Covid-19 vaccination – public discussion around the world

Nitika Sharma
ITEC 5205W(2021), Design and development of Data-Intensive Application
(*nitikasharma3@cmail.carleton.ca*)

Submitted to:
Dr. M. Omair Shafiq, PhD
Assistant Professor (School of Information Technology)
Faculty of Engineering and Design, Carleton University
(*omair.shafiq@carleton.ca*)

Abstract—**The study aims to achieve the public discussion about the Covid-19 vaccination around the world through Twitter and display the comparison of public discussion from two different time duration i.e. December 2020 vs April 2021. Also, the study focuses on deriving a think-aloud summary table with the necessary details such as, percentage of positive and negative tweets, total number of tweets and total number of re-tweets of different cities and countries around the world, for giving a better visualization to obtain the difference in the public discussion of different regions. . Method: Extracted the Twitter data using Tweepy and comparing the same with the already available open dataset through Kaggle. Further, the research performed sentiment analysis using TextBlob to analyze the textual data. Conclusion: Identified and created a think-aloud summary table, using tweets from an open dataset of 2021, with the important keywords, hashtags and most frequently used words along with the percentage of positive & negative Tweets. Also, the table included the total number of actual tweets and the re-tweets. Obtained the comparison and difference of the datasets of two different time-periods (Dec 2020 and initial few weeks of April 2021) and displayed the results using scatter plot and the word cloud for efficient data visualization. Future Work: More analysis on the discussion in different languages from Twitter.**

Index Terms— **Covid-19, Covid-19 vaccination, datasets, open dataset, public discussions, python, sentiment analysis, Tweepy, Twitter**

I. INTRODUCTION

COVID-19(CORONAVIRUS) is discovered as one of the most infectious diseases around the globe [22]. Since November 2019, the entire world has been terrified and severely affected by covid-19. After the confirmation of the first positive coronavirus case in Wuhan (China), it got transmitted to almost every country resulting in a large number of infected populations and eventually, an increasing death rate [21]. For the prevention and to reduce the transmission of the virus, WHO recommended maintaining the social distancing, washing hands frequently, or use an alcohol-based rub [22]. As per WHO, there are confirmed cases: 137,541,598 confirmed deaths 2,960,777, and 223 countries and territories are affected by the virus to date (April 14th, 2021, 1:40 pm CET) which is changing at every minute [6]. The segregated data is displayed in Figure1. [6].

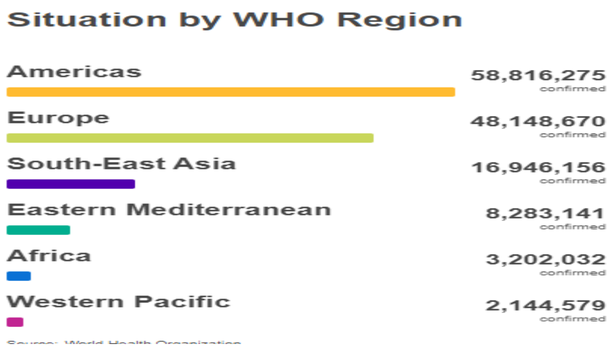


Figure 1. [6]

Vaccination of Covid-19 is a much-awaited thing in this pandemic. After a long waiting tenure, a set of companies was finally successful in launching the vaccination which is a stepping stone to ease out this pandemic situation. A total of 249,160,837 doses of vaccine have been registered as per WHO [6]. But, as this is a global pandemic, the entire population around the world is impacted by this and the availability of vaccination for every one of them is the need of the hour. It is crucial to recruit a strategic process so that the vaccine efficacy can be demonstrated properly and at a quicker rate [15]. The public reaction to the vaccination trials, their myths regarding its side-effects, its after-effects, and many more queries regarding the procedure needs to be treated with immense importance for the smooth vaccination drive. Similarly, the health and government officials need to know the public's reaction so that they could do the needful. For serving this purpose, social media is used as a frequent medium to analyze the public opinion trend. And Twitter is considered one of the most crucial platforms for public discussion of the pandemic. As per information of the US, 68% of the American adults claim to get the news about Covid-19 from social media. Therefore, this paper includes all the necessary data about the public discussions which are important for tracking the Covid-19 vaccination progress, as it is of utmost importance to analyze and calculate the estimated level of protection against the virus, and how the vaccination is distributed and utilized by the global population. Also, with the help of proper data extraction, data analysis, and data visualization, it will become easier to keep a track of the fulfillment and further requirement of the vaccination procedure. As a result, it can be quickly identified which section of the population is at more risk and can face severe consequences of the infection and need the vaccination on a prior basis, such as frontline health workers and elderly people [15].

II. MOTIVATION

a) Covid-19 has such a great impact on the world that it forced people to lock themselves up in their respective houses. The only possible solution to get rid of this problem is the vaccination so that life can get back to its normal schedule. Also, there is a lot more data over social media that is vital and can be attained as that's the place where most of the discussion takes place; especially about health care concerns (Twitter is one of the examples). Therefore, the development and management of a suitable secure dataset are essential for the regulatory approval and acceptance by the public of any new vaccine [15]. It becomes important to monitor and capture the trends on social media to get a good grip in this area. The decision about which sector and which age group of people will need the vaccination on priority and how its availability can be easily achieved can be determined conveniently. As the process of vaccination has begun in phases, with the updated public opinion about it will help the government and health officials to answer several public queries, put the desired information on social media that can align their approach towards the vaccination procedure, and help them build an efficient plan to carry on the process. As, the more the acceptance rate of the vaccination by the public, the more the world will step closer to become a coronavirus-free world.

b) The existing literature review highlights the importance of public opinion that turned out to be essential for fighting the pandemic. As the vaccination process is initiated in the entire world in phases, several concerns of the population are to be analyzed for an efficient campaign. This research is useful because the collected, analyzed and updated data will provide valuable insights into this entire vaccination field and will also help in predicting the future activities that are essential for its development. The aim is to not just only control the current global pandemic situation but also provide the entire population of the work with effective long-term and stable immunization strategies against this virus. The discussion about the success rate of the vaccination process will give a piece of complete information about many important factors, such as current infected rate, death rate, recovery rate, number of vaccinations given, etc.

.

III. LITERATURE SURVEY AND COMPARATIVE ANALYSIS

Overview: A broad review of different types of research and analysis methods is conducted to identify the impact of Covid-19 on the world through public discussion on social media (especially Twitter). Since the period of last year and a half, various public reaction trends have been analyzed that helped the health and government officials to act accordingly so that the difficult situation of the pandemic could be handled with efficiency.

The existing research focuses on the current number of cases and deaths due to Covid-19 and gathered the epidemiological data and used different analysis methods to visualize public perception trends [8]. Tsao et al., [19] used the machine learning method to identify the change in public attitude and the government's attitude toward infodemics to help the health care agencies to design desired interventions that are necessary during this pandemic. It also displayed the importance of accurate information that helps in handling the spreading of unwanted and incorrect information. The advantage of social media and its real-time information is also mentioned in Depoux et al., [7]. As, nowadays, social media is also used for spreading inaccurate data. Different methods have been opted by different works of literature to uniquely identify the patterns and trends of public opinions and discussions. Taking this into consideration, Li et al., [16] performed a comprehensive analysis of Twitter and Weibo for understanding the approach of the public towards the policies and actions of the government. This will help the officials to create better solutions for fighting the pandemic. The current data and latest topics of discussion will contribute to making amendments for developing better procedures [5]. On Twitter, people tend to use a variety of unigrams and bigrams to express their views. The authors also perform sentiment analysis to track public opinion [24]. Hamoui et al. [11] and Valdez at al., [20] are the two pieces of literature that explain the pandemic situation and the public reactions to it in two different regions i.e. Arabic countries and the US. Bracci et al.

[4], explains the impact of Covid-19 on the economy and also explained the aspects due to the shadow economy. A large-scale dataset of geotagged tweets is also analyzed in Lamsal, 2020 [14]. A comparative analysis of the literature review is conducted and illustrated in Table 1.

Table 1. Comparative Analysis

Reference number & Information and year of publication	Topic	Problem addressed	Strengths	Methods	Key findings	Weaknesses
[19] ***The Lancet Digital Health*** **2021**	What social media told us in the time of COVID-19: a scoping review	Identification of infodemics, public attitude, and their mental health towards Covid-19. Also, analyzing government reactions to the pandemic through social media.	Directing the policymakers and the health care agencies for intervention designing to provide accurate knowledge translation to the public.	Exploratory searches on Covid-19 via open research dataset challenge and google scholar. Usage of machine learning analysis like LDA &RF. Screening of peer-reviewed literature. Surveillance and monitoring of social media (Twitter, Weibo)	46% of data used from Twitter and 18% of data used from Weibo for analysis. Demonstrated methods to detect and predict the Covid-19 cases via social media. Observed the need for vaccinations with rapidly increasing cases. Reliable information for the public by the government is critical for infodemics. Correct and accurate information plays a crucial role in handling infodemics and misinformations.	Though the research analyzed the languages, locations, timings, hashtags, and keywords through social media certain shortcomings are observed such as the biased selection of data and retrospective designs.
[7] ***Journal of Travel Medicine*** **2020**	The pandemic of social media panic travels faster than the COVID-19 outbreak	The spread of misleading information and conspiracy theories about the pandemic, through social media.	Combating the panic of social media regarding the pandemic.	Analyzing the discussions on social media both geographically and over time. Spatiotemporal analysis to obtain the connectivity and disconnectivity with the epidemiological situation. Intervention campaigns to be conducted by the authorities related to public health.	Rapid assessment of the changes in the public attitudes and perceptions regarding the pandemic. Sharing of real-time information to avoid any misunderstandings and rumors with the fast-evolving speed of social media.	More examples and emphasis on communication strategies are needed.
[2] ***Journal of Medical Internet Research*** **2020**	The Impact of Social Media on Panic During the COVID-19 Pandemic in Iraqi Kurdistan: Online Questionnaire Study	Determine the impact of Covid-19 in the Kurdistan region of Iraq and understanding its effects on the mental health of the public via social media.	To avoid the spreading of inaccurate information, such as food shortages, health professionals and media experts are to be contacted for ensuring only well-vetted information is provided to the public.	An online questionnaire study which including the details about the gender, source, etc. SPSS software is used to analyze data of several online platforms such as Facebook, Snapchat, WhatsApp, Twitter, YouTube, etc., the age and gender of the user, and the duration of the news.	A positive correlation between social media and the spreading of pandemic panic has been observed. For the better future of Kurdistan, the educators and the media experts have to make sure that the information which is good and reliable must be disseminated.	Due to lockdown, it became difficult to find the participants who could contribute to the research. This was observed as one of the constraints to collect the representative data.
[16] ***arXiv:2005.14464*** **2020**	Analyzing COVID-19 on Online Social Media: Trends, Sentiments and Emotions	After the application of the mandatory rules by the government, it caused a severe impact on society.	Keeping track of the sentiments and emotions of the public signifies the evolution of their attitude towards the global crisis.	Comprehensive analysis on the data obtained from Twitter and Weibo from Jan 20th, 2020 to May 11th, 2020. Differentiate the emotions of the public of both China and the United States to identify the differences between the public reactions of different countries.	A computational approach for understanding the public approach towards the pandemic effects and government policies for the creation of better solutions for fighting the pandemic.	The gap in the domain of research of the dataset due to multiple meanings of the same emotion.

