Analyzing The Interaction with Monkeypox Vaccine in Twitter Using Customized Sentiment Analysis Dashboard

Senior Project

by

Fatima Bailoun

Hiba Dawi

Submitted to the Data Science Department of the

Lebanese University

Beirut, Lebanon

In partial fulfillment of the requirements for the degree of

BACHELOR OF DATA SCIENCE

Spring 2022

Approved by:

Supervisor Dr. Hussein Y. Hazimeh

Committee Members Dr. Mageda Sharafeddine, Dr. Hussein Hazimeh,

Dr. Amal Kchour

ACKNOWLEDGMENT

First and foremost, I would like to praise and thank God, the Almighty, who has granted us countless blessings, knowledge, and opportunity to us, so that we have been finally able to accomplish this project.

We owe our gratitude to Dr. Hussein Hazimeh, the senior project supervisor, for the successful pursuit of our project so far and his kind patronage.

To my committee, Dr. Mageda Sharafeddine, Dr. Hussein Hazimeh, and Dr. Amal kchour, we are extremely grateful for your assistance and suggestions throughout our project.

Our thanks to the Faculty of Information at Lebanese University, for feeling proud of being members of its students, which helps improve our educational level and skills.

We sincerely appreciate our caring parents for functioning as a constant source of motivation and inspiration. We also acknowledge the important roles that our friends have played in our lives.

ABSTRACT

This proposal is prepared for the Data science senior project Committee at the Lebanese university. Before April 2022, human cases of the endemic monkeypox virus outside of Africa were infrequently reported. Currently, cases are occurring worldwide. Infection outcomes, risk factors, clinical presentation, and transmission all remain unclear. To guide the World Health Organization (WHO) Secretariat on the use of smallpox and monkeypox vaccinations, an Ad-hoc Strategic Advisory Group of Experts (SAGE) on Immunization Working Group was established in April 2022.

Globally, people were suffering from a fatal virus which is the coronavirus, which affects each country economically, psychologically, and sociologically. These consequences turned people to be worried about any ineffective virus that may furbish this misery. Unfortunately, a few months ago another viroid diffused, called "Monkeypox", synchronistically its vaccine existed.

To know the global feedback about this vaccine, we scraped Twitter to collect tweets that form our dataset. After that, sentiment analysis is applied to reach our target.

Keywords: Monkeypox, vaccine, coronavirus, Twitter, dataset; tweets, Data Science, World Health Organization.

TABLE OF CONTENTS

TABLE OF FIGURES

LIST OF TABLES

Chapter One: Introduction

1.1 Introduction:

Monkeypox (MPX) may be a disease caused by an infection with the monkeypox virus. This virus is an element of the identical family to smallpox but usually has milder symptoms than smallpox. Monkeypox is called so because it had been first discovered in 1958 in colonies of research monkeys. Despite being called "monkeypox", the source of the disease remains unknown. the primary human case was recorded in 1970 within the Democratic Republic of the Congo and has since been reported in several central and western African countries.

Monkeypox has gained prominence this year because of a global outbreak that began in May 2022, with over 1000 cases within the US as of July 2022.

Health experts say that vaccines used during the smallpox eradication program also protect monkeypox. However, newer vaccines are developed, one in every of which has been approved specifically for the prevention of monkeypox, per the globe Health Organization (WHO).

1.2 Problem Statement:

Raising awareness of risk factors and educating people about the measures they will go to reduce exposure to the virus is the main prevention strategy for monkeypox. Scientific studies are now underway to assess the feasibility and appropriateness of vaccination for the prevention and control of monkeypox. Some countries have or are developing, policies to supply vaccines to persons who could also be in danger like laboratory personnel, rapid response teams, and medical experts. After that, this vaccine is available to the entire community. Therefore, this project is completed to check the people's point of view about the monkeypox vaccine. Since Twitter is the best platform to share thoughts on a classy topic, thus we depend on it to urge our data.

The main aim of this project is to succeed in a transparent result about the world's view regarding the monkeypox vaccine, and whether or not they can take it[2].

1.3 Motivation

Interest in vaccines against monkeypox is growing because the outbreak spreads around the world, but on the opposite hand, people are terrified of any new health catastrophe, after the covid-19 era. Thus, we shop around for people's opinions to research them sentimentally and predict how ready they're to induce the monkeypox vaccine[10].

As a result, there has been an incredible increase in the use of social media platforms in the recent past. Twitter, one such social media platform, is employed by people of just about all age groups from all parts of the planet. At present, there are about 450 million monthly active users on Twitter. Hence, mining social media conversations, for example, tweets, to develop datasets has been of great interest to the scientific community within the previous few years.

Studying such a subject can help every government to make your mind up on the acceptable policies that ought to be followed. Moreover, health organizations can get pleasure from this study by taking into consideration the overall world's feedback about the monkeypox vaccine, thus they'll know the way to figure for creating people more satisfied with this vaccine. additionally, by reading this project, students can take a replacement idea about scraping Twitter and the way sentiment analysis is often done[11].

1.4 Contribution

The outcome of this research showed that clustered geo-tagged Twitter posts are accustomed better analyzing the dynamics in sentiments toward community–based infectious diseases-related discussions, like Monkeypox. this will provide additional city-level information to health policy in planning and decision-making regarding vaccine hesitancy for future outbreaks.

Collecting data from Twitter is wiped out several ways, in our study we used the *snscrape python library* to gather tweets. This dataset which is made from the monkeypox tweets is then undergone preprocessing to scrub it, after that, sentiment analysis is completed for it using the *Textblob library* in python. The results are plotted in graphs using the *seaborn python library*[12].

Chapter Two: State of the Art

2.1 Introduction:

In this chapter, we'll introduce some projects just like our project ("Monkeypox Vaccine Feedbacks Sentiment Analysis Dashboard Using Twitter"). from these related projects, we might take pleasure in their advantages and forestall repeating their errors.

2.2 Similar Projects:

2.2.1 Similar Project about monkeypox:

In this report *“MonkeyPox2022Tweets: The First Public Twitter Dataset on the 2022 MonkeyPox Outbreak”*, The monkeypox virus has affected 16,836 people in 74 different nations since the first incidence on May 7, 2022, and the number of cases is rising. The World Health Organization just proclaimed monkeypox a worldwide health emergency following a recent "emergency meeting." As a result, several nations are putting in place various types of safeguards, regulations, and guidelines to stop the virus from spreading. Due to these regulations and the rise in cases around the world, there have been increasing discussions regarding monkeypox information sharing and searching on social media, particularly Twitter. Huge volumes of Big Data are being produced as a result of these discussions. The scientific community in the domains of Big Data, Data Mining, Natural Language Processing, and its connected areas has been very interested in mining Twitter conversations on certain subjects, infections, or global concerns to develop datasets in recent years. Even though several Twitter datasets on various subjects have been created in the past, none of them contain a collection of tweets about the present monkeypox outbreak. Additionally, none of the studies about the breakout of monkeypox in 2022 have, to now, concentrated on the creation of any such datasets. The scientific community in the domains of Big Data, Data Mining, Natural Language Processing, and its connected areas has been very interested in mining Twitter conversations on certain subjects, infections, or global concerns to develop datasets in recent years. Even though several Twitter datasets on various subjects have been created in the past, none of them contain a collection of tweets about the present monkeypox outbreak. Additionally, none of the studies about the breakout of monkeypox in 2022 have, to now, concentrated on the creation of any such datasets[6].

Supplementary Materials: Not applicable.

Funding: This research received no external funding

Institutional Review Board Statement: Not applicable

Informed Consent Statement: Not applicable

Data Availability Statement: The data presented in this study are publicly available at https://doi.org/10.5281/zenodo.6898178

Conflicts of Interest: The author declares no conflict of interest

2.2.2 Similar Project about Twitter Sentiment Analysis:

In this report "*Public sentiments toward COVID-19 vaccines in South African cities: An analysis of Twitter posts*", the scientific community in the domains of Big Data, Data Mining, Natural Language Processing, and its connected areas has been very interested in mining Twitter conversations on certain subjects, infections, or global concerns to develop datasets in recent years. Even though several Twitter datasets on various subjects have been created in the past, none of them contain a collection of tweets about the present monkeypox outbreak. Additionally, none of the studies about the breakout of monkeypox in 2022 have, to now, concentrated on the creation of any such datasets. The scientific community in the domains of Big Data, Data Mining, Natural Language Processing, and its connected areas has been very interested in mining Twitter conversations on certain subjects, infections, or global concerns to develop datasets in recent years. Even though several Twitter datasets on various subjects have been created in the past, none of them contain a collection of tweets about the present monkeypox outbreak. Additionally, none of the studies about the breakout of monkeypox in 2022 have, to now, concentrated on the creation of any such datasets[7].

Supplementary Materials: The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpubh.022.987376/full#supplementary-material

Funding: This research was funded by Canada's International Development Research Centre (IDRC) and Swedish International Development Cooperation Agency (SIDA) (Grant No. 109559-001).

Institutional Review Board Statement: Not applicable

Informed Consent Statement: Not applicable

Data Availability Statement: The dataset used for this study can be found in the online repository at: https://www.kaggle.com/datasets/ogbuokiriblessing/tweetdatasa.

Conflicts of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

2.2.3 Similar Project about mining social media to understand user opinions:

This research "*Mining Social Media to Understand User Opinions On IoT Security and Privacy*" reveals the perspectives of Twitter and Reddit users regarding how they see the security and privacy of IoT devices. IoT security and privacy are generally seen favorably on Twitter, whereas they are often viewed negatively on Reddit. The consumer impression of IoT security and privacy is rapidly shifting, and users on Twitter and Reddit have recently switched from unfavorable opinions to good ones. This is a great finding. Some topics, such as smart TVs, drones, speakers, voice assistants, fitness trackers, and smartwatches, have drawn criticism from both Twitter and Reddit users.

By gathering information on devices like Alexa, Fitbit, apple watch, and many more, the research can be expanded to discover customers' sentiments linked to specific IoT items in the aforementioned domains.

The collected data is itself considered a dataset that is labeled using a flare pre-trained model, thus there are some limits. However, additional research can try labeling data using VADER and Textblob to observe how the results differ. A pre-trained BERT base uncased model is employed in this study, however, we can also train the BERT model with our data and compare the outcomes. This has not been attempted in this research because doing so might result in overfitting [33]. LDA model parameters can also be changed. The stop words for each LDA model were expanded for this study to concentrate more on topics relating to security and privacy, which may also be changed. The year-by-year sentiment study for each field could not be completed due to time constraints, but it would be beneficial to identify the sentiment for each topic and track changes. Instead, extensive data from Twitter and Reddit is used to do year-wise sentiment analysis[8].

Supplementary Materials: Not applicable

Funding: This research received no external funding

Institutional Review Board Statement: Not applicable

Informed Consent Statement: Not applicable

Data Availability Statement: The query search keywords used for data collection in this research are shown in the excel sheet: Google Doc Link https://docs.google.com/spreadsheets/d/1OfEnKi2PZZlmpSTs2y1CuIHXzdh2HEfN0ZQCxNInljU/edit?usp=sharing

Conflicts of Interest: The author declares no conflict of interest.

