Capstone Project
on
Coronavirus Tweet Sentiment Analysis

by

Muktesh Singh

Problem Statement

- **Analysing various sentiments of COVID_19 tweets during the period March 2020 to April 2020.**

- **Build a classification model to predict the sentiment of COVID-19 tweets.**

Sentiment Analysis

- Sentiment analysis is a natural language processing technique to find emotions related to the public/customers opinion (text data). It may be positive, negative, neutral etc.

- It helps different stake holders to understand the public/customers mindset and their requirement.

- For example government can make policies based on public reaction on new strain, food scarcity, panic attacks etc during COVID.

Data Summary

Details of dataset – coronavirus tweets.csv

- Number of rows – 41156
- Number of columns – 5
- Datatypes - int64 and object
- Only Location column has some null values

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 41157 entries, 0 to 41156
Data columns (total 6 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   UserName       41157 non-null  int64
 1   ScreenName     41157 non-null  int64
 2   Location       32567 non-null  object
 3   TweetAt        41157 non-null  object
 4   OriginalTweet  41157 non-null  object
 5   Sentiment      41157 non-null  object
dtypes: int64(2), object(4)
memory usage: 1.9+ MB
```

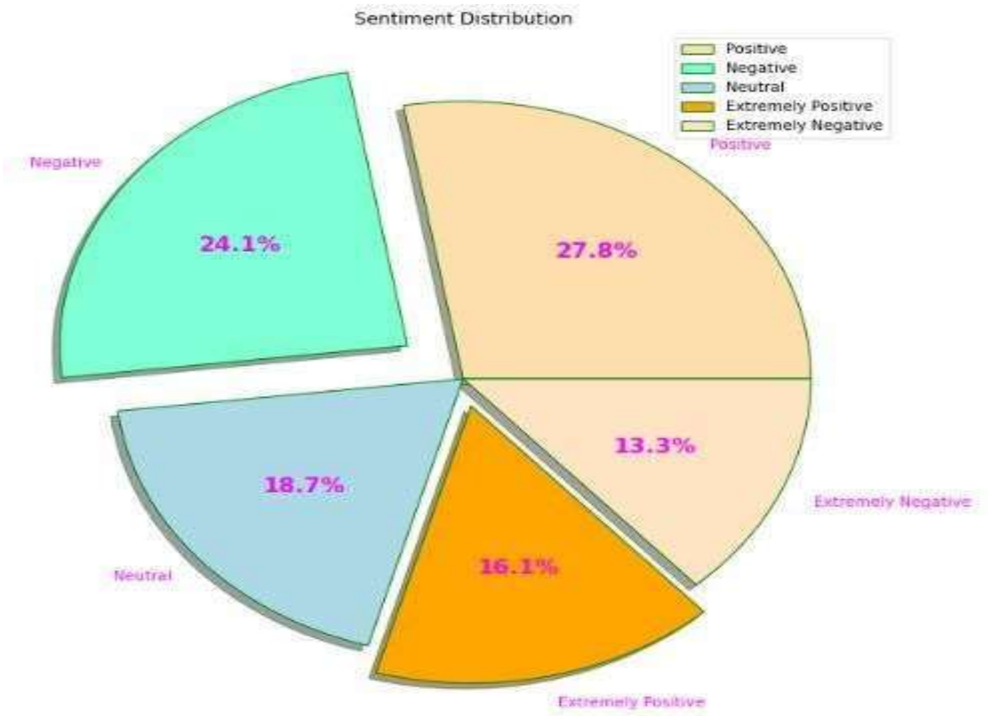
Null Values

