 'fear' when the discussion topic was about Covid-19 new cases and mortality rate.

[9] analyzed sentiments from tweets during the Covid-19 outbreak in Nepal. Text mining and sentiment analysis was performed using the Google Colab platform. Python was utilized as the programming language. The results of the study indicated a mix of emotions like hope, fear, sadness, and disgust across Twitter data during the last ten days of May 2020.

[14] performed a study that targeted the identification of key topics posted by the public on Twitter related to Covid-19 starting from February until the mid of March 2020. The researchers employed Twitter API, Python libraries, and Postgre SQL database for data collection and analysis. Initially, tweet data and detailed metadata were obtained using Twitter API. Then, the tweets were analyzed by unigram and bigram word frequencies. Finally, the topics discussed in the tweets were identified by applying Latent Dirichlet Allocation (LDA). The results of this study recognized 12 topics under the main themes of virus origin, source, impact, and mitigation strategies. Out of a total of 12 topics, 10 topics exhibited a positive sentiment while the remaining 2 topics exhibited a negative sentiment.

III. Methodology and Experiment Details

The methodology for this work as presented in Fig. 1, starts with the collection of tweets from the Twitter API followed by tweet cleaning and analysis processes and finally ends with predicting the impact of the Covid-19 vaccine on public health. The keyword used for tweet collection is a combination of four words and is specified as '#covid-19+vaccine+public+health'. It results in the collection of tweets containing all of these words. For this experiment, three datasets of tweets have been collected. The first dataset has tweets from October 2020 – December 2020, while the second dataset contains tweets from January 2021 – June 2021. The third dataset is updated with tweets from July 2021 – October 2021. The tweets are analysed in four different ways: word frequency analysis, bigram and trigram analysis, sentiment analysis, and topic modeling.

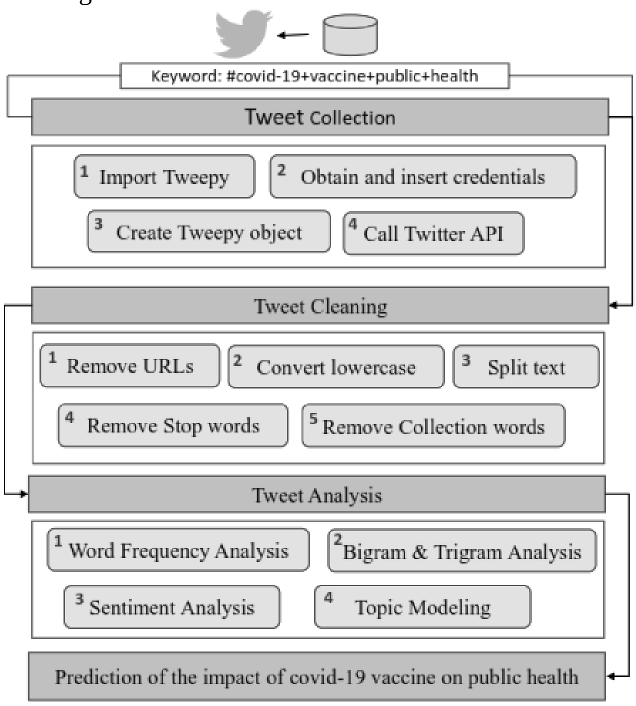

Fig. 1. Methodology for Tweet Collection and Analysis

The word frequency analysis, bigram, and trigram analysis have been conducted on the first dataset. Sentiment analysis has been accomplished on the first and second datasets. Topic modeling has been done with the second and third tweet dataset. On the second dataset, topic modelling has been implemented using Latent Dirichlet Allocation (LDA).On the third dataset, LDA and LSA (Latent Semantic Analysis), both approaches have been used to find the topic keywords.

A. Tweet Collection

The process of tweet collection requires importing the Tweepy Python package that provides a simple and easy approach to access Twitter API through Python programming language. The next step is to provide the required credentials for fetching data from Twitter. These credentials were obtained by creating an application in the Twitter Development Portal. After setting the credentials, the Tweepy object is created and used to call Twitter API for data collection relative to the specified keyword.

B. Tweet Cleaning

Tweet cleaning follows the tweet collection process. In the tweet cleaning process that is data pre-processing, first URLs were removed using a regular expression (re) package in Python, then tweets were converted to lowercase followed by splitting tweet text and removing stop words and collection words from the tweets. In the text splitting, the tweets are converted into a split list of strings. This is done by using the split () method in python. In a language, stop words are those words that can easily be ignored without affecting the meaning and information in a sentence. The stop word removal process makes tweet text more suitable and meaningful for analysis. The collection words are the keywords specified for tweet collection. We also removed collection words due to the reason that we know that these words are surely part of all the collected tweets so by removing these words from tweets we can focus on other important words that occurred in tweets to obtain useful insights.

C. Tweet Analysis

Finally, the clean tweets were analyzed for word frequency, bigrams and trigrams, and most importantly, the sentiments to predict the impact of the Covid-19 vaccine on public health. The word frequency analysis of tweets is done by flattening the clean list of tweet words and creating a counter for the unique words across the tweet words list. The bigram and trigram analysis extracts and counts the co-occurring words in clean tweets. A bigram is two co-occurring words sequence while a trigram is a triple co-occurring words sequence. The analysis is done by using the bigram and trigram functions from NLTK (Natural Language Tool Kit) library in Python. The sentiment analysis on both sets of tweets is performed using the TextBlob package in Python. TextBlob helps to identify the polarity of the tweets. The polarity value may be positive, neutral, or negative. The positive polarity indicates a positive sentiment towards the topic while a negative polarity indicates negative sentiment.

D. Topic Modeling

For a large corpus of textual data, topic modeling provides a way to organize and group the data in terms of topics identified from the data itself. A topic model is a model that is trained to identify and discover the topics in a document automatically. Topics are made up of a group of words that best describe the information contained in the data. Topics can also be defined as the repeating patterns of simultaneously occurring terms in the data. There are various methods available for obtaining topic models. One of the well-known methods is *Latent Dirichlet Allocation* (LDA) proposed by [15]. LDA is a generative probabilistic model that identifies various topics present in a corpus and also determines the degree of presence of certain topics in the documents contained in the corpus. To use LDA for topic modeling, the number of topics to be discovered in the corpus is specified. The algorithm then finds the hidden topics in the form of a probability distribution of a grouping of words over the vocabulary of the topic in the corpus. In this experiment, topics in tweets are modeled using the LDA technique in the Python programming language. The number of topics is specified as 10. pyLDAvis has been used to visualize the topic model and interpret the topics obtained from our clean tweet data. LDAvis [16] is an interactive visualization of topics discovered using the LDA method. An important characteristic of LDAvis is the relevance attribute that allows users to explore the relationships between topics and terms to get the most from the LDA fitted topic model. *Latent Sematic Analysis* (LSA) is a technique to identify groups of words with same semantics from the input text. The classification part is done using TF/IDF (term frequency/inverse document frequency) and the dimensions are reduced using Singular Value Decomposition (SVD). The basic intuition behind LSA is that the words that share similar meaning tend to share similar contexts [17].

