Capstone Project-3

Sentiment Analysis : Predicting sentiment of COVID-19 tweets

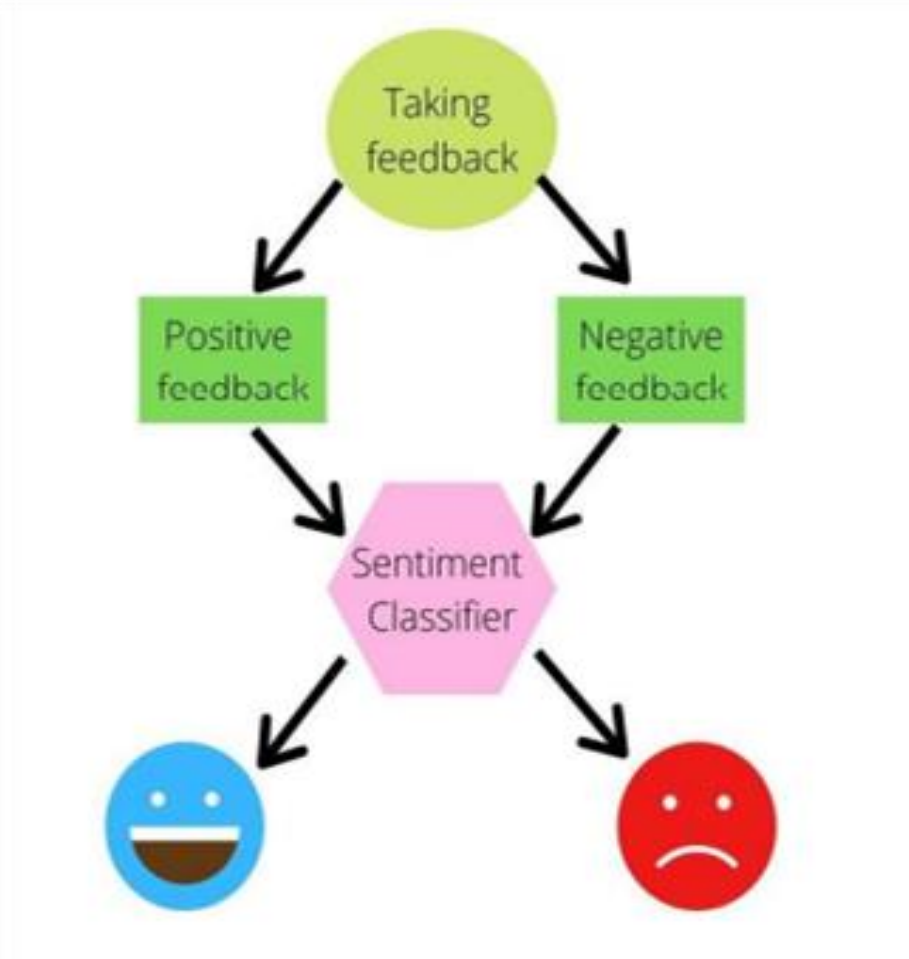
By – Sachin Yallapurkar

Points for Discussion

- Problem Statement
- Introduction & Data Summary
- Exploratory Data Analysis
- Text Preprocessing
- Feature Engineering
- Vectorization
- Model Training
- Evaluation
- Challenges
- Conclusion

Problem Statement

- The challenge is 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.This is a supervised ml classification problem.



Introduction

- Sentiment Analysis is the process of computationally identifying and categorizing opinions expressed in a piece of text, especially in order to determine whether the writer's attitude towards a particular topic is positive, negative, or neutral.
- COVID-19 originally known as Coronavirus Disease of 2019, has been declared as a pandemic by World Health Organization (WHO) on 11th March 2020.
- The study analyzes various types of tweets gathered during the pandemic times hence can be useful in policy making to safeguard the countries by demystifying the pertinent facts and information.

Data Summary

- We were provided with Coronavirus_Tweets.csv dataset.
- Shape of the dataset is (41157,6).
- The **Sentiment column** is the dependent variable which consists of 5 different labels which are positive, negative, neutral, extremely positive and extremely negative.