				Retrieval of the post related to Covid -19 through social media and analyzing the result intensity trends.		
[12] ***JMIR Public Health and Surveillance*** **2020**	Assessment of Health Information About COVID-19 Prevention on the Internet: Infodemiological Study	Investigating Covid-19 prevention information on the internet.	Analyzing the websites and the information that people wanted to research about the pandemic and its prevention helps in easily identifying the information accuracy and determining what information they seek. Help in improving the availability of information by WHO which is only available in 32.5% to 81.3% of the links.	A descriptive analysis on the weblinks of google search based on the factors like: Type of authorship, language, country of publication, and recommendations to avoid Covid-19.	Identification of ambiguous information that did not follow WHO recommendations. Over 80 weblinks were analyzed.	The analysis of the constantly changing information is the limitation of the research. The study is deduced as the intrinsic one to the concept of the internet.
[9] ***Computers in Biology and Medicine*** **2020**	A scoping review of the use of Twitter for public health research	Reviewing the literature on how the Twitter application is accessed for gathering the information regarding public health and current research.	Understanding the concept of physical and emotional well-being and also the measurement of the burden of diseases with the help of existing literature and identification of gaps will help in improvement for further research.	Scoping review methodology which is the mapping of concepts based on the pieces of evidence available using the Twitter platform. Map the key problems and the tackling concepts followed by the analysis of the collected data based on the aim of the research, the focused disease, methods used, the country, and year of research.	Twitter helps in extracting powerful, easy to access, and real-time data. Trends of research are identified over time according to the current illness.	As the study is conducted under the terms of a broad search and it didn’t consider the topics that were still under progress, there is a possibility of relevant studies to be missed.
[23] ***PLoS ONE*** **2020**	Framing COVID-19: How we conceptualize and discuss the pandemic on Twitter	Analysis of large tweets on Covid-19 during March to April 2020.	Exploring the Twitter database is increasingly bringing a productive field of research. Issues about the pandemic are discussed on Twitter and also about the treatment and containment of the virus which will help in framing the treatment methods of the disease.	Identification of the topics that were discussed about Covid-19 on Twitter. (Using Twitter API and Tweepy Python library) Followed by the analysis of the topic model.	As per Twitter, The most common topics that were discussed in Covid-19 were politics, reacting to the epidemic, community, and social compassion. Treatment and diagnostics are the frequently discussed topics.	Comparison of the metaphorical usage of words and the literal use of words needs proper analysis. The actual distribution of topics derived from Twitter is not reflected properly.
[5] ***JMIR Public Health and Surveillance*** **2020**	Tracking Social Media Discourse about the COVID-19 Pandemic: Development of a Public Coronavirus Twitter Data Set	Description of Covid-19 Twitter datasets including multiple languages using Covid-19- TweetIDs GitHub repository.	Analysis of the online conversations of the Twitter dataset about Covid-19 will provide the researcher the ability to study the impact of social media during the global health crisis.	Using specific keywords and accounts, tweets are collected using Twitter API. Continuously monitoring the trending topics on Twitter.	Identified the number of verified accounts and their information tweets related to the pandemic. These are also referred to as authentic accounts. Tweets in different languages such as Spanish, Italian, and Japanese over time, are also collected.	The data collected from the free streaming API only included 1% of the total volume of the dataset of Twitter. And, the collected data is limited to the certain filters that are created for the study
[17] ***IEEE Transactions on Network and Service Management*** **2020**	Critical Impact of Social Networks Infodemic on Defeating Coronavirus COVID-19 Pandemic: Twitter-	A large-scale study of all Covid-19 communications other than those from trusted sources such as WHO and other authorized government entities with the help of data mining from Twitter.	The concerns of the government organizations are the abundance of fake news on social media about the pandemic. The article focuses on exploring the large dataset qualitatively to identify the	Analysis of over 1 million Covid-19 related tweets that are collected over two months. 288,000 users profiles were analyzed Analyzing the content of messages to identify the	The article investigated the negative impact of infodemics related to Covid-19 via a large-scale Twitter-based study using real-life experiments. The importance of legitimate information is a priority.	The building of Covid related dictionaries manually may eventually result in assumption-based conclusions due to lack of experimental study.

	Based Study and Research Directions		shortcomings and provide the direction to future research for limiting the impact of the misinformation spread.	misleading post-detection.		Analysis of unverified profiles may result in false claims. It's important to identify the authenticity of the accounts from where the data is collected.
[24] ***Journal of Medical Internet Research*** **2020**	Twitter discussions and emotions about the COVID-19 pandemic: a machine learning approach	Examining the concerns, discussions, and sentiments of users via their tweets about Covid-19.	Analyzing the intensity of the fear amongst the public during the new cases and death counts due to Covid-19, expressed through social media, is an important aspect to study so that proper health-related planning and responses can be conveyed.	Purposive sampling approach for Twitter data collection. (Sampling then data collection followed by pre-processing of raw data) Used the machine learning approach to analyze the Twitter data and 4 million Twitter messages using 25 hashtags in 2020 from March 1 to April 21. Identified using unigrams, bigrams, and the topics related to sentiments during the pandemic. NLP (natural language processing) approach is used for sentimental analysis.	The public uses different types of words while referring to Covid-19 such as coronavirus, virus. As per the new emerging topic, the increase in the discussion regarding the connection between Covid-19 and politics is increasing as the situation evolves. Fear and anticipation of the pandemic is a frequent topic.	New keywords that come in as the situation evolves are not taken into consideration. Less exploration on the public trust on the existing measures taken by the health organizations and the government.
[8] ***Applied Network Science*** **2020**	Public risk perception and emotion on Twitter during the Covid-19 pandemic	The deciding factor between the public behavior, their cooperation towards the social restrictions, and their perception about the risk of the pandemic are analyzed.	With the evolution of the pandemic, the change in emotion in its response is not understood properly. Hence, the analysis of the real-time data can give a better picture.	A set of over 80 million original tweets(not repeated) is gathered and maintained using Twitter free Stream API. Gathering the epidemiological data, i.e., the number of cases and deaths due to Covid-19. Followed by the analysis of the public's perception of Covid-19 and observing the trends.	Twitter is a powerful platform for monitoring public risk perceptions during a time of large-scale crisis. The real-time public's attitude could be tracked by the health officials such as the degree of outrage and their reaction and focus on the threat, via Twitter.	More analysis on the reaction of the public in different countries can enhance the risk perception. With this, the role of emotions makes the information about public health during crisis more helpful.
[20] ***Journal of Medical Internet Research*** **2020**	Social Media Insights Into US Mental Health During the COVID-19 Pandemic: Longitudinal Analysis of Twitter Data	The increase in usage of social media and change in sentiments during the pandemic in the US by analyzing the corpus of US tweets.	Social media is to be analyzed to study the societal mood of the users to get an idea about the public's well-being and mental health at the time of crisis.	Initiated with the collection of English language US tweets. Followed by characterizing the hashtags that were used over time and then assessed 20 major US studies to identify the changes in social media usage. Analyzing the mood shifts and sentiments variation of the public with the evolution of the pandemic.	Pandemic has negatively impacted the sentiments of the people of the US. The analysis revealed that the content about Covid-19 on social media has an abrupt shift in the lifestyle and the sentiments of the public relative to the time before the pandemic.	Due to the geotagging feature provided by the information, some of the data regarding the user location can be misrepresented
[3] ***arXiv:2004.03688*** **2020**	A large-scale COVID-19 Twitter chatter dataset for open scientific research - an international collaboration	Mapping and measuring the role of social dynamics during the pandemic.	The availability of open Twitter dataset will allow the researchers to identify the source of misinformation about the pandemic that leads to real-time emotional responses by the public on social media.	Raw data collection is done through Tweepy and Python package from January 2020 to November 8th, 2020. Tweets are hydrated with the help of the Social Media Mining Toolkit Data validation is done using different CSV and zip files.	The tweet IDs are made available using Zenodo so that they can be re-used for analysis purposes. The details of the Covid-19 datasets, that were released, are also available.	No discussion about the vaccination in the extracted keywords.

[10] ***arXiv:2006.05783*** **2020**	Tracking the Twitter attention around the research efforts on the COVID-19 pandemic	Investigation on online discussions about Covid-19 with the involvement of Twitter users. Consideration of attention of Twitter towards scientific research.	The analysis of the activities of Twitter, during the time of PHEIC (Public Health Emergency of International Concern), helps in getting a broader perspective on science and society interaction and also helps in identifying that how the scientific findings regarding the pandemic are communicated to the population of the world.	Data collection is done by Twitter with the help of the specific mentioned data of Covid-19 publications. Data is categorized based on the original and re-tweets. The titles of the topics about the pandemic that are discussed are also analyzed.	Sharing and discussion of Twitter trend information behavior related to Covid-19. A long-term impact of Covid-19 on the Twitter discussion is observed.	Due to the limitations of bibliometric information that were retrieved, only the titles of the most discussed topics are taken into consideration. Whereas, important factors such as abstracts, keywords, etc were not included.
[1] ***Journal of Medical Internet Research*** ***2020***	Top Concerns of Tweeters During the COVID-19 Pandemic: Infoveillance Study	Identification of the important topics that are posted during Covid-19	To help the health officials with the advanced surveillance system that can help in analyzing the current situation of the pandemic.	Twitter search API Tweepy Python Library PostgreSQL database	The situational information about the pandemic is extracted on a real-time basis. The analysis helped in identifying the number of different communities that used social media for expressing their views about the ongoing pandemic.	As Twitter only allows access to the data of the past 1 week, Covid-19 related data before that period is not used. Due to the unavailability of data from private accounts, some of the findings of the topics relevant to COVID-19 are missing.
[11] ***10.20944/preprints202007.0172.v1*** **2020**	What are Covid-19 Arabic tweeters talking about?	Analysis of the content by Arabic users on Twitter.	As the usage of Twitter in Arabic countries increased to 17 million tweets during the pandemic, it is a potential data mining platform to provide valuable information during the crises. Identification of the main concerns of the public regarding will be useful for health professionals and social scientists.	Data preparation is done by collecting the raw data initially and after filtering the essential data, the formation of datasets is done. Non-negative Matrix Factorization (NMF) is used for the topic modeling.	Trends of unigrams, bigrams, and trigrams displayed the change in their usage over time. The NMF provided meaningful topics that are used for conversations and discussions by the Arabic users. Such as epidemics and pandemic, country-related discussion, methods for a decrease in the coronavirus spread, news, and reports.	More analysis and identification of the sentiments of the Arabic users should be considered. (Sentiment Analysis)
[4] ***EPJ Data Science*** **2020**	Dark Web Marketplaces and COVID-19: before the vaccine	The impact of dark web marketplaces on the economy and the public during Covid-19.	Analysis of the public opinions by Twitter and Wikipedia search visits is beneficial for understanding the shadow economy.	Dataset collection of DWM from January 2020 to November 2020 by using Flashpoint Intelligence. Listings of the important goods for Covid-19 such as masks, ventilators, and tests on 13 DWMs. Tweets samples related to Covid-19. Analyzing the usage of certain keywords with the evolution of the pandemic. Collection of data in English from Wikipedia API for analyzing the most searched topics.	The requirement of in-depth monitoring of the online shadow economy that displays the correlation of public attention and DWMs is highlighted. The findings can be used by policymakers and public agencies for a better understanding of the shadow economy.	Data is biased towards the English language.
[14] ***Applied Intelligence***	Design and analysis of a large-scale COVID-19 tweets	Provides a large-scale dataset of over 310 million tweets.	Twitter has been observed as one of the most important platforms for the extraction of awareness	Data collection is done from search API and streaming API. Data filtration based on geo-	The unigrams and bigrams that are trending are identified.	***Dataset to be used for analysis purposes. (No weakness)***