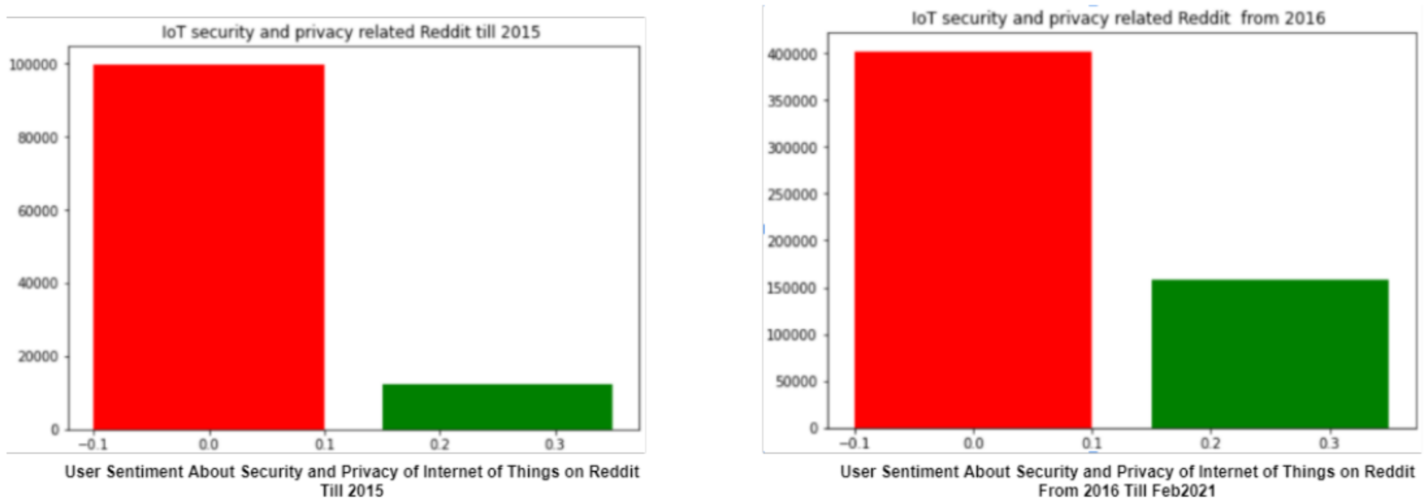


Figure 1: User Sentiment About Security and Privacy of IoT on Reddit

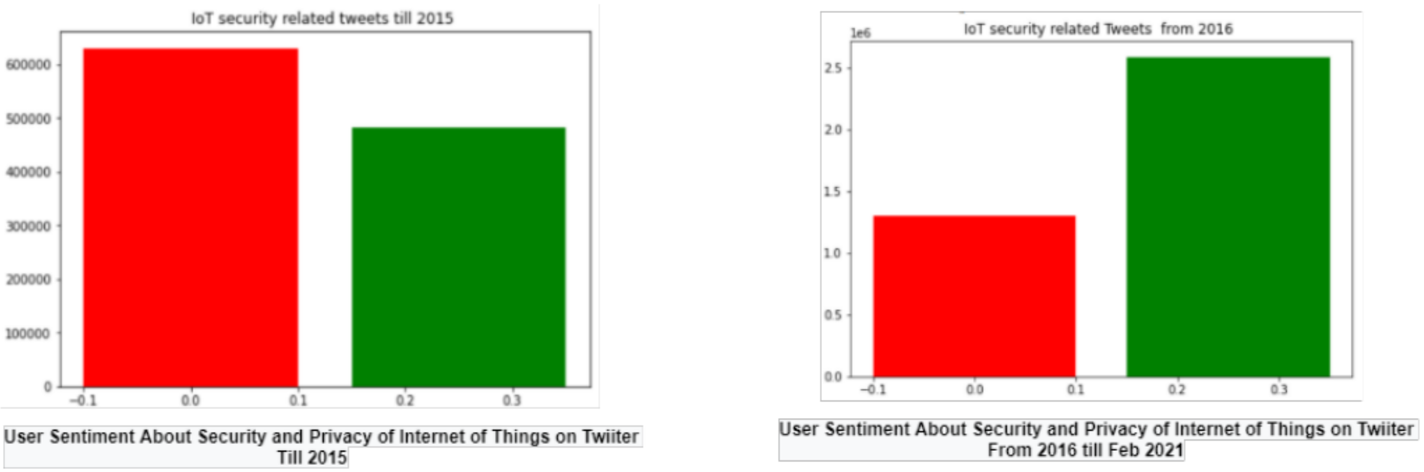


Figure 2:User Sentiment About Security and Privacy of IoT on Twitter

Chapter Three: Implementation

3.1 Introduction:

This chapter discusses the steps involved in the implementation and the tools and libraries that were utilized.

3.2 Implementation Tools:

The language used for implementation in this project is: Python
The libraries in our code are snscrape, pandas, NumPy, matplotlib.pyplot, and pandas.plotting, seaborn, neattext.functions, textblob, word cloud, nltk, PIL, plotly_express, statsmodels, sklearn, pmdarima.

3.3 Implementation steps:

1st, to **scrape** the needed tweets from Twitter, we install and import the *snscrape* library, then we write a query to select the specific characteristics of our tweets. We defined the language of the tweets we want to scrape, it was English[3]. The tweets are posted "since:2022-04-01 until:2022-08-18" and this period was chosen because monkeypox goes as a public topic since the first of April, and we stopped scraping on the 18 of August because we scraped the tweets on that day.

Then, we detect the attributes we needed to study from our tweets, the attributes were: Date, Username, Location, Verified, Hashtag, and text of the Tweet. After that, we store them in a data frame and save them as a CSV file[2].

2nd, Text **Preprocessing** was done in which we installed and imported the *neattext* library to clean the CSV columns from any noise. Noise in our dataset means the unneeded characters in the text tweet such as mentions (user handles), hashtags, URLs (hyperlinks), multiple white spaces, and punctuations[2].

3rd, we applied **Sentiment Analysis** to our cleaned data using the *TextBlob* library. We get our tweets' polarity, subjectivity, and sentiment. In addition to that, we used the *pandas* library to import the function json_normalize () to normalize the results in a table. Then we plotted each polarity, subjectivity, and sentiment on a graph concerning the number of tweets[5].

4th, we applied **Keyword Extraction** for neutral, positive, and negative sentiment. After dividing the tweets into neutral, positive, and negative, we removed stop words and converted them to tokens to apply tokenization to plot some histograms for the 3 data frames (pos_df, neg_df, neut_df) using the *seaborn* library.

5th, using the *wordcloud* library we plotted four **WordCloud graphs** that show the most repeated tokens in the tweets. The first graph was for all tweets, and the other three graphs are for positive, negative, and neutral respectively.

6th, using the *FreqDist* () function from the *NLTK* library, we got the repeated hashtags and the value of repetition of each one and stored them in a data frame using the *pandas* library. After that, we plotted the **Top 5 Hashtags** in our data as a bar graph, via the *seaborn* library.

7th, via the *matplotlib.pyplot* library we applied the **Pie Chart** on the verified column in our CSV file to show the percentage of the verified accounts that tweeted about the monkeypox vaccine topic.

8th, we created a data frame that contains the **sentiment data concerning each month**, and the count value of each sentiment type (pos, neg, neut) then we plotted the “Distribution of Sentiment Classes Among Months” in the form of a bar graph via the *matplotlib.pyplot* library. In addition to that, we plotted the **Distribution of All Tweets Among Months**.

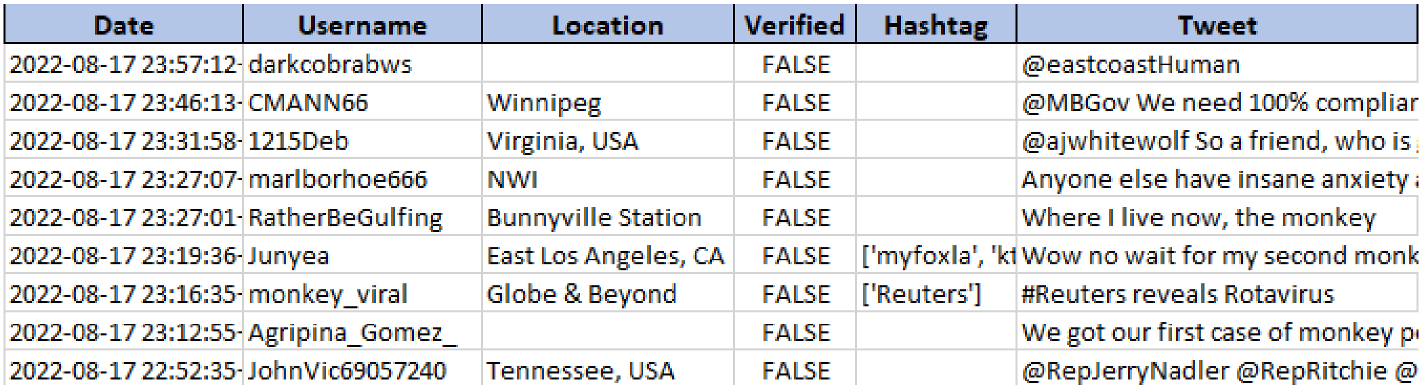
9th, completing with the above data frame we used the function *groupby ()* to group the tweets by months, and we got the polarity mean in each month. We visualized this result in a **line chart** for more emphasis on the mean changing from one month to another. Working with polarity, we plotted a figure with “Polarity” on the y-axis and “Time” on the x-axis. The tweets are distributed concerning the day they are tweeted, to finally have a whole figure that can be analyzed soon.

10th, dealing with location attribute, we built a function named *plot_map (),* which takes two arguments: data frame name, and column name. We called our cleaned data frame with the “polarity” column, to be distributed concerning the “Location” column. Our **Map** appeared successfully, thus we can know now the polarity in each country, which will give us a general glace about the public opinion in that country.[6]

Chapter Four: Data Exploration & Preprocessing

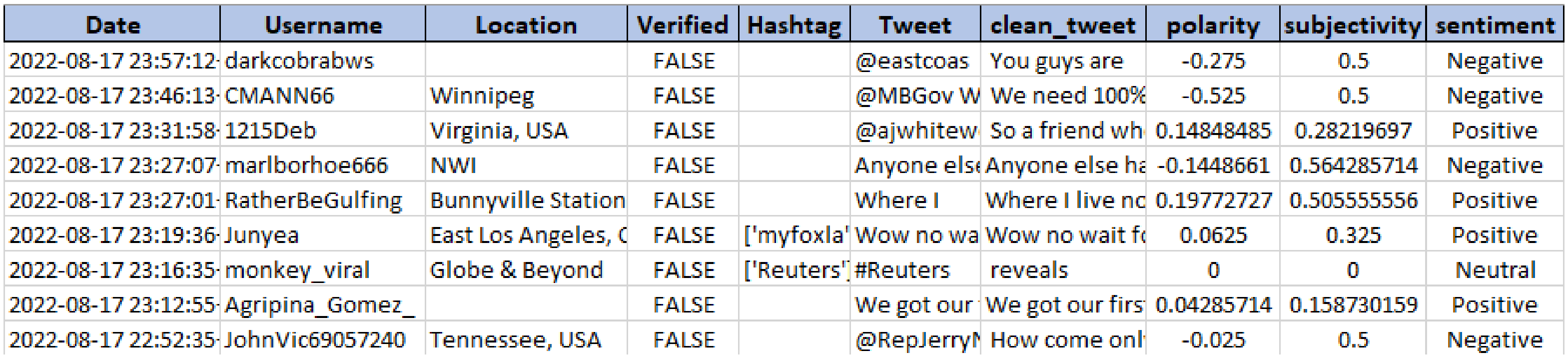
4.1 Introduction:

In this chapter, we will introduce the analysis part done on the collected data. The data was collected through the *snscrape* library from Twitter. More than 33,000 tweets were collected. Different steps were done starting from data explorations, data preprocessing, etc...



