Sentiment Analysis of COVID-19 Vaccination

Using the TextBlob API and word cloud visualizations

King Abdullah II School of Information Technology, University of Jordan

Ribal Ali Attoun
Computer Information Systems
University of Jordan
Amman, Jordan
ribalattoun@gmail.com

Anas Tamem Al-momani
Computer Science
University of Jordan
Amman, Jordan
anasmomani253@gmail.com

Majdi Sawalha (Supervisor)
Computer Information Systems
University of Jordan
Amman, Jordan
sawalha.majdi@ju.edu.jo

Reem Al-fayes (Supervisor)
Computer Information Systems
University of Jordan
Amman, Jordan
r.alfayez@ju.edu.jo

Heba Saadeh (Supervisor)
Computer Science
University of Jordan
Amman, Jordan
heba.saadeh@ju.edu.jo

Bilal Abu Salih (Supervisor)
Computer Science
University of Jordan
Amman, Jordan
b.abusalih@ju.edu.jo

Abstract**—** With a substantial proportion of the population currently hesitant to take the COVID-19 vaccine, it is important that people have access to accurate information. However, the bulk of the population has optimistic feelings about these vaccinations, there are also negative feelings about them, according to the analysis. In this paper, we present our dataset, a growing collection of English-language Twitter posts about COVID-19 vaccines. We provide statistics regarding the tweets over time, the hashtags used. We also illustrate how these data might be utilized by performing an analysis of the prevalence over time, topic groups of hashtags, geographical and chronologically distributions. Additionally, we develop and present a dashboard, allowing people to visualize the sentiment analysis of the COVID-19 vaccine for all Countries geolocated posts in our dataset.

Keywords**—***Data Science, Sentiment Analysis, COVID-19, Corona, Vaccine, COVID-19 Vaccination.*

I. Introduction

With the global continuity of the COVID-19 pandemic, widespread vaccination with the COVID-19 vaccine is critical for achieving herd immunity and preventing the virus from spreading further. Significant concerns around vaccination willingness remain as policymakers seek to authorize and administer healthy and reliable vaccinations: What are the public's views and expectations about these vaccines, and how do they impact vaccine uptake? These issues are crucial for developing an education and outreach strategy that will result in optimal vaccine penetration and herd immunity.

Twitter, a microblogging platform with over 187 million daily monetizable active users, is a useful tool for gaining a better understanding of public opinion about the COVID-19 vaccine on a large scale. #COVID19 was the most popular hashtag on Twitter in 2020, with over 400 million mentions. The rapid distribution of information (whether accurate or not) and dispersion of emotion across geographic and social borders have become a hallmark of social media. Real-time shifts and adaptation of population-level behaviors can be inferred from social media text analysis. Twitter has been an especially important data tool in public health and healthcare-related studies and has been used to study public opinion and recognize patterns during the COVID-19 pandemic. [1]

We aimed to do sentiment analysis on COVID-19 vaccine-related tweets, as well as analysis of the accountable, originating user accounts, to gain insight into the evolution of public attitudes toward the COVID-19 vaccines over time in each country.

II. Literature review

A. *Sentiment analysis of COVID-19 vaccine in Indonesia using Naïve Bayes Algorithm*

This research assesses the opinion of the Indonesian people through a social network analysis of the COVID-19 vaccine in January 2021. They used sentiment analysis by crawling Twitter data with 'Vaccine Covid-19' as keywords. This method is achieved by writing up Twitter opinions about an event or issue so that positive or negative opinions can be concluded using the Naive Bayes Algorithm.[2]

B. *Public Sentiment Analysis of COVID19 Vaccination Drive in India*

This research work has been conducted to analyze the sentiments in the tweets posted in India regarding these two vaccines. The analysis shows that while a majority of the population is posting with positive sentiments towards these vaccines, there are also negative sentiments associated with them, associated with the emotions such as fear and anger.[3]

In summary, the work presented in this paper aimed to do sentiment analysis on COVID-19 vaccine-related tweets, as well as analysis of the accountable, originating user accounts, to gain insight into the evolution of public attitudes toward the COVID-19 vaccines over time in each country around the world. While related works focused on sentiment analysis on COVID-19 vaccine-related tweets in a particular country, and they use different algorithms to do sentiment analysis. Further, we are able to study at a much larger scale than previously possible.