2020	dataset		information connected to any crisis. The datasets can be used for a better understanding of the public dissertation to the pandemic.	tweets and sentiment-tweets. Data visualization of geotagged tweets		
[13] ***https://doi.org/10.5281/zenodo.3735015*** **2020**	Coronavirus Twitter Data: A collection of COVID-19 tweets with automated annotations	Dataset of Twitter regarding coronavirus public discussion.	Researchers will get a benefit from this data for conducting various analyses, especially the ones with geographic components.	Zenodo The Twitter IDs can be used to obtain the original data for analysis purposes.	The essential data is present that includes the dates, keywords regarding Covid-19, and also the geolocations.	***Dataset to be used for analysis purposes. (No weakness)***
[18] ***https://doi.org/10.21227/781w-ef42*** **2020**	Coronavirus (COVID-19) Tweets Dataset	The dataset includes the IDs and the sentiment scores tweets in regards to Covid-19.	The tweets related to coronavirus can be used for further research and design.	Different keywords and hashtags are used for monitoring the real-time Twitter feeds.	The dataset is used in other papers.	***Dataset to be used for analysis purposes.*** **(No weakness)**

IV. PROBLEM IDENTIFICATION

GAPS ANALYSIS OF EXISTING LITERATURE:

After analyzing multiple research papers, various gaps have been observed that gave a direction to this research. While coming across with the literature, several limitations and gaps are detected, that is stated as follows:

- One common gap that has been extracted from all the reviewed research papers is that the public opinions and discussions about the vaccination are not highlighted and discussed enough.
- In the study of Edo-Osagie et al., [9] the broad search method is adopted due to which, it only included the completed topics and did not consider the ones that were still under process. As a result, some topics related to relevant studies are missed.
- While accessing the Twitter data from free search API, it only allows limited access to the huge volume of data. As the research has its filters for the identification of the data, some of the crucial data is left unattended [5].
- Also, the data and information keep changing continuously therefore; the analysis of the same was also found missing [12].
- The gaps related to the extracted keywords, their meanings, lack of research about changing public interests as time evolves are also identified [24].
- It is also not certain that the tweet IDs that are used during the data collected are verified and genuine. The data from unverified IDs may result in claiming false results [17].

PROBLEM STATEMENT:

The vaccination process has already begun in phases. Therefore, this paper attempts to answer the following questions:

1. To what extent we have acquired the information from the existing literature about the importance of public opinion and how it turns out to be essential for fighting the pandemic?
2. What new information of the public opinions, discussions, and concerns, through Twitter, about the vaccination process that is currently being conducted in the entire world should be crucially analyzed for an efficient campaign?
3. Why it is important to analyze the data extracted from the Twitter using Tweepy, its comparison with the older tweets about the vaccination, and the displaying of the analysis of a few cities and countries to identify the difference in public opinion from different sections of the world? Also, how it can help the researchers, readers and health officials for the effective alignment of the vaccination process?

Also, the paper sought to offer some recommendations and highlight the existing gaps in the research for future investigation.

METHODS OF PROBLEM SOLVING

The research focuses on carrying out a Think-Aloud study for better understanding of the most frequently used words, keywords and the hash tags that are used in the public discussions [28].

INITIATION OF THINK-ALOUD METHOD

1. Comparing the two datasets of different time duration:

Extracted dataset from Twitter using Tweepy (in python)
[April 2021] [26]

VS

Online dataset available through Kaggle
[December 2020] [27]

2. Performing the data analysis of different cities and different countries in the world to understand the public discussion in different parts of world using the open dataset.

RESEARCH DESIGN

This research paper will conduct qualitative research using purposive sampling for the collection of tweets related to the phases of Covid-19 vaccination. The entire process involves four steps and is displayed in Figure 2.: (1) Data extraction, (2) Data preprocessing, (3) Data analysis and Data visualization, and (4) Results.

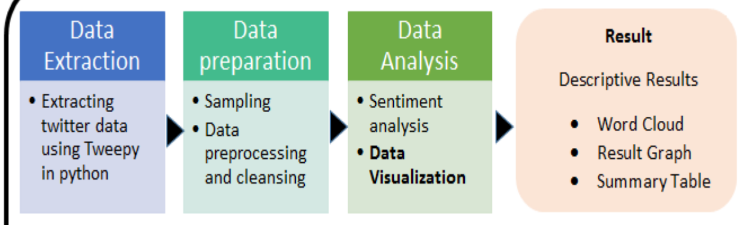


FIGURE 2. FLOWCHART OF ENTIRE PROCESS

DATA EXTRACTION

The Twitter data will be extracted using Tweepy (Python will be used for extraction). As the data keeps on changing continuously, the recent data of public discussions and necessary information by the health officials will be collected from Twitter.

DATA PREPRATION

The data preparation will be done in two steps:

I. Data sampling and Data collection

A set of keywords such as “Covid-19 vaccination, Pfizer, Covid vaccine dose” will be used as search terms for fetching the Covid-19 tweets. As, the Twitter API only gives the access to the data of past few weeks, the latest 500 tweets are collected for preprocessing. Considering a particular period, the data will be collected in the form of raw data. The English-tweets are also filtered and collected. Also, to save time and maintain the authenticity of the collected data, the duplicate data i.e the re-tweets (RT) will also be removed.

II. Data Preprocessing [24]

After the collection of data, the cleaning of raw data is of utmost importance. The data that is not crucial for carrying out of the research is filtered in this step. Therefore, we will use Python to clean the data.

i. Removing the ‘@’ mentions and the the characters after it, such as A-Z, a-z, 0-9.
ii. Removing the hashtags ‘#’ from the raw data.
iii. Removing the hyperlinks ‘http?:’ from the collected data
iv. Removing the punctuations
v. Removing the shortwords for refining the data
vi. Tokenization of data
vii. Removal of stopwords

DATA ANALYSIS AND DATA VISUALIZATION

As the research is about the public discussions about the Covid-19 vaccine and its reactions, opinions, and emotions, therefore, the analysis and visualization will be done using the below mentioned step:

- Sentiment Analysis and data visualization

  This is the process that focuses on defining the opinions, attitudes, and emotions of the public through their texts. With the help of sentiment analysis, the research will determine the reaction of the public about the vaccination trials, depending on the current situation. For this, textblob, which is natural language analysis(NLP) approach, will be used as this will help in classifying the important and exact sentiments of the Twitter message. The output of the TextBlob analysis is delivered in the form of polarity and subjectivity. Negative polarity is stated by ‘-1.0’ and the positive polarity is defined by ‘+1.0’. And the polarity score ‘0’ stands for neutral. The subjectivity identifies the subjective sentence as ‘1.0’ and the objective subject as 0.0.

  Data visualization will help in better understanding of the graphically represented data. This research will use the scatter plot to provide an accessible way for the understanding of trends and patterns of the information [25].

CREATION OF THINK-ALOUD SUMMARY TABLE

Ghenia et al., 2020 [28], studied how the think-aloud helps in better understanding of the public opinions. Similarly, this study also gives a table with all the details such as:

- Tweets and re-tweets count
- Percentage of positive and negative tweets
- Description of frequently used words in the discussion, keywords and hash tags.

The data is available from an open dataset and the filtration of the crucial data is done as per the requirement of the research.

- The collected tweets are from verified accounts to maintain the authenticity of the final result.
- The collected tweets are the original tweets and the re-tweets count has been calculated separately.
- The entire process of gathering the tweets is done filtering the date, countries and cities as per the research requirement.
- For the comparison of the Twitter extracted messages (initial few weeks of April 2021), the tweets dated December 2020 from open dataset are collected.
- For different countries and cities, the tweets from January, 2021 to April initial few weeks, 2021, are filtered and collected.

V. PROPOSED SOLUTION

DETAILS OF PROPOSED SOLUTION

A lot of analysis and research has already been performed on public discussions about the Covid-19 pandemic. But, public opinion about the vaccination process is still an important and crucial topic to be discussed. And keeping in mind the rising cases of Covid-19, an efficient vaccination campaign must be conducted in the entire world. So, with this research, we will be able to deduce what exactly does the public thinks about this vaccination trial/procedure and how their myths and misinformation can be answered and resolved.

The extended use of Twitter during times of crisis is a good step towards the solution to this problem. The execution steps of the process are as follows:

- Twitter API search method is adopted and the data is extracted using Tweepy in python.
- There are four essential that are necessary for getting the access of the Twitter API. The keys are:
  i. Consumer_key
  ii. Consumer_secret_key
  iii. Access_token
  iv. Access_token_secret

  All these keys are supposed to be confidential due to privacy issues of the API. Then the twitter

OAuthHandler is used for the authentication purpose.

- Use the google colab for uploading the already existing open dataset.
- The search terms and their amount are decided as per the research, before initiating the data extraction process. Cursor object of tweepy is used for searching the terms.
- After the successful data extraction process, data preprocessing is done to get rid of all the unwanted components that are not required in the research.
- Data analysis (sentiment analysis) and Data visualization (graph plot and word cloud) are done for portraying the clear picture of the public discussions regarding the Covid-19 vaccination and their opinion about the same.

The entire process of extraction, preprocessing and analysis of the real-time Twitter data and the existing open dataset will accelerate the process of the research. The valuable data of Twitter will help in analyzing the public discussion and the sentiments related to multiple topics about Covid-19.

ARCHITECTURE

The research includes two different architectures for both, the Twitter extracted dataset and the open dataset. Both of them include the detailed flow of the entire process conducted.

- Architecture of the analysis process for Twitter extracted data using Tweepy [April 2021] is displayed in Figure 3.

