Omicron Sentimental Analysis
Twitter Data Analysis

MIS64061 Advanced Machine Learning

Noorah Alkhaniny

Abstract

Omicron is a raising virus that has started in the end of 2021. People all over the world have been worried about it specially after it speeding and so many cases are diagnosed. The objective of this project is to show people opinions regarding Omicron on Twitter and what are the highest countries tweeting about it.

Contents

Introduction

Omicron is a raising virus in the ending of 2021 and it is spreading all over the world and it effected many life aspects. On November 24, 2021, a new it was reported to the World Health and on December 1, 2021 the first confirmed U.S. case of Omicron was identified. People and news platform has been talking about is all over the social media such as Twitter and, Facebook and etc. In this project, I will focus on conducting Neural Network Sentimental Analysis on people opinions as positive, negative and neutral. Finding the geographical description of tweets and what are the top countries tweeting about Omicron.

Data Source

This project is built on a collected tweets from Twitter, using Twitter API tool to gather more than 30,000 tweets that contain the word "Omicron". The dataset is available at Kaggle and updated on a daily bases and it has different data such as usernames, tweet date, user location and etc.

Data Cleaning

The data has so many additional information that is not needed in this project such as hashtags, links, mentions and links. To clean the data I have used Text_clean tool and:

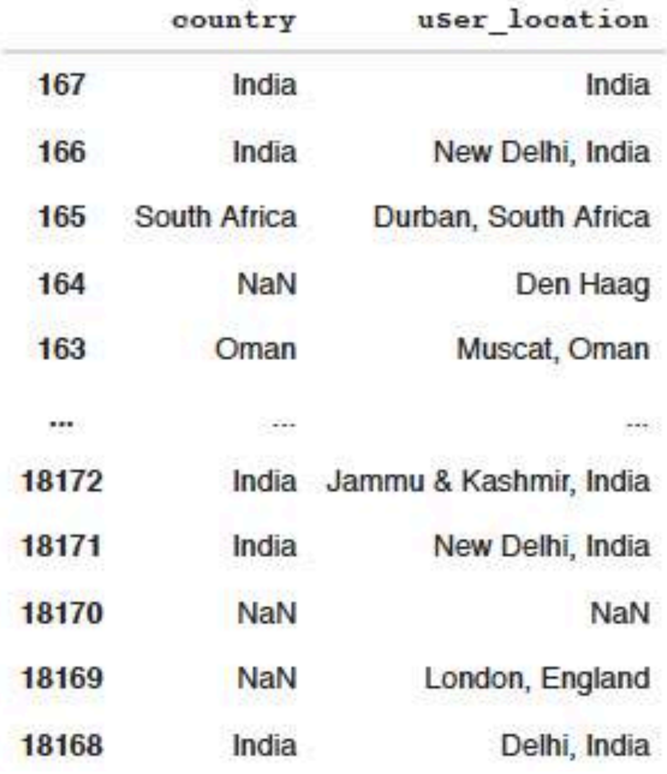
- *Strip_emoji* → to remove all emojies in the dataset
- *Strip_all_entities* → to remove all "@" that will delete mentions
- *Clean_hashtags* → to remove all hashtags in the dataset

Models

For creating the models, I will divide this section into two parts first is the geographical analysis and then the sentimental analysis.

Geographical Analysis:

For finding the geographical description, first I used count *User_location* but unfortunately it was not giving the full information since there are so many missing users locations. To solve the problem, I have used the *Pycountry* tool to provide the geographical information and then we can see what are the most countries tweeting about Omicron.



	country	user_location
167	India	India
166	India	New Delhi, India
165	South Africa	Durban, South Africa
164	NaN	Den Haag
163	Oman	Muscat, Oman
...	...	...
18172	India	Jammu & Kashmir, India
18171	India	New Delhi, India
18170	NaN	NaN
18169	NaN	London, England
18168	India	Delhi, India

As shown above, India is the highest country tweeting about the virus. By using *Pycountry* we can see it help to identify NaN for the city name since some users only show city not countries.