```
UserName            0
ScreenName          0
Location         8590
TweetAt             0
OriginalTweet       0
Sentiment           0
```

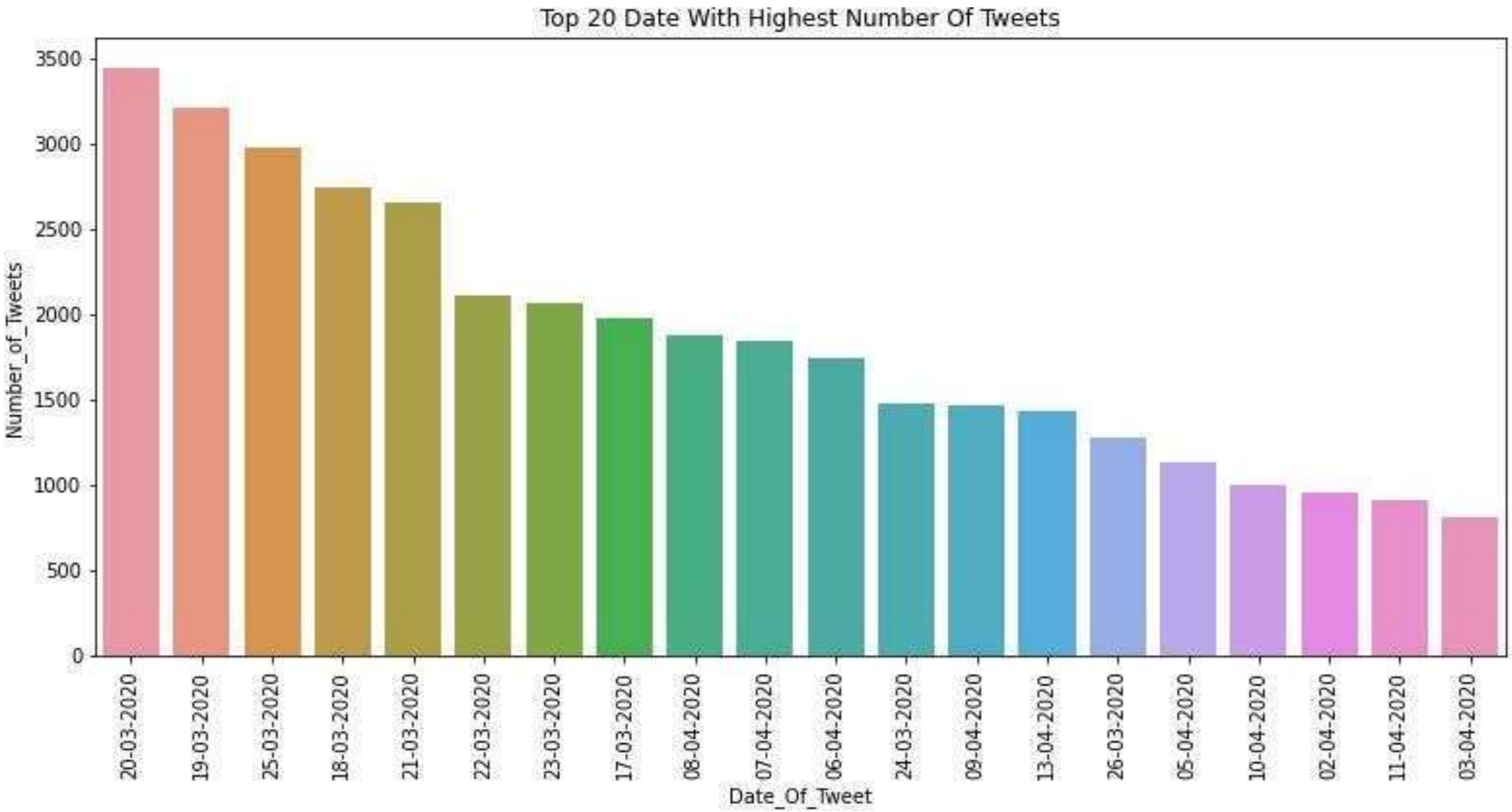
Exploratory Data Analysis

Sentiment distribution of tweets

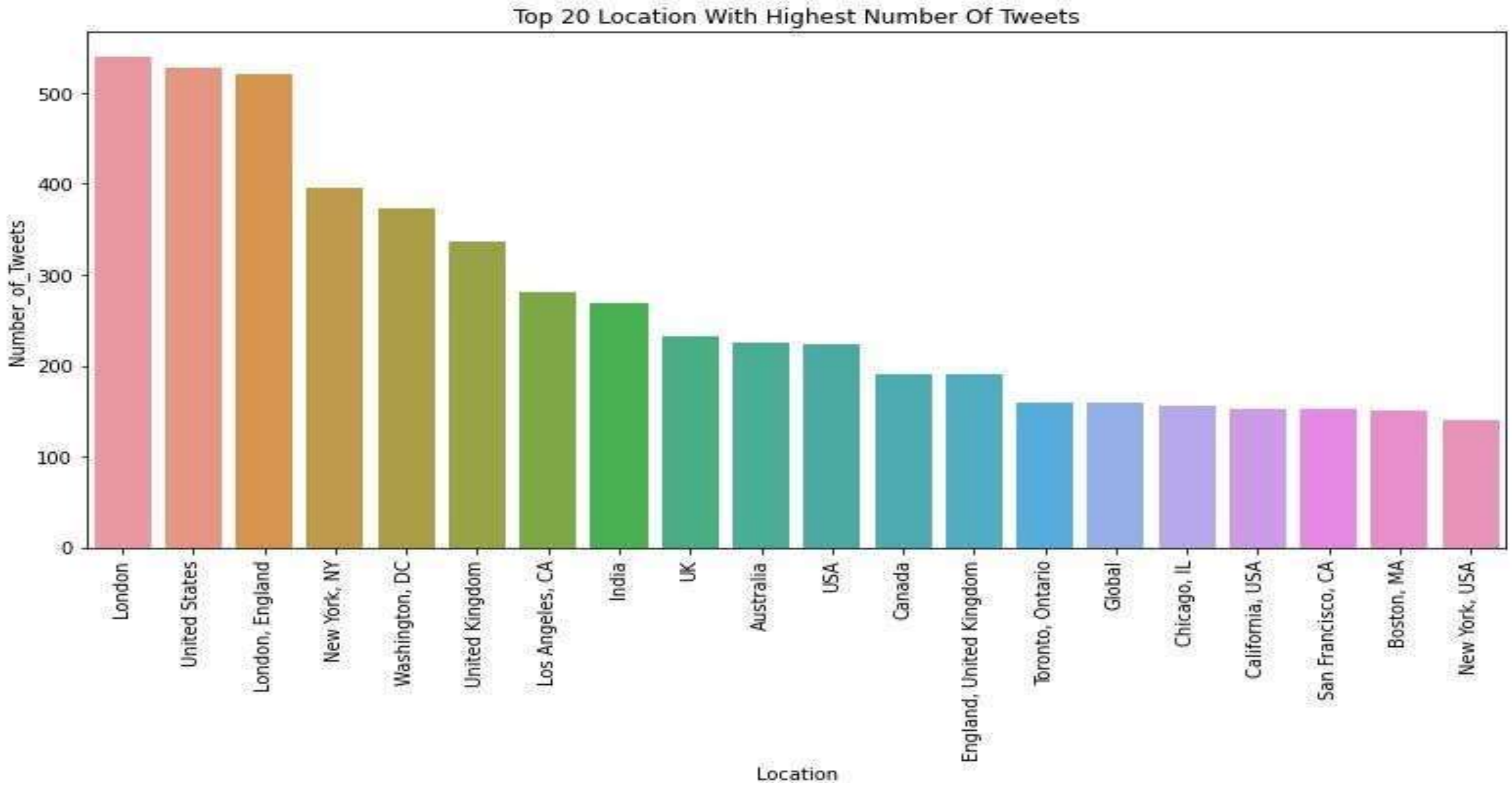
	Sentiment	Number_of_Tweets
0	Positive	11422
1	Negative	9917
2	Neutral	7713
3	Extremely Positive	6624
4	Extremely Negative	5481



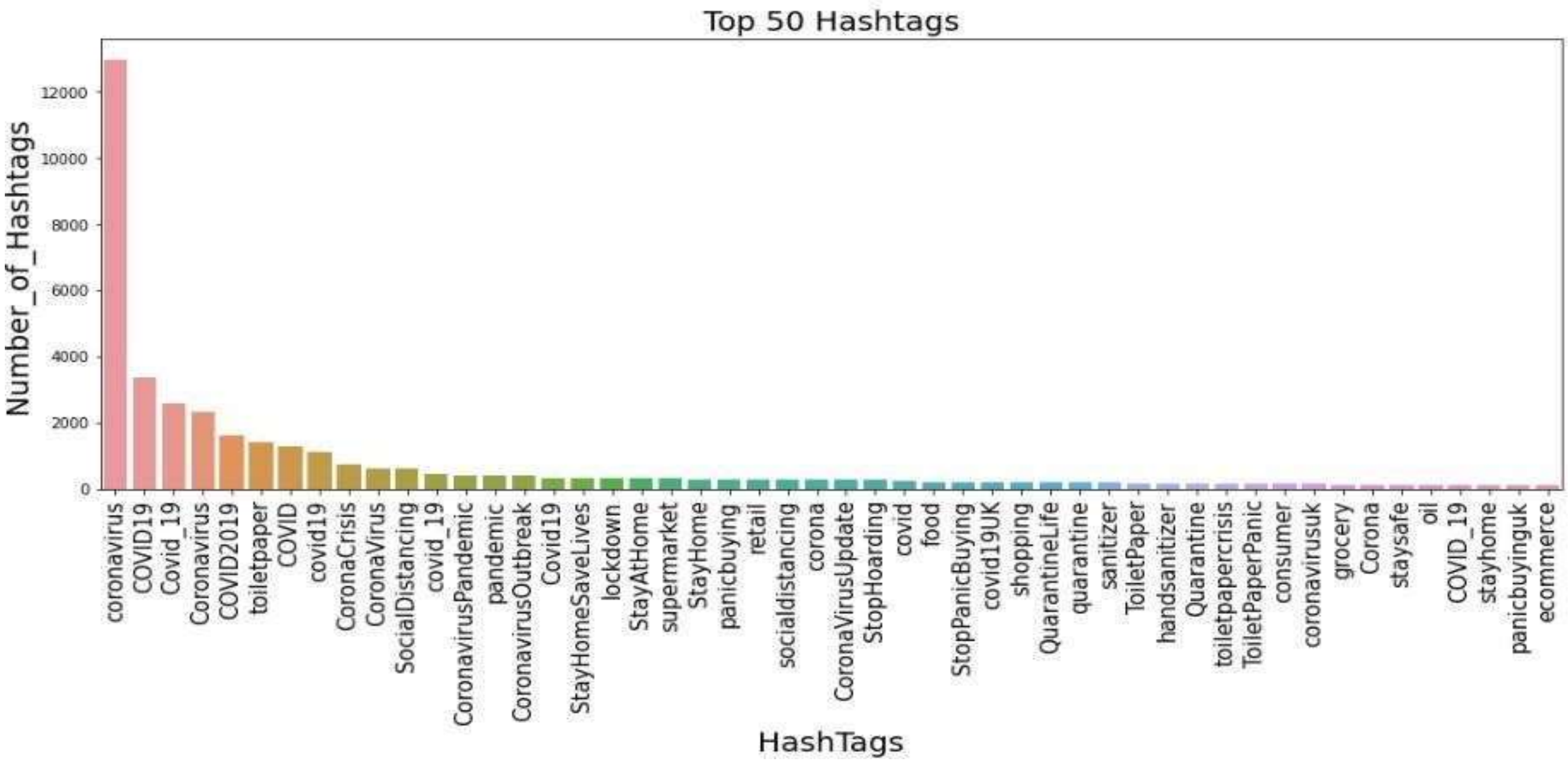
Top 20 Date with Highest Number of Tweets



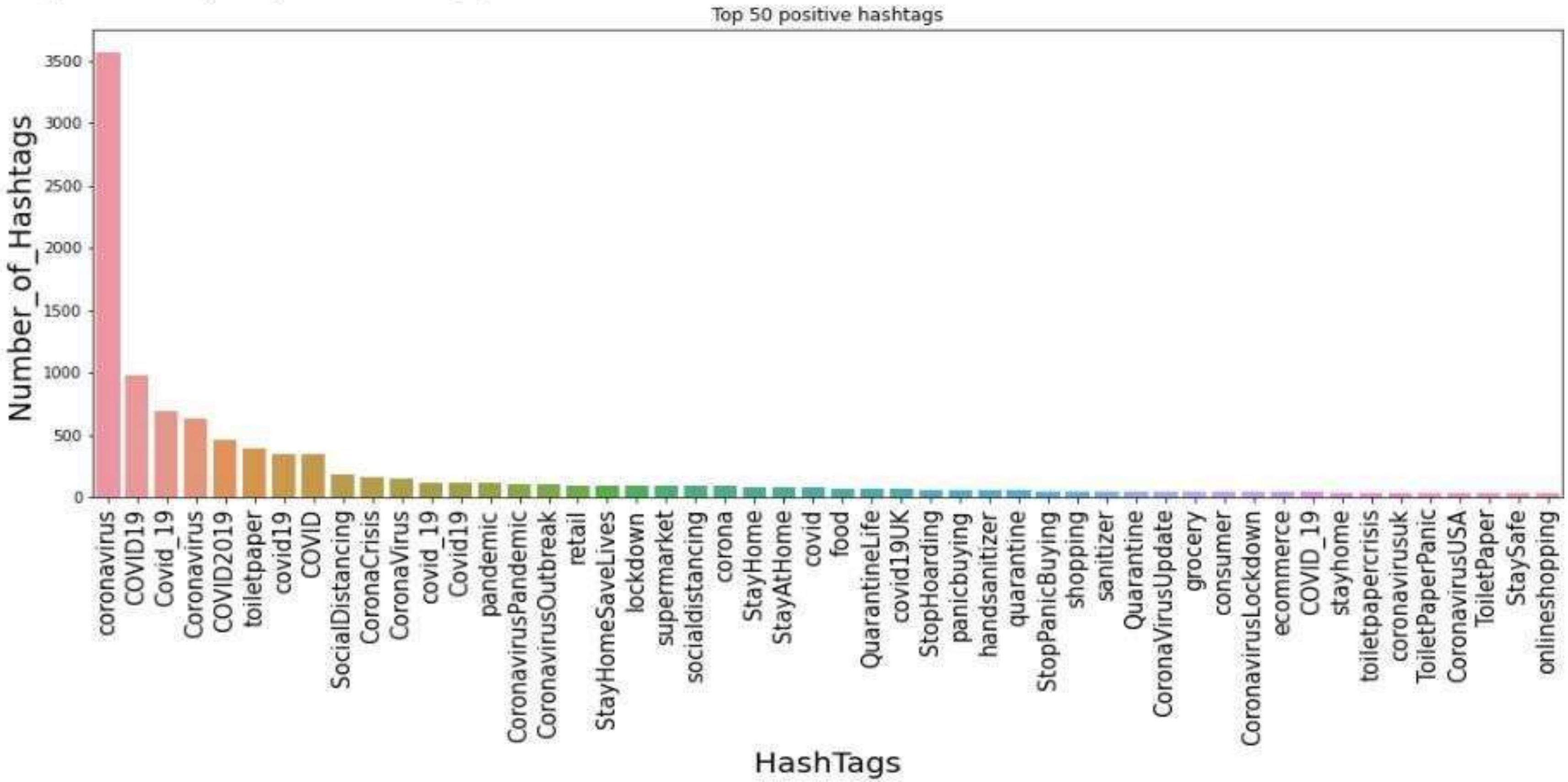
Top 20 Location with Highest Number of Tweets



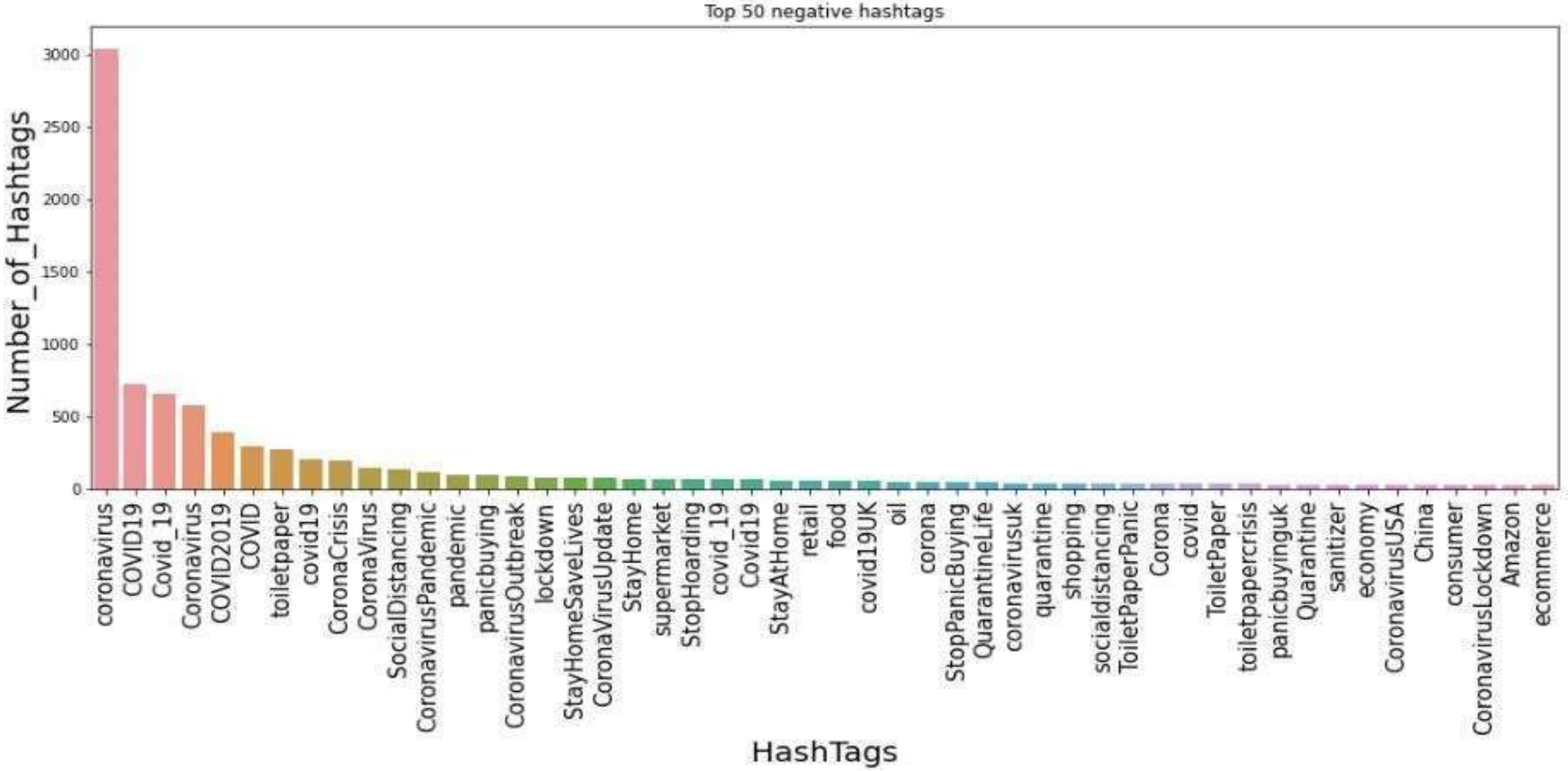
Top 50 Hashtags



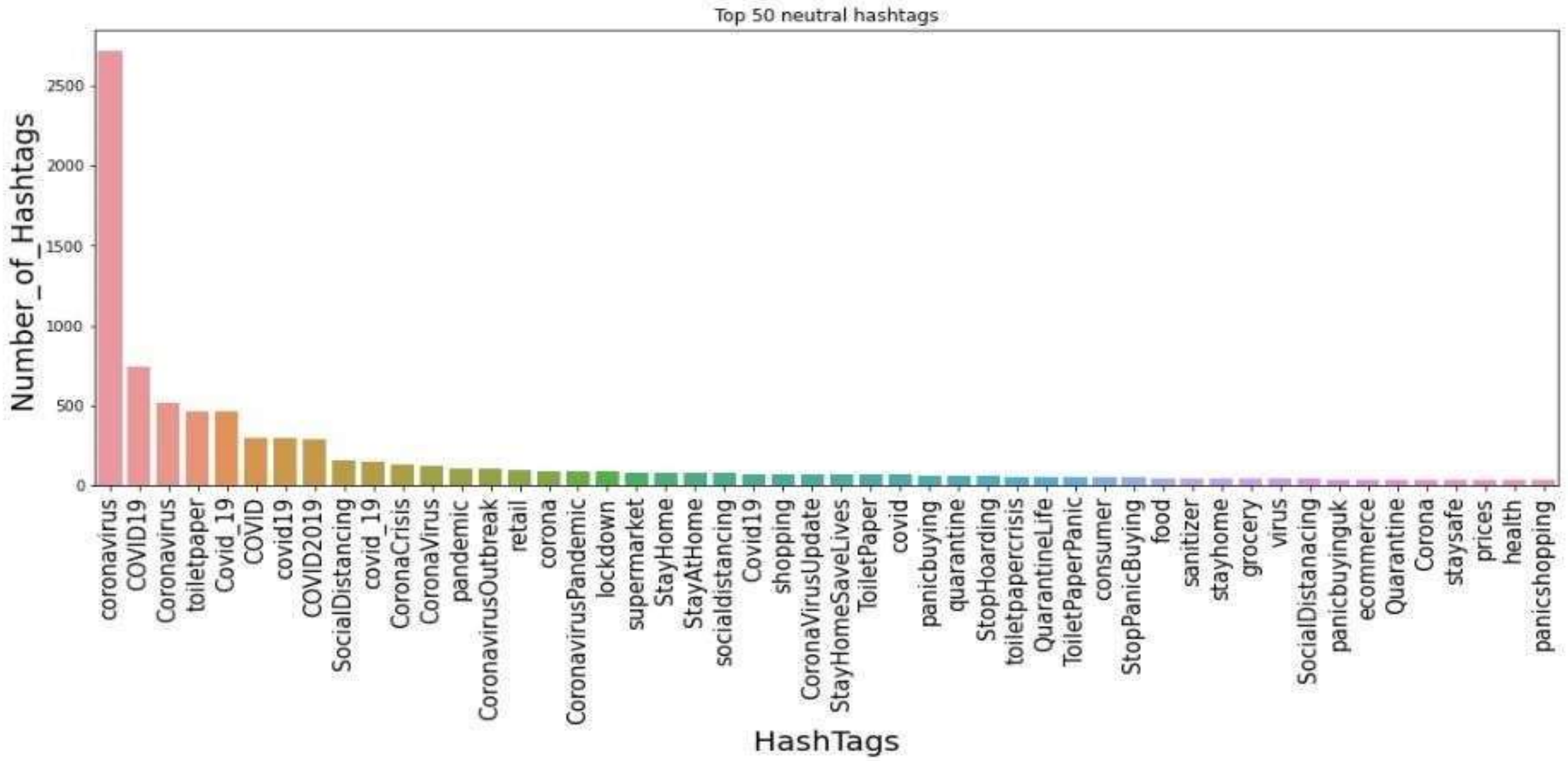
Top 50 Positive Hashtags



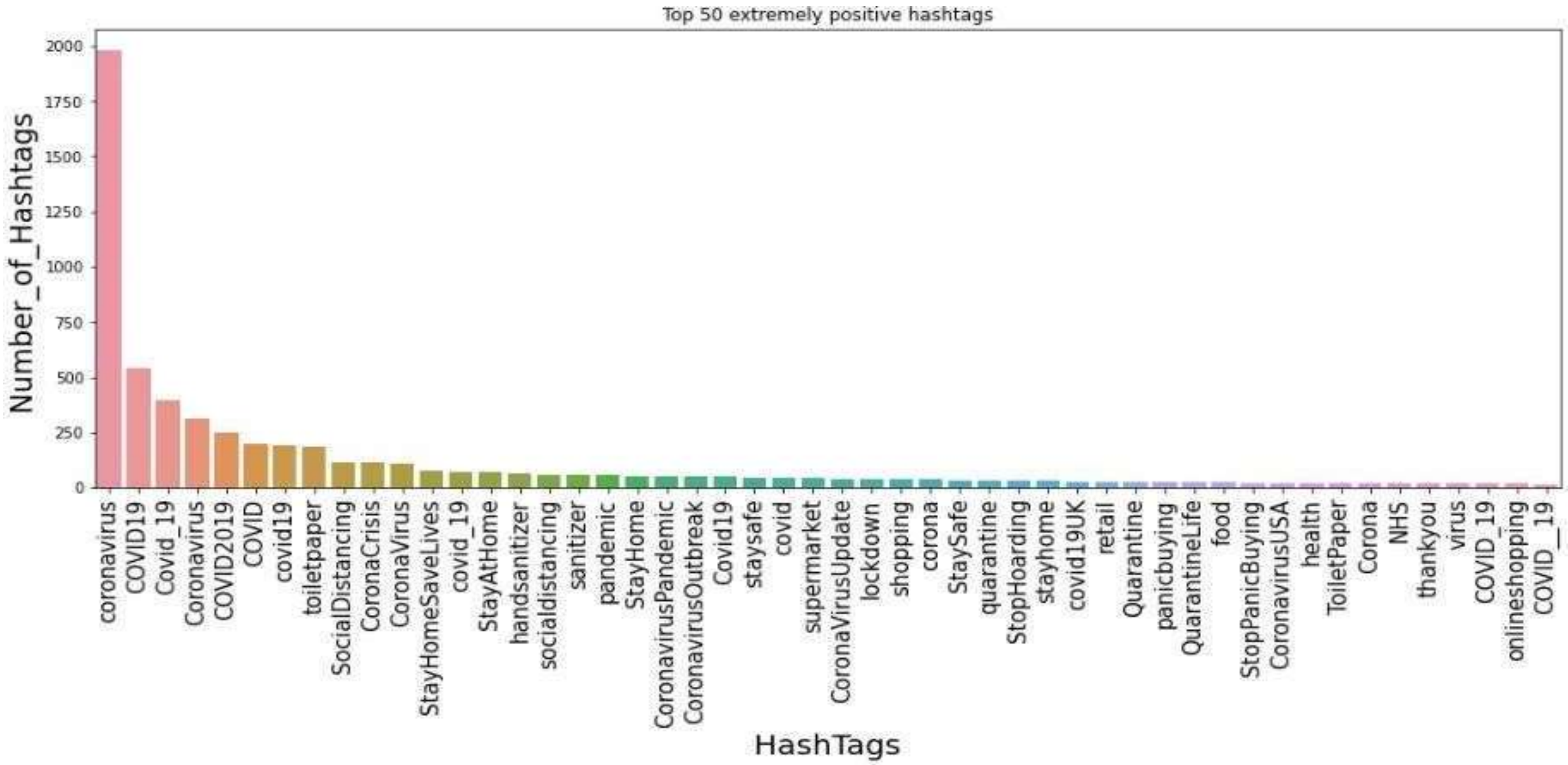
Top 50 Negative Hashtags



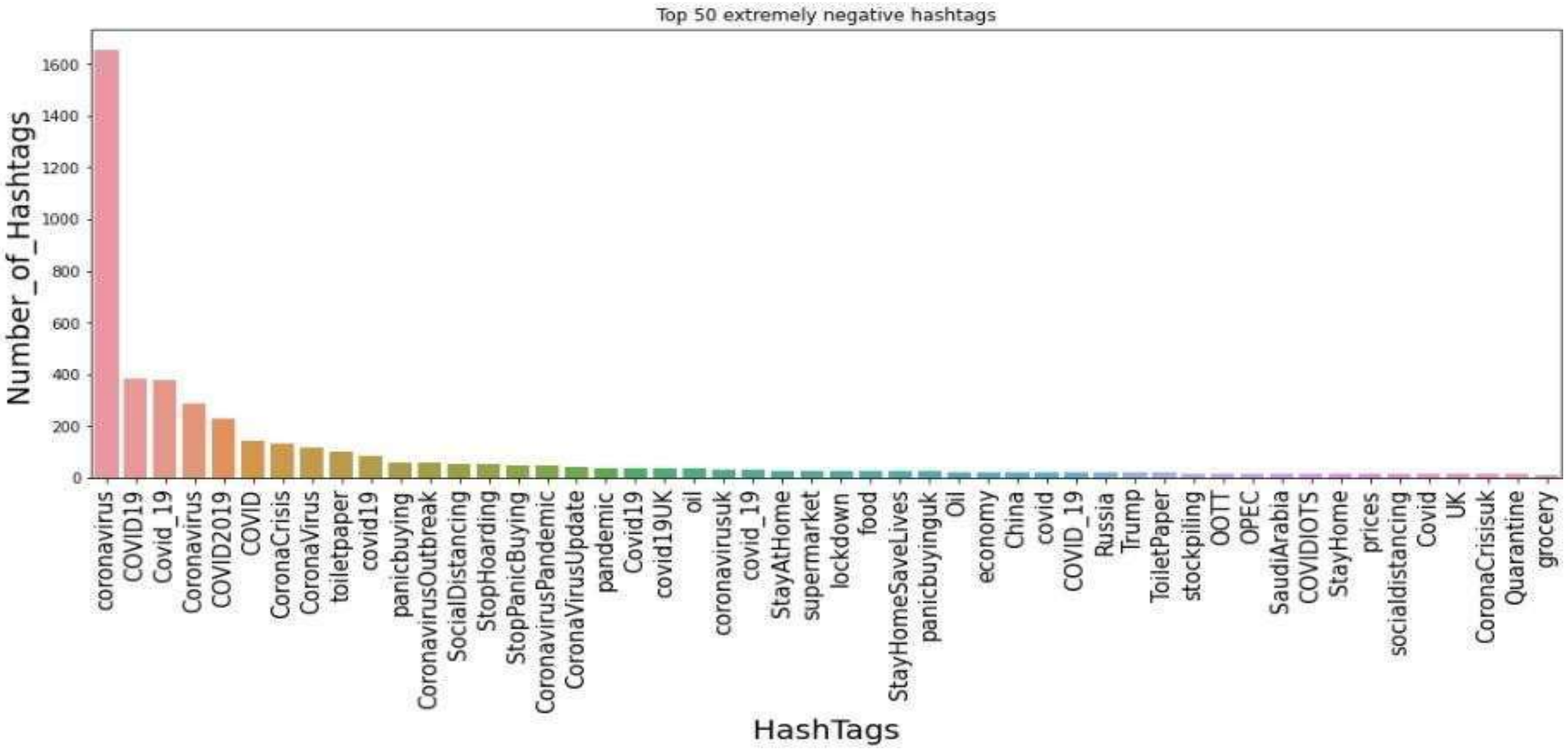
Top 50 Neutral Hashtags



Top 50 Extremely Positive Hashtags



Top 50 Extremely Negative Hashtags