IV. RESULTS OF ANALYSIS

The word frequency analysis of tweets gives an illustrative overview of what is being discussed by the public. Secondly, it provides an opportunity for an in-depth analysis of the commonly occurring data frequencies. Fig. 2 presents the frequencies of the 25 most commonly occurring words in public discussions on Twitter. Here, we note that the words 'the', 'covid-19', 'vaccine', 'public', and 'health' have the highest frequencies of occurrence.

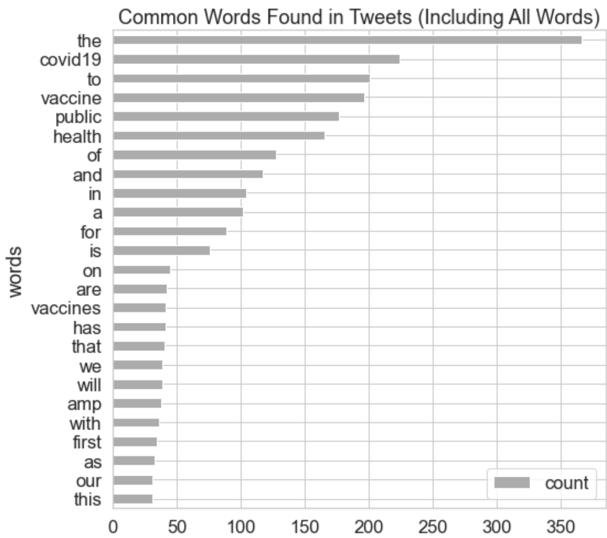

Fig. 2. Word Frequency Analysis of top 25 common words including all words in tweets

The other commonly occurring words are mostly the stop words that are not very useful in this analysis so, in the next step, we removed stop words and then analyzed the word frequencies. Fig. 3 shows word frequencies when the stop words (like the, to, an, and, of, etc) are excluded. Now, we observe that in addition to collection words, people are discussing vaccine doses, Moderna, Pfizer, experts, distribution, etc. Next, we excluded the collection words (like covid-19, vaccine, public, health) from the analysis because we know that these words are found in all the collected tweets and it is useful to know other words. The result of this analysis is presented in Fig. 4, where we find further vocabulary in the discussion by the public, for example, pandemic, healthcare, doses, received, Pfizer, Moderna, amp, coronavirus, etc.

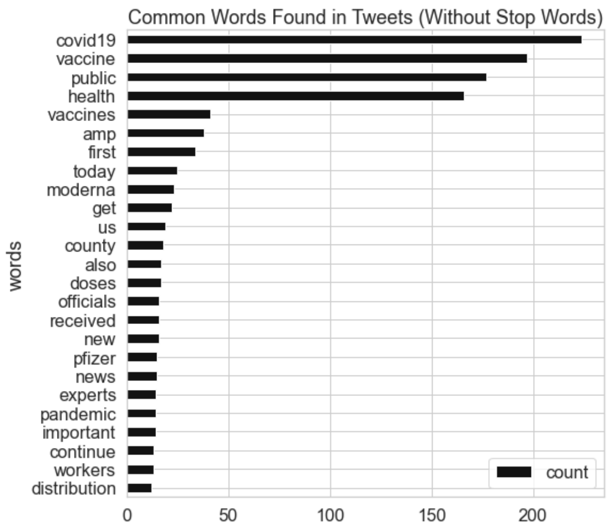

Fig. 3. Word Frequency Analysis of top 25 common words excluding stop words

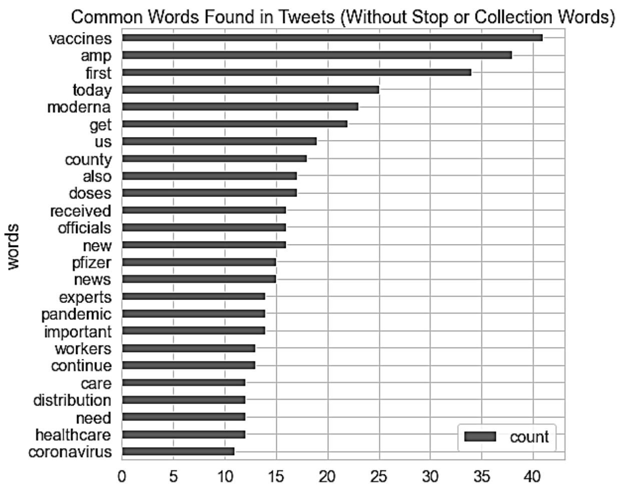

Fig. 4. Word Frequency Analysis of top 25 common words excluding stop words and collection words

Table I presents top 10 preprocessed and clean tweets from 1000 collected tweets. The topics about Covid-19 vaccine delivery, distribution, effects, public health are being talked about. Table II exhibits 19 commonly occurring bigrams in the 1000 collected tweets and Fig. 6 gives a visualization of a network of 30 bigrams.

TABLE I
TOP 10 FROM 1000 COLLECTED TWEETS AFTER PREPROCESSING

No	Tweet
1	Its arrived Taking delivery of the Pfizer BioNTech vaccine into our Cold Chain Storage this morning We can now ha
2	This is the head of the Croatian Institute for Public Health posing with the first case of COVID19 vaccine that ar
3	This is the Equity Matrix that is refd in the COVID19 vaccine rollout recommendations I hope all provinces
4	this nprfreshair on corona covid19 w edyong209 discusses 1 first response public health is severely underfun
5	COVID19 The hype amp hullabaloo over the availability amp distribution of the COVID19Vaccine in India have been gat
6	These rushed vaccines and their campaigns will go down in history as a great disasterOur first responders shoul
7	So far only seven states have followed public health officials advice to specifically prioritize incarcerated peo
8	Maximizing the Uptake of a COVID19 Vaccine in People With Severe Mental IllnessA Public Health Priority
9	Resources for healthcare professionals about receiving the vaccine and communicating confidently about it to patien
10	Riverside University Health System Public HealthIf you have a concern or have been contacted for what you belie

