Capstone Project-3

Coronavirus Tweet Sentiment Analysis

Team members

KAMALUDDIN SHAIKH

AMOL RASAM

SHUBHAM JHA

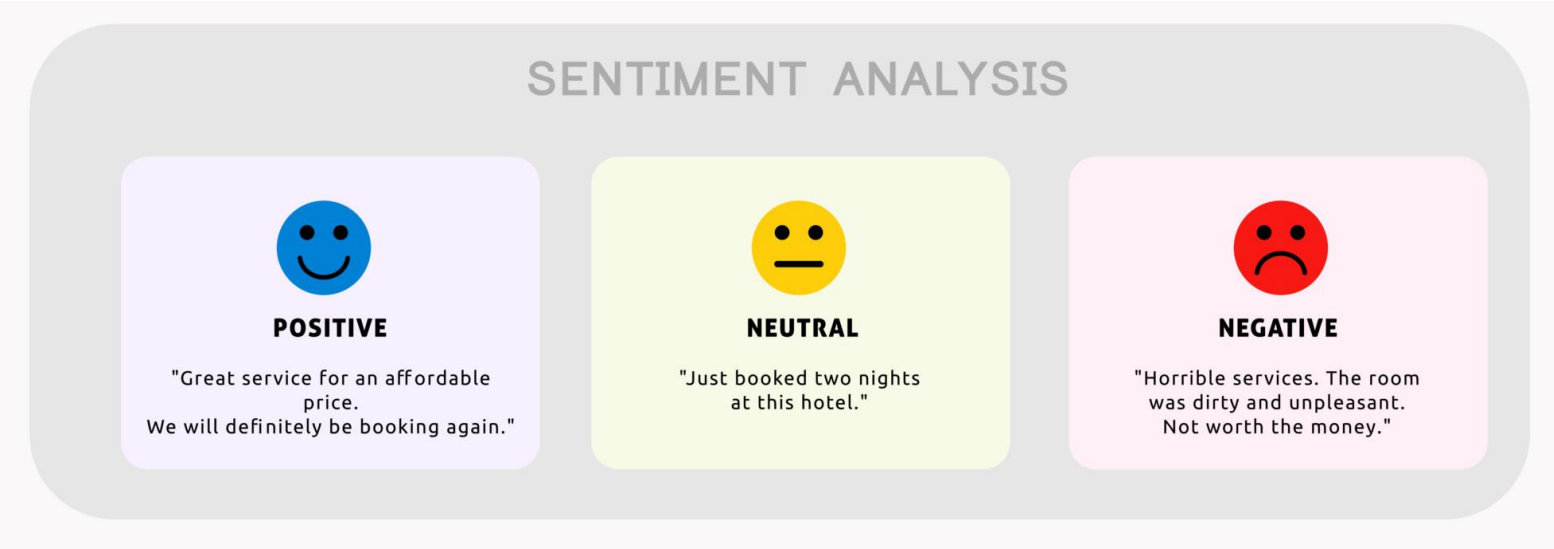
PRETESH

Content

- Problem Statement
- Data Summary
- EDA- Location Analysis
- Sentiment Analysis
- Daily Tweet count
- Monthly and weekly Tweet count
- Data Pre-Processing
- Top Words before and after Stemming
- Label Encoding
- ML Models and Metrics
- Challenges
- Conclusion

Problem Statement

- This challenge asks you to build a classification model to predict the sentiment of COVID-19 tweets. The tweets have been pulled from Twitter and manual tagging has been done then.
- The names and usernames have been given codes to avoid any privacy concerns. We are given information like Location, Tweet At, Original Tweet, and Sentiment.