Insights from EDA

- From sentiment distribution of tweets it is clear that 27.8% of the tweets are positive followed by negative(24.1%).

- 20^{th} March 2020 was the date with highest number of tweets.

- London was the city with most of number of tweets twitted by people on twitter.

- #coronavirus is the most used hashtags by a large margin in all the sentiments.

- Few hashtags are common in all the sentiments but many hashtags are different for different sentiments.

Data Preprocessing

- The raw data extracted from twitter contains so much noise. If we apply machine learning algorithm to this data the model will give inaccurate results. To resolve this problem we need to perform the following steps

1) Remove usernames
2) Remove URLs
3) Remove punctuation, special characters
4) Remove stop words
5) Lemmatization

Original tweet v/s Clean tweet

	OriginalTweet	Clean_Tweet
0	@MeNyrbie @Phil_Gahan @Chrisitv https://t.co/iFz9FAn2Pa and https://t.co/xX6ghGFzCC and https://t.co/I2NlzdxNo8	
1	advice Talk to your neighbours family to exchange phone numbers create contact list with phone numbers of neighbours schools employer chemist GP set up online shopping accounts if poss adequate supplies of regular meds but not over order	advice talk neighbour family exchange phone number create contact list phone number neighbour school employer chemist gp set online shopping account po adequate supply regular med order
2	Coronavirus Australia: Woolworths to give elderly, disabled dedicated shopping hours amid COVID-19 outbreak https://t.co/bInCA9Vp8P	coronavirus australia woolworths give elderly disabled dedicated shopping hour amid covid outbreak
3	My food stock is not the only one which is empty...\r\r\n\r\r\nPLEASE, don't panic, THERE WILL BE ENOUGH FOOD FOR EVERYONE if you do not take more than you need. \r\r\nStay calm, stay safe.