Date	Username	Location	Verified	Hashtag	Tweet
2022-08-17 23:57:12-	darkcobrabws		FALSE		@eastcoastHuman
2022-08-17 23:46:13-	CMANN66	Winnipeg	FALSE		@MBGov We need 100% compliar
2022-08-17 23:31:58-	1215Deb	Virginia, USA	FALSE		@ajwhitewolf So a friend, who is
2022-08-17 23:27:07-	marlborhoe666	NWI	FALSE		Anyone else have insane anxiety
2022-08-17 23:27:01-	RatherBeGulfing	Bunnyville Station	FALSE		Where I live now, the monkey
2022-08-17 23:19:36-	Junyea	East Los Angeles, CA	FALSE	['myfoxla', 'kt	Wow no wait for my second monk
2022-08-17 23:16:35-	monkey_viral	Globe & Beyond	FALSE	['Reuters']	#Reuters reveals Rotavirus
2022-08-17 23:12:55-	Agripina_Gomez_		FALSE		We got our first case of monkey p
2022-08-17 22:52:35-	JohnVic69057240	Tennessee, USA	FALSE		@RepJerryNadler @RepRitchie @

Table 1: The Initial Dataset



Date	Username	Location	Verified	Hashtag	Tweet	clean_tweet	polarity	subjectivity	sentiment
2022-08-17 23:57:12-	darkcobrabws		FALSE		@eastcoas	You guys are	-0.275	0.5	Negative
2022-08-17 23:46:13-	CMANN66	Winnipeg	FALSE		@MBGov W	We need 100%	-0.525	0.5	Negative
2022-08-17 23:31:58-	1215Deb	Virginia, USA	FALSE		@ajwhitew	So a friend wh	0.14848485	0.28219697	Positive
2022-08-17 23:27:07-	marlborhoe666	NWI	FALSE		Anyone els	Anyone else ha	-0.1448661	0.564285714	Negative
2022-08-17 23:27:01-	RatherBeGulfing	Bunnyville Station	FALSE		Where I	Where I live nc	0.19772727	0.505555556	Positive
2022-08-17 23:19:36-	Junyea	East Los Angeles, C	FALSE	['myfoxla'	Wow no wa	Wow no wait fc	0.0625	0.325	Positive
2022-08-17 23:16:35-	monkey_viral	Globe & Beyond	FALSE	['Reuters']	#Reuters	reveals	0	0	Neutral
2022-08-17 23:12:55-	Agripina_Gomez_		FALSE		We got our	We got our firs	0.04285714	0.158730159	Positive
2022-08-17 22:52:35-	JohnVic69057240	Tennessee, USA	FALSE		@RepJerryN	How come onl	-0.025	0.5	Negative

Table 2: Dataset After Cleaning and Sentiment Analysis

4.2 Data Exploration

The data was saved in two different separated CSV files. The first one was the data before processing and cleaning and was saved in a CSV file named "monkeypox_1.csv". while the second was the cleaned data that will be analyzed, it is in a CSV file named "polarity_monkeypox.csv". These CSV files were imported using pandas read_csv () function[5].

4.2.1 Attributes Description:

Name of Attribute	Data Type	Description	After/Before processing
Date	Datetime	gives the day-month-year and the time the tweet was posted	Before
Username	Text	the name of the user that writes the tweet	Before
Location	Text	from where the tweet was posted	Before
Verified	Boolean	True: Verified Account False: Unverified Account	Before
Hashtag	Text	Gives the hashtags that appeared in each tweet	Before
Tweet	Text	Shows the text tweeted	Before
clean_tweet	Text	Same as Tweet attribute, but missing the hashtags, user handles (mentions), punctuations, and links	After
polarity	Float	Lies in the range of [-1,1], helps in understanding the opinion expressed by the tweet. From the sign of the polarity score, we can detect the sentiment type	After
subjectivity	Float	lies in the range [0,1], which quantifies the amount of personal opinion and factual information contained in the text. The higher subjectivity means that the text contains personal opinions rather than factual information.	After
Sentiment	Text	If polarity > 0 sentiment = positive If polarity > 0 sentiment = negative If polarity = 0 sentiment = neutral	After

Table 3: Attribute Description

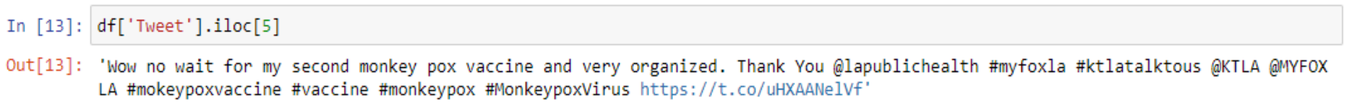
4.3 Data Preprocessing:

One of the crucial phases of every machine learning research is data preprocessing. Before feeding the data into a machine learning system, it involves cleaning and formatting the data. For preprocessing we install and imported the *neattext library*, which includes various helpful functions to apply better cleaning. The updated cleaned tweets were saved in a new column named "clean_tweet".

The following activities make up the NLP preprocessing steps:

- removing hashtags
- removing mentions (user handles)
- Removing Multiple White Spaces
- Removing Links and URLs
- Removing Punctuations

- Let's take one tweet as an example to confirm that the procedure is being performed correctly:



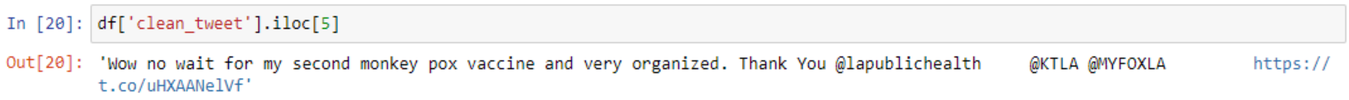
```
In [13]: df['Tweet'].iloc[5]

Out[13]: 'Wow no wait for my second monkey pox vaccine and very organized. Thank You @lapublichealth #myfoxla #ktlatalktous @KTLA @MYFOX
         LA #mokeypoxvaccine #vaccine #monkeypox #MonkeypoxVirus https://t.co/uHXAANelVf'
```

Figure 3: The unprocessed tweet

Removing Hashtags:

To have more trusted results in polarity, we decided to remove hashtags. Thus we used the *extract_hashtags ()* function[4].



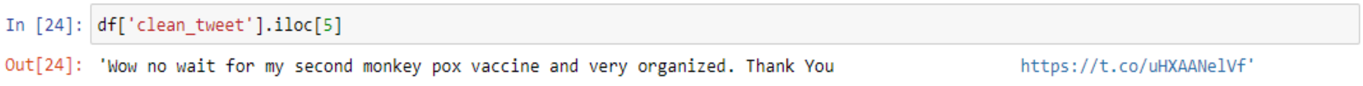
```
In [20]: df['clean_tweet'].iloc[5]

Out[20]: 'Wow no wait for my second monkey pox vaccine and very organized. Thank You @lapublichealth     @KTLA @MYFOXLA         https://
         t.co/uHXAANelVf'
```

Figure 4: Tweet After Removing Hashtags

Removing Mentions (user handles):

Mentions in tweets are useless in analysis, so to get rid of them we used the *remove_userhandles ()* function[4].



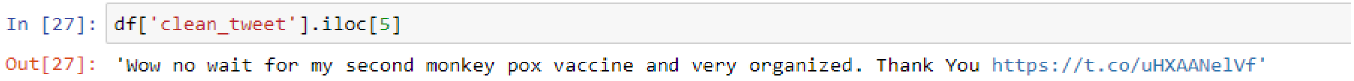
```
In [24]: df['clean_tweet'].iloc[5]

Out[24]: 'Wow no wait for my second monkey pox vaccine and very organized. Thank You                  https://t.co/uHXAANelVf'
```

Figure 5: Tweet After Removing Mentions

Removing Multiple White Spaces:

Due to the extraction of hashtags and mentions, extra white spaces appeared in the cleaned tweets. Therefore, removing them is a must, so we used the remove_multiple_spaces () function[4].



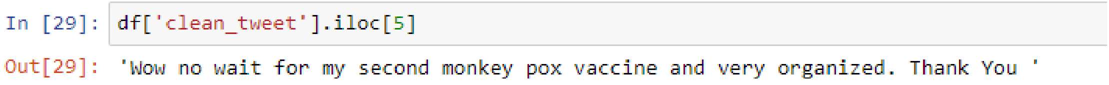
```
In [27]: df['clean_tweet'].iloc[5]

Out[27]: 'Wow no wait for my second monkey pox vaccine and very organized. Thank You https://t.co/uHXAANelVf'
```

Figure 6: Tweet After Removing White Spaces

Removing Links and URLs:

The URL (or Uniform Resource Locator) in a text refers to a site on the internet but offers no other details. As a result, we also eliminate these using the remove_urls() function provided by the neattext library[4].



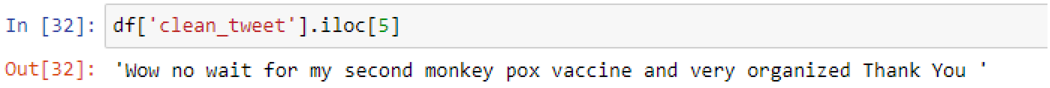
```
In [29]: df['clean_tweet'].iloc[5]

Out[29]: 'Wow no wait for my second monkey pox vaccine and very organized. Thank You '
```

Figure 7: Tweet After Removing Urls

Removing Punctuations:

The removal of punctuation marks is an essential NLP preprocessing step because these marks, which are used to separate text into sentences, paragraphs, and phrases, have an impact on the outcomes of any text processing approach, particularly those that depend on the frequency of occurrences of words and phrases. So to get rid of them we used the remove_puncts () function[4].



```
In [32]: df['clean_tweet'].iloc[5]

Out[32]: 'Wow no wait for my second monkey pox vaccine and very organized Thank You '
```

Figure 8: Tweet After Removing Punctuations

Now, we have accomplished the necessary text pre-processing, and our NLP task now has meaningful text.

Chapter Five: Data Analysis & Interpretation

5.1 introduction:

This section presents the results and findings of this work. After all the preprocessing part was done, exploring the data is a must to see the diversity in the data collected which will help in understanding it more.

5.2 Sentiment Analysis:

Twitter Sentiment Analysis is the usage of advanced text mining techniques to analyze the sentiment of the tweet in the form of positive, negative, and neutral. It is often referred to as "opinion mining" and is largely used to analyze discussions, opinions, and viewpoints (all expressed in the form of tweets) to determine commercial strategy, conduct political analysis, and evaluate societal behavior.

In our study, we used the *TextBlob* library to achieve Sentiment Analysis. Then, we built a function named *get_sentiment ()* that takes a text as an argument and gives its polarity and subjectivity. According to the polarity value, we ran an if statement that checks the sign of this value and prints down the type of sentiment. If the polarity is under 0, it prints “Negative”, while if it is above 0 it prints “Positive”, else it will print “Neutral”.

After that, we plotted a histogram that shows the distribution of tweets according to their polarity value[1].

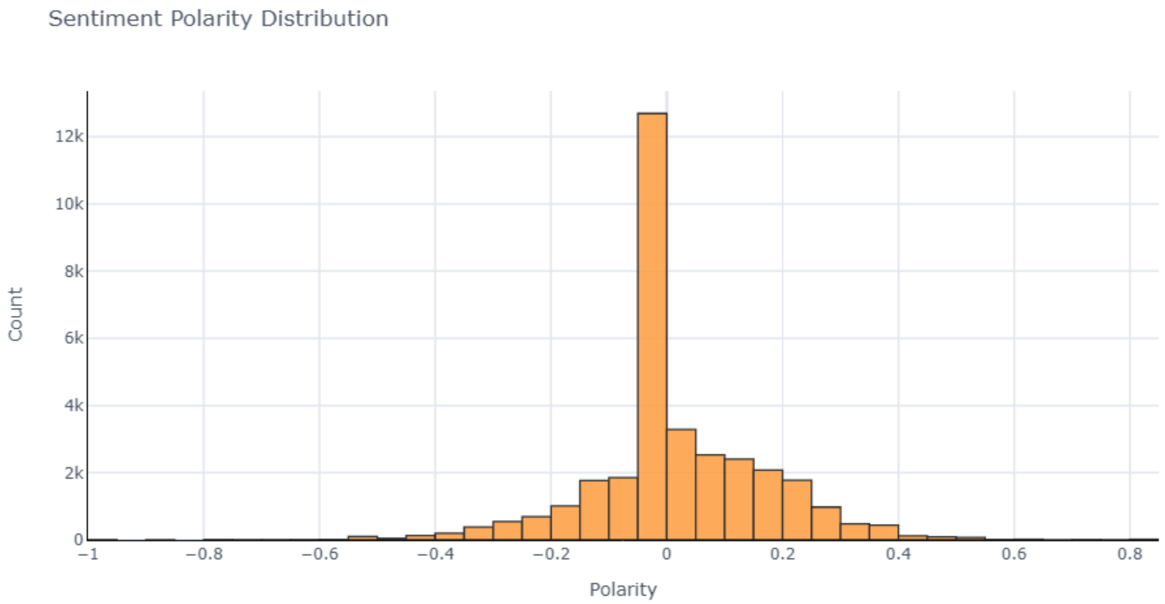


Figure 9: Tweets Distribution According to Polarity

- This histogram shows that the most frequent value is scored on the negative side, which is -0.05 (more than 12,000 tweets). This score demonstrates that, although being negative, it is extremely close to zero, indicating that these users' opinions are not particularly strongly negative.