We are given the following columns in our data:

1. **Location** - name of the location from which the tweet was shared.
2. **TweetAt** - time of the tweet at which the tweet was shared.
3. **OriginalTweet** - the original tweet shared by the user.
4. **UserName** - identification given by twitter to the user.
5. **ScreenName** - name projected on the screen of the user.
6. **Sentiment** - defined sentiment or label from the shared tweet.

EXPLORATORY DATA ANALYSIS

LOOKING FOR MISSING VALUES & UNIQUE VALUES

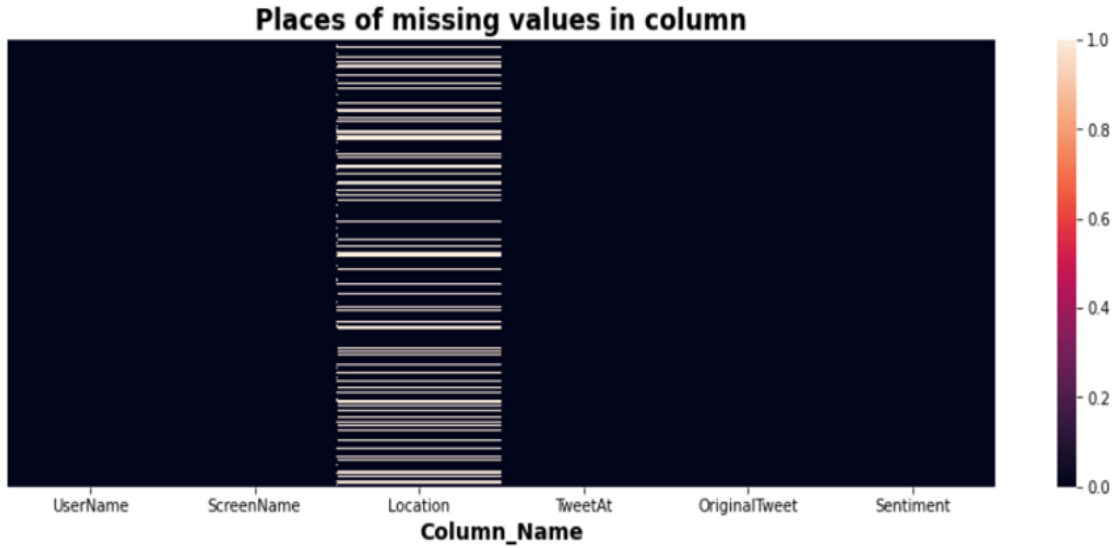
- On looking at the dataset we can see that only ‘Location’ column shows the missing value.
- So, to get rid-off from the missing values, we just left the feature as usual, as don’t know the exact location from where the tweets has been done.
- Also the dataset contain 5 - unique values in ‘Sentiment’, and 30- unique dates of tweeting (indicating all tweets has been tweeted within the period of