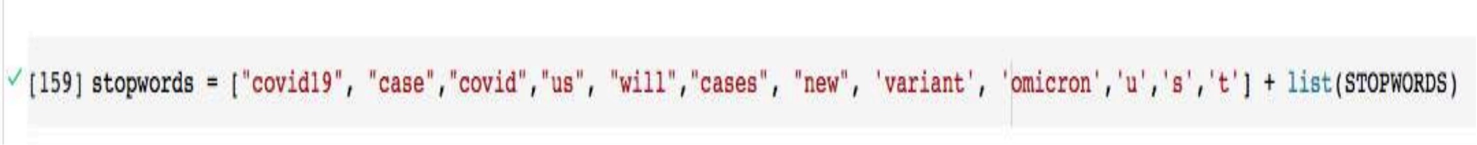
Sentential Analysis:

For creating the sentimental analysis, I will be using tools that help to show the positive, negative and neutral tweets. Here are all the tools I will be using:

- *Stopwords* → to exclude some words in the sentimental analysis like Covid-19
- *NLTK* → tool to use for tweets sentimental analysis
- *TextBLOB* → tool to tweets use for sentimental analysis
- *Wordcloud* → To visualize the what are the most used positive/negative words

Stopwords:

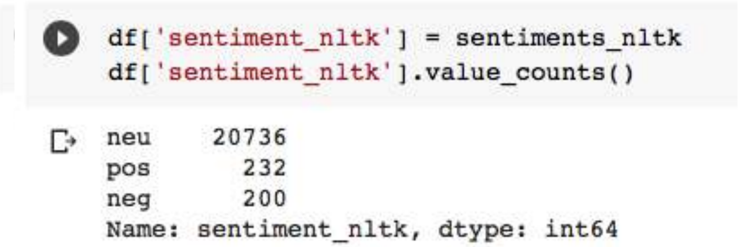
This toll will help me to show analysis without the words that are going to affect my dataset because they are so common in Omicron subject. For example, the Omicrom is excluded because it going to affect my analysis since it is by default is the most common used word. Here is the code I have used and all the words excluded:



```
[159] stopwords = ["covid19", "case","covid","us", "will","cases", "new", 'variant', 'omicron','u','s','t'] + list(STOPWORDS)
```

NLTK:

I have download the library for *NLTK* to show positive, negative and neutral tweets and here is what I got:



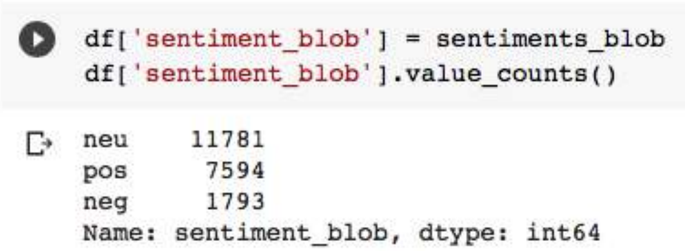
```
df['sentiment_nltk'] = sentiments_nltk
df['sentiment_nltk'].value_counts()

neu    20736
pos      232
neg      200
Name: sentiment_nltk, dtype: int64
```

As we can see from using NLTK tool, we have 232 positive tweets and 200 negative tweets and 20736 neutral tweets.

TextBLOB:

have download the library for *TextBLOB* to show positive, negative and neutral tweets and here is what I got:



```
df['sentiment_blob'] = sentiments_blob
df['sentiment_blob'].value_counts()
```