We created two kinds of visualization, to determine the number of each sentiment type (Positive, Negative, and Neutral). The first one is a pie chart that gives the result as percentages, whereas the second one is a bar graph that gives the count of each type[1].

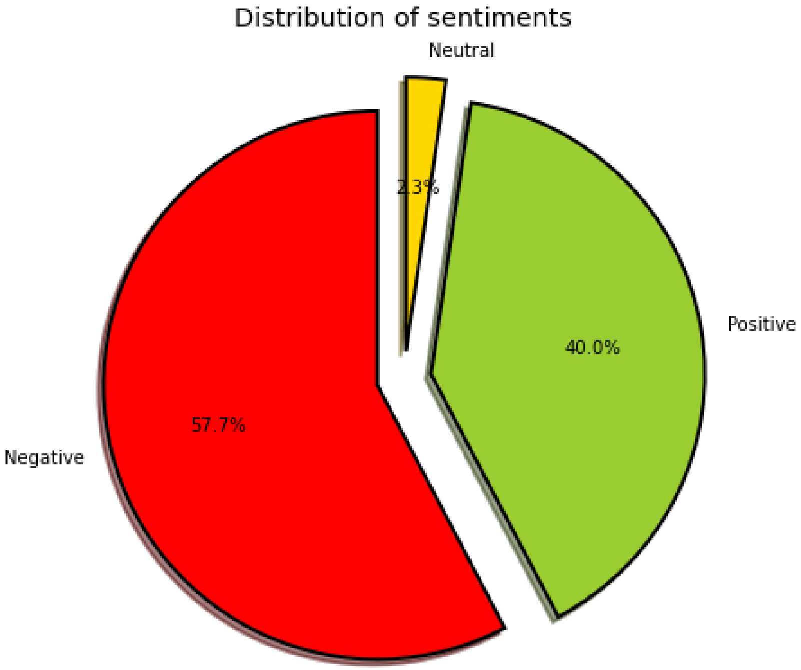


Figure 10: Pie Chart of Sentiment Types

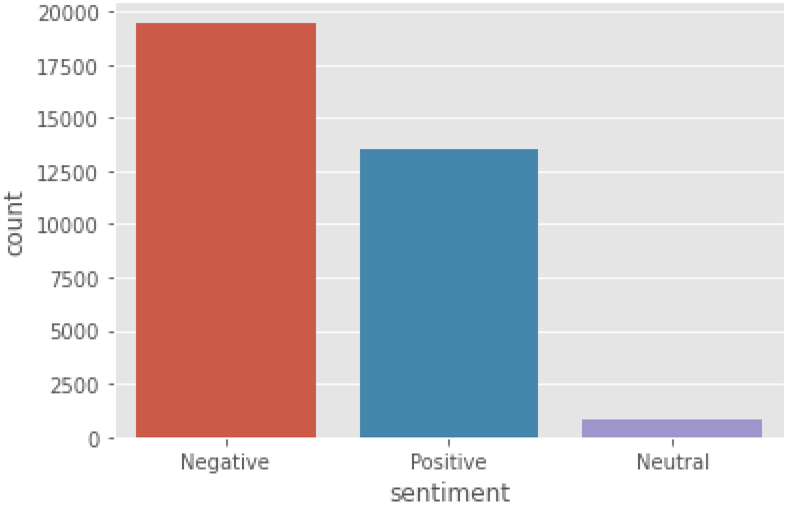


Figure 11: Bar Graph of Sentiment Types

- These visualizations indicate that we have more than 19,000 tweets (57.7%) for type “Negative”, more than 13,000 tweets (40%) for “Positive”, and less than 800 tweets (2.3%) for “Neutral”. This very low percentage in the “Neutral” type makes sense that the collected tweets can be identified as “Negative” or “Positive”, which allows for a more precise analysis.

5.3 Tokenization:

A token is a piece of a whole, so a word is a token in a sentence, and a sentence is a token in a paragraph. Tokenization is the process of splitting a string into a list of tokens. After writing a query to apply tokenization on the positive, negative, and neutral sentiment data frames, 3 lists of tokens will appear for each data frame respectively. These lists will be used to count the most frequently tweeted word.

Moreover, we have done a barplot for each list to get clearer results for analysis.