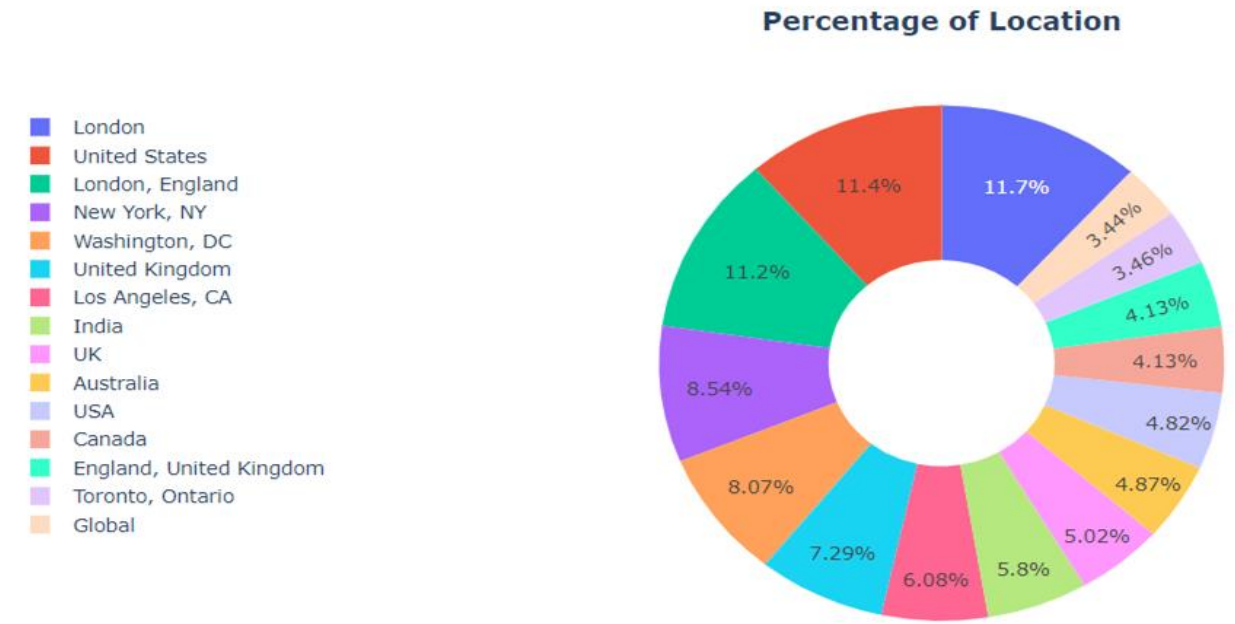
```
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   UserName       41157 non-null  int64
 1   ScreenName     41157 non-null  int64
 2   Location       32567 non-null  object
 3   TweetAt        41157 non-null  object
 4   OriginalTweet  41157 non-null  object
 5   Sentiment      41157 non-null  object
```

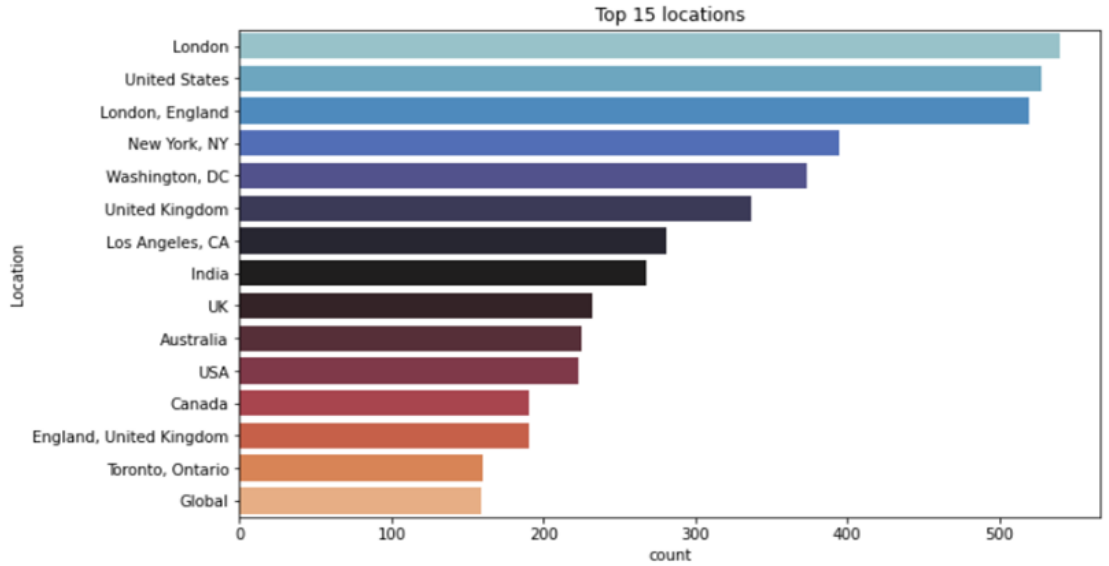
```
Total Unique Values in UserName - 41157
Total Unique Values in ScreenName - 41157
Total Unique Values in Location - 12221
Total Unique Values in TweetAt - 30
Total Unique Values in OriginalTweet - 41157
Total Unique Values in Sentiment - 5
```

LOOKING INTO: ‘Location’ COLUMN

- There are **20.87%** (8567) of missing values or null values of various places in location column.
- Most of the tweets has been tweeted from **London** (11.7% among top 15 locations).