Data Summary

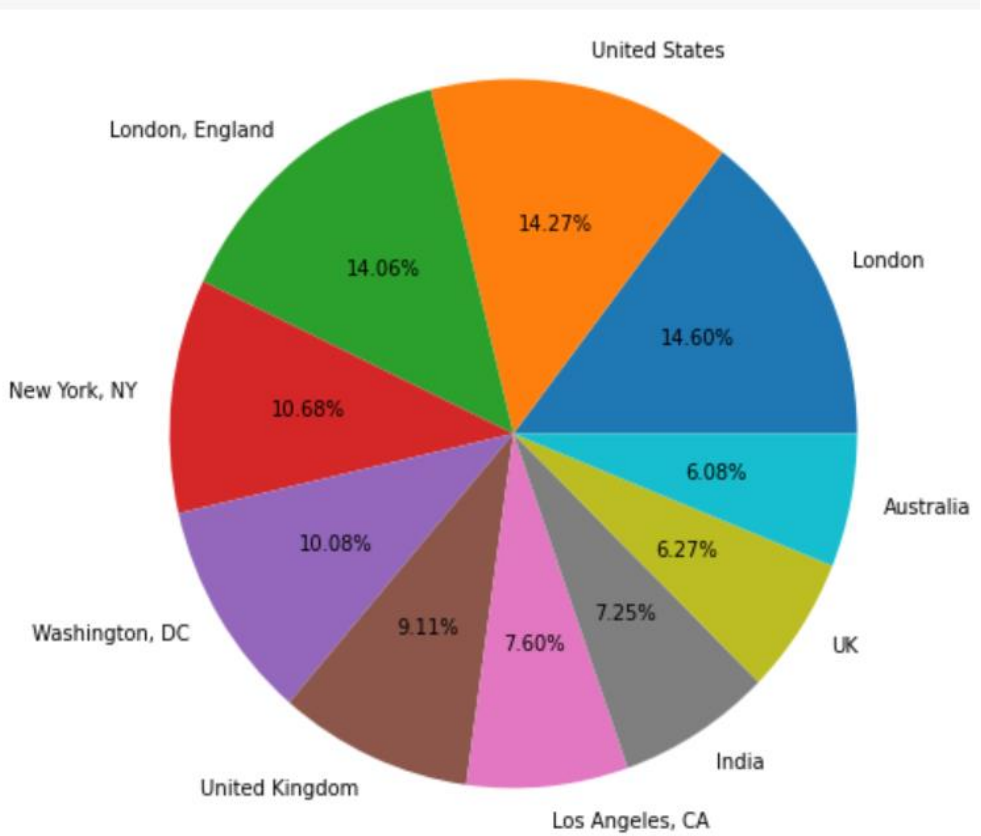
The task is to build a classification model to predict the sentiment of Covid-19 related tweets. The dataset contains following columns:

- **Location:** The location at which the tweet was made.
- **Tweet At:** The date on which the tweet was made.
- **Original Tweet:** This is the actual text of the tweet.
- **Sentiment:** This is the sentiment of the tweet, which is manually tagged 41157 Rows Multiclass classification with 5 classes: Extremely Positive, Positive, Neutral, Negative, Extremely Negative.

