Text Analysis

Tweeter Sentiment on Covid 19

Introduction

- COVID-19 pandemic has affected lives of people around the world and has also disrupted economic activities and created tough times.
- This project is specifically about a special COVID variant B 1.1.529 termed as Omicron.
- We are interested in using social media (Twitter) to collect the data regarding the tweets related to omicron using specific hashtags.
- Here, we analyze the sentiments of the users on the Twitter platform based upon the hashtags.

About the Dataset

- Using the Twitter API and a Python script, this query collected about 17000 tweets using the hashtag #Omicron, #COVID19, etc.
- We have multiple fields in our data such as name of the user, location of the user, user tweets, and if the user is verified or not.

#	Variable	Description	Type
1	Id	Unique identifier for each tweet	Categorical
2	User_Name	Twitter profile name	Text
3	User_Location	Twitter profile location	Text
4	User_Description	Twitter profile bio	Text
5	User_Created	Date and time Twitter account was created	Date/Time
6	User_Followers	Number of Twitter followers	Numeric
7	User_Friends	Number of friends followed on Twitter	Numeric
8	User_Favorites	Number of tweets marked as a favorite	Numeric
9	User_Verified	Twitter account of public interest that is authentic	Binary
10	Date	Date and time tweet was posted	Date/Time
11	Text	Message posted to Twitter	Text
12	Hashtags	Keyword or phrase to group conversations on Twitter	Text
13	Source	Specific Twitter application used	Text
14	Retweets	Number of times the tweet was reposted	Numeric
15	Favorites	Number of times the tweet was marked as a favorite	Numeric
16	Is_Retweet	If this is a reposted tweet	Binary

Methodology

- Sentiment Analysis is a natural language processing (NLP) technique to determine if text data is positive, neutral, or negative.
- An important tool in business intelligence to better understand customer trends and experiences.
- We plan to use Python code and implement Valence Aware Dictionary for sentiment Reasoning (VADER) to determine the overall tone shared by Twitter users about the Omicron variant.

- Clustering an unsupervised machine learning algorithm that forms groups in the dataset consisting of data points that are very similar to each other.
- Effective at organizing data to create meaningful structure, identify hidden patterns, and gain better insights.
- We plan to use the Python library Sciki (SKlearn) to apply clustering on the dataset to group Twitter users and find any trends among who is more positive, neutral, or negative in their reactions about Omicron.

Research Questions

This dataset will help to better understand how Twitter users reacted to the rise of Omicron. Specifically, we want to know:

- Are there more positive or negative reactions to Omicron?
- Are there any associations between positive and negative tweets about Omicron?
- Are there groups of Twitter users more likely to react positively? Or negatively?
- Are there are any geographic trends?
- Where are more users tweeting about Omicron from?

Sentimental Analysis:

- This dataset is neither too positive nor negative but on an average tending toward a neutral review for omicron variant of COVID-19.
- Sensitivity analysis depicts that the reviews are more of personal opinions, emotion or judgment.
- The dataset has following distribution among the dataset:
- Positive review count: 5356
- Negative review count: 2075
- Neutral review count: 9615

```
1 positive_topics #Headlines with positive sentiments
```

```
['covid give case would last high omicron always know straight promise take variant corona disease',
 'omicron covid vaccine people infection mask live india pandemic could better effective million still delta',
 'omicron variant coronavirus come latest mild keep update study look virus lead find today available',
 'case report omicron death covid today time booster news many update daily year even right',
 'omicron test positive good wave health covid evidence show strong rate first child week early']
```

```
1 negetive_topics #Headlines with positive sentiments
```

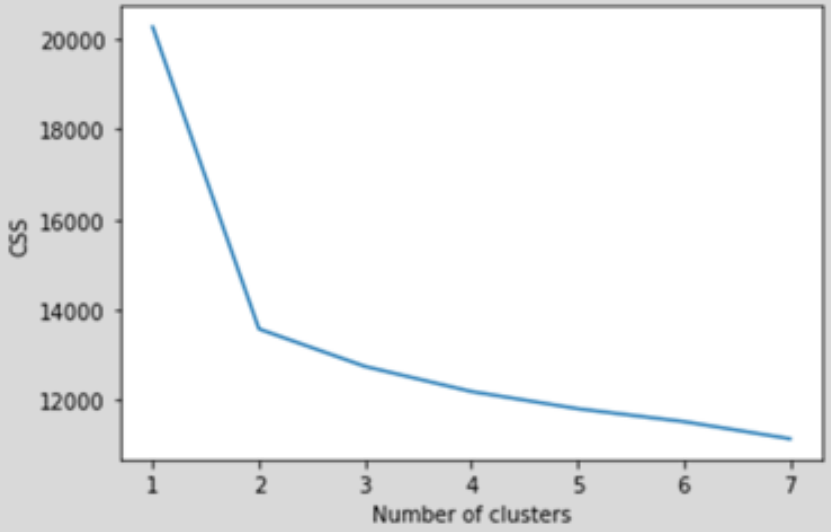
```
['omicron covid previous long tell infection know news truth pandemic could fake medium time report',
 'record late vaccine lockdown track another early health trace didoing glib meal rishi unlock avoid',
 'omicron case coronavirus virus world report mean take bring around india cold sadly military death',
 'omicron case past death infect population worse covid hour mask thing serious total vaccination booster',
 'omicron variant come still base detrick fort year need test study truth people dangerous wave']
```

```
1 neutral_topics #Headlines with positive sentiments
```

```
['omicron people pandemic fort detrick delta virus come still country government india minister take would',
 'covid insight analytics county team death case population confirm growth daily total distribution state life',
 'omicron need drericding report help cdcgov erictopol know mtosterholm covidwatch scottgottliebmd truth emergency time tell',
 'omicron vaccine variant covid show infection update study report news enough world restriction vaccination like',
 'omicron coronavirus give mask health case wave believe milder inevitable surrender surge variant today subvariant']
```