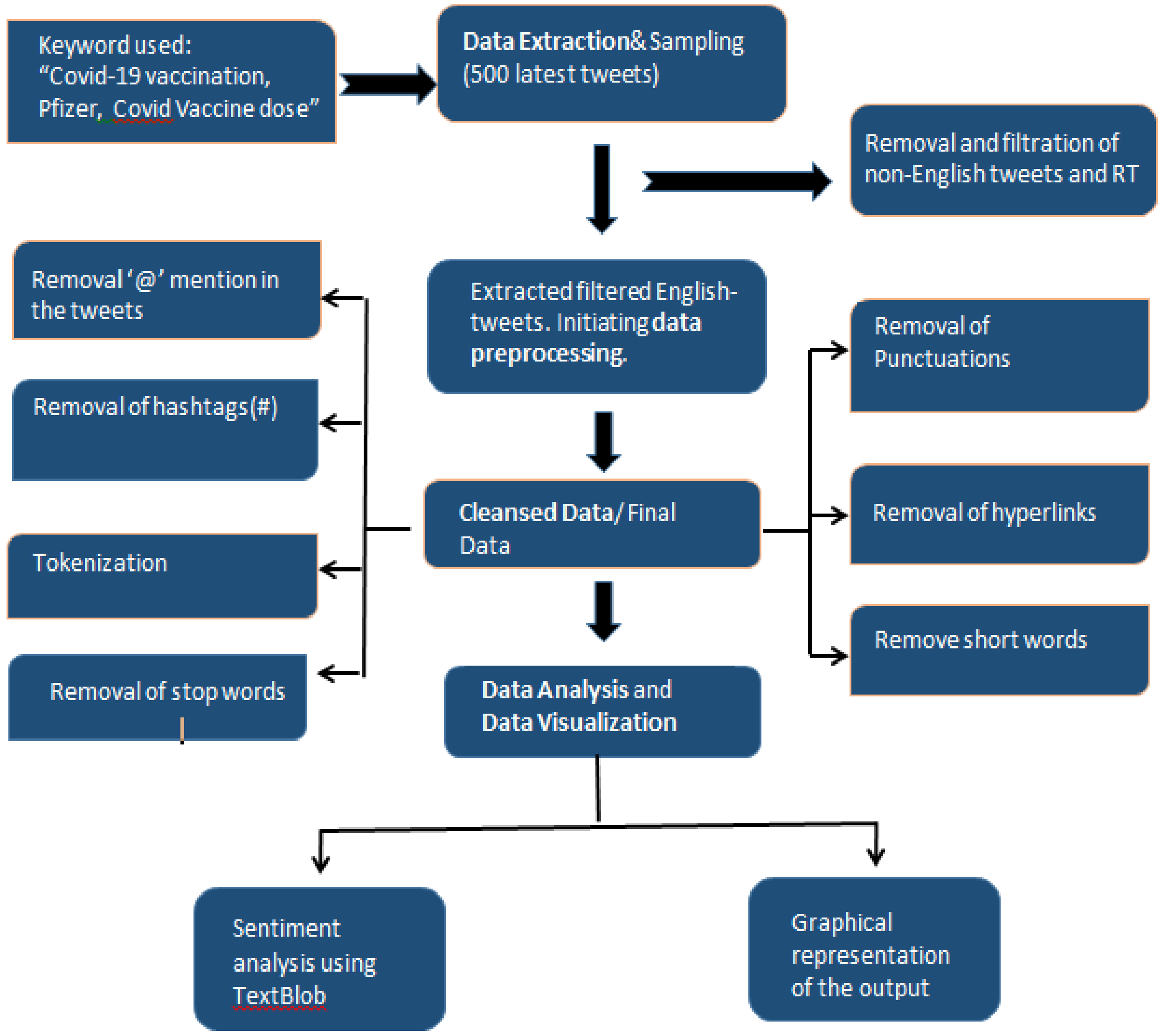

Keyword used: "Covid-19 vaccination, Pfizer, Covid Vaccine dose"
Data Extraction& Sampling (500 latest tweets)
Removal and filtration of non-English tweets and RT
Removal '@' mention in the tweets
Removal of hashtags(#)
Tokenization
Removal of stop words
Extracted filtered English-tweets. Initiating data preprocessing.
Cleansed Data/ Final Data
Data Analysis and Data Visualization
Removal of Punctuations
Removal of hyperlinks
Remove short words
Sentiment analysis using TextBlob
Graphical representation of the output

Figure 3. Architecture of Twitter extracted data through Tweepy

- Architecture of the analysis process for online available dataset [December 2020] is displayed in Figure 4.

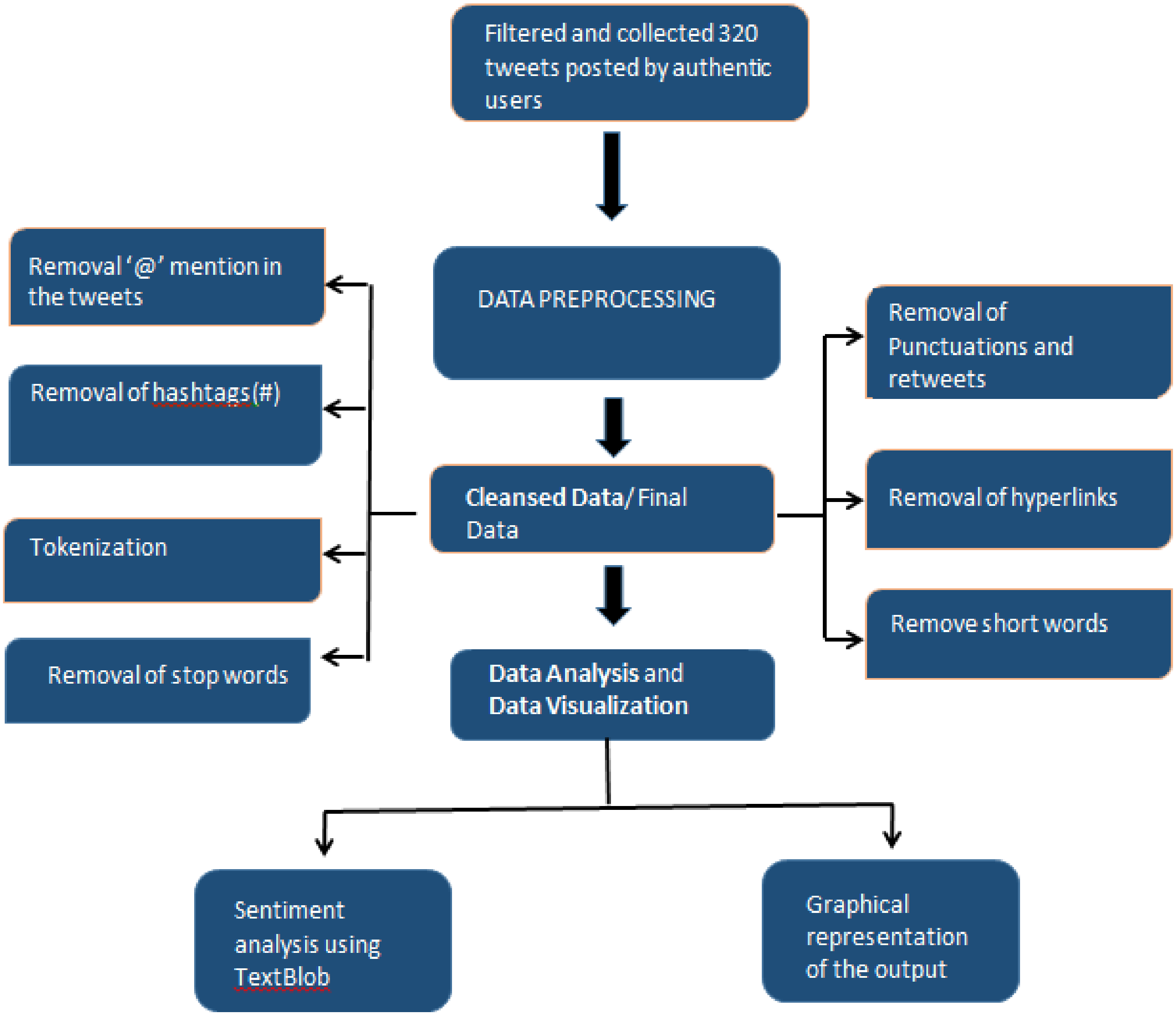

Filtered and collected 320 tweets posted by authentic users
DATA PREPROCESSING
Removal '@' mention in the tweets
Removal of hashtags(#)
Tokenization
Removal of stop words
Removal of Punctuations and retweets
Removal of hyperlinks
Remove short words
Cleansed Data/ Final Data
Data Analysis and Data Visualization
Sentiment analysis using TextBlob
Graphical representation of the output

Figure 4. Architecture of analysis of online available dataset

VI. Pseudo code of the Solution

A. *Data Extraction*

```
#Initiating with the importing of the libraries and package
Import the tweepy package from tweepy library
 #for twitter API search
Import TextBlob package from textblob library
#for performing sentiment analysis
Import WordCloud
#for displaying the frequent words
Import numpy
#for handling numerical data
Import re
#replacing strings
Import the matplotlib.pyplot package
#for plotting graphs for end results
from nltk.corpus import stopwords
# for removing stopwords

#Passing the keys for authorizing with the tweepy OAuthhandler
#Authentication with the consumer_key and consumer_secret
tweepy .OAuthHandler(consumer_key, consumer_secret
#Authentication for setting the access token
authen.set_access_token(access_key, access_secret)
#Pass the above authorizations to tweepy
tweepy.API(authen, wait_on_rate_limit=True)

#For loading the dataset
Use the google colab upload function for loading the csv file of the dataset

#extract the data using search key word and creation of a dataframe
Specify the keyword as a search_term = ‘Covid vaccination, Pfizer, covid vaccine dose’
Use “ –filter:retweets” for filtering all the RT to avoid repetition in data  in the
search_term variable and store the new data in the new variable
Number of tweets to be extracted, tweet_amount = ‘500’
#Use cursor for searching the tweets in twitter and put in the variable
Variable=tweepy.Cursor(api.search, q=new_var, language = ”en", tweet_mode = “extended“).items(tweet_amount)

#Extracting and printing the raw tweets into a dataframe
Pandas. DataFrame [final_text for tweet in variable] , column name = [‘Message’]
Call pandas.set_option to display all the columns and rows to none
Print the DataFrame
```

B. *Data Preprocessing*

```
#Data filtration and cleansing process
#Create a function called cleanTxt()
Remove ‘@’ mentions using re.sub(). Declare the pattern to be found as a raw string.
Remove ‘unwanted characters such as A-Z, a-z, 0-9 using re.sub()
re.sub(r’pattern to be removed or replaced’,  ‘ ‘ )
#removing @mentions
Remove the hashtags ‘#’ using re.sub()
#removing hashtags
Remove the hyperlinks ‘https?:’ using re.sub()
#removing hyperlinks
Remove the punctuation ‘[^\w\s]’ using re.sub()
#removing punctuation marks
Apply the cleanTxt function to the extracted tweets
Print dataframe

#removal of short words whose length is less than 3 or 4 based on the research requirement
Defining a function.
Using the split function:
    if length of the word is less than 3/4
    remove the word
Print the dataframe

#data tokenization (consideration of individual words as tokens)
Using the split function and storing in the variable tokenized_tweets and applying it on the existing dataframe
Print tokenized_tweet

#removing the stopwords using the ntlk corpus and display all tokenized words in a single sentence
For i in the range (length of the tokenized_tweets):
   Join the tokenized_tweets
Store the tokenized_tweets in dataframe’message’
Print the final dataframe
```

C. *Data analysis and data visualization*

Sentiment analysis using TextBlob

```
(Defining subjectivity and polarity)
Defining the function ‘getSubjectivity’ and ‘getPolarity’
Apply both on the message dataframe
Display the resulting dataframe
#with two more columns displaying the subjectivity and the polarity of the tweets

#function building for the computation of the positive, negative and neutral analysis
Define a function named ‘getAnalysis’ where
       If score(polarity) is less than 0 then
              Print ‘Negative’
       Else if score(polarity) is equal to 0 then
              Print ‘Neutral’
       Else
              Print ‘Positive’
Display the Analysis after applying the result of the function ‘getAnalysis’ to the polarity dataframe

#Print all the positive and negative tweets in a sorted manner
```

#after calculating the percentage of both positive and negative tweets and plotting the count on the graph using the functions of matplotlib.pyplot
Result of positive tweets in percent
Result of negative tweets in percent

Word Cloud formation

#WordCloud formation of frequently used words.
Define the width, height, font_size and generate(all_words)
Plot the wordcloud using the matplotlib.pyplot package.

```
plt.figure          #creation of a figure object
plt.imshow          #for displaying the data as an image
plt.axis            #for displaying axis of the image
plt.show()          #print the wordcloud
```

VII. Implementation Details

After the successful data extraction from twitter, loading the dataset and preprocessing the data on the basis of carrying out the research, the sentiment analysis is performed. Figure 5. and Figure 6. display the comparison of the discussions of two different time period i.e. December, 2020 vs April,2021 with the percentage of positive and negative tweets.