\r\r\n\r\r\n#COVID19france #COVID_19 #COVID19 #coronavirus #confinement #Confinementotal #ConfinementGeneral https://t.co/zrlG0Z520j	food stock one empty please panic enough food everyone take need stay calm stay safe covid france covid covid coronavirus confinement confinementotal confinementgeneral
4	Me, ready to go at supermarket during the #COVID19 outbreak.\r\r\n\r\r\nNot because I'm paranoid, but because my food stock is litteraly empty. The #coronavirus is a serious thing, but please, don't panic. It causes shortage...\r\r\n\r\r\n#CoronavirusFrance #restezchezvous #StayAtHome #confinement https://t.co/usmuaLq72n	ready go supermarket covid outbreak paranoid food stock litteraly empty coronavirus serious thing please panic cause shortage coronavirusfrance restezchezvous stayathome confinement

Model Training

Converting Text to Matrix

We cannot pass the textual data directly to the ML algorithm. These words need to then be encoded as integers, or floating-point values. We can do it using following methods

1) Count Vectorizer Method

Count vectorizer convert a collection of text documents to matrix of integers. Where each integer represents the frequency of the word token in that document.

2) TF-IDF Method

TF-IDF method represents not only the count of the word token in the document it also reflect how important a word is to a document in collection of corpus.

- TF = (Number of times term t appears in a document)/(Number of terms in the document)
- IDF = log(N/n), where, N is the total number of documents and n is the number of documents the term t has appeared in.
- TF-IDF = TF*IDF

Different Models Used

1. **Naive Bayes Classifier**
2. **Random Forest Classifier**
3. **Logistic Regression**
4. **XGBOOST**
5. **Support Vector Machine Classifier**

Naive Bayes Classifier

Binary Classification