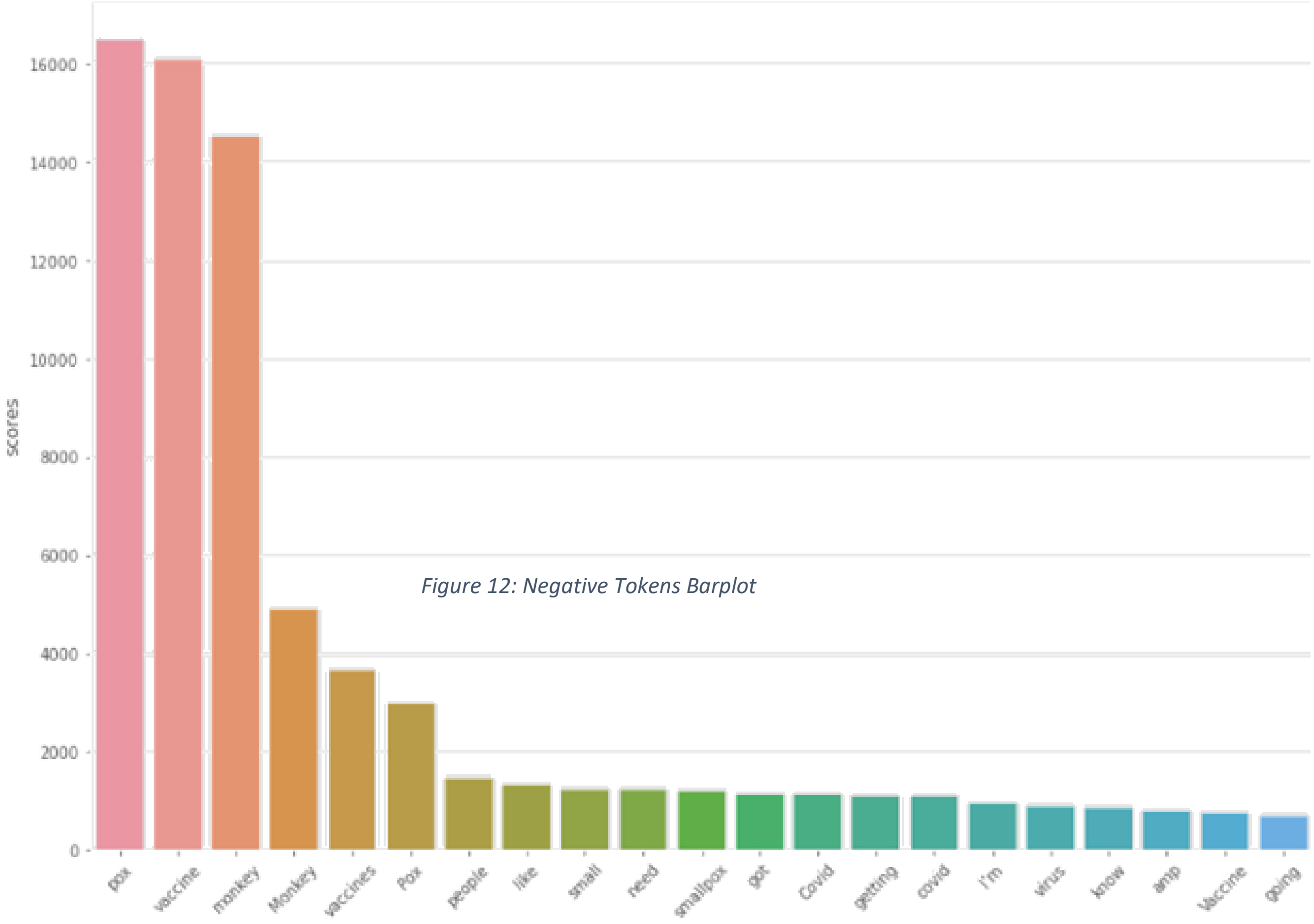


Figure 12: Negative Tokens Barplot

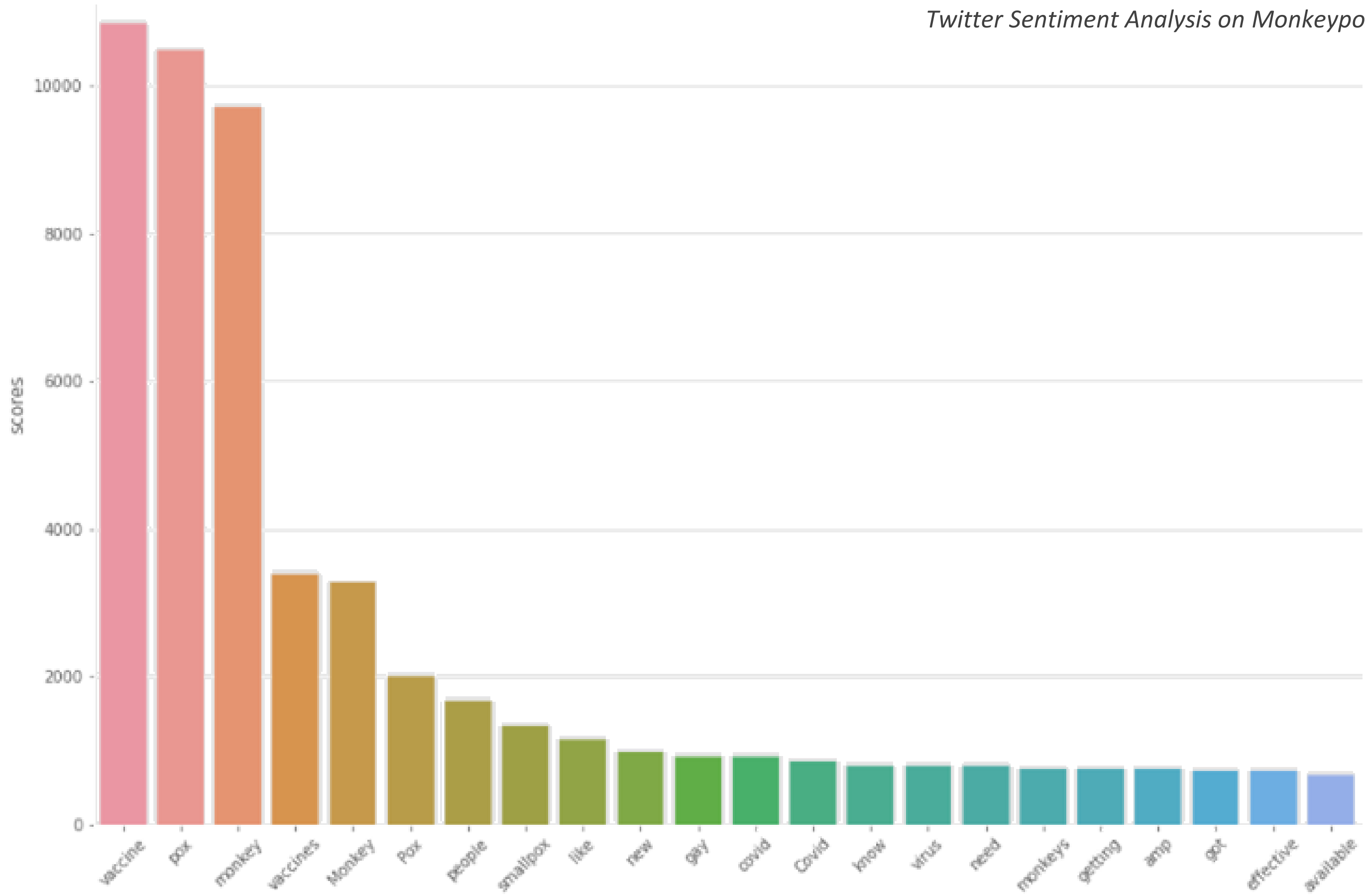


Figure 14: Positive Tokens Barplot

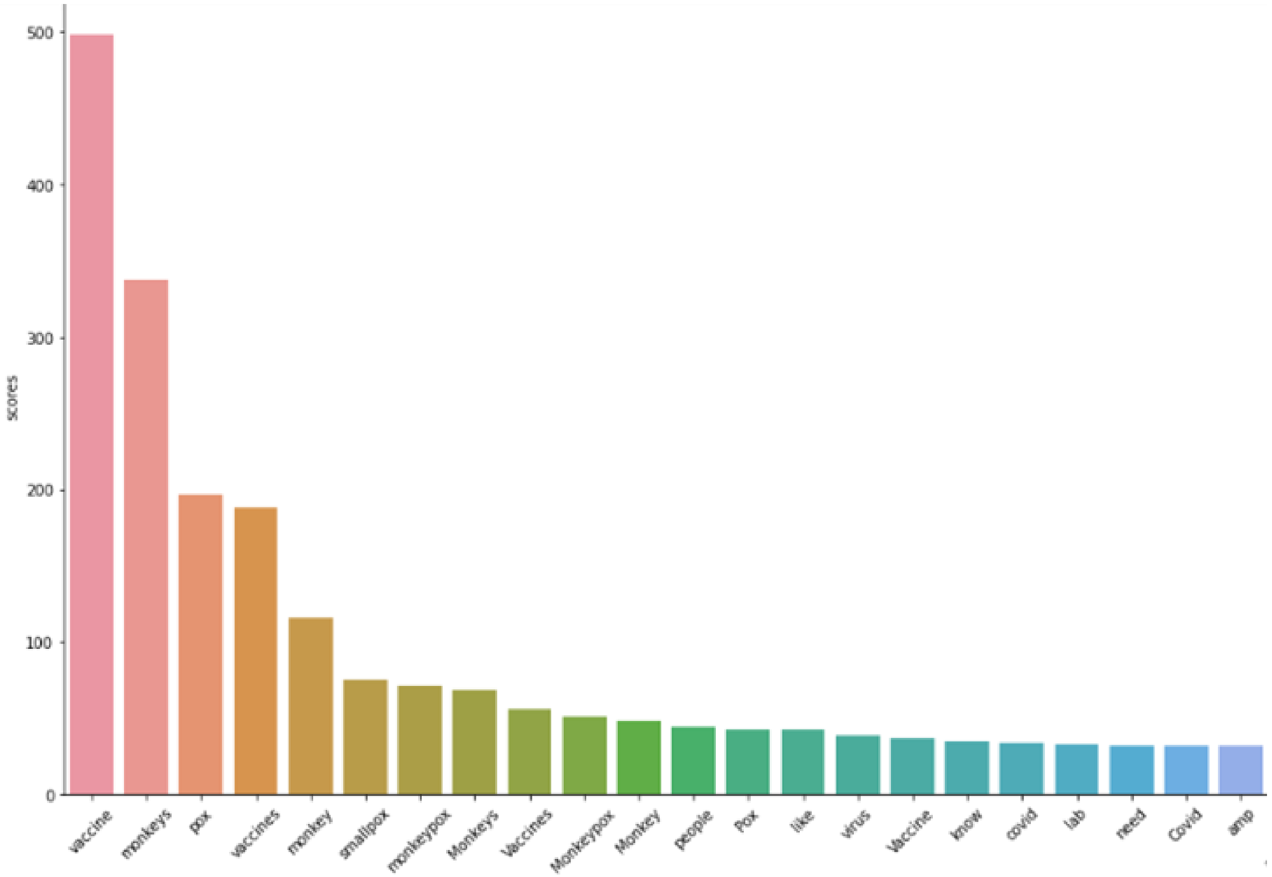


Figure 13: Neutral Tokens Barplot

- These bar plots identify the count of the most frequent tokens in each sentiment type. This step in coding facilitates the drawing of a WordCloud visualization.

5.4 Word Cloud

We can use a Word Cloud, a highly popular visual representation, to get a broad sense of the tone of our data. More than 33,000 tweets about monkeypox vaccination gathered between April 2022 and August 2022 were used to create a Word Cloud. Each word is written in a font size that corresponds to how often it appears in the text (i.e. bigger term means higher frequency).