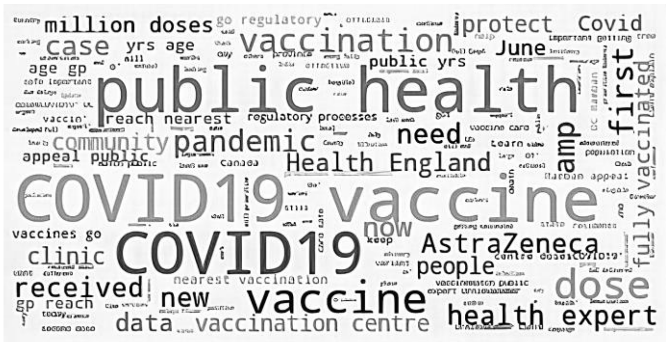

Fig. 5. Word Cloud plot of tweet data

TABLE II
BIGRAMS FROM THE COLLECTED TWEETS

No	Bigram	Count
1	('two', 'doses')	30
2	('said', 'aware')	21
3	(‘aware’,’handful’)	21
4	(‘state’,’department’)	20
5	(‘department’,’said’)	20
6	(‘handful’,’teens’)	20
7	(‘teens’,’staffers’)	20
8	(‘staffers’,’received’)	20
9	(‘doses’,’needed’)	17
10	('distribution', '#covid19vaccine')	4
11	('moderna', '#covid19')	10
12	('doses', '#covid19')	9
13	('supply', 'arrive')	6
14	('first', 'doses')	8
15	('findings', 'important')	8
16	('important', 'vaccines,')	8
17	('vaccines,', 'therapeutics')	8
18	('therapeutics', 'strategies')	8
19	('strategies', 'combat')	8

Table III reveals 20 trigrams and their counts from the collected tweets. The network of 30 trigrams can be visualized in Fig. 7. Table IV lists the polarity value along with the corresponding tweet. It can be noted that out of 15 listed tweets, 10 tweets have a positive polarity value and therefore, exhibit a positive sentiment while 5 tweets have a negative polarity revealing negative sentiments.

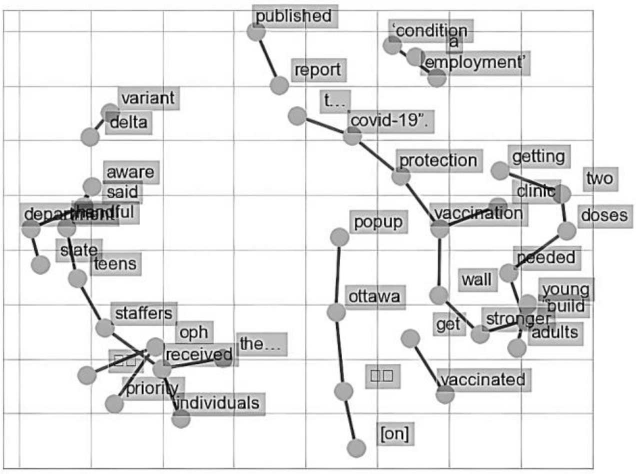

Fig. 6. Bigram Network Visualization

TABLE III
TOP 20 TRIGRAMS FROM THE COLLECTED TWEETS

No	Trigram	Count
1	('said', 'aware', 'handful')	21
2	('state', 'department', 'said')	20
3	('department', 'said', 'aware')	20
4	('aware', 'handful', 'teens')	20
5	('handful', 'teens', 'staffers')	20
6	('teens', 'staffers', 'received')	20
7	('getting', 'two', 'doses')	16
8	('two', 'doses', 'needed')	16
9	('doses', 'needed', '"build')	16
10	('needed', '"build', 'stronger')	16
11	('doses', '#covid19', 'vaccine')	8
12	('findings', 'important', 'vaccines,')	8
13	('important', 'vaccines,', 'therapeutics')	8
14	('vaccines,', 'therapeutics', 'public')	8
15	('therapeutics', 'public', 'health')	8
17	('health', 'strategies', 'combat')	8
18	('strategies', 'combat', 'covid-19')	8
19	('combat', 'covid-19', 'pandemic…')	8
20	('first', 'doses', '#covid19')	7

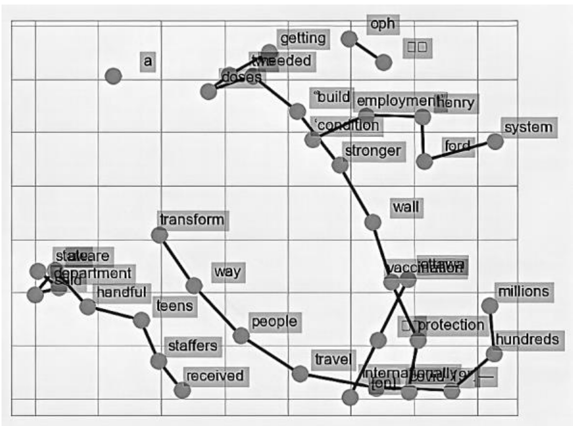

Fig. 7. Trigram Network Visualization

TABLE IV
TWEETS POLARITY/SENTIMENT FROM TOP 15 TWEETS

No	Polarity Value	Tweet
1	-0.231	vaccine crisis in LMICsCOVID19 has disrupted the care of other diseases much more common and devastating through
2	0.5	Its been tested its safe and when its my turn I will be taking the COVID19 vaccinePremier Joe Savik
3	-0.6	Its arrived Taking delivery of the Pfizer BioNTech vaccine into our Cold Chain Storage this morning We can now ha
4	0.125	This is the head of the Croatian Institute for Public Health posing with the first case of COVID19 vaccine that ar
5	0.125	this nprfreshair on corona covid19 w edyong209 discusses 1 first response public health is severely underfun
6	0.25	Reuters Actions speak louder than wordsCrown prince Mohammed bin Salman taking his first dose of the COVID19
7	0.5	613 HotSpotting is dramatically more efficient than uniform allocation We conclude hotspotting could enable pu
8	0.25	Actions speak louder than wordsCrown prince Mohammed bin Salman taking his first dose of the COVID19 vaccine
9	0.033	So far only seven states have followed public health officials advice to specifically prioritize incarcerated peo
10	-0.05	Maximizing the Uptake of a COVID19 Vaccine in People With Severe Mental IllnessA Public Health Priority
11	-0.25	The Department of Health reminds the public that it is illegal to sell COVID19 vaccine doses COVID19VaccinePHR
12	0.05	If you are making wildly disparaging comments about the vaccine and have no public health expertise you may be r
13	-0.25	The UnitedStates has been Bangladeshs closest partner for the past five decades working to improve
14	0.25	As wealthy nations claim the majority of vaccines public health experts warn of a COVID19 surge in Africa
15	0.06	Delta variant is not as deadly as the original CV19 Data from Public Health England show that there were 117 de

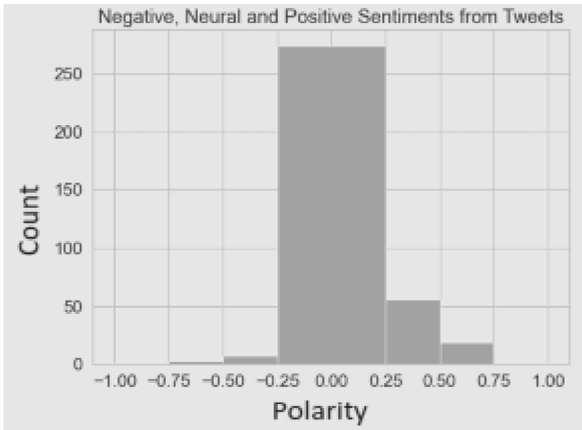

Fig. 8. The plot of Sentiment Analysis with tweet polarity values across the x-axis and tweet count across the y-axis (Tweet set 1 – October 2020 to December 2020)