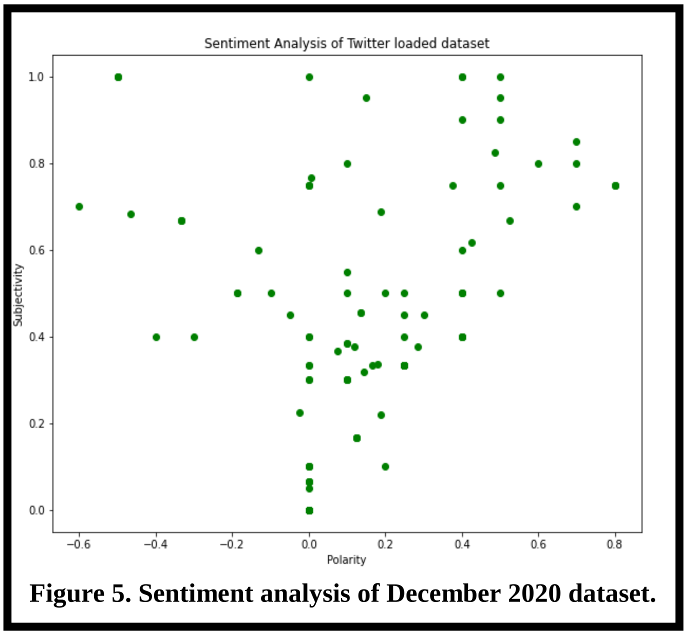


Figure 5. Sentiment analysis of December 2020 dataset.

Percentage of Positive Tweets – 30.6
Percentage of Negative Tweets- 5.0

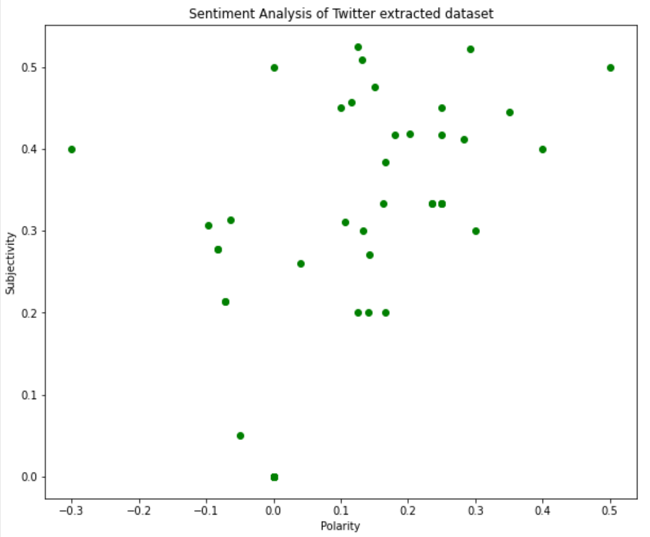


Figure 6. Sentiment analysis of beg. April 2021 dataset.

Percentage of Positive Tweets – 52.7
Percentage of Negative Tweets- 14.5

Word cloud comparison of December 2020 [Figure 7] and initial few weeks of April 2021 dataset [Figure 8].

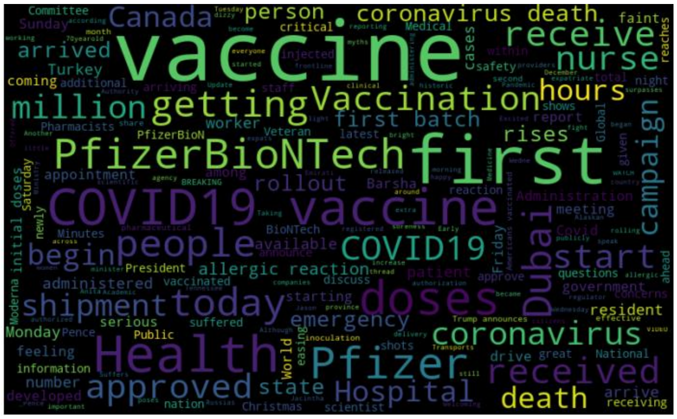


Figure 7. Word Cloud of December 2020 dataset

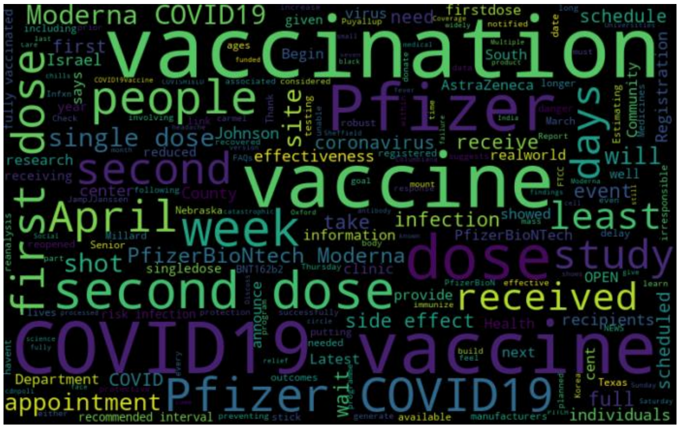


Figure 8. Word Cloud of April 2021 dataset

For extracted tweets from Twitter:

- Total number of tweets extracted = 500
- Total number of re-tweets = 445
- Original tweets = 55

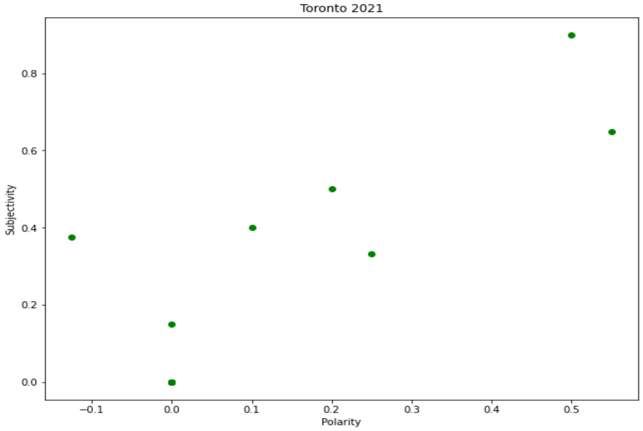
For loaded dataset:

- The tweets used are filtered as per the requirement for conducting the research.
- Total number of tweets obtained = 320

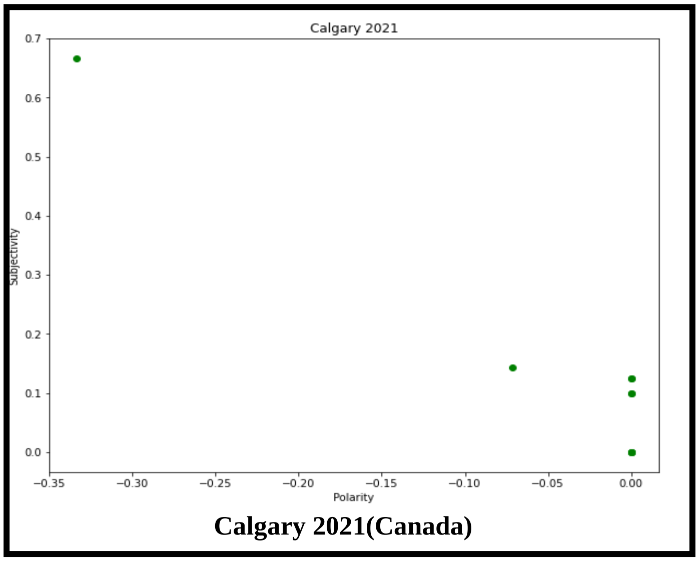
Furthermore, some of the analyses of the public discussion over the tweets in 2021 (January to April, 2021) to identify the difference in the thought process of the public of different cities and different countries around the world is carried out. The open dataset is filtered on the basis of several factors mentioned below for maintaining the authenticity and accuracy of the analysis and results.

- Only extracting the tweets from ‘true and verified users’
- Identifying the tweets from January, 2021 and to the initial few weeks of April, 2021
- Only extracting the original tweets.