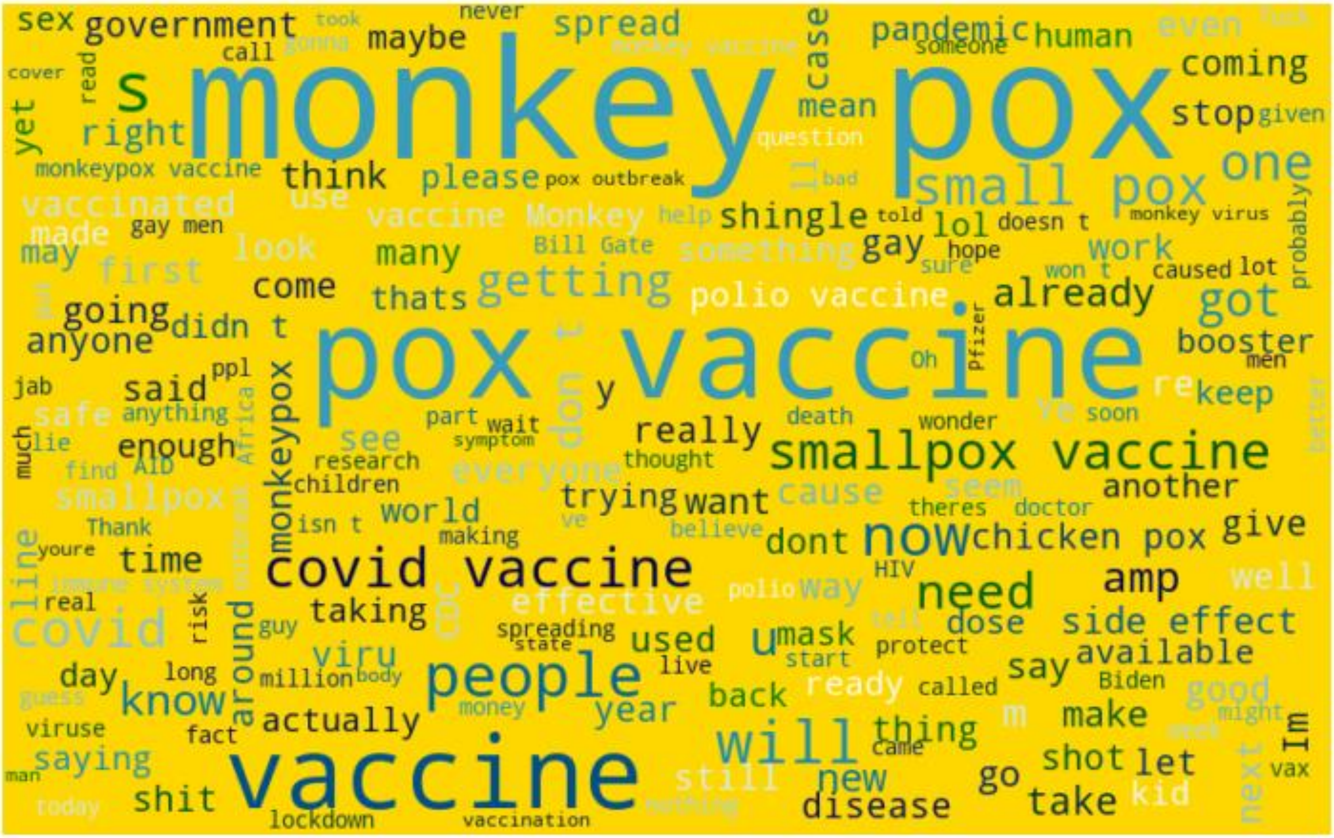


Figure 15: All Tweets Word Cloud

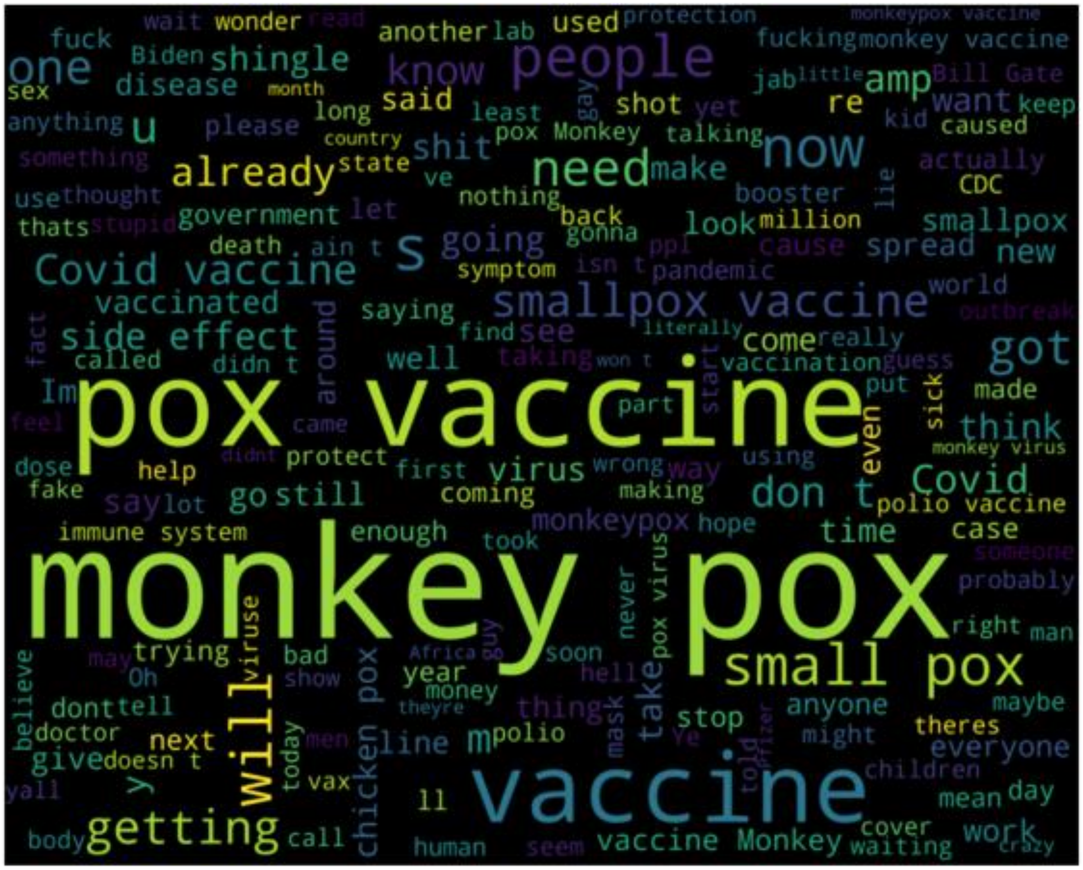


Figure 19: Negative Tweets Word Cloud

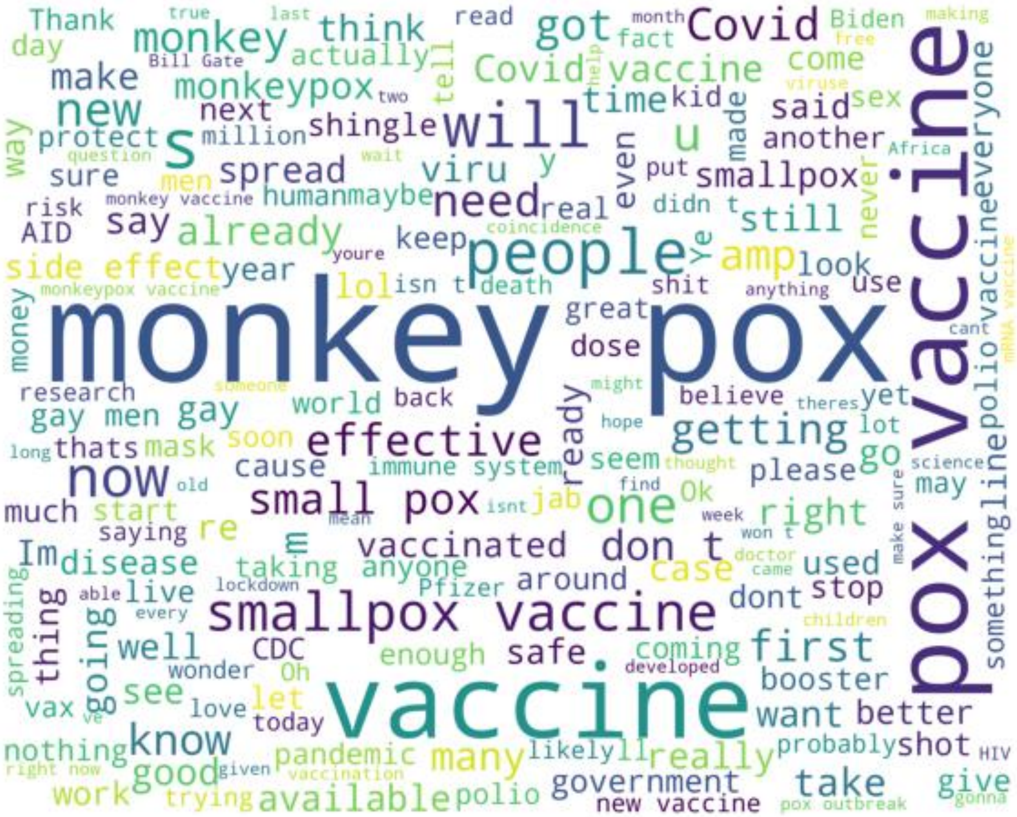


Figure 18: Positive Tweets Word Cloud

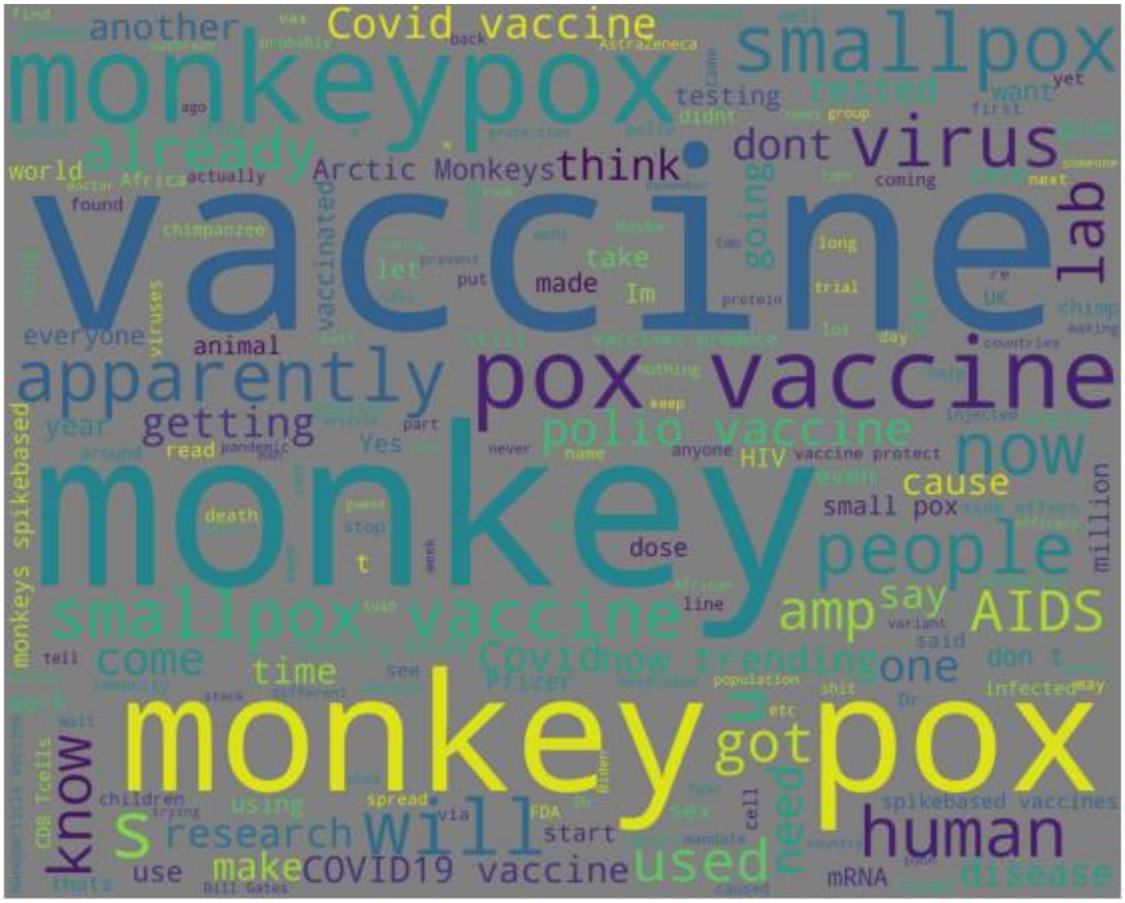


Figure 20: Neutral Tweets Word Cloud

- By looking at these figures, we can make some considerations:

  In Fig.16, in the negative word cloud, there are terms like polio, bad, stupid, and side effect. While in Fig.17 in the positive word cloud, we read words like believe, good, effective, and protect. Thus, even with this preliminary inspection, we can observe some differences between the corpus of positive and negative tweets.

  As expected, some very frequent terms such as monkeypox, vaccine, and a pandemic are in common with the word clouds, since they can be used in both contexts

5.5 Verified VS Unverified tweeters:

Verified accounts identify the most important individuals, including the World Health Organization, government politicians, influencers, and such. These persons affect the public's perception because they have a broad influence. Additionally, the fact that these accounts tweet about the monkeypox vaccine indicates that it is a widely discussed topic in all societies.

The number of verified accounts that tweeted about the monkeypox vaccination between April 2022 and August 2022 is represented in a pie chart that we've made to analyze these points of view.

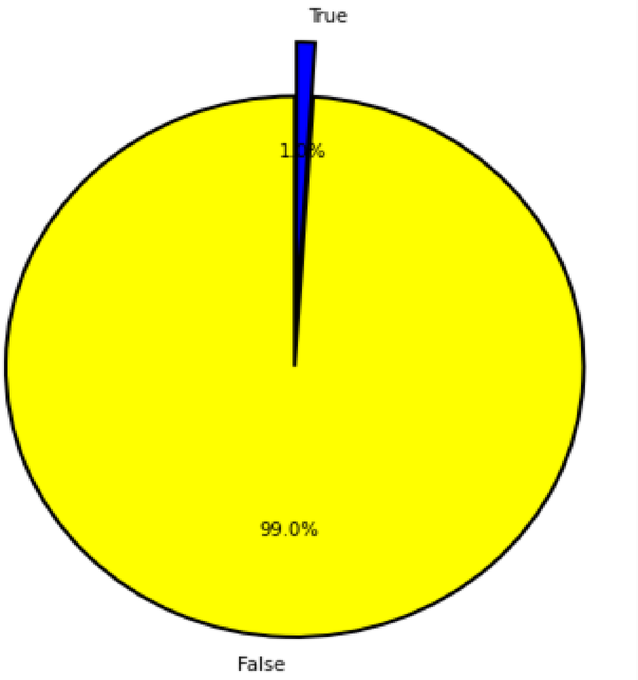


Figure 21: Quantity of Verified VS Unverified Accounts (True: verified, False: Unverified)

- According to the pie chart, approximately only 1% of verified accounts have tweeted about the monkeypox vaccine; This indicates that means the topic is not public in all states and countries; it is limited to specific locations.

5.6 Top 5 Hashtags:

Using hashtags is essentially a way to group conversations or content around a certain topic, making it easy for people to find content that interests them. Therefore, we wrote a couple of code that illustrates to us the top 5 hashtags used about the monkeypox vaccine topic in a bar graph.

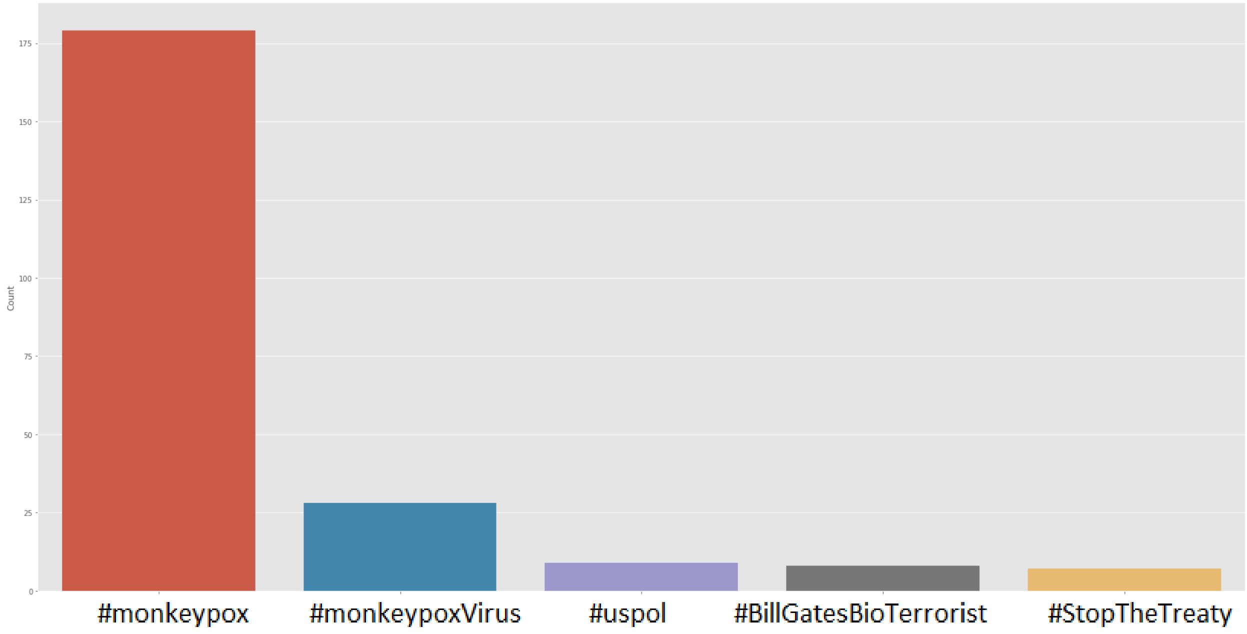


Figure 22:Top 5 Hashtags

- As shown in the above graph, the top 5 hashtags tweeted (in decreasing order) are monkeypox, monkeypoxVirus, uspol, BillGatesBioTerrorist, and StopTheTreaty.
- What can be interpreted from the following is that the last 3 hashtags are negatively expressing while the first two hashtags are so logical since it is the topic the tweets are talking about. This means that the general tone of the tweets is negative.

5.7 Sentiment Analysis with respect to Date:

In this section, we started dealing with both the "Date" and "Sentiment" columns to find the number of tweets as well as sentiment types scored in each month.