III. Development environment

Jupyter notebooks strike a balance between simple text editors, which are fast to start and simple and easy to manipulate, and IDE's which tend to start slower and be feature-rich and complex. Simple text editors typically can only edit code, and cannot run the code. A full IDE can edit code, run the code, debug code, provide syntax highlighting and context help. In the context of problem-solving, Jupyter notebooks are quite handy. Jupyter notebooks open quickly and quickly produce output. Data exploration, data cleaning, and plot building are accomplished in Jupyter notebooks easier and quicker than in a text editor or an IDE.[4]

IV. METHODOLOGY

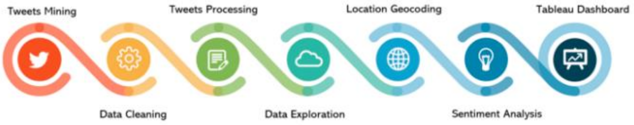


Figure 1: Twitter Sentiment Analysis Project Flowchart [5]

A. *Tweets Mining*

We had to build our dataset from scratch, unlike many other projects where we used pre-existing datasets. Using tweepy, a Python library for accessing the Twitter API, we can collect tweets. We used the Twitter API to mine our dataset in two ways:

- Search API – allows us to collect old tweets.
- Streaming API – allows us to collect tweets in real-time based on keywords, User ID’s and location.
- Online datasets for old Tweets

We had scraped **276,724 unique tweets** from 9th August 2020 to 27th April 2021, Access to historical tweets is extremely limited. We can retrieve the last 3,200 tweets from a user timeline and search the last 7-9 days of tweets. Therefore, we had to search for online dataset. [6][7]

Search Query: we passed 17 different phrases – ['covid vaccine', 'covid', 'corona', 'covid19', 'WearAMask', 'StayHomeStaySafe', 'SocialDistancing’, 'Comirnaty', 'Moderna', 'AstraZeneca', 'SputnikV', 'Janssen', 'CoronaVac', 'BBIBP_CorV', 'EpiVacCorona', 'Convidicea', 'Covaxin'] to the API so that it returns tweets containing them. Twitter requires a specific syntax to recognize that you want an “exact phrase” to match. Also, we only mined tweets created in **English** for this analysis.

Information Returned: We specified that the API returns the following data for each tweet – [text, time, place, user_location, user_description, language].

B. *Data Preparation*

We flattened the tweet JSON into a single level to analyze tweets at scale. This allows us to store the tweets in a DataFrame format.

Json before flattening:

```
{
  "created_at": "Thu Apr 06 15:24:15 +0000      2017",
  "id_str": "850006245121695744",
  "text": "1\/ Today we\u2019re sharing our vision for
https:\/\/t.co\/XweGngmxlP",
  "user": {
    "id": 2244994945,
    "name": "Twitter Dev",
    "screen_name": "TwitterDev",
    "location": "Internet",
    "url": "https:\/\/dev.twitter.com\/",
    "description": "Your official source for Twitter
Platform newshttps:\/\/twittercommunity.com\/
\u2328\ufe0f #TapIntoTwitter"
  },
  "place": {  ...  } ...
}
```

Json after flattening:

```
[
  {
           "Text":"RT @MinhazMerchant: The  glee
           with which old rotted cosystem journos have
           greeted China\u2019s 18.3% GDP growth in
           Jan-March 2021 (without point\u2026",
           "Location":"india",
           "Time":"Sat Apr 17 20:06:43 +0000 2021”
  },

  {        "Text":"RT @devilsxblessing: Covid hospital
           beds in pune\nhttps:\/ \/t.co\ /a2Exuednf1",
           "Location":"india",
           "Time":"Sat Apr 17 20:06:43 +0000 2021"
  },
           …
]
```

C. *Data Cleaning*

Cleaning up our data, we looked for duplicate tweets, checked for empty rows and try to identify the Tweet's source by replace “NaN” or Null values for the “Location” column with a location in the child JSON "User" Tweets['user']['location'], and if was null, we look in another child JSON "place," Tweets['place']['country']. Finally, we look at the "Bio" section of the user's profile and extract the country name if one is present Tweets['user']['description’], by searching in a dataset for Countries and Cities to see if there are any terms that match the user's description and the Countries dataset.