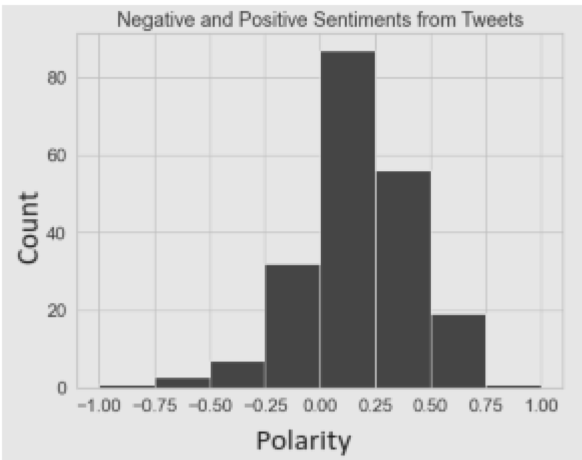

Fig. 9. The plot of Sentiment Analysis with tweet polarity values across the x-axis and tweet count across y-axis (Neutral tweets are excluded for clear visualization of negative and positive sentiments) (Tweet set 1 – October 2020 to December 2020)

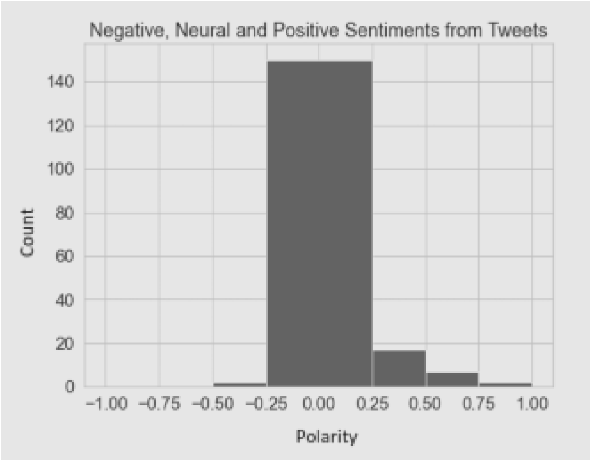

Fig. 10. The plot of Sentiment Analysis with tweet polarity values across the x-axis and tweet count across the y-axis (Tweet set 2 – January 2021 to June 2021)

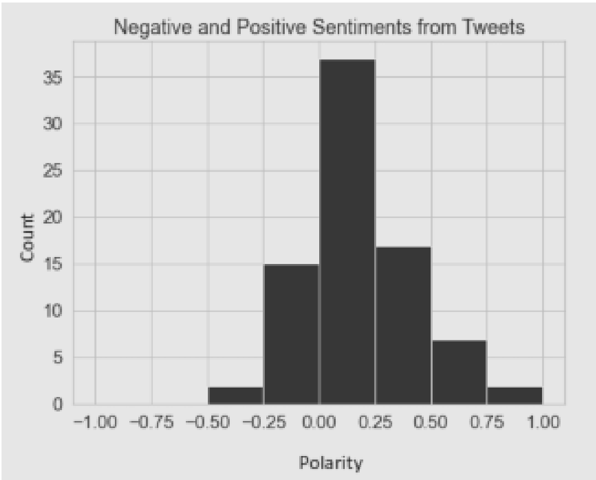

Fig. 11. The plot of Sentiment Analysis with tweet polarity values across the x-axis and tweet count across y-axis (Neutral tweets are excluded for clear visualization of negative and positive sentiments) (Tweet set 2 – January 2021 to June 2021)

Fig. 8 and Fig. 9 show the plots of sentiments from tweet set 1. From Fig. 8, it can be observed that the number of neutral tweets ranges between 500 and 600, positive between 50 and 100, and negative tweets are less than 20. In Fig. 9, neutral tweets have been eliminated to allow for a clear picture of the negative and the positive sentiment. Fig. 10 and Fig. 11 depict the plots of sentiments for tweet set 2. It is observed that in both sets of tweets, positive sentiment is dominant. Overall, we conclude that the opinions and discussions of the people are positive about the Covid-19 vaccine in relation to public health.

In Fig. 12, the global topical view of the tweet data is presented. Since the number of topics to be discovered is specified equal to 10 while writing the Python code, LDA has discovered 10 topics represented by the circles plotted along a two-dimensional plane. The area of the circle defines the prevalence of the topics. For example, in Fig. 12, topic 1 is the most commonly discussed topic while topic 10 is the least. The circle colored dark represent topic 1. Fig. 13 to Fig. 18 present bar charts for interpreting topic 1 with varying values of the relevance metric λ. The bar charts have overlaid bars. The light grey bars give the corpus-wide frequency of the terms while the dark grey bars give the topic-specific frequency.

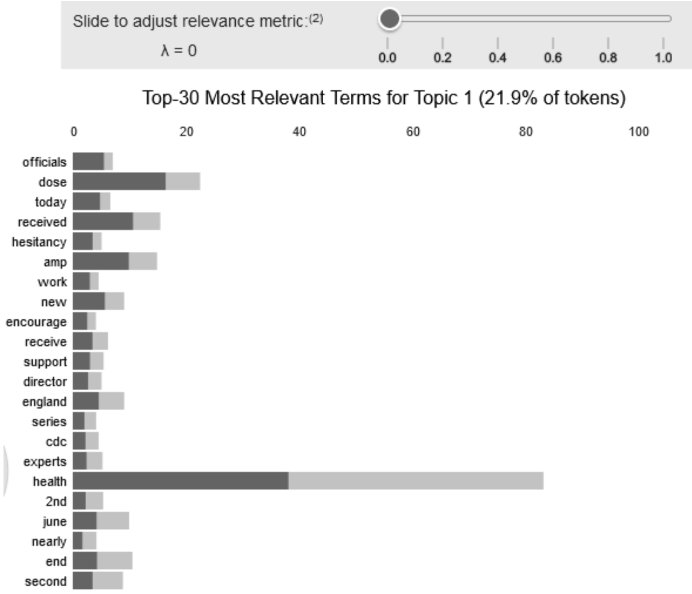

Fig. 13. Most useful terms for interpreting topic 1 with the relevance metric λ=0

The relevance metric λ determines the ranking of the usefulness of the terms exclusive to the topic. [16] found the optimal value of λ to be 0.6 based on an experiment that resulted in 70 percent chance of correct topic identification. For values of λ close to 0, the estimated probability of correct topic identification was found to be 53 percent while for values of λ close to 1, it was around 63 percent. The ratio of the width of the dark grey and light grey bars indicate clearly whether a term is a common term or it is more relevant to the selected topic.