```
training accuracy Score    :  0.868883826879271
Validation accuracy Score :  0.7916666666666666
              precision    recall  f1-score   support

           0       0.70      0.73      0.72      2955
           1       0.85      0.83      0.84      5277

    accuracy                           0.79      8232
   macro avg       0.77      0.78      0.78      8232
weighted avg       0.79      0.79      0.79      8232
```

Multi Class Classification

```
training accuracy Score    :  0.7303264996203492
Validation accuracy Score :  0.4866375121477162
                    precision    recall  f1-score   support

Extremely Negative       0.41      0.58      0.48       784
Extremely Positive       0.43      0.57      0.49       982
          Negative       0.51      0.44      0.48      2303
           Neutral       0.40      0.65      0.49       942
          Positive       0.59      0.42      0.49      3221

          accuracy                           0.49      8232
         macro avg       0.47      0.53      0.49      8232
      weighted avg       0.51      0.49      0.49      8232
```

Random Forest Classifier

Binary Classification

```
Training accuracy Score    :  0.9998785117691723
Validation accuracy Score :  0.8358843537414966
              precision    recall  f1-score   support

           0       0.72      0.82      0.77      2721
           1       0.90      0.84      0.87      5511

    accuracy                           0.84      8232
   macro avg       0.81      0.83      0.82      8232
weighted avg       0.84      0.84      0.84      8232
```

Multi Class Classification

```
training accuracy Score    :  0.9997873955960517
Validation accuracy Score :  0.565597667638484
                    precision    recall  f1-score   support

Extremely Negative       0.39      0.69      0.49       615
Extremely Positive       0.36      0.73      0.48       646
          Negative       0.53      0.51      0.52      2047
           Neutral       0.81      0.61      0.69      2054
          Positive       0.64      0.51      0.57      2870

          accuracy                           0.57      8232
         macro avg       0.54      0.61      0.55      8232
      weighted avg       0.61      0.57      0.58      8232
```

Logistic Regression

Binary Classification

Training accuracy Score : 0.9555353075170843
Validation accuracy Score : 0.8654033041788144

	precision	recall	f1-score	support
0	0.77	0.85	0.81	2794
1	0.92	0.87	0.90	5438
accuracy			0.87	8232
macro avg	0.85	0.86	0.85	8232
weighted avg	0.87	0.87	0.87	8232

Multi Class Classification

training accuracy Score : 0.929081245254366
Validation accuracy Score : 0.6137026239067055

	precision	recall	f1-score	support
Extremely Negative	0.61	0.67	0.64	1006
Extremely Positive	0.61	0.69	0.65	1162
Negative	0.55	0.57	0.56	1921
Neutral	0.72	0.65	0.68	1712
Positive	0.60	0.57	0.58	2431
accuracy			0.61	8232
macro avg	0.62	0.63	0.62	8232
weighted avg	0.62	0.61	0.61	8232

XGBOOST

Binary Classification

Training accuracy Score   : 0.741199696279423
Validation accuracy Score : 0.7396744412050534

	precision	recall	f1-score	support
0	0.37	0.84	0.52	1361
1	0.96	0.72	0.82	6871
accuracy			0.74	8232
macro avg	0.67	0.78	0.67	8232
weighted avg	0.86	0.74	0.77	8232

Multi Class Classification

training accuracy Score   : 0.49281700835231584
Validation accuracy Score : 0.47922740524781343

	precision	recall	f1-score	support
Extremely Negative	0.39	0.59	0.47	716
Extremely Positive	0.40	0.68	0.51	784
Negative	0.38	0.45	0.41	1666
Neutral	0.58	0.46	0.52	1948
Positive	0.58	0.43	0.49	3118
accuracy			0.48	8232
macro avg	0.47	0.52	0.48	8232
weighted avg	0.51	0.48	0.48	8232

Support Vector Machine Classifier

Binary Classification

```
Training accuracy Score    :  0.9590888382687928
Validation accuracy Score :  0.8380709426627794
              precision    recall  f1-score   support

           0       0.67      0.86      0.76      2403
           1       0.94      0.83      0.88      5829

    accuracy                           0.84      8232
   macro avg       0.81      0.85      0.82      8232
weighted avg       0.86      0.84      0.84      8232
```

Multi Class Classification

```
training accuracy Score    :  0.9100075930144267
Validation accuracy Score :  0.5998542274052479
                    precision    recall  f1-score   support