EDA – Location Analysis

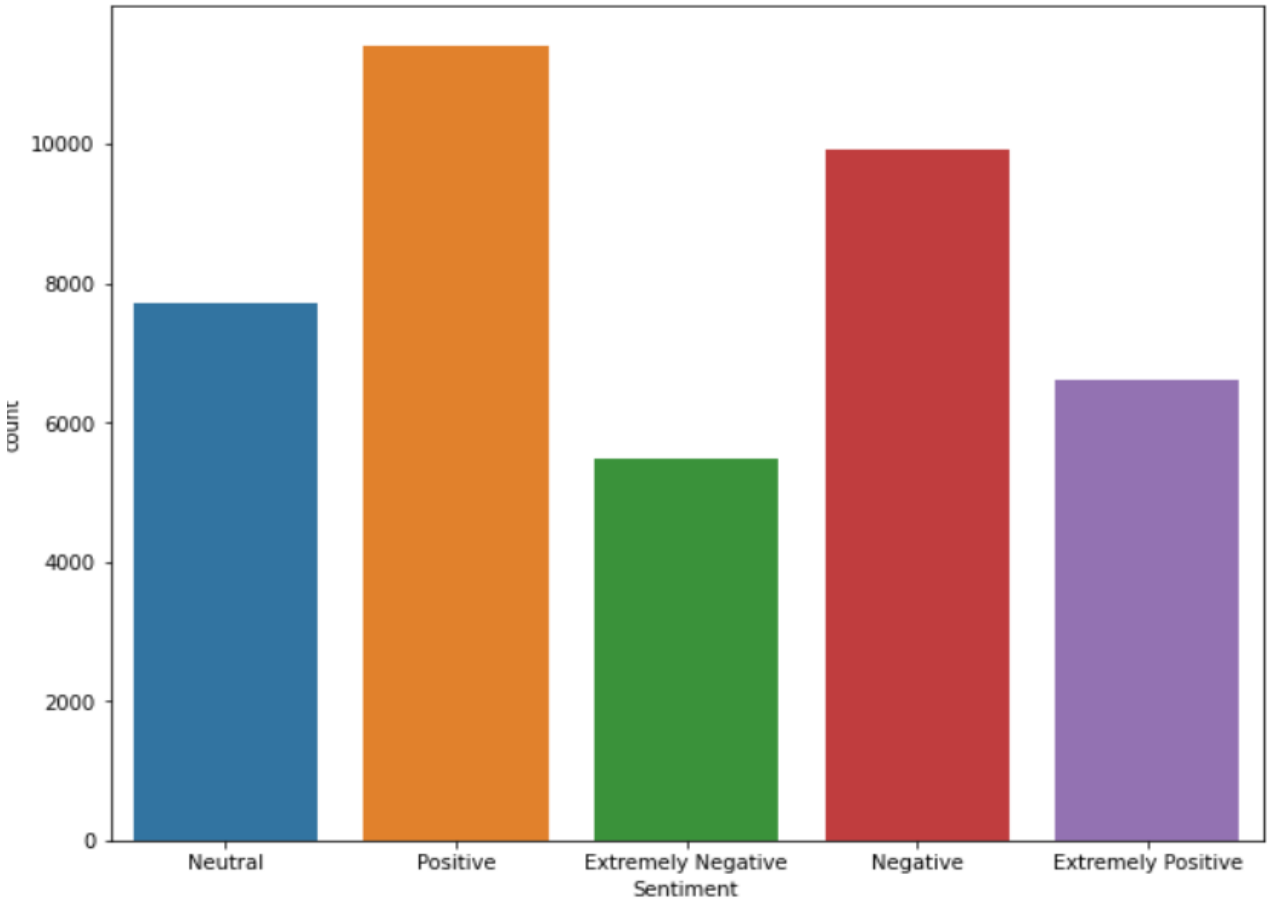
```
Top_Location_Of_tweet

London                 540
United States          528
London, England        520
New York, NY           395
Washington, DC         373
United Kingdom         337
Los Angeles, CA        281
India                  268
UK                     232
Australia              225
Name: Location, dtype: int64
```



We can see Most Tweets are coming from London. almost 28.66%.

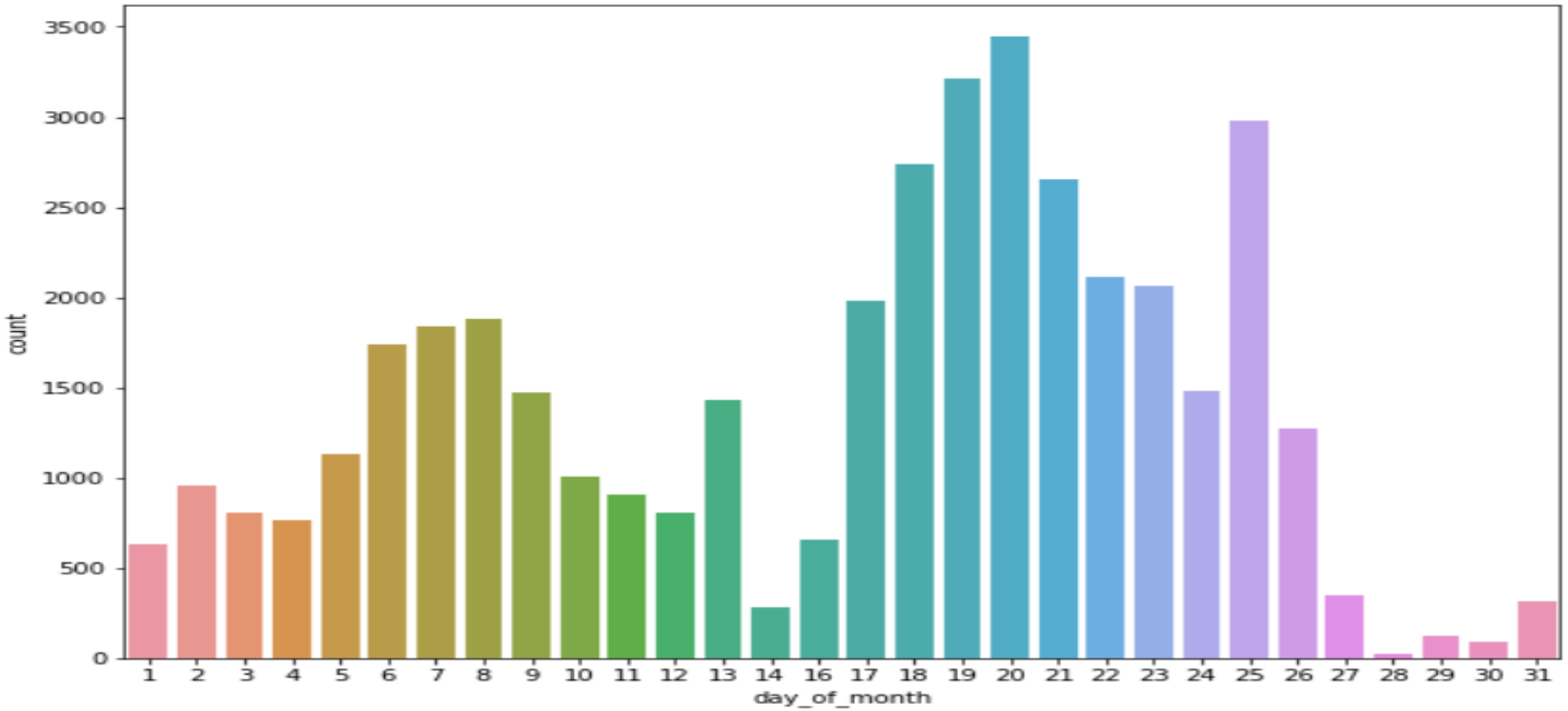
Sentiment Analysis




```
Positive              11422
Negative               9917
Neutral                7713
Extremely Positive     6624
Extremely Negative     5481
Name: Sentiment, dtype: int64
```