We observe in Fig. 14, with λ=0.2, the top three terms are ‘dose’, ‘received’, ‘officials’. Comparing these terms, the term ‘officials’ is most exclusive to this topic. The topic can be interpreted as related to the vaccine administration.

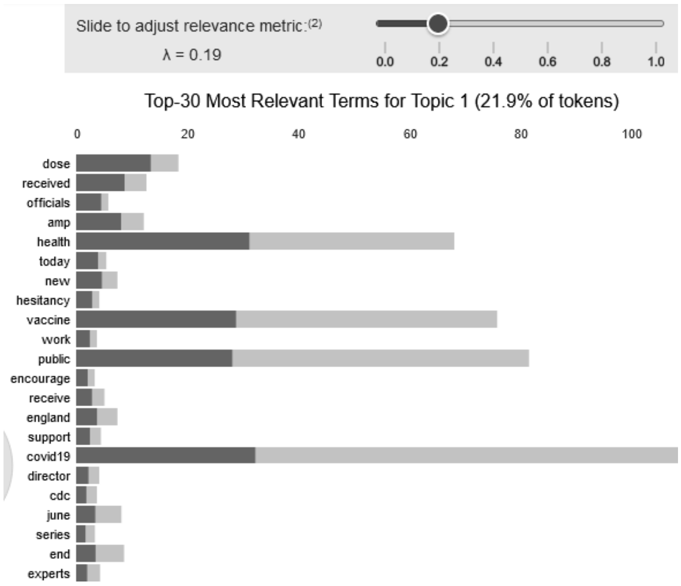

Fig. 14. Most useful terms for interpreting topic 1 with the relevance metric λ=0.2

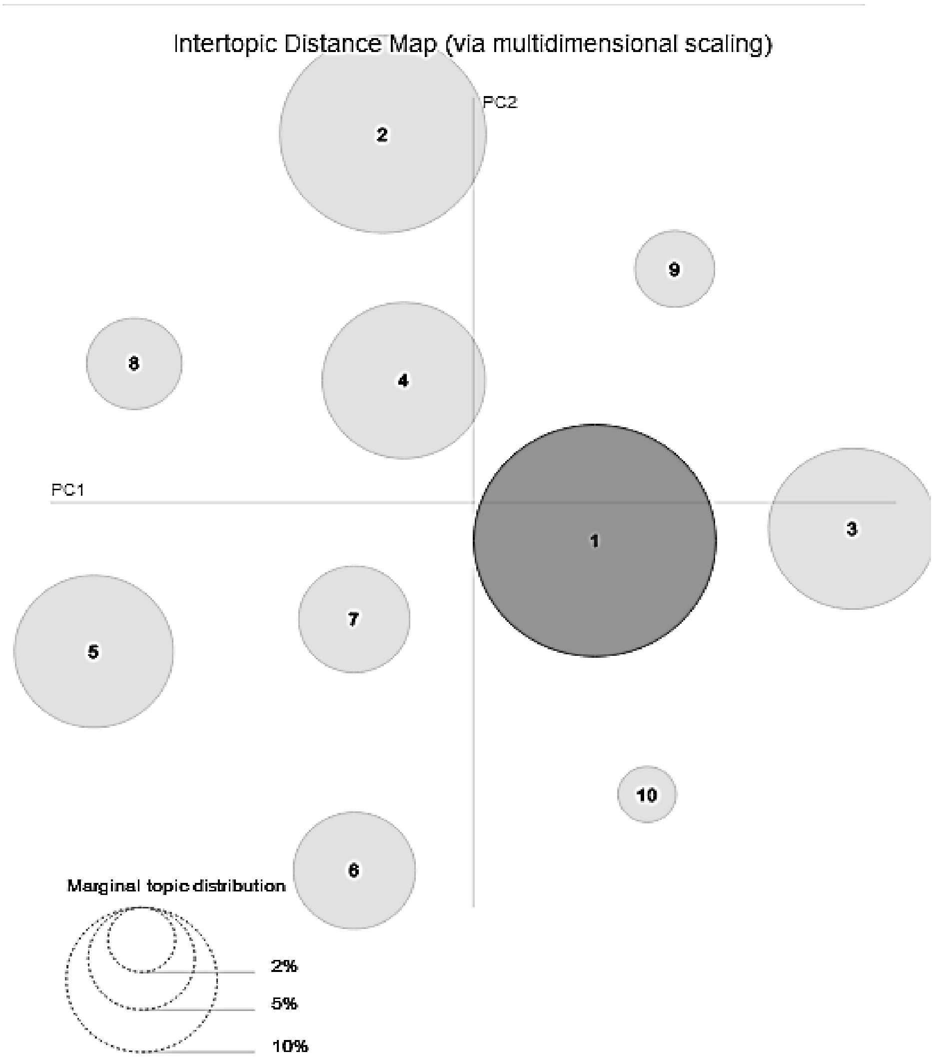

Fig. 12. The global topic view using pyLDAvis

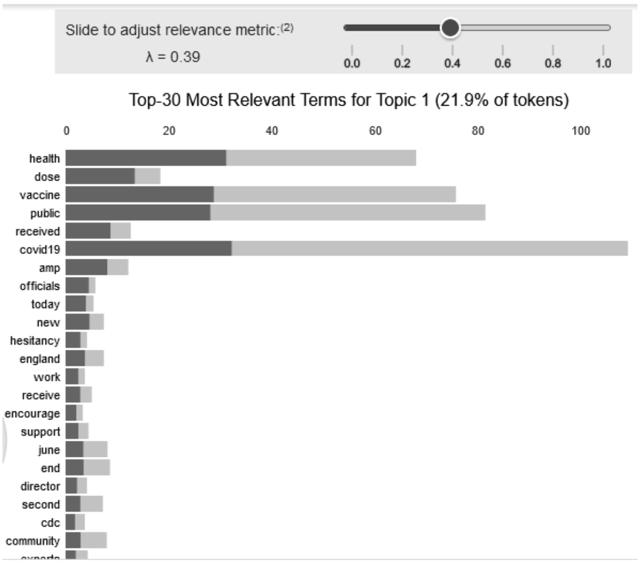

Fig. 15. Most useful terms for interpreting topic 1 with the relevance metric λ=0.4

From Fig. 15, λ=0.4 and the top three terms are ‘health’, ‘dose’, ‘vaccine’. As appears from the bar chart, the terms ‘health’ and ‘vaccine’ are common terms across the tweet dataset that are also belonging to topic 1 while the term ‘dose’ best describes this topic. So, we can interpret that the discussion is related to ‘vaccine doses’.