D. *Tweets Processing*

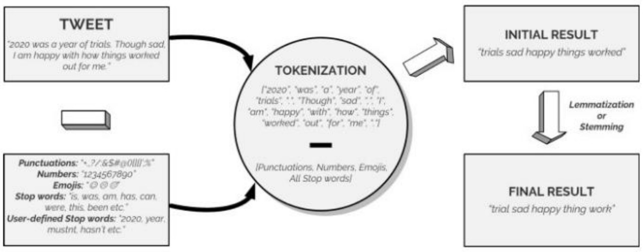
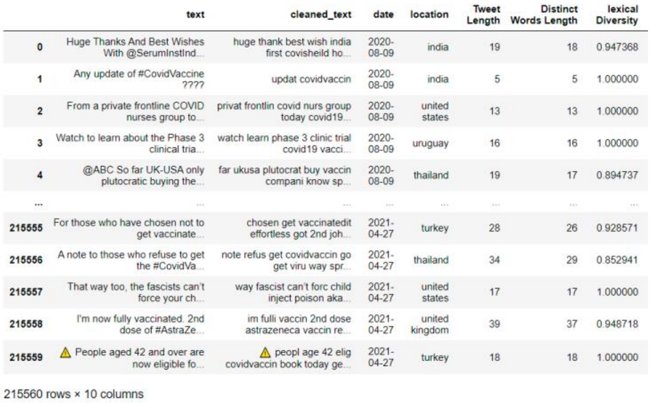

TWEET
"2020 was a year of trials. Though sad, I am happy with how things worked out for me."
TOKENIZATION
["2020", "was", "a", "year", "of", "trials", ".", "Though", "sad", ",", "I", "am", "happy", "with", "how", "things", "worked", "out", "for", "me", "."]
[Punctuations, Numbers, Emojis, All Stop words]
INITIAL RESULT
"trials sad happy things worked"
Lemmatization or Stemming
FINAL RESULT
"trial sad happy thing work"
Punctuations: "-.?/:&$#@()[]|'%"
Numbers: "1234567890"
Emojis:
Stop words: "is, was, am, has, can, were, this, been etc."
User-defined Stop words: "2020, year, mustnt, hasn't etc."

Figure 2: Tweet Processing using Tokenization in NLP [8]

To achieve the ultimate goal, Sentiment Analysis, there was a need to clean up the individual tweets. We convert the text in lower case for further work on tweet text, removed the stop words (of, a, in, etc.) from the statement because these words are not useful to support the labels of sentiments data, remove the punctuations because these are the noise in the data and not meaningful, remove repeating characters in the words, remove emails, remove URL’s from the tweets in a single run. Additionally, we used a concept known as “Tokenization” in NLP. It is a method of splitting a sentence into smaller units called “tokens” to remove unnecessary elements. Another noteworthy technique is “Lemmatization”. This is a process of returning words to their “base” form.

E. *Data Exploration*

We Created a pandas data frame that contains the attributes of Tweets' JSON file after Flattening and "Cleaned Text" which contains the tweet text after applying the previous step "Data Cleaning". We gained more insights about what we're working with by adding "Tweet Length", "Distinct Words Length" and "lexical Diversity".



	text	cleaned_text	date	location	Tweet Length	Distinct Words Length	lexical Diversity
0	Huge Thanks And Best Wishes With @SerumInstInd...	huge thank best wish india first covisheild ho...	2020-08-09	india	19	18	0.947368
1	Any update of #CovidVaccine ????	updat covidvaccin	2020-08-09	india	5	5	1.000000
2	From a private frontline COVID nurses group to...	privat frontlin covid nurs group today covid19...	2020-08-09	united states	13	13	1.000000
3	Watch to learn about the Phase 3 clinical tria...	watch learn phase 3 clinic trial covid19 vacci...	2020-08-09	uruguay	16	16	1.000000
4	@ABC So far UK-USA only plutocratic buying the...	far ukusa plutocrat buy vaccin compani know sp...	2020-08-09	thailand	19	17	0.894737
...	...	...	...	...	...	...	...
215555	For those who have chosen not to get vaccinate...	chosen get vaccinatedit effortless got 2nd joh...	2021-04-27	turkey	28	26	0.928571
215556	A note to those who refuse to get the #CovidVa...	note refus get covidvaccin go get viru way spr...	2021-04-27	thailand	34	29	0.852941
215557	That way too, the fascists can't force your ch...	way fascist can't forc child inject poison aka...	2021-04-27	united states	17	17	1.000000
215558	I'm now fully vaccinated. 2nd dose of #AstraZe...	im fulli vaccin 2nd dose astrazeneca vaccin re...	2021-04-27	united kingdom	39	37	0.948718
215559	⚠ People aged 42 and over are now eligible fo...	⚠ peopl age 42 elig covidvaccin book today ge...	2021-04-27	turkey	18	18	1.000000

215560 rows × 10 columns

Figure 3: Pandas DataFrame Plot with new Features

Plot the 30 most frequent words in our data set.