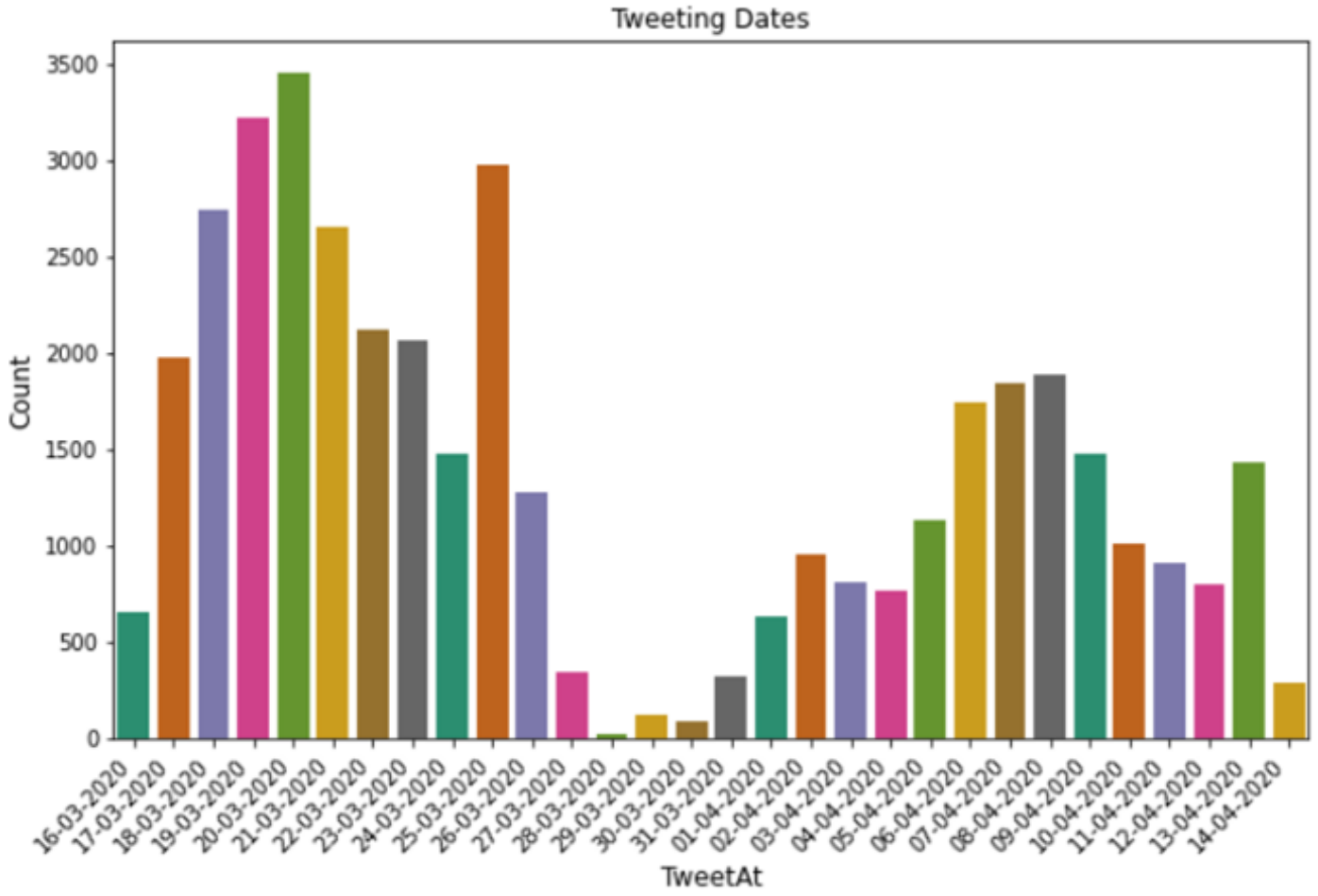