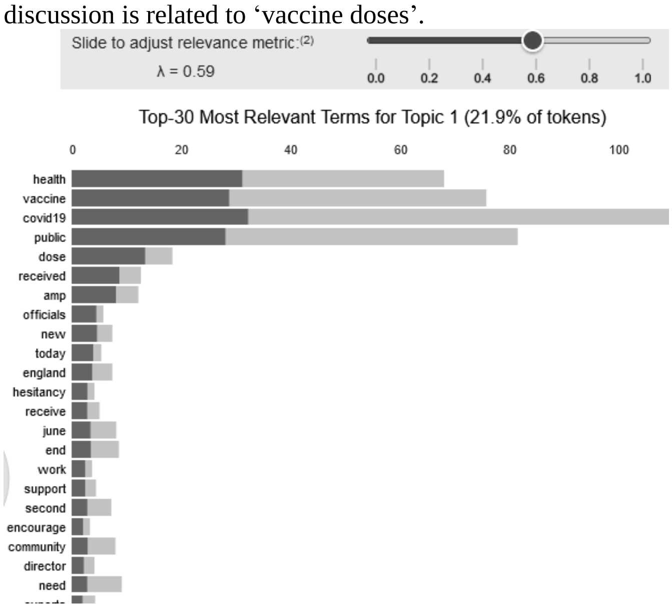

Fig. 16. Most useful terms for interpreting topic 1 with the relevance metric λ=0.6

In Fig. 16, the value of λ is set to 0.6, here we exclude the first four terms in our interpretation because these terms represent the keywords used to collect the tweet dataset. Following these four terms, the top 3 are ‘dose’, ‘received’, ‘amp’. Again, we interpret the topic as related to ‘administration of vaccine doses’.

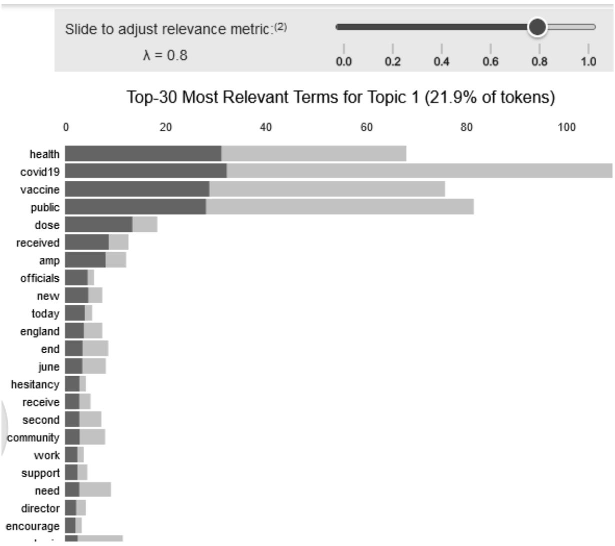

Fig. 17. Most useful terms for interpreting topic 1 with the relevance metric λ=0.8

It is observed from Fig. 17 and Fig. 18 that with λ=0.6 and 0.8. There is not much difference in the top topic terms when compared with Fig. 16 with λ=0.6, so topic 1 is interpreted to be about ‘administration of vaccine doses’. The rest of the topics determined by LDA are interpreted by plotting the word cloud.

Table V summarizes the topics discovered from tweet dataset 2 and Table VI compares five keywords from each of the five topics computed using LDA and LSA on tweet dataset 3.

TABLE V
TOPICS DISCOVERED FROM TWEETS (TWEET SET 2)

No.	Topic
1	Administration of vaccine doses
2	Vaccination of young residents to combat virus
3	Experts opinion on covid 19 immunization
4	Vaccination of state population
5	Health clinics in Ramban 2021
6	Expert opinions on public health during June
7	Administration of the second dose
8	Scientific discussion on covid 19 and vaccine
9	Vaccine delivery in England
10	Myocarditis and pericarditis

TABLE VI
TOPIC KEYORDS USING LDA & LSA (TWEET SET 3)

Topic	Latent Dirichlet Allocation (LDA)	Latent Sematic Analysis (LSA)
1	vaccine, achieve, milestone, govaxx,10k	vaccine, Texas, kids, Friday, share
2	vaccinated, weekly, holding, leaders, reminder	public, October, national, community, information
3	cases, positive, tests, daily, updateoct	health, Texas, minister, leaders, govt
4	vaccines, booster, today, clinic, new	doses, local, refusing, good, information
5	mandates, govabbott, Texas, republican, says	govaxx, partner, 10k, achieve, milestone

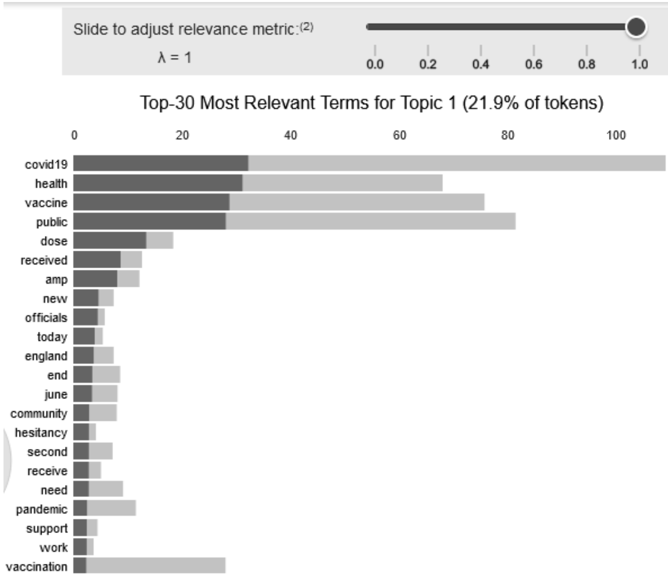

Fig. 18. Most useful terms for interpreting topic 1 with the relevance metric λ=1

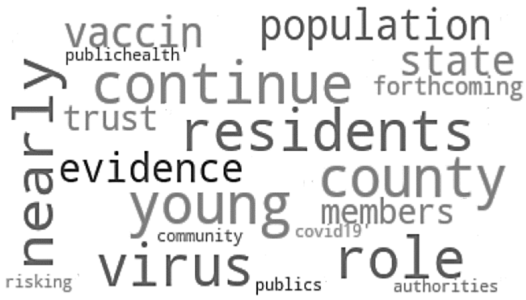

Fig. 19. Word cloud for 20 frequent terms in topic 2

From Fig. 19, we interpret the focus of discussion in topic 2 as ‘vaccination of young residents to combat virus’.