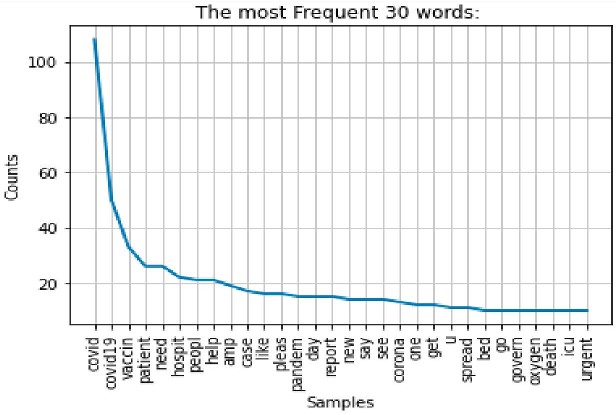


Figure 4: 30 Most Frequent Words

Using the WordCloud library, we generated a Word Cloud based on the frequency of words and superimpose these words on this image:

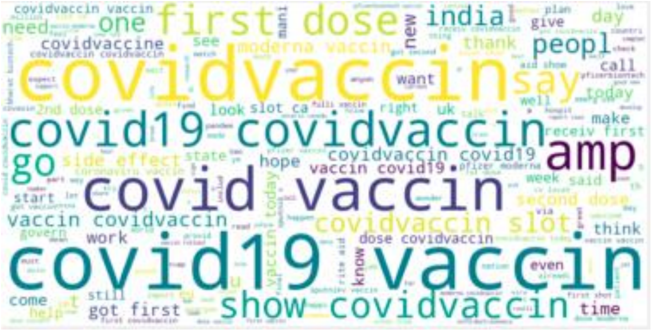


Figure 5: Sentiment Analysis of COVID-19 Vaccination WordCloud

F. *Location Geocoding*

For our final dashboard, we wanted to add a map that shows the average polarity per country. To do this, we needs basic geographic information such as the country's name, Longitude and Latitude for each country. So, we got a dataset that contains ["CountryName", "CapitalName", "CapitalLatitude", "CapitalLongitude", "CountryCode", "ContinentName"] for all countries around the world.

G. *Sentiment Analysis*

From the words we use in our statements, one can tell whether they are Positive, Negative or Neutral.

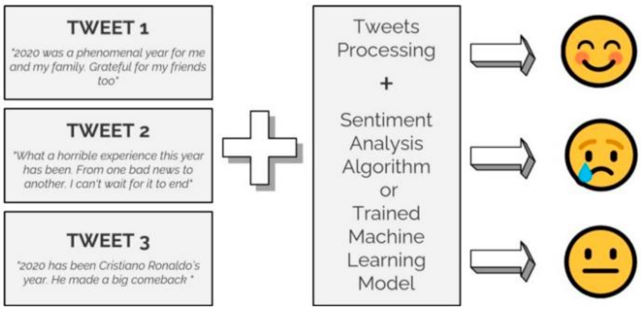

TWEET 1
"2020 was a phenomenal year for me and my family. Grateful for my friends too"
TWEET 2
"What a horrible experience this year has been. From one bad news to another. I can't wait for it to end"
TWEET 3
"2020 has been Cristiano Ronaldo's year. He made a big comeback "
Tweets Processing
+
Sentiment Analysis Algorithm or Trained Machine Learning Model

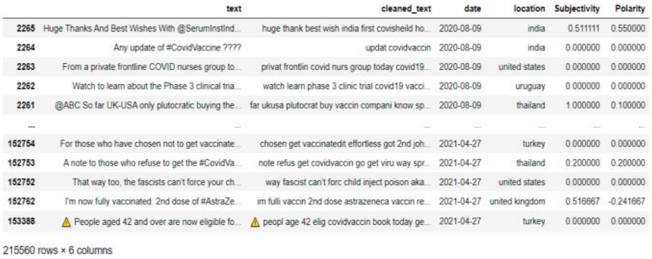
Figure 6: Simple Sentiment Analysis Illustration [9]

Through use libraries such as TextBlob and VADER, we can analyze text and return their Sentiment score. This is a part of Unsupervised Machine Learning (UML).