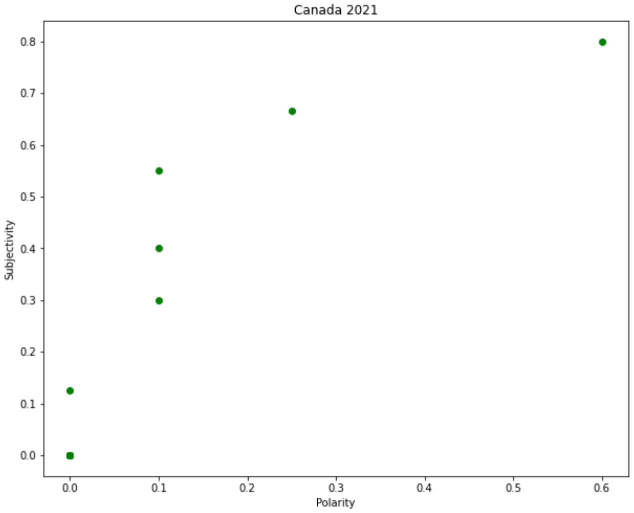
SENTIMENT ANALYSIS OF FEW CITIES AND COUNTRIES



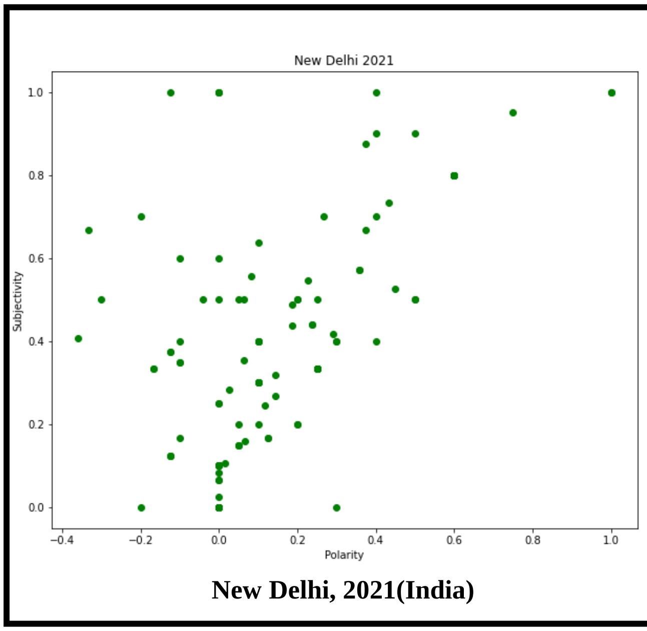
Toronto 2021(Canada)



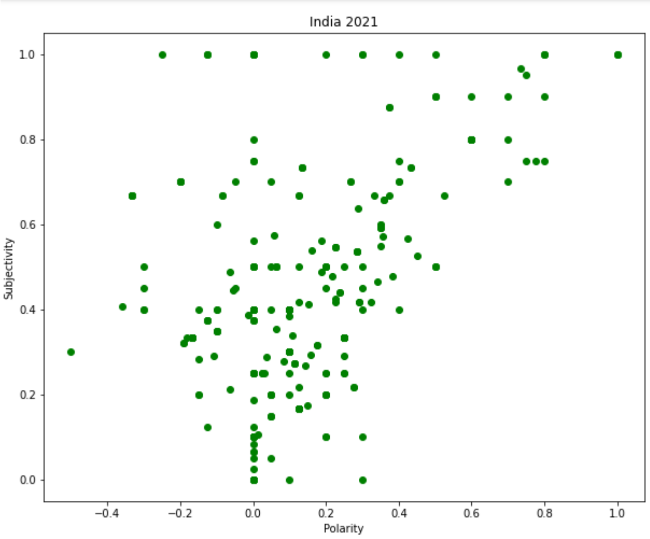
Calgary 2021(Canada)



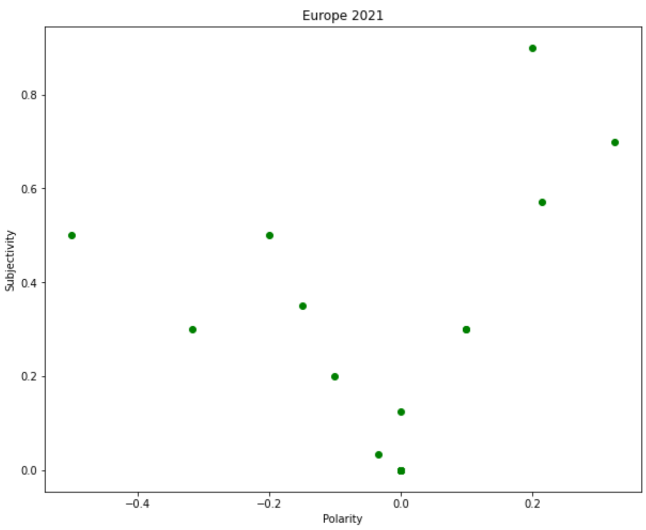
Canada (2021)



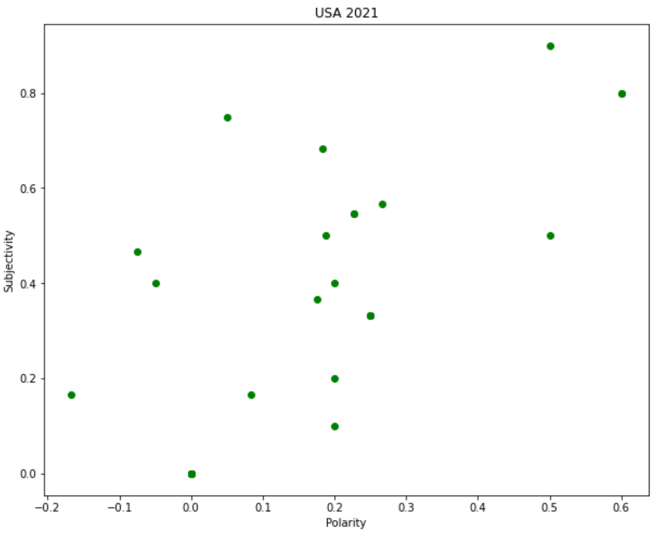
New Delhi, 2021(India)



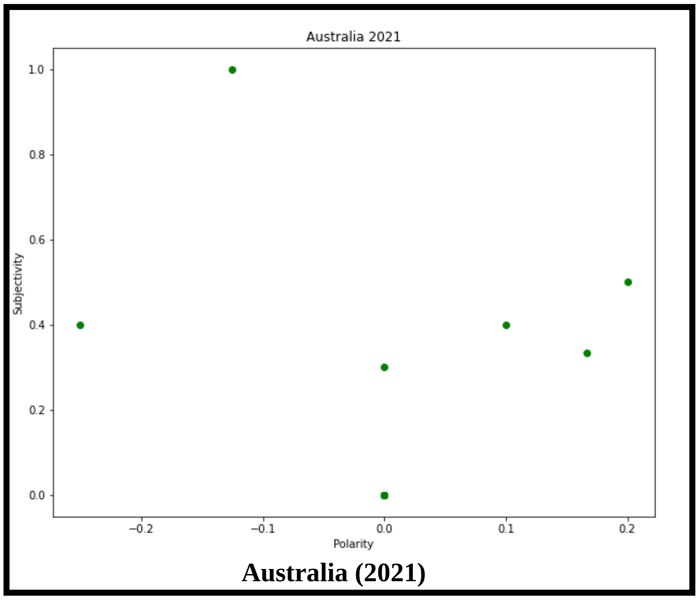
India (2021)



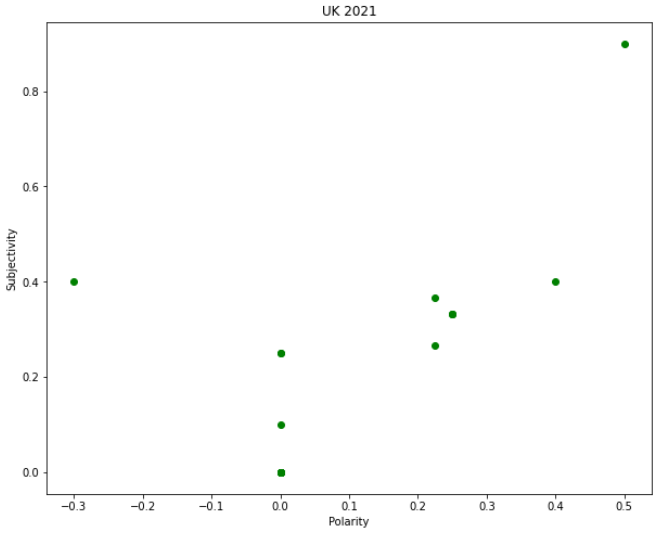
Europe (2021)



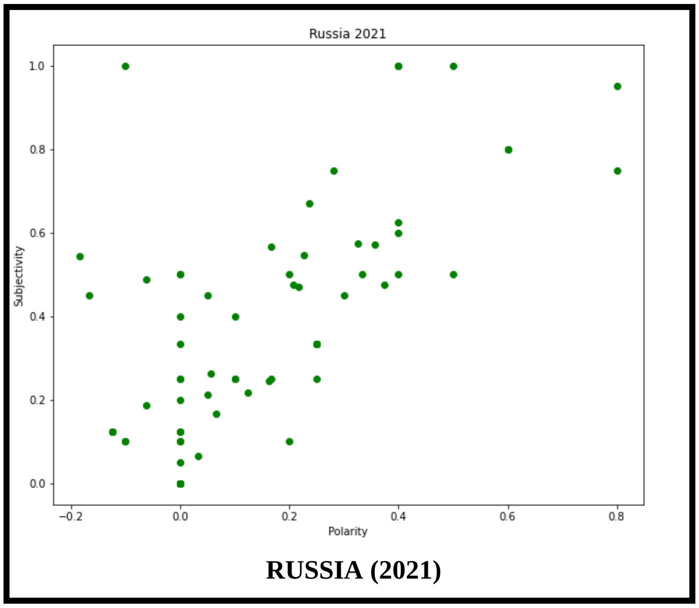
USA (2021)



Australia (2021)

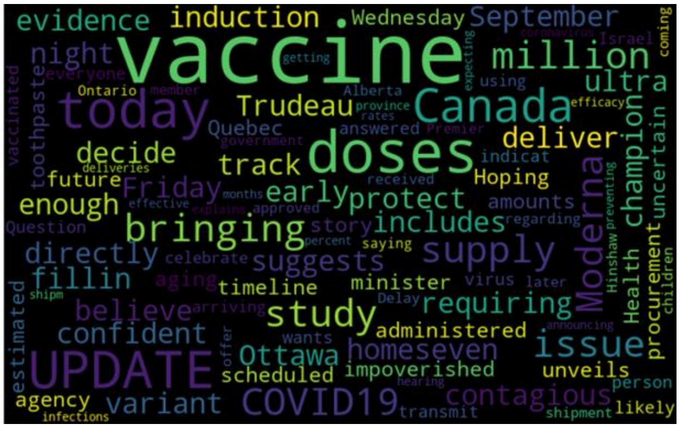


UK (2021)

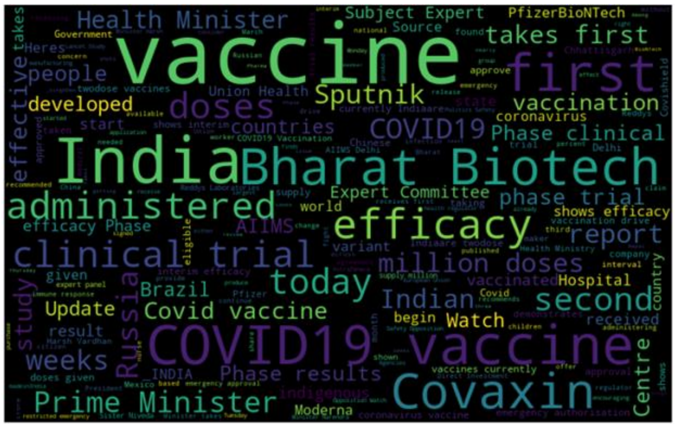


RUSSIA (2021)

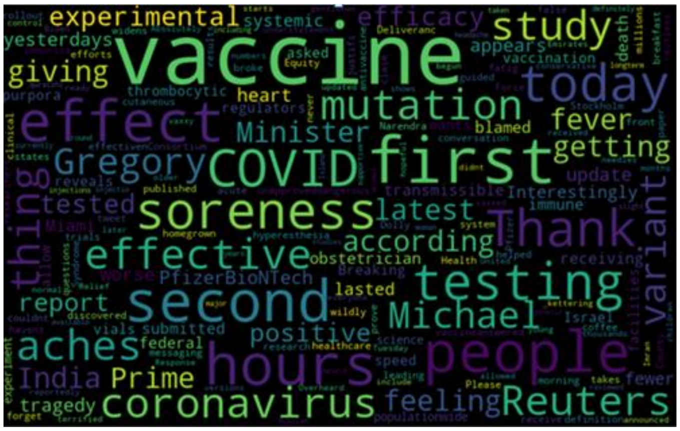
Word Cloud of the discussions through tweets by some countries around the globe.