Clustering

Analyzing the elbow plot for the optimal cluster count we get that having _ clusters is a good choice.



- On performing clustering on hashtags to form various cluster of the hashtags and view them as one.
- Looking at the elbow curve we see that just 2 clusters will be sufficient to differentiate the hashtags

```
CLUSTER #1
Key Features: ['news', 'democracy', 'save', 'save democracy', 'dr', 'tomthunkit', 'newsonline', 'medical', 'marie', 'druider']
------------------------------------------------------------------------------
CLUSTER #2
Key Features: ['joyner', 'nathan joyner', 'nathan', 'erwin', 'explorer', 'express', 'feiglding', 'frank', 'frank west', 'frase
r']
------------------------------------------------------------------------------
```

Geographical Analysis

To see how people are reacting in the specific areas.

Tweet distribution:

- From the distribution it's clear that most people tweeting are from Los Angeles, CA and is nearly 5X compared to that of India.
- Also, its visible that there is no specific format for location making it difficult to give any exact insight out of the data.

- Sensitivity
  - Polarity=0.3805, depicting that people have a positive response towards omicron compared to previous COVID-19 strains.
  - Subjectivity=0.9384, the dataset is highly subjective

```
1 loc_rev['user_location'].value_counts()
```

```
Los Angeles, CA                 2658
India                            474
USA                              325
In Your Mind Now                 211
Chandigarh                       210
                                ...
North Kingstown, RI                1
Global                             1
Kiev                               1
Pompano Beach, FL                  1
Auckland Region, New Zealand       1
Name: user_location, Length: 2474, dtype: int64
```

```
1 df_doc_LaTopic['dominant_topic'].value_counts()
```

```
2    522
3    476
1    329
0    279
4    254
Name: dominant_topic, dtype: int64
```

```
1 LA_topics
```

```
['deaths 2022-02-10 population usa… growth tea… cases usafacts… confirmed 2022-02-14 daily state distribution total death',
 'distribution state 2022-02-14 total population 2022-02-08 death usafacts… usafac… deaths usa… daily 2022-02-10 growth cases',
 '2022-02-14 population growth cases tea… confirmed deaths daily usa… usafacts… 2022-02-10 distribution state death usafac…',
 'usafacts… 2022-02-14 cases confirmed daily deaths total 2022-02-08 growth distribution population 2022-02-10 tea… death usafa
c…',
 'confirmed growth cases 2022-02-10 2022-02-14 deaths population tea… state distribution death daily 2022-02-08 usafacts… tota
l']
```

User Analysis

To see how people are reacting on the topic of COVID variant B 1.1.529 termed as Omicron.

- Most active user on the taken set of tweeters is Nathan Joyner also he is from Los Angeles, CA so choosing him for analysis will provide us with same results as that of that. So, choosing the user “save DEMOCRACY” for analysis.
- Polarity=0.0703, subjectivity=0.5267, he is more, or a neutral user have subjective tweets but is neither negative nor positive towards the omicron variant of COVID-19

```
1 user_rev['user_name'].value_counts()
```

```
Nathan Joyner                  2632
save DEMOCRACY                  282
Tomthunkit™                     211
Newsonline                      205
bron druider                    156
                               ...
Time Of India                     1
RestaurantOwner                   1
Pandemic-Aid Networks             1
Lieutenant General Ron Place      1
Kuldip Patel                      1
Name: user_name, Length: 6012, dtype: int64
```

```
1 df_doc_NJTopic['dominant_topic'].value_counts()
```

```
0    87
1    69
2    41
Name: dominant_topic, dtype: int64
```

```
1 NJ_topics
```

```
['that inevitable surrender give milder just omicron will believe fighti… cases promised enough indicator going',
 'what cases that would always covid straight going give promised conti… said… cases/capita keep indicator',
 'enough omicron that believe will indicator what come going leading inevitable would cases said… milder']
```

Thank You