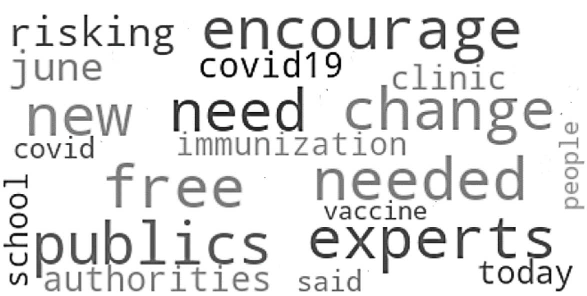

Fig. 20. Word cloud for 20 frequent terms in topic 3

From Fig. 20, we interpret topic 3 as opinions of ‘experts on covid 19 immunization’.

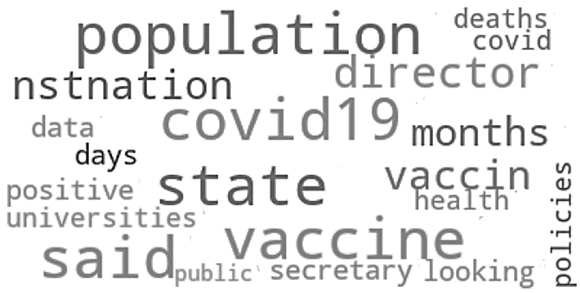

Fig. 21. Word cloud for 20 frequent terms in topic 4

From word cloud in Fig. 21, topic 4 is interpreted to be ‘vaccination of state population’.

Fig. 22. Word cloud for 20 frequent terms in topic 5

From word cloud in Fig. 22, topic 5 is interpreted to be ‘health clinics in Ramban 2021’

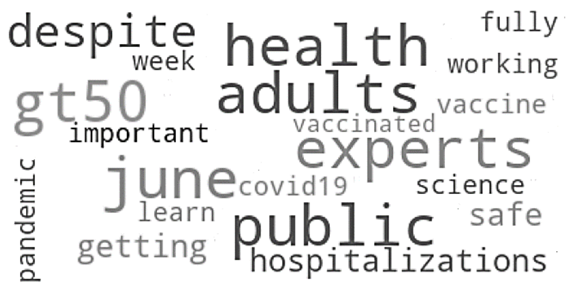

Fig. 23. Word cloud for 20 frequent terms in topic 6

From word cloud in Fig. 23, topic 6 is interpreted to be ‘expert opinions on public health in June’.

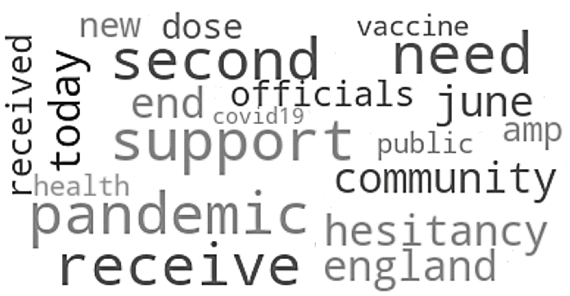

Fig. 24. Word cloud for 20 frequent terms in topic 7

From word cloud in Fig. 24, topic 7 is interpreted to be ‘administration of the second dose’

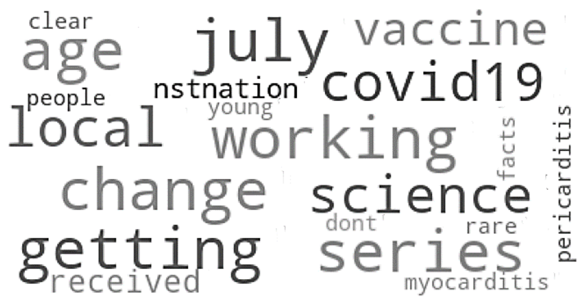

Fig. 25. Word cloud for 20 frequent terms in topic 8

From word cloud in Fig. 25, topic 8 is interpreted to be ‘scientific discussion on covid 19 and vaccine’

Fig. 26. Word cloud for 20 frequent terms in topic 9

From word cloud in Fig. 26, topic 9 is interpreted to be ‘vaccine delivery in England’.

Fig. 27. Word cloud for 20 frequent terms in topic 10

From word cloud in Fig. 27, topic 10 is interpreted to be ‘myocarditis and pericarditis.

V. Discussion

The word frequency analysis helps in identifying the patterns of words that are trending in the discussions on Twitter and other media and can also reflect the Twitter users’ psychology [18] under particular circumstances, for example, pandemic, natural disasters, fire, etc. In this research, the word frequency analysis gives an insight into the most common words uttered by the people about the Covid-19 vaccine and public health. From the analysis, the most common words found in tweets included Covid-19, vaccine, public, health, pandemic, Pfizer, Moderna, healthcare, etc.

Bigrams and trigrams provide further insight into the topic by finding out two and three co-occurring words in the tweets. The three common bigrams were found to be ('distribution', '#covid19vaccine'), ('moderna', '#covid19') and, ('doses', '#covid19'). The three common trigrams were ('doses', '#covid19', 'vaccine'), ('findings', 'important', 'vaccines') and, ('important', 'vaccines', 'therapeutics'). In [16], the word frequencies including unigram, bigram and, trigram were represented using power law distribution statistical technique.

Sentiment analysis is a well-known technique in Natural Language Processing (NLP) research. Sentiment analysis starts with text data as input followed by the identification phase, feature selection process, sentiment classification, and finally sentiment polarity and subjectivity determination [11]. In this research, sentiment analysis is done using Python library package TextBlob. TextBlob uses the ‘Naïve Bayes’ classification algorithm [9]. In addition to some statistical techniques, several studies utilized ‘Support Vector Machines’ and ‘Artificial Neural Networks’ for analyzing sentiments [18]. [19] applied deep learning to the tweet classification task and inferred that tweet classification using Convolutional Neural Network (CNN) yielded a superior result as compared to the Naïve Bayes classifier.

Topic Modeling has many implications including information retrieval, summarization, discussion analysis, etc. The most commonly used techniques for topic modeling are Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA). [20] evaluated various topic modeling algorithms on tweet data including LSA, LDA, LDA-U, Biterm Topic Model (BTM), and word2vec Gaussian Mixture Model. They determined that LDA might not perform very well on short text documents and BTM gains performance improvement on LDA when specifically working with tweets. [21] compared LSA and LDA performance on tweet datasets and indicated that the results of LDA are better as compared to LSA.

The results of topic modeling in this experiment indicate the topics that are currently prevailing like administration of vaccine doses, second doses, vaccination for the young population, vaccine delivery, etc. The data obtained from Twitter about the Covid-19 vaccine and public health can be utilized by the concerned agencies in the healthcare sector in several useful ways. For example, negative sentiments can be further analyzed to identify the problems and reasons for peoples' negative attitudes. When the problems are identified, concerned agencies can be informed, and steps can be taken in real-time to overcome those difficulties. In this case, if the people are talking negatively about the vaccine, there might be a possibility that vaccine education and communication are lacking to those people or there may be financial issues with them. The government can take appropriate measures to deal with such issues efficiently and effectively. Secondly, due to data limitations, user location is not considered in this analysis. The user location together with tweet polarity can be used to identify the regions from where positive and negative sentiments are being received.