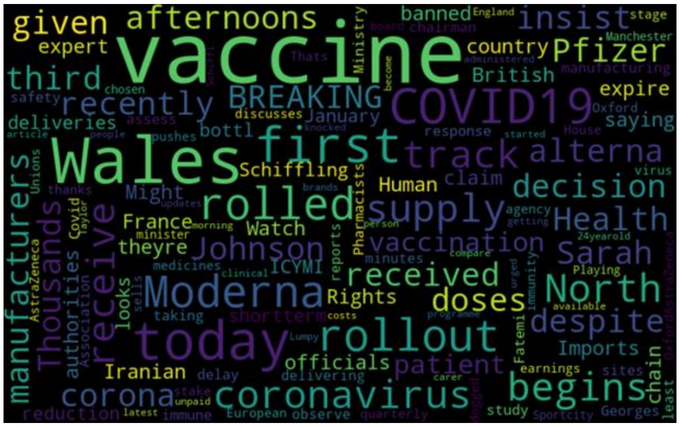
CANADA (2021)



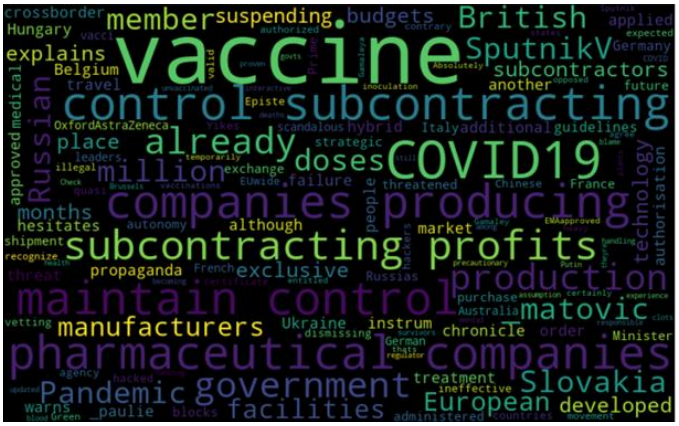
INDIA (2021)



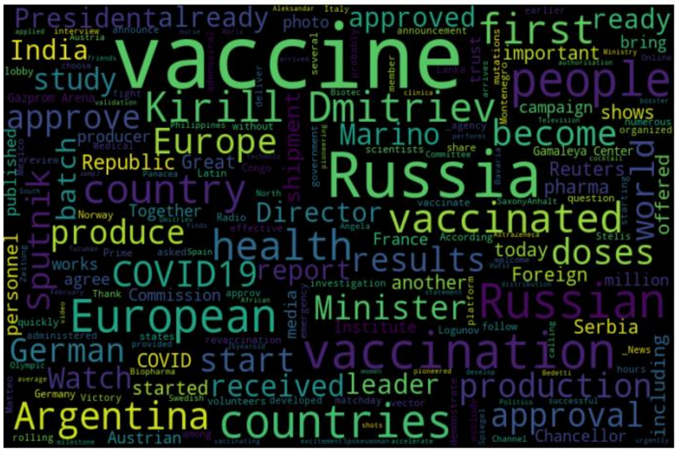
USA (2021)



UK (2021)



EUROPE (2021)



RUSSIA (2021)

VIII. DESIGN DISCUSSION

The entire design of the research made this process a lot more convenient to carry out. The data extraction step is the one where the collection of 500 latest tweets is done successfully. The advantage of this step was that the re-tweets, that are one of the reasons for data duplicity, are filtered while the collection process of the raw data. Comparing the extracted twitter data with the online available dataset, which contained the tweets of the time when the vaccination process initiated, helped in giving the clearer picture of the change in trends and the discussions over the period of time and how the people are responding to the process.

The preprocessing step for the data made the research more easy and quick. After data preprocessing step, only the crucial information is obtained at the end, that will help in conducting the better data analysis and visualization process.

TextBlob, an NLP (Natural language process), is used for the sentiment analysis as this package as it helps in the performing the analysis of the textual data with accuracy. It will give the better understanding about the public discussions and their opinions about Covid-19 vaccination.

The word cloud and the scatter plot, of different cities and countries around the globe, gave a better visualization of data. It signified the trends, patterns and the frequently used words, hash tags and keywords about the Covid-19 vaccination discussions by the public.

IX. EVALUATION

As, it has been observed that for the research areas that involves a large amount of daily posts on social media related to the public opinion, sentiment mining is an important factor to carry out the output of the research [29]. Twitter has a dataset of large texts, therefore, to summarize the data it is given some weight i.e. is it a positive, negative or a neutral text.

The complex analysis and operations on the textual data is supported by the TextBlob library. For natural language processing, TextBlob uses NTLK. By using the concept of polarity and subjectivity, the research analysis is carried out.

- Polarity helped in the quantification of the sentiments with a positive, negative and neutral value.
- Subjectivity is used for analyzing the public opinion.

The entire process is carried out for the utmost authenticity and the accuracy of the result.

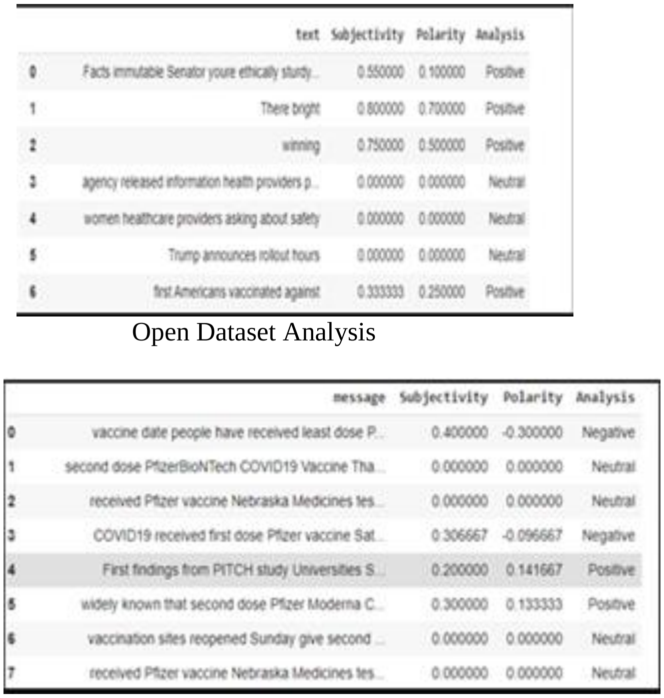
Table 2. displays the crux or summary of think-aloud discussion of the public about the Covid-19 vaccination of different cities and countries from January, 2021 to initial few weeks of April,2021.

RESEARCH OPERATIONAL ANALYSIS

The extraction of data using Tweepy helps in comparing the initial scenario with the current scenario about the public discussions regarding the Covid-19 vaccination drive and their opinions regarding the same. And since, the current scenario is an important aspect for this research, and the extracted data included several re-tweets, so to maintain the authenticity and the accuracy of the comparison and the end results the calculation and removal of re-tweets is done. The actual tweets are obtained and analyzed at the end. The python code done for the research eventually resulted in:

- Extraction of the current tweets from Twitter that eventually gave the actual tweets with the removed re-tweets.
- Further, comparison of the extracted tweets with the tweets present in the open dataset from Kaggle by performing the sentiment analysis and identifying the variation in patterns and the difference in public opinion over a period of time.
- Data visualization of the total number of positive and negative tweets of different cities and countries around the world to identify and understand the variation in the public opinions of different regions. Along with this, calculation of the total number of original tweets and re-tweets is also done.
- Word Cloud formation and comparison of different cities and countries for better understanding of the analysis results.

Figure 9 includes and displays the subjectivity, polarity and analysis of two different time period datasets, open dataset (December, 2020) and Twitter extracted dataset using Tweepy (Initial few weeks of April, 2021)



	text	Subjectivity	Polarity	Analysis
0	Facts immutable Senator youre ethically sturdy...	0.550000	0.100000	Positive
1	There bright	0.800000	0.700000	Positive
2	winning	0.750000	0.500000	Positive
3	agency released information health providers p...	0.000000	0.000000	Neutral
4	women healthcare providers asking about safety	0.000000	0.000000	Neutral
5	Trump announces rollout hours	0.000000	0.000000	Neutral
6	first Americans vaccinated against	0.333333	0.250000	Positive

Open Dataset Analysis

	message	Subjectivity	Polarity	Analysis
0	vaccine date people have received least dose P...	0.400000	-0.300000	Negative
1	second dose PfizerBioNTech COVID19 Vaccine Tha...	0.000000	0.000000	Neutral
2	received Pfizer vaccine Nebraska Medicines tes...	0.000000	0.000000	Neutral
3	COVID19 received first dose Pfizer vaccine Sat...	0.306667	-0.096667	Negative
4	First findings from PITCH study Universities S...	0.200000	0.141667	Positive
5	widely known that second dose Pfizer Moderna C...	0.300000	0.133333	Positive
6	vaccination sites reopened Sunday give second ...	0.000000	0.000000	Neutral
7	received Pfizer vaccine Nebraska Medicines tes...	0.000000	0.000000	Neutral

Twitter Extracted Data Analysis using Tweepy

Figure 9. Analysis snippets of two datasets i.e. Twitter extracted and open datatset

Table 2. Think-Aloud summary table of discussion

NO.	**City/ Country**	**Description (most frequently used words, keywords and hashtags)**	**Positive Tweets(%)**	**Negative Tweets(%)**	**Tweets (original)**	**Re-tweets**
1	Toronto, (Canada) (2021)	Covid-19 Vaccine, #COVID19VACCINES, #PfizerBioTech, #ModernaVaccine, cdn health	38.5	7.7	13	164
2	Calgary, (Canada) (2021)	Vaccine, #COVID19AB, Moderna, doses	0.0	20.0	27	2262
3	Vancouver, (Canada) (2021)	Covid-19, #COVID19, vaccines, #PfizerVaccine, #Moderna, AstraZeneca	20.0	0.0	10	180
4	Canada (2021)	#PfizerBioNTech , Long Term Care, #COVID19, AstraZeneca, #Moderna, Oxford AstraZeneca, Go Science, # PfizerVaccine, COVID Vaccines, #COVID19ab	25.0	0.0	20	118
5	Delhi, (India)(2021)	Vaccine, Covid-19 vaccine, #COVID19Vaccine, Coronavirus Strain, China, US, #COVISHIELD, # COVAXIN, covaxine, India, SputnikV, AIIMS, Covid-19, Vaccination Drive, #BharatBioTech, AstraZeneca, Narendra Modi, #AatmaNirbharBharat, Cricket, Vaccination, #Pfizer, Covid_19, #WHO, Pandemic, #UK, Night Curfew, #IndiaFightsCovid19, Russia	32.2	6.5	322	9432
6	India (2021)	BioNTech, WHO, #PfizerBioNTech, Covid Vaccine, #COVAXIN, Oxford vaccine, Moderna, #COVID19, Mutation, UK, single shot, vaccine, #oxfordastrazeneca, #BharatBiotech, SputnikV, #IndiaWillWin, immunity, coronavirus, #Modi, #CoronavirusStrains,	29.0	6.4	947	12068