As a first step, we used the *group_by* () function to group the "Date" column by months. Then, we counted the number of tweets in each month and plotted the result in a bar graph for better visualization.

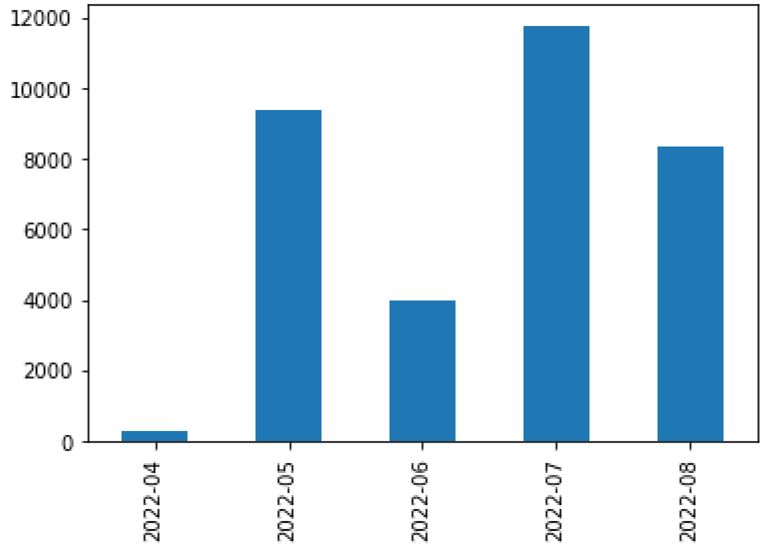


Figure 23: Distribution of Tweets Among Months

- This bar graph reveals the number of tweets each month. We have extracted tweets from the very beginning of month 4 (April), nevertheless, it scores a very low number of tweets (less than 300 tweets). This indicates that people weren't interested in the monkeypox vaccine at that time. But when looking to the next months, we can notice that the monkeypox vaccine started to be a trend on Twitter, and people begin talking about it as it reaches its peak (more than 11,000 tweets) in July (7th month).
- Note that we can't consider August (the 8th month) since we extracted tweets till 18 August only.

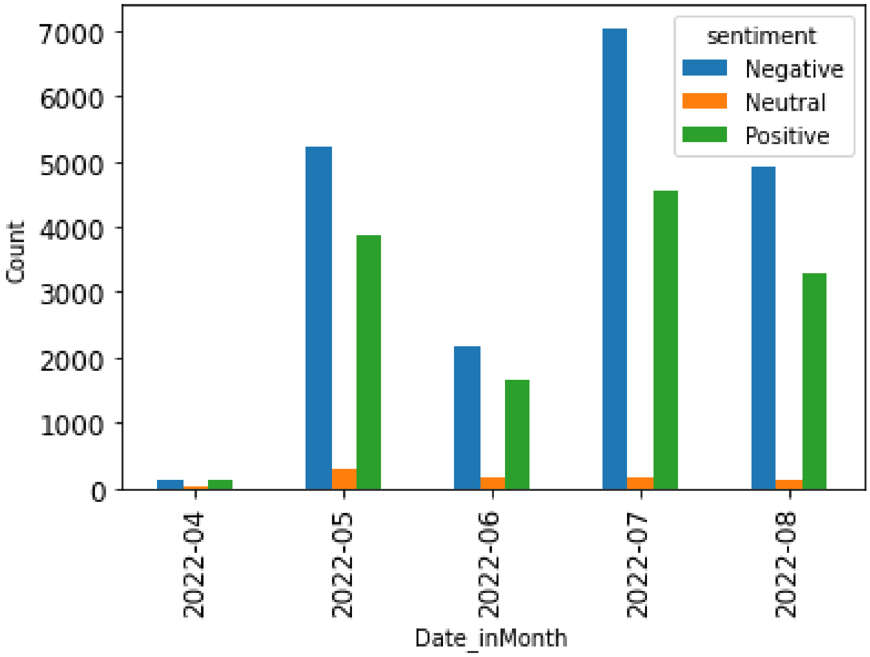


Figure 24: Distribution of Sentiment Classes Among Months

- We decided to broaden our attention and determine the distribution of each sentiment type throughout the months. Every month, we discovered, there were more negative than positive tweets, but if the negativity rose, the positives did too.

5.8 Polarity:

Due to the importance of the polarity value, and its impact on the sentiment result and analysis, we constructed a plot that shows the polarity value of each tweet every single day.

- Using the below graph, we found that the majority of the tweets fall between the range of +0.05 and -0.05, indicating that they are closest to zero. As a result, we can't say with certainty whether we have a generally positive or negative opinion. Consequently, some people's perspectives may evolve.

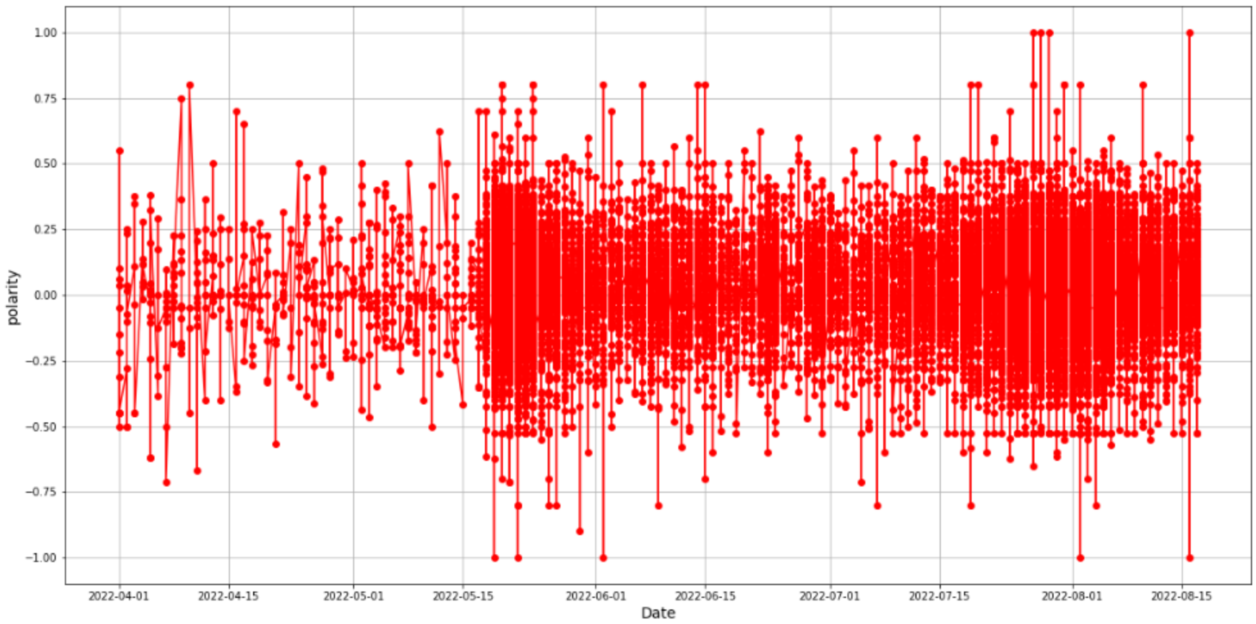


Figure 25: Polarity vs Time

For more accurate results, we choose to get the mean polarity in each month. Hence, we grouped the "Date" column by months and called the mean () function for the "polarity" column. After that, we plotted a line chart with “Polarity Mean” on the y-axis, and “Date in months” on the x-axis

- Through the below line chart, we've discovered that all the means are positive except in April. This result makes sense in our analysis where we have found above that number of negative tweets is higher than positive ones. This finding makes sense mathematically because, despite the higher quantity of negative tweets, the positive tweets have higher positive values. Thus, we can consider that positive tweets' users are more convinced with their opinion.

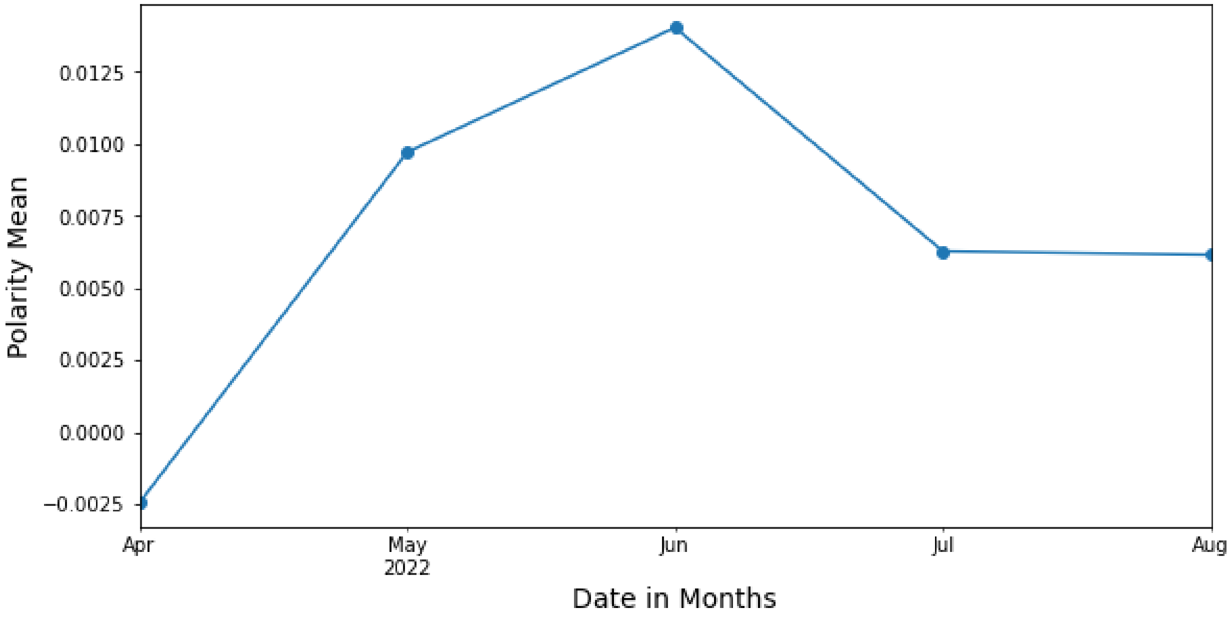


Figure 26: Polarity Mean of tweets among Months

5.9 Map:

Taking the "Location" column into consideration, we tried to plot a web map that distributes the locations according to the polarity in each country. Thus, we plotted it using the *plotly_express* library.