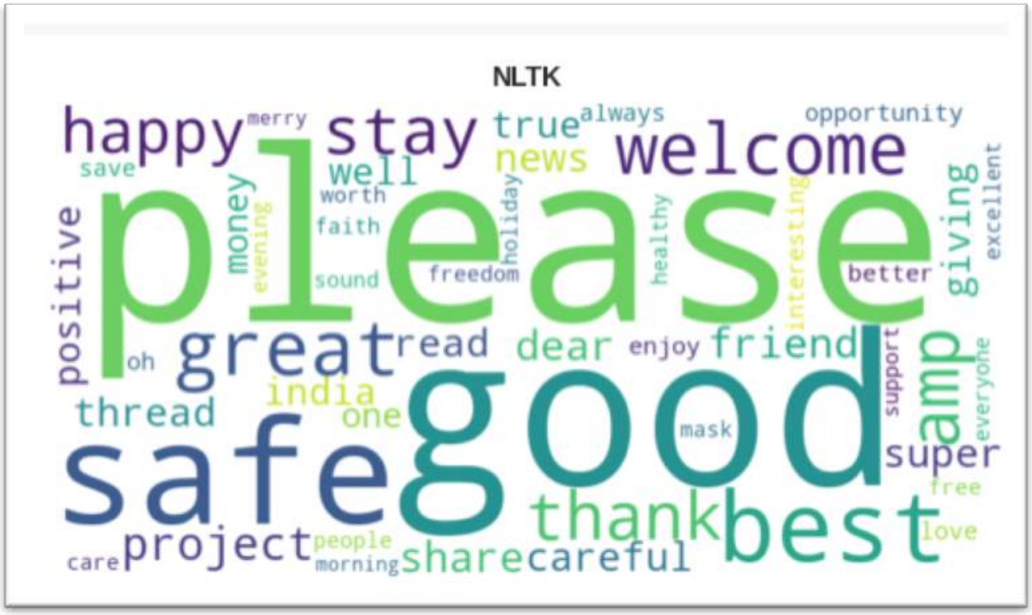
```
neu    11781
pos     7594
neg     1793
Name: sentiment_blob, dtype: int64
```

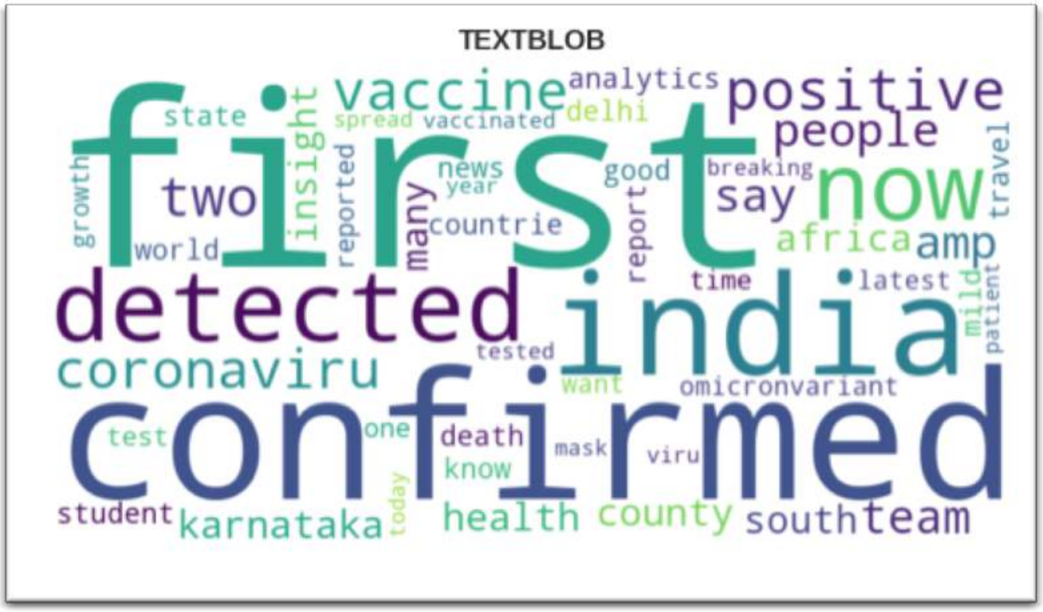
As we can see from using *TextBLOB* tool, we have 7594 positive tweets and 1793 negative tweets and 11781 neutral tweets.

Wordcloud:

Wordcloud tool allow me to visualize the sentimental analysis I got form the *NLTK* and *TectBLOB* tools. This tool will show the most common positive and negative words. Here is the output:

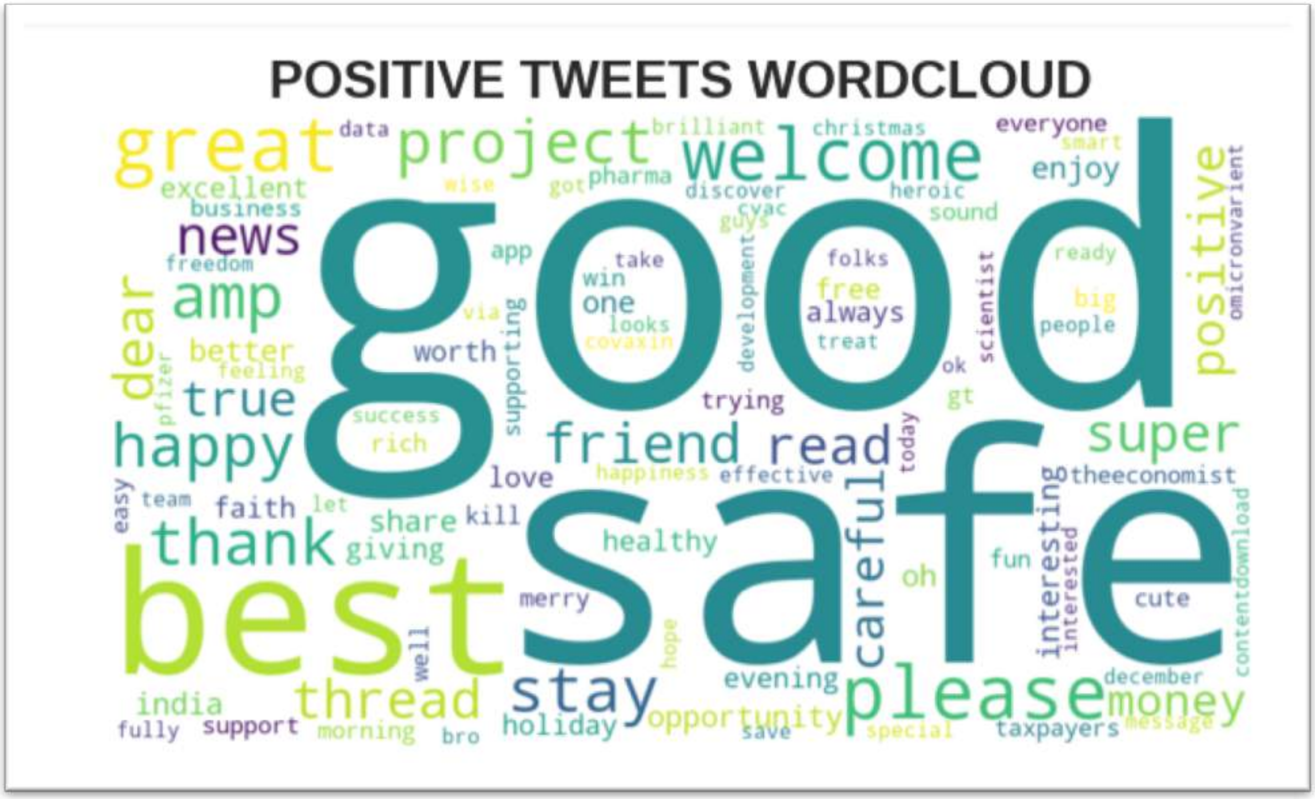
Positive:



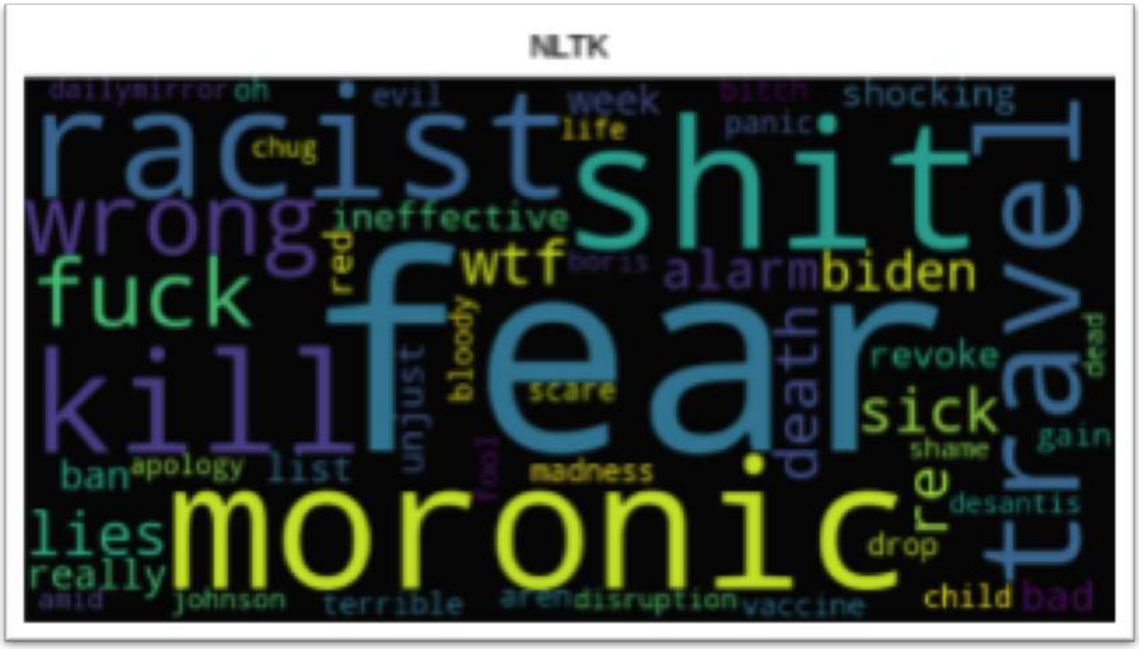


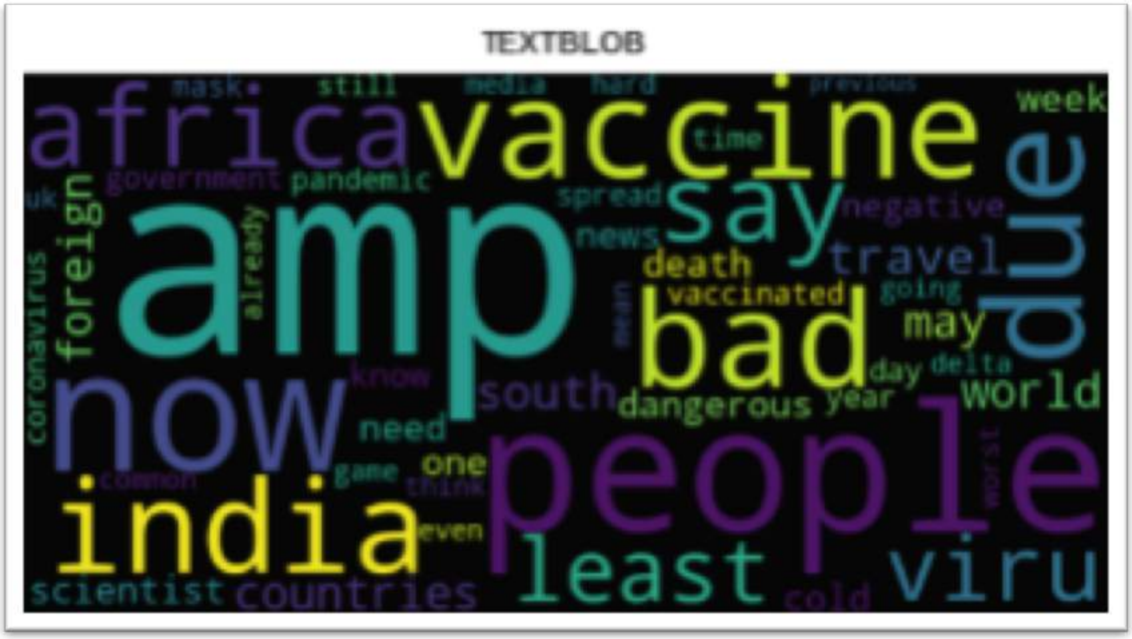
As we can see, *NLTK* is showing more emotional positive words than *TextBLOB* which is really interesting!