Since the context for the trained model differs from the context for these tweets, we must emphasize that these algorithms have error margins. For this analysis, we went with TextBlob.

TextBlob analyzes sentences by giving each sentence a Subjectivity and Polarity score. Based on the Polarity scores, one can define the tweets' sentiment category. A Polarity score of < 0 is Negative, 0 is Neutral, while > 0 is Positive.

We used the Pandas "apply" method on the "Polarity" column in our dataframe to return the respective Sentiment score for each tweet:



	text	cleaned_text	date	location	Subjectivity	Polarity
2265	Huge Thanks And Best Wishes With @SerumInstInd...	huge thank best wish india first covisheild ho...	2020-08-09	india	0.511111	0.550000
2264	Any update of #CovidVaccine ????	updat covidvaccin	2020-08-09	india	0.000000	0.000000
2263	From a private frontline COVID nurses group to...	privat frontlin covid nurs group today covid19...	2020-08-09	united states	0.000000	0.000000
2262	Watch to learn about the Phase 3 clinical tria...	watch learn phase 3 clinic trial covid19 vacci...	2020-08-09	uruguay	0.000000	0.000000
2261	@ABC So far UK-USA only plutocratic buying the...	far ukusa plutocrat buy vaccin compani know sp...	2020-08-09	thailand	1.000000	0.100000
...	...	...	...	...	...	...
152754	For those who have chosen not to get vaccinate...	chosen get vaccinatedit effortless got 2nd joh...	2021-04-27	turkey	0.000000	0.000000
152753	A note to those who refuse to get the #CovidVa...	note refus get covidvaccin go get viru way spr...	2021-04-27	thailand	0.200000	0.200000
152752	That way too, the fascists can't force your ch...	way fascist can't forc child inject poison aka...	2021-04-27	united states	0.000000	0.000000
152762	I'm now fully vaccinated. 2nd dose of #AstraZe...	im fulli vaccin 2nd dose astrazeneca vaccin re...	2021-04-27	united kingdom	0.516667	-0.241667
153388	⚠ People aged 42 and over are now eligible fo...	⚠ peopl age 42 elig covidvaccin book today ge...	2021-04-27	turkey	0.000000	0.000000

215560 rows × 6 columns

Figure 7: Pandas DataFrame Plot with Subjectivity and Polarity Features

Because of, The combination of the negative emoji, the positive emoji and multi-mixed sentiment emoji affects negatively of the sentiment of the tweet as a whole, we didn't do sentiment analysis for Emojis.

V. RESULTS

A heat map that highlights the Sentiment category and score for all countries to visualize our results. We generated a JSON file for each Country that contained all tweets that we had collected for this specific Country. Then we got the AVG_Polarity() for each file and created a new JSON file that contained "Country name", "Capital Name", "Average Polarity", "Latitude" and "Longitude" for all Countries.

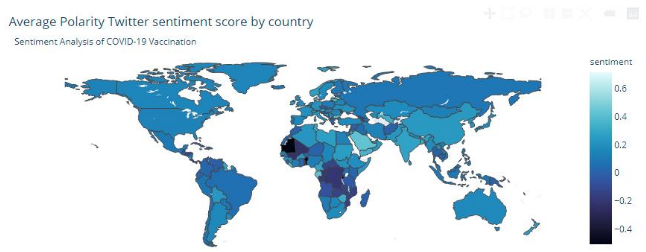

Average Polarity Twitter sentiment score by country
Sentiment Analysis of COVID-19 Vaccination
sentiment
0.6
0.4
0.2
0
−0.2
−0.4

Figure 8: Average Polarity Twitter sentiment score by country

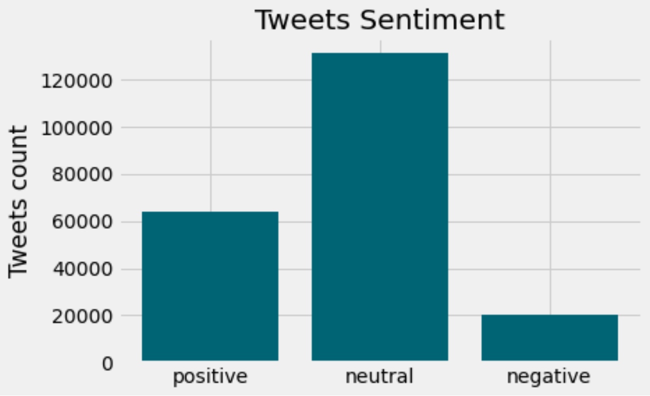


Figure 9: Tweet breakdown by sentiment Category Bar Chart

Among all of the tweets in the dataset, about 60.8% of them were designated by TextBlob as neutral in sentiment (polarity = 0.0), and the rest of the data consisting of 29.7% positively charged tweets (polarity > 0.0) and 9.5% negatively charged tweets (polarity < 0.0).