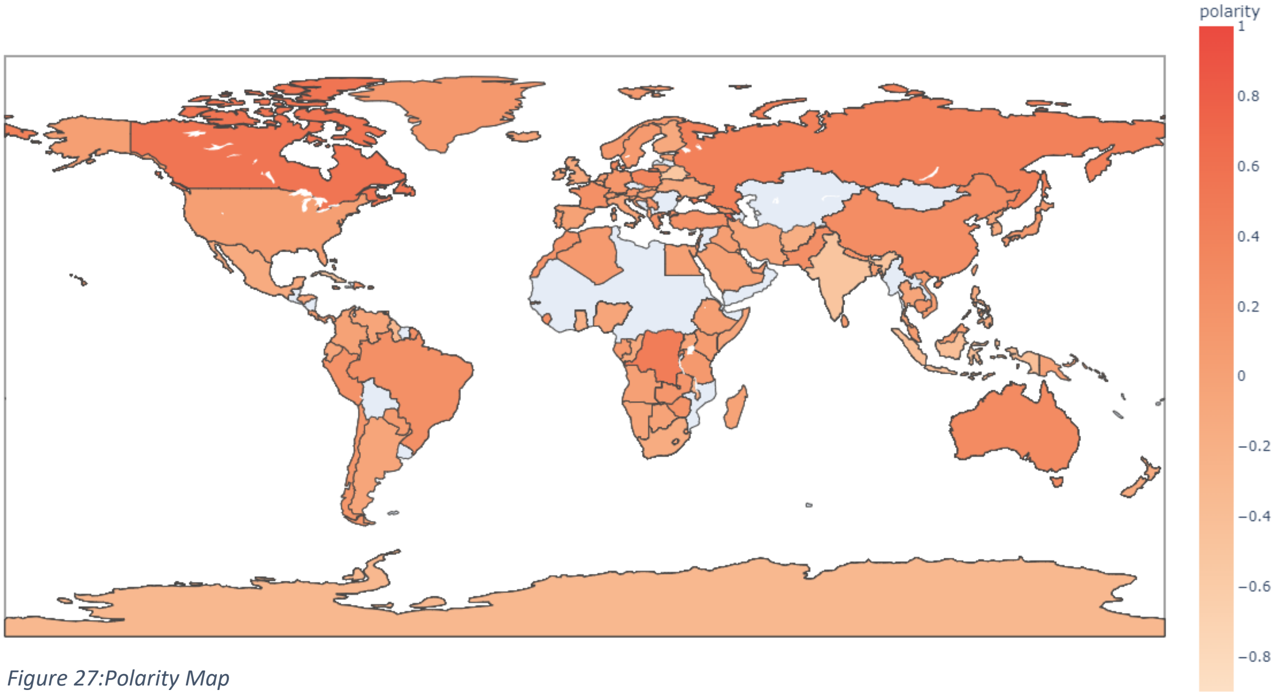


Figure 27:Polarity Map

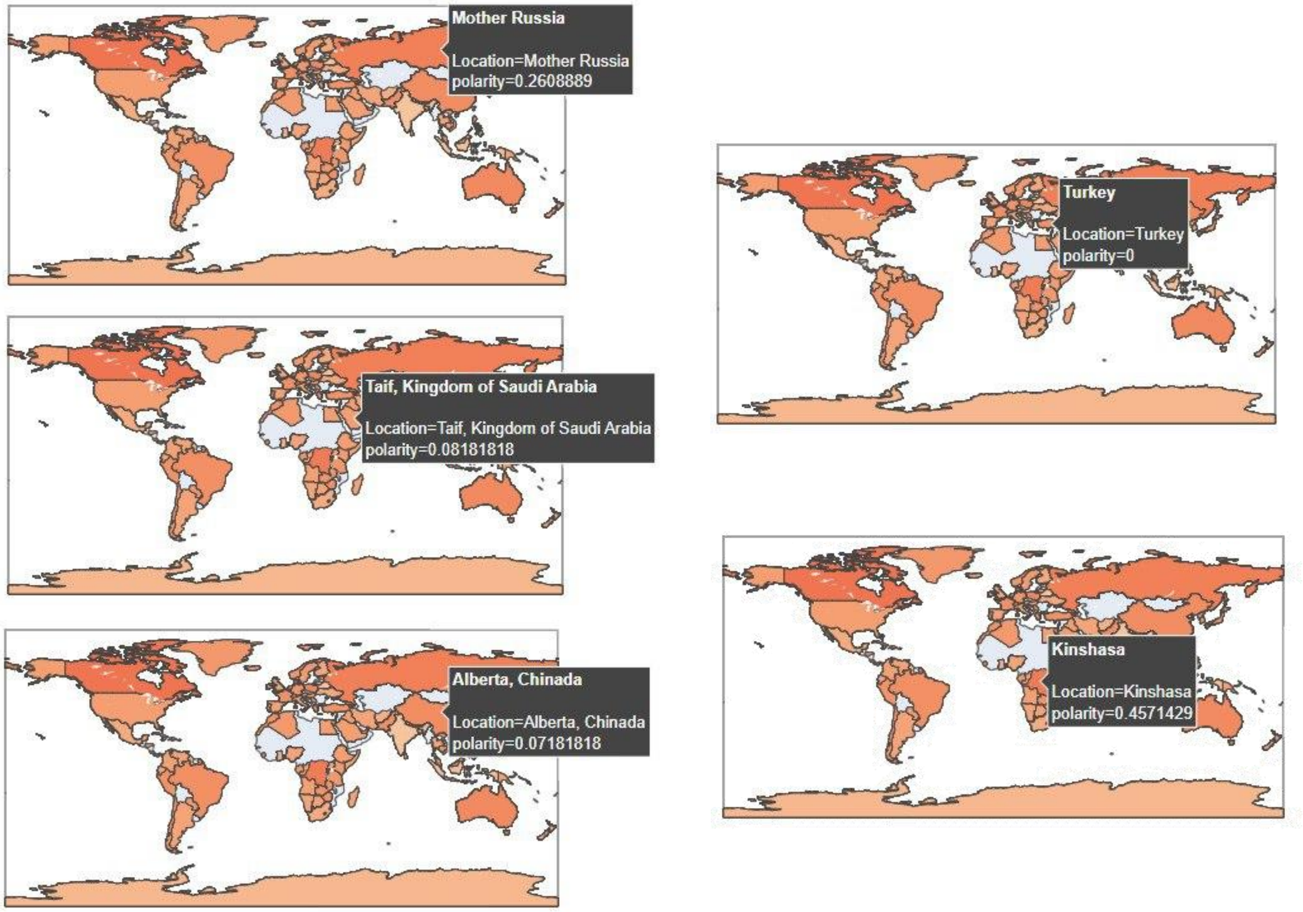


Figure 29: Countries with Positive Polarity

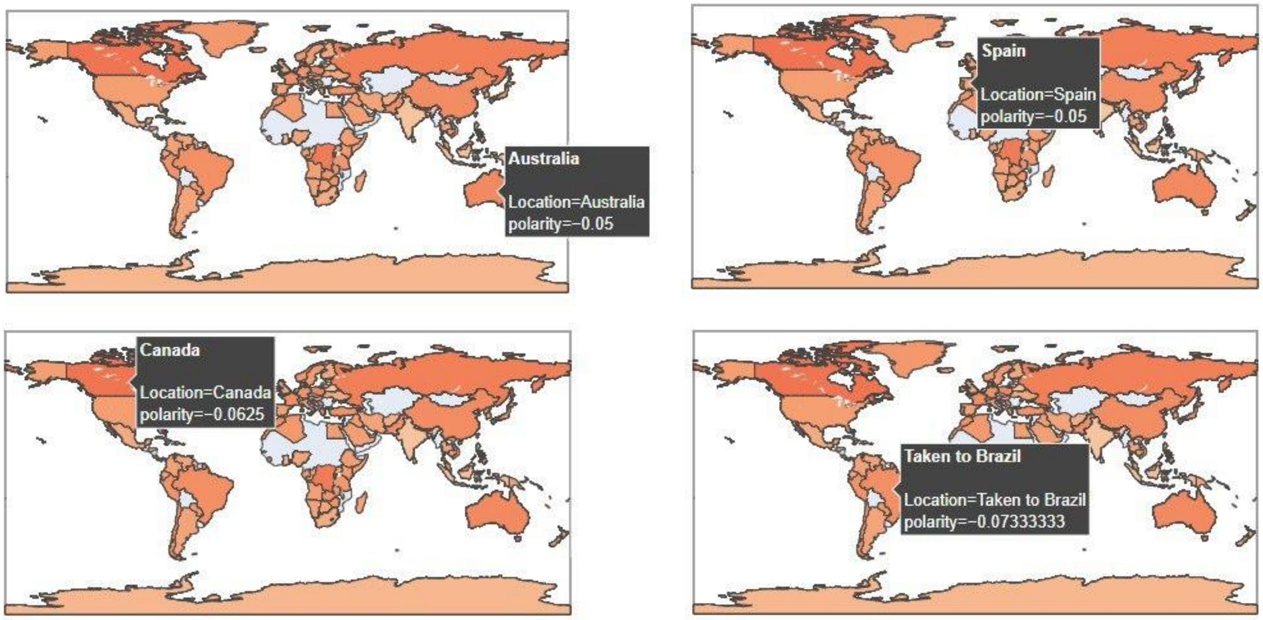


Figure 28: Countries with Negative Polarity

- We've merely highlighted a few of the major countries on our map. As a result, we discovered that most of the countries in Africa (Alberta and Kinshasa) and Asia (Russia, Turkey, and Saudi Arabia) have a positive polarity. In contrast, the majority of the countries in Europe (Spain), America (Canada and Brazil), and Australia have negative polarity.

  This outcome demonstrates how different countries' perspectives on the monkeypox vaccine differ.

Chapter Six: Conclusion and Future Work

6.1 Conclusion:

In this thesis, we examine the sentiments expressed in tweets about the monkeypox vaccine and display the findings using bar graphs, line charts, histograms, and maps at certain times. Sentiment analysis has considered the most significant resource for making decisions. The majority of individuals rely on it to produce an effective product. Microblogging sites are a desirable source of data for sentiment analysis and opinion mining due to the enormous volume of sentiment-rich data they contain.

In this study, we performed sentiment analysis on tweets from April 1 to August 18, without identifying a specific location. These tweets were collected using the snscrape library. We used the neattext library to filter the tweets. Then, we utilized the textblob library to do sentiment analysis. To determine the polarity of the tweets, which were afterward divided into positive, negative, and neutral expressions, Textblob, an open-source text processing package built in Python, was helpful.

We plot and create a web map to represent the outcome. In both cases, we observed how the differences between positive, negative, and neutral attitudes relied on the time and place. We also looked at the evolution of the attitudes over a period of months.

Any decision-making process can be managed with Twitter sentiment analysis. Additionally, visualization facilitates quicker assimilation of knowledge. Twitter sentiment research can assist a firm in understanding the attitudes of people regarding the company, its products, or a project in any location or time they want if they are aiming to capitalize on a particularly hot issue.

Our research leads us to the conclusion that as monkeypox has evolved, people are becoming more and more aware of it as a result of the numerous cases that have been reported. Therefore, public knowledge of monkeypox increases but over 2 in 5 individuals are unlikely to get vaccinated if exposed, as we statistically calculated the results above.

Additionally, the general public is inclined to receive this vaccination. Even if there are still negative tweets, they are very close to zero, therefore they might soon turn positive as the monkeypox virus spreads over the world.

Furthermore, we observed that viewpoints toward vaccination varied between cities. This is maybe due to individual rights, religion, and mistrust.

6.2 Future Work:

To make the most of the time available, we can refine our approaches for future work. This will prevent the analysis from being restricted by a lack of time. We should classify sentiments in more ways than just "positive," "negative," and "neutral" to increase the efficiency of our sentiment analysis. We should employ more categories, such as "angry," "happy," "frustrated," "sad," and "neutral," among others.

Because the medical field is so wide and always evolving, a lot of diseases and viruses could spread over the world and catch people's attention. We, therefore, hope that students in the rising generation would explore our study and learn how to use Twitter Sentiment Analysis to determine public opinion globally on any topic.

REFERENCES

1- Data Analysis of Covid19 Tweets (Sentiment Analysis & KE), Jesse E. Ague, 25 May 2021

2- Natural Language Processing for Beginners: Using TextBlob, Shubham Jain , 23 December 2020

3- Monkeypox Key facts, WHO, 19 May 2022

4- Public sentiments toward COVID-19 vaccines in South African cities: An analysis of Twitter posts, Public Health, 12 August 2022

5- Vaccines and immunization for monkeypox, WHO, 14 June 2022

6- Sentiment Analysis in Geo-Social Streams by using Machine Learning Techniques, Roberto Henriques, February 2017

7- Assessing the Impact of the Economic and COVID-19 Crises in Lebanon, WFP, 30 June 2020

8- Monkeypox Vaccine: U.s. Orders 500,000 Jynneos Doses as Cases Rise, Bahl R. , 12 June 2022

9- An augmented multilingual Twitter dataset for studying the COVID-19 infodemic, Christian E. Lopez, 20 October 2021

10- Text Mining and Sentiment Analysis: Analysis with R, Sanil Mhatre, 13 May 2020

11- How to analyze Twitter data, Alex York**,** 10 December 2020

12- Twitter Sentiment Analysis using NLTK, Python, Mohamed Afham, 25 Sep 2019