Our tweet sentiment analysis on both tweet datasets indicated a large number of neutral sentiments, a moderate number of positive sentiments, and a relatively small number of negative sentiments. Taking only the negative and positive sentiments into account, from this analysis we predict a positive impact of the Covid-19 vaccine on public health because more positive tweets as compared to the negative, indicate a healthy discussion and show public awareness and acceptance of the Covid-19 vaccine.

VI. Conclusion

The research concludes by discovering topics in tweets related to the Covid-19 vaccine and public health as well as predicting a positive impact of the Covid-19 vaccine on public health employing sentiment analysis using Twitter social media as the data collection platform. The research also provides useful suggestions for using Twitter data for better management of strategies relating to Covid-19 vaccine awareness, information, education, and public acceptance. Twitter, therefore proves to be a useful source of knowledge and information for public healthcare officials that makes it possible to formulate better healthcare policies for an overall improvement of public health.

In the future, the research work can be extended by employing deep learning methods or other classification algorithms for sentiment analysis. This work utilized LDA for topic modeling. In the future, other methods like Biterm Topic Model (BTM) that have been proved to be effective for tweet topic modeling might be utilized for better results.

References

[1] J. Roozenbeek, C. R. Schneider, S. Dryhurst, J. Kerr, A. L. J. Freeman, G. Recchia, A. M. Bles, and S. Linden, "Susceptibility to misinformation about COVID-19 around the world," Royal Society Open Science, vol. 7, no. 10, pp. 1-15, 2020, doi: 10.1098/rsos.201199

[2] S. C. L. Kamerlin and P. M. Kasson, "Managing Coronavirus Disease 2019 Spread With Voluntary Public Health Measures: Sweden as a Case Study for Pandemic Control," Clinical Infectious Diseases, pp. 1-8, 2020, doi: 10.1093/cid/ciaa864

[3] M. Sheek-hussein and F. M. Abu-zidan, "COVID-19 Vaccine : Hope and reality," vol. 20, no. 4, pp. 1507–1509, 2020.

[4] W. S. Chou and A. Budenz, "Considering Emotion in COVID-19 Vaccine Communication: Addressing Vaccine Hesitancy and Fostering Vaccine Confidence," Health Communication Journal, vol. 35, no. 14, pp. 1718–1722, 2020, doi: 10.1080/10410236.2020.1838096

[5] S. M. Jungmann and M. Witthöft, "Health anxiety, cyberchondria, and coping in the current COVID-19 pandemic: Which factors are related to coronavirus anxiety?," *Journal of Anxiety Disorders*, vol. 73, 2020, doi: 10.1016/j.janxdis.2020.102239

[6] M. Poljak and R. Norrby, "The impact of vaccines on public health: The role of ESCMID," Clinical Microbiology and Infection, vol. 20, no. S5, 2014, doi: 10.1111/1469-0691.12598

[7] W. Budiharto and M. Meiliana, "Prediction and analysis of Indonesia Presidential election from Twitter using sentiment analysis," Journal of Big Data, vol. 5, no. 1, pp. 1–10, 2018, doi: 10.1186/s40537-018-0164-1

[8] O. Y. Adwan, M. Al-Tawil, A. M. Huneiti, R. A. Shahin, A. A. Abu Zayed, and R. H. Al-Dibsi, "Twitter sentiment analysis approaches: A survey," International Journal of Emerging Technologies in Learning, vol. 15, no. 15, pp. 79-93, 2020, doi: 10.3991/ijet.v15i15.14467

[9] B. P. Pokharel, "Twitter Sentiment analysis during COVID-19 Outbreak in Nepal," Nepal Open University, 2020.

[10] E. Psomakelis, K. Tserpes, D. Anagnostopoulos, and T. Varvarigou, "Comparing methods for twitter sentiment analysis," Proceedings of the International Conference on Knowledge Discovery and Information Retrieval, pp. 225–232, 2014, doi: 10.5220/0005075302250232

[11] K. H. Manguri, R. N. Ramadhan, and P. R. Mohammed Amin, "Twitter Sentiment Analysis on Worldwide COVID-19 Outbreaks," Kurdistan Journal of Applied Research, pp. 54–65, 2020, doi: 10.24017/covid.8.

[12] J. Xue, J. Chen, R. Hu, C. Chen, C. Zheng, Y. Su, and T. Zhu, "Twitter discussions and emotions about COVID-19 pandemic: a machine learning approach," vol. 22, no. 11, 2020, Available:http://library1.nida.ac.th/termpaper6/sd/2554/19755.pdf

[13] G. Eysenbach, "Infodemiology and infoveillance: Tracking online health information and cyberbehavior for public health," American Journal of Preventive Medicine, vol. 40, no. 5, pp. S154–S158, 2011, doi: 10.1016/j.amepre.2011.02.006

[14] A. Abd-Alrazaq, D. Alhuwail, M. Househ, M. Hai, and Z. Shah, "Top concerns of tweeters during the COVID-19 pandemic: A surveillance study," Journal of Medical Internet Research, vol. 22, no. 4, pp. 1-9, 2020, doi: 10.2196/19016

[15] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," Journal of Machine Learning Research, vol. 3, pp. 993–1022, 2003, doi: 10.1016/b978-0-12-411519-4.00006-9

[16] C. Sievert and K. Shirley, "LDAvis: A method for visualizing and interpreting topics," Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces, pp. 63–70, 2014, doi: 10.3115/v1/w14-3110

[17] T. Cvitanic, B. Lee, H. I. Song, K. Fu, and D. Rosen, "LDA v. LSA: A comparison of two computational text analysis tools for the functional categorization of patents," CEUR Workshop Proceedings, vol. 1815, pp. 41–50, 2016.

[18] N. K. Rajput, B. A. Grover, and V. K. Rathi, "Word Frequency and Sentiment Analysis of Twitter Messages During Coronavirus Pandemic," *arXiv*, 2020. Available: https://arxiv.org/pdf/2004.03925.pdf

[19] S. Hashida, K. Tamura, and T. Sakai, "Classifying Tweets using

Convolutional Neural Networks with Multi-Channel Distributed Representation," IAENG International Journal of Computer Science, vol. 46, no. 1, pp. 68-75, 2019, doi: 10.1109/SMC.2018.00041

[20] E. Jonsson and J. Stolee, "An Evaluation of Topic Modelling Techniques for Twitter," *arXiv*, 2016. Available: https://www.cs.toronto.edu/~jstolee/projects/topic.pdf

[21] S. Qomariyah, N. Iriawan, and K. Fithriasari, "Topic modeling Twitter data using Latent Dirichlet Allocation and Latent Semantic Analysis," AIP Conference Proceedings, vol. 2194, 2019, doi: 10.1063/1.5139825