Wordcloud for all the *NLTK* and *TextBLOB positive* words together:



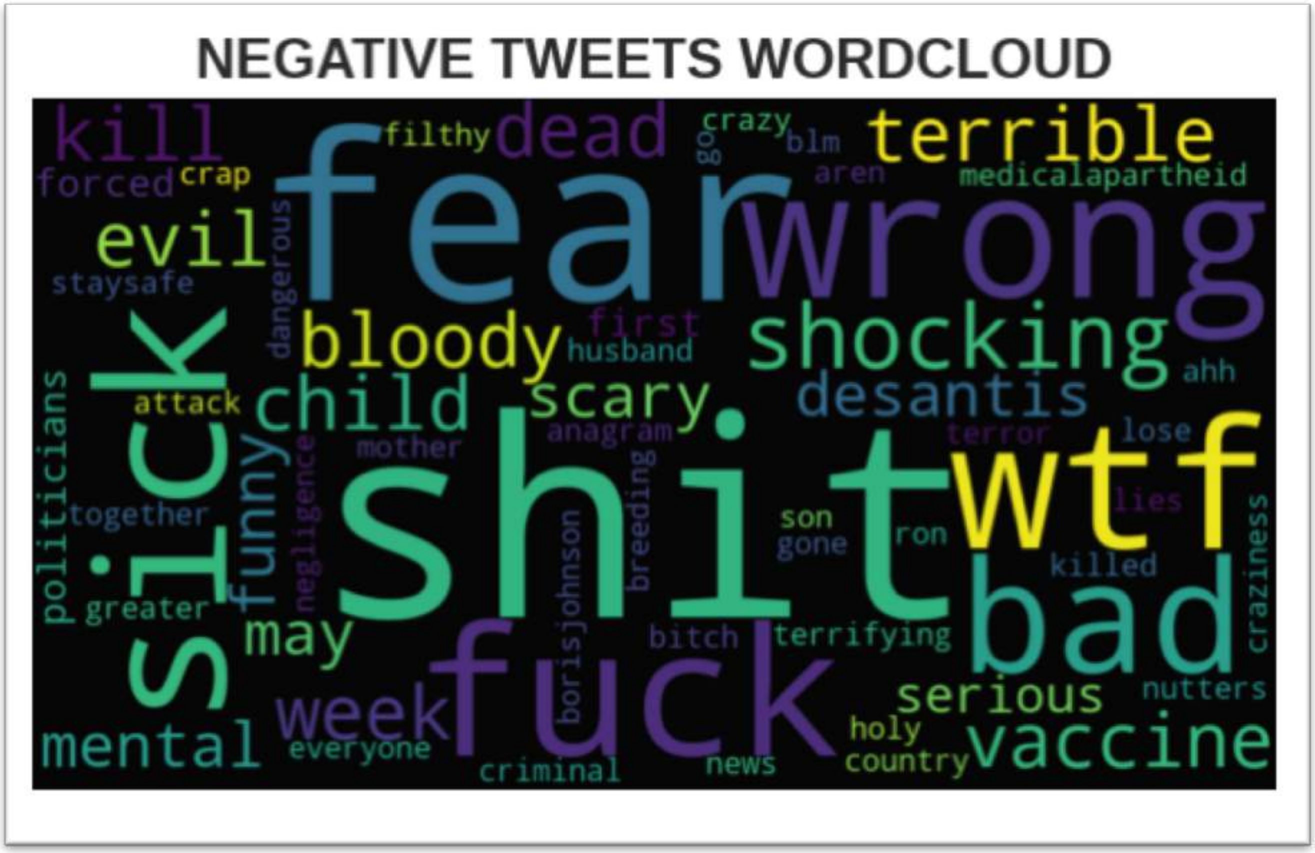
Negative:





As we can see, *NLTK* is showing more emotional negative words than *TextBLOB* which is really interesting!

Wordcloud for all the *NLTK* and *TextBLOB positive* words together:



Conclusion:

In conclusion, this project was developed on making analysis for a rising virus and it is effecting so many people's lives; I do hope these hard times will pass with no harm and sadness. The project is still on the making since the data is updating on a daily basis and I might have different outputs and analysis.

References:

1- Gpreda. (2021, December 1). *Omicron is rising*. Kaggle. Retrieved December 13, 2021, from https://www.kaggle.com/gpreda/omicron-is-rising/data#:~:text=calendar_view_week-,omicron,-.csv.

2- *Pycountry*. PyPI. (n.d.). Retrieved December 13, 2021, from https://pypi.org/project/pycountry/.

3- *Python word clouds: How to create a word cloud*. DataCamp Community. (n.d.). Retrieved December 13, 2021, from https://www.datacamp.com/community/tutorials/wordcloud-python.