Extremely Negative       0.47      0.70      0.56       732
Extremely Positive       0.53      0.77      0.62       909
          Negative       0.55      0.54      0.54      2024
           Neutral       0.72      0.63      0.67      1748
          Positive       0.67      0.55      0.60      2819

          accuracy                           0.60      8232
         macro avg       0.59      0.64      0.60      8232
      weighted avg       0.62      0.60      0.60      8232
```

Models in terms of Test Accuracy

Binary Classification

Model	Test accuracy
Logistic Regression	0.865403
Support Vector Machines	0.838071
Random Forest	0.835884
Naive Bayes	0.791667
XGBoost	0.739674

Multi Class Classification

Model	Test accuracy
Support vector machine	0.616861
Logistic Regression	0.613703
RANDOM FOREST CLASSIFIER	0.560253
Extreme Gradient Boosting	0.551020
Stochastic Gradient Descent-SGD Classifier	0.509840
Naive Bayes Classifier	0.489189

Hyperparameter Tuning for Top Model

Binary Classification for Logistic Regression

training accuracy Score : 0.9656492027334852
Validation accuracy Score : 0.8652818270165209

	precision	recall	f1-score	support
0	0.78	0.85	0.81	2837
1	0.92	0.87	0.89	5395
accuracy			0.87	8232
macro avg	0.85	0.86	0.85	8232
weighted avg	0.87	0.87	0.87	8232

Tuned Parameter C=1.623
Where C is Regularization strength

Multi Class Classification for Support Vector Classifier

training accuracy Score : 0.8137281700835232
Validation accuracy Score : 0.6168610301263362

	precision	recall	f1-score	support
Extremely Negative	0.53	0.69	0.60	838
Extremely Positive	0.57	0.75	0.65	1016
Negative	0.56	0.56	0.56	1980
Neutral	0.78	0.62	0.69	1937
Positive	0.63	0.59	0.61	2461
accuracy			0.62	8232
macro avg	0.61	0.64	0.62	8232
weighted avg	0.63	0.62	0.62	8232

Tuned Parameter C=3, gamma=0.01
Where C is Regularization strength
and gamma is Kernel Coefficient

TF-IDF for Top Model

Binary Classification for Logistic Regression

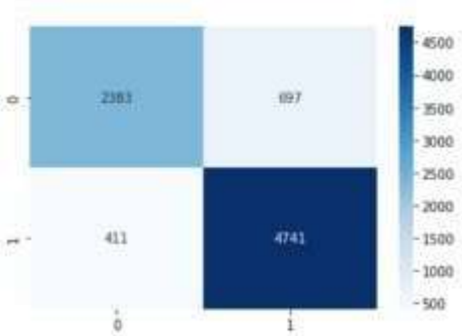
```
training accuracy Score    :  0.8865299924069856
Validation accuracy Score :  0.8454810495626822
              precision    recall  f1-score   support

           0       0.70      0.87      0.77      2474
           1       0.94      0.84      0.88      5758

    accuracy                           0.85      8232
   macro avg       0.82      0.85      0.83      8232
weighted avg       0.86      0.85      0.85      8232
```

Multi Class Classification for Support Vector Classifier

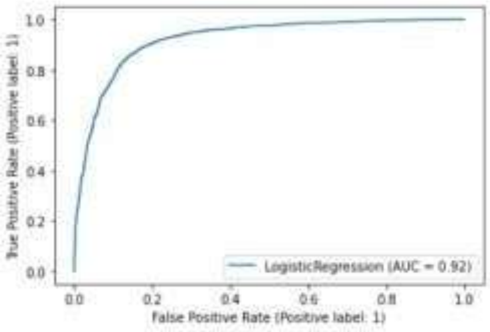
```
training accuracy Score    :  0.9624601366742597
Validation accuracy Score :  0.6058066083576288
                    precision    recall  f1-score   support

Extremely Negative       0.47      0.74      0.57       698
Extremely Positive       0.50      0.78      0.61       845
          Negative       0.61      0.54      0.58      2234
           Neutral       0.66      0.68      0.67      1497
          Positive       0.69      0.54      0.61      2958

          accuracy                           0.61      8232
         macro avg       0.59      0.65      0.61      8232
      weighted avg       0.63      0.61      0.61      8232
```

Confusion Matrix and ROC Curve for Top 2 Binary Classification Model

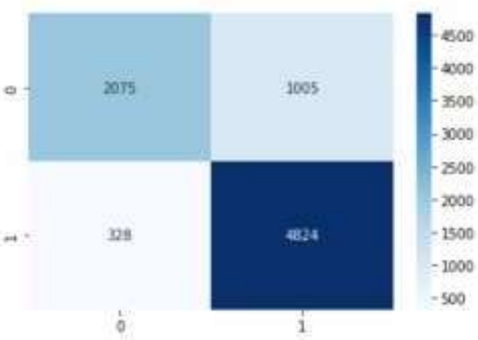
Confusion matrix for Logistic Regression



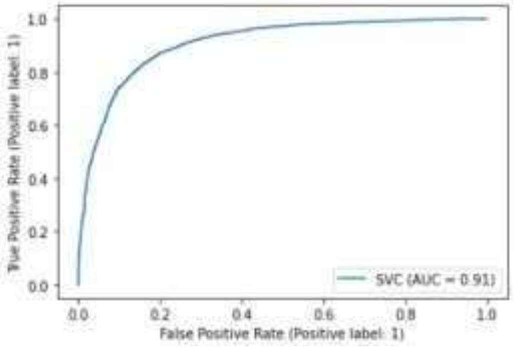
ROC Curve for Logistic Regression



Confusion matrix for Support Vector Machine



ROC Curve for Support Vector Machine



Conclusion

- Started with loading the dataset, followed by EDA which gives important insights of the data and helps in feature selection.
- After EDA, we extracted and cleaned the important features and pre-process it to a matrix of numbers so that it can be passed to ML algorithms.
- We manipulated the multiclass target variable to binary variable.
- We applied multiple ML algorithms for both multiclass and binary classification models and evaluated it with different matrices like accuracy score, precision, Recall, f1 score etc.
- Finally, we got SVC model as best multiclass classifier model with 61.1% test accuracy and logistic regression model as best binary classifier model with 86.5% test accuracy.

Challenges

- The dataset contained lots of noise or irrelevant data such as usernames, URLs etc.
- Since it is a multi class classification problem with 5 classes, model becomes more complex then binary classification.
- The number of observations of all the five classes are not balanced due to which the accuracy of multi class classification is baised towards the majority class.
- After manipulating the multi class target variable to a binary class variable the accuracy is increased but information about the various class is lost.