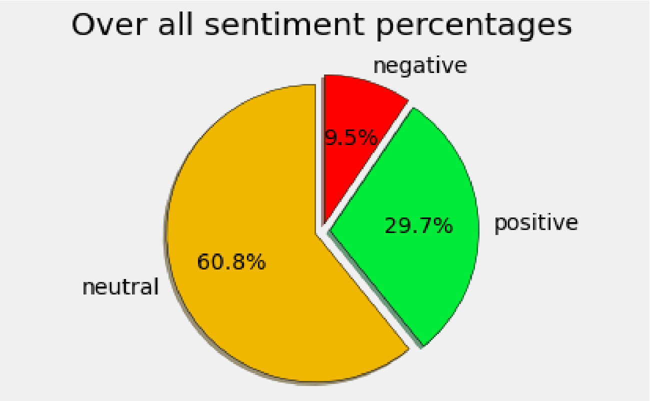


Figure 10: Tweet breakdown by sentiment Category Pie Chart

VI. DISCUSSION

In this paper, we introduce a new public dataset that tracks Twitter conversation about COVID-19 vaccines. We use influential keywords, geographic distribution of tweets, and clusters of associated hashtags to classify the results. We also have a data dashboard that displays insights and observations derived from the data.

In the future, we intend to investigate the connection between online discussion of COVID-19 vaccines and public health outcomes such as COVID-19 mortality and vaccine uptake . We want to see how myths about vaccines and anti-vaccine sentiment spreads on social media. Finally, We would attempt to structure our code in such a way that we will be able to use it to perform sentiment analysis on other possible patterns and visualize our findings geographically and chronologically through a website.

VII. CONCLUSION

This research using the TextBlob sentiment classification from Twitter data with the keyword 'COVID-19' filtered by the keyword 'vaccine'. The tweets collected for this study ranged from 9th August 2020 to 27th April 2021. Analysis in that period showed 23.67% positive sentiment, 9.49% negative sentiment, and 60.83% neutral sentiment.

ACKNOWLEDGMENT

Our superiors, Dr Majdi Sawalha, Dr Reem Al-Fayez, Dr Heba Saadeh, and Dr Belal Abu-Saleh, were instrumental in making this paper and the study behind it possible. Their enthusiasm, knowledge, and meticulous attention to detail served as an inspiration and kept us on track to complete this paper's final draft. The generosity and expertise of each and every one of you has changed this report in many respects and rescued us from many mistakes; those that eventually remain are solely our own responsibility.

REFERENCES

[1] McGraw, T. Spending 2020 Together on Twitter. https://blog.twitter.com/en_us/topics/insights/2020/spending-2020-together-on-twitter.html.

[2] Pristiyono, Mulkan Ritonga, Muhammad Ali Al Ihsan, Agus Anjar and Fauziah Hanum Rambe "Sentiment analysis of COVID-19 vaccine in Indonesia using Naïve Bayes Algorithm", 2021.

[3] Akash D Dubey "Public Sentiment Analysis of COVID19 Vaccination Drive in India", 2021.

[4] "Why Jupyter Notebooks?", Why Jupyter Notebooks? - Problem Solving with Python.

[5] "Twitter Sentiment Analysis Project", Image developed by Jessica Uwoghiren via Slide Model.

[6] All COVID-19 Vaccines Tweets, https://www.kaggle.com/gpreda/all-covid19-vaccines-tweets.

[7] Covid Vaccine Tweets, https://www.kaggle.com/kaushiksuresh147/covidvaccine-tweets.

[8] "Tweet Processing using Tokenization in NLP", Image developed Jessica Uwoghiren using Microsoft Visio.

[9] "Simple Sentiment Analysis Illustration", Image developed by Jessica Uwoghiren using Microsoft Visio.

[10] Corey Schafer, https://www.youtube.com/channel/UCCezIgC97PvUuR4_gbFUs5g

APPENDICES

To reach our code:

https://github.com/RibalAttoun

https://github.com/anas1509

Link for our heat map:

https://plotly.com/~RibalA.Attoun/5/