		Covishield, #Coronil, #Covid19pandemic, Covid-19 vaccination, #PMModi, Largest vaccine drive, India Fights Covid-19, #ElonMusk, #MadeinIndia, #VaccineUpdate, ZyCoVD, #CNBCTV18Exclusive, #FactsVsMyths, Canada, #NationalVaccinationDay2021, #OxfordAstraZeneca, Government Hospital, second dose, Immunization, #China, AIIMS, #lockdownMaharashtra, Stock Market				
7	USA(2021)	#PfizerBioNTech, #USVaccine, Covid vaccine, #Covid19UK, Fauci, Moderna, Vaccine, #BlackDoctors, India, #COVAXIN, #JohnsonAndJohnsonVaccine, vaccinated	40.9	6.8	44	2384
8	UK(2021)	Covid Vaccinations, #PfizerBioNTech, Moderna, AstraZeneca, Covid Vaccines, #OxfordAstraZeneca, WALES, vaccination, England, immune	23.3	3.3	30	49
9	Europe (2021)	Pharmaceuticals companies, European, vaccination, SputnikV, doses, Covid-19, #EU, #PfizerBioNTech, #AstraZeneca, #Covid_19, vaccine	17.2	20.7	29	428
10	Australia (2021)	#COVAXIN, Oxford AstraZeneca, #PFizerBioNTech, Russia, Vienna, SputnikV, registered, second, variant, testing	30.0	20.0	7	53
11	Russia(2021)	#Venezuela, #RussiaHelps, Dmitriev, approved, Sputnik, vaccine, Europe, Russia, Covid vaccination, Russian, doses, #SputnikVaccinated, health results, vaccination	29.2	7.3	137	18230

RELATED WORK COMPARISON

After the comparative analysis of 20 related papers, it has been observed that the public vaccination discussion of the Covid-19 is yet to be examined and analyzed. Few of the comparison with the related work are:

1. The article, 'The Impact of Social Media on Panic during the COVID-19 Pandemic in Iraqi Kurdistan: Online Questionnaire Study' by Ahmad & Murad (2020), studied the impact of Covid-19 in Kurdistan via social media whereas this research paper analyzes the public discussions around the world.
2. The article, 'Twitter discussions and emotions about the COVID-19 pandemic: a machine learning approach' by Xue et al. (2020), analyzed the public discussion about coronavirus on Twitter.
3. The article, 'Tracking the Twitter attention around the research efforts on the COVID-19 pandemic' by Fang & Costas (2020), observed a long-term impact of Covid-19 on Twitter discussion.

X. CONCLUSION AND FUTURE WORK

The research concluded in sentiment analysis of several public discussion comparisons about the vaccination drive of the Covid-19 crisis. The key takeaways of the research are:

- The comparison of the Twitter extracted dataset and open dataset with the help of the scatter plot of analysis and word cloud for better data visualization (Figure 5, Figure 6, Figure 7, Figure 8).
- The comparison of the public discussions of a few different cities and countries is also obtained using scatter plot and word cloud for better differentiation of public opinions throughout the duration of Jan, 2021 till initial few weeks of April 2021.
- The think-aloud summary table (Table 2) provides the entire details about the most frequently used words, keywords and hashtags that are used in different cities and countries around the world for discussing about the Covid-19 vaccination. Along with that, the table also includes the count of the original tweets and the re-tweets. It also displays the percentage of positive and the negative tweets for the better comparison. To maintain the accuracy and the authenticity of data, the available dataset is filtered and only the data from the verified account is taken into consideration.

This will not just help as a base study for the upcoming research about the pandemic but will also help the readers to understand the variation in the public opinion around the globe.

As a future addition and extension of the work, other languages can also be taken into consideration for the research. The analysis of the tweets of languages other than English could also provide an impactful research.

REFERENCES

[1] Abd-Alrazaq, A., Alhuwail, D., Househ, M., Hamdi, M. and Shah, Z., 2020. Top Concerns of Tweeters during the COVID-19 Pandemic: Infoveillance Study. *Journal of Medical Internet Research*, 22(4), p.e19016.

[2] Ahmad, A. and Murad, H., 2020. The Impact of Social Media on Panic During the COVID-19 Pandemic in Iraqi Kurdistan: Online Questionnaire Study. *Journal of Medical Internet Research*, 22(5), p.e19556.

[3] Banda, J., Tekumalla, R., Wang, G., Yu, J., Lui, T., Ding, Y., Artemova, K., Tutubalina, E., Chowell, G., 2020. A large-scale COVID-19 Twitter chatter dataset for open scientific research - an international collaboration. arXiv:2004.03688.

[4] Bracci, A., Nadini, M., Aliapoulios, M., McCoy, D., Gray, I., Teytelboym, A., Gallo, A. and Baronchelli, A., 2021. Dark Web Marketplaces and COVID-19: before the vaccine. *EPJ Data Science*, 10(1).

[5] Chen, E., Lerman, K. and Ferrara, E., 2020. Tracking Social Media Discourse about the COVID-19 Pandemic: Development of a Public Coronavirus Twitter Data Set. *JMIR Public Health and Surveillance*, 6(2), p.e19273. Doi: 10.2196/19273.

[6] Covid19.who.int. 2021. *WHO Coronavirus Disease (COVID-19) Dashboard*. [online] Available at: <https://covid19.who.int/> [Accessed 15 April 2021].

[7] Depoux, A., Martin, S., Karafillakis, E., Preet, R., Wilder-Smith, A. and Larson, H., 2020. The pandemic of social media panic travels faster than the COVID-19 outbreak. *Journal of Travel Medicine*, 27(3).

[8] Dyer, J. and Kolic, B., 2020. Public risk perception and emotion on Twitter during the Covid-19 pandemic. *Applied Network Science*, 5(1).

[9] Edo-Osagie, O., De La Iglesia, B., Lake, I. and Edeghere, O., 2020. A scoping review of the use of Twitter for public health research. *Computers in Biology and Medicine*, 122, p.103770.

[10] Fang, Z., Costas, R., 2020. Tracking the Twitter attention around the research efforts on the COVID-19 pandemic. arXiv:2006.05783.

[11] Hamoui, B., Alashaikh, A. and Alanazi, E., 2020. What are Covid-19 Arabic tweeters talking about? DOI:10.20944/preprints202007.0172.v1.

[12] Hernández-García, I. and Giménez-Júlvez, T., 2020. Assessment of Health Information About COVID-19 Prevention on the Internet: Infodemiological Study. *JMIR Public Health and Surveillance*, 6(2), p.e18717.

[13] Huang, X., Jamison, A., Broniatowski, D., Quinn, S., and Dredze, M., 2020. Coronavirus Twitter Data: A collection of COVID-19 tweets with automated annotations. Doi:https://doi.org/10.5281/zenodo.3735015.

[14] Lamsal, R., 2020. Design and analysis of a large-scale COVID-19 tweets dataset. *Applied Intelligence*.

[15] Le, T., Cramer, J., Chen, R. and Mayhew, S., 2020. Evolution of the COVID-19 vaccine development landscape. *Nature Reviews Drug Discovery*, 19(10), pp.667-668.

[16] Li, X., Zhou, M., Wu, J., Yuan, A., Wu, F., Li, J., 2020. Analyzing COVID-19 on online social media: trends, sentiments and emotions. arXiv preprint arXiv:2005.14464.

[17] Mourad, A., Srour, A., Harmanai, H., Jenainati, C. and Arafeh, M., 2020. Critical Impact of Social Networks Infodemic on Defeating Coronavirus COVID-19 Pandemic: Twitter-Based Study and Research Directions. *IEEE Transactions on Network and Service Management*, 17(4), pp.2145-2155.

[18] Rabindra Lamsal, 2020. Coronavirus (COVID-19) Tweets Dataset. Doi: https://dx.doi.org/10.21227/781w-ef42.

[19] Tsao, S., Chen, H., Tisseverasinghe, T., Yang, Y., Li, L. and Butt, Z., 2021. What social media told us in the time of COVID-19: a scoping review. *The Lancet Digital Health.*

[20] Valdez, D., ten Thij, M., Bathina, K., Rutter, L. and Bollen, J., 2020. Social Media Insights Into US Mental Health During the COVID-19 Pandemic: Longitudinal Analysis of Twitter Data. *Journal of Medical Internet Research*, 22(12), p.e21418.

[21] Who.int.2021. [online] Available at: <https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200423-sitrep-94-covid-19.pdf> [Accessed 5 March 2021].

[22] Who.int.2021.Coronavirus. [online] Available at: <https://www.who.int/health-topics/coronavirus#tab=tab_1> [Accessed 5 March 2021].

[23] Wicke, P., Bolognesi, M.M., (2020). Framing COVID-19: How we conceptualize and discuss the pandemic on Twitter . *PLoS ONE,* 15(9): e0240010. Doi: https://doi.org/10.1371/journal.pone.0240010.

[24] Xue, J., Chen, J., Hu, R., Chen, C., Zheng, C., Su, Y. and Zhu, T., 2020. Twitter Discussions and Emotions About the COVID-19 Pandemic: Machine Learning Approach. *Journal of Medical Internet Research*, 22(11), p.e20550.

[25] Stack Abuse. 2021. *Sentiment Analysis in Python With TextBlob*. [online] Available at: <https://stackabuse.com/sentiment-analysis-in-python-with-textblob/> [Accessed 23 March 2021].

[26] Authentication Tutorial — tweepy 3.10.0 documentation. [online] Available at: https://docs.tweepy.org/en/latest/auth_tutorial.html [Accessed 19 Mar. 2021].

[27] Kaggle.com. 2021. All COVID-19 Vaccines Tweets. [online] Available at: <https://www.kaggle.com/gpreda/all-covid19-vaccines-tweets> [Accessed 9 April 2021].

[28] Ghenia, A., Smucker, D. M. & Clarke, L, A, C., (2020). A Think-Aloud study to understand factors affecting online health search. *CHIIR '20:* ***Conference on Human Information Interaction and Retrieval*****.** Doi: https://dl.acm.org/doi/abs/10.1145/3343413.3377961?casa_token=f2kJdrTsJ2EAAAAA:q7XYjIFK-NpsfqrkrejXENbIjFoQj-czg6urhIa5YdKPWTNiFa8VYlNl_0VBv7alNq9UhqgfBT4 .

[29] Bagheri, H. & Islam, J. M., (2020). Sentiment analysis of twitter data. DOI: https://arxiv.org/abs/1711.10377.