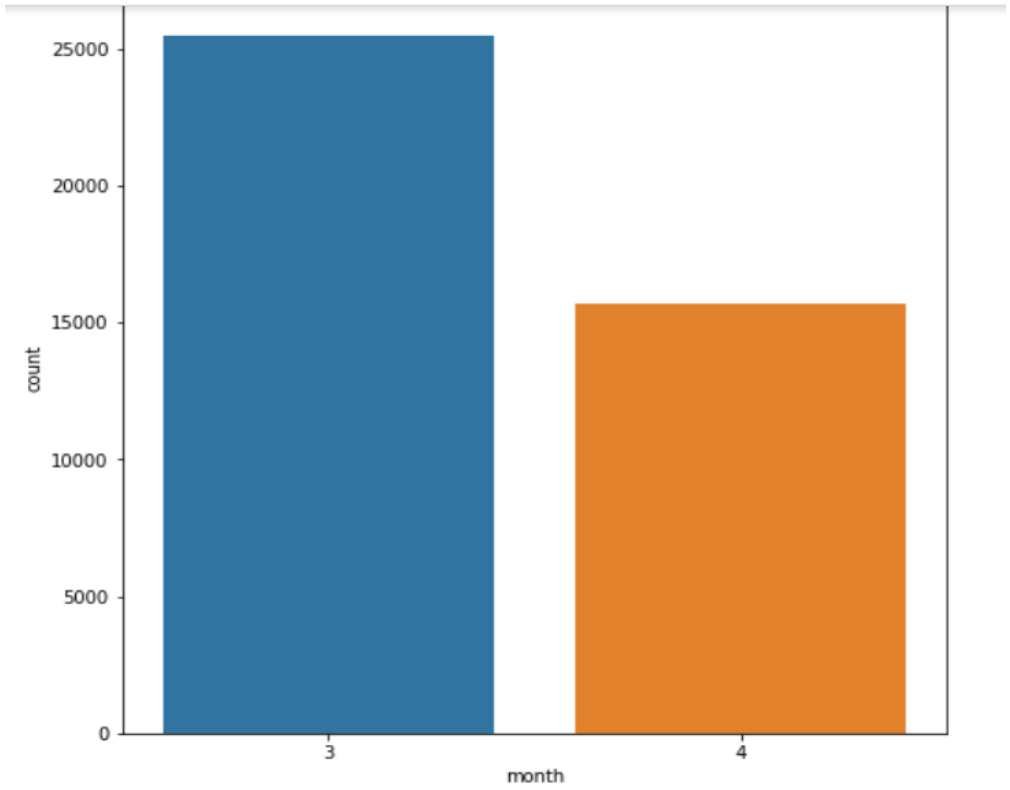
we know that most of the peoples are having positive sentiments about various issues shows us their optimism during pandemic times.

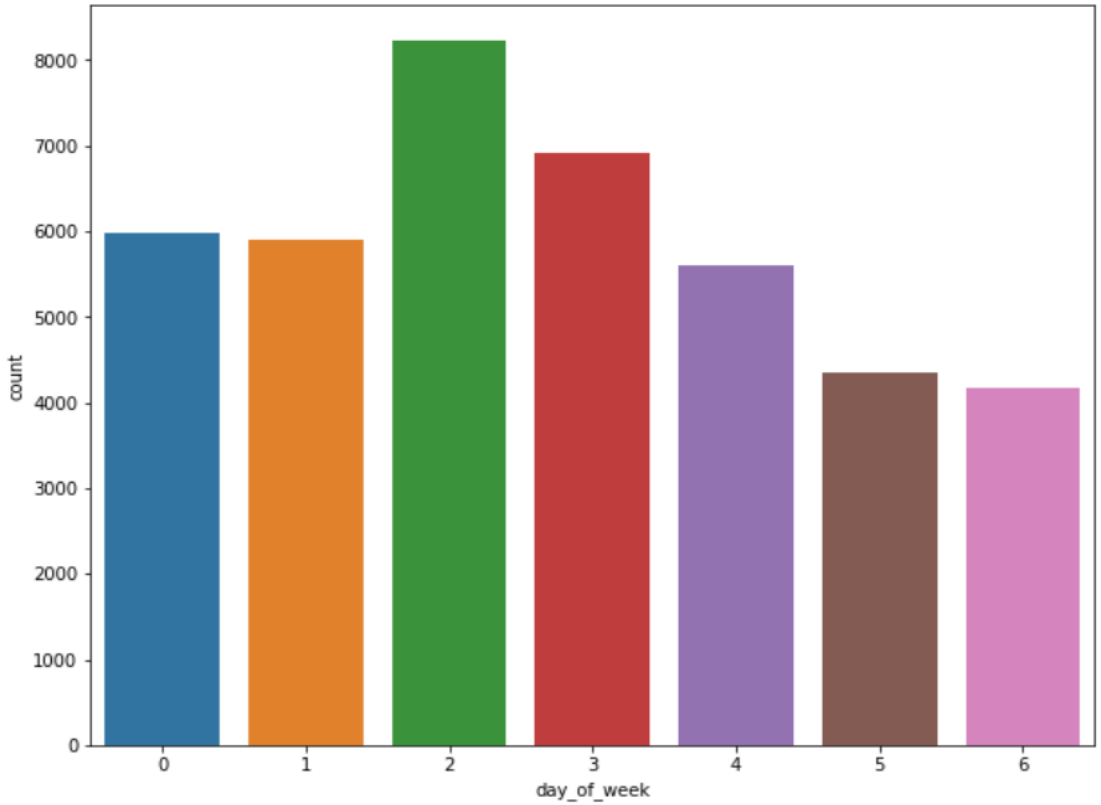
Daily Tweet Analysis



Most people tweeted during 18th to 25th in the month of March 2020.
We know the corona was spreading during this particular time and it was declared as National emergency by western countries and pandemic by WHO in the month of march.

Monthly and Weekly Tweet Count





March is the month in which most number of tweets were tweeted – 25499 tweets.

Most people tweeted mostly during Tuesday and Wednesday of every week.

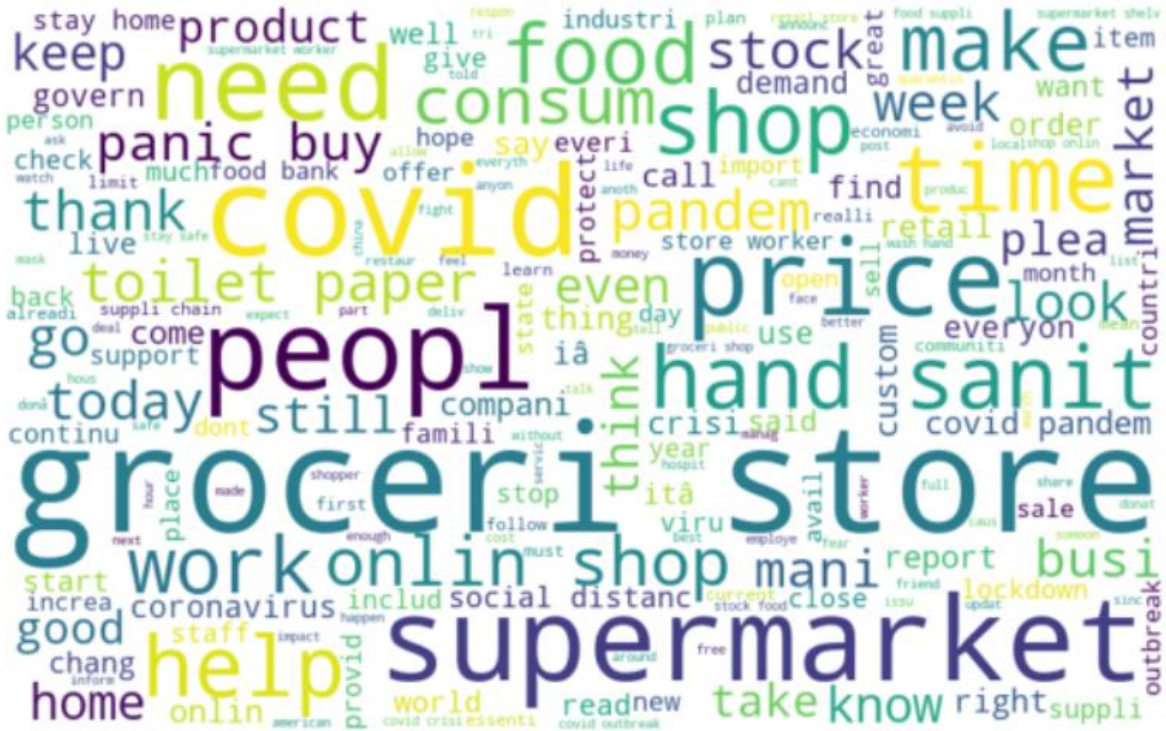
Data Pre-Processing

- **Steps taken to prepare the data**
  - Removing Twitter Handles/ Usernames
  - Removing URL links
  - Removing # symbols and retaining the tags
  - Removing Punctuations and stop words
  - Removing short words
  - Tokenization and stemming

Top Word - Stemming

Top words – Before Stemming

Top words – After Stemming



Label Encoding

- We had 5 classes in our Sentiment column “ Extremely Positive, Positive, Neutral, Negative and Extremely Negative”. We converted it into 3 class “Positive, Neutral and Negative”.
- It will give us better Accuracy and better overall understanding.

- **Extremely Positive = 2**
- **Positive= 2**
- **Neutral= 1**
- **Negative= 0**
- **Extremely Negative= 0**

```
dataset.Sentiment.value_counts()
```

```
2    18046
0    15398
1     7713
Name: Sentiment, dtype: int64
```

ML Models and Metrics

No	Model	Test accuracy	Precision	Recall
1	Logistic Regression	0.794866	0.794407	0.794866
2	Random Forest	0.771299	0.772827	0.771299
3	Naive Bayes	0.686751	0.692846	0.686751
4	XGBoost	0.686751	0.692846	0.686751
5	Decision Tree	0.499028	0.899430	0.499028
6	K Nearest Neighbour	0.340217	0.794275	0.340217

Logistic Regression model performed the best out of all these Models.

Challenges

- **Text Pre-Processing**
    1. Removing @users
    2. Removing urls
    3. Removing #tags
- **Vectorization**
- **Label Encoding**
- **Model Selection**

Conclusion

- As we have implemented six different models to predict the sentiment of COVID-19 Tweets. Logistic Regression, Random Forest Classifier, Decision Tree, Naive Bayes, K Nearest Neighbor and Xgboost Classifier.
- Logistic Regression model performed the best among them. In this way, we can explore more from various textual data and tweets. Our models will try to predict the various sentiments correctly.
- Our Model will help Government to make use of this information in policymaking as they can able to know how people are reacting to this new strain, what all challenges they are facing such as food scarcity, panic attacks, etc. Various profit organizations can make a profit by analyzing various sentiments as one of the tweets telling us about the scarcity of masks and toilet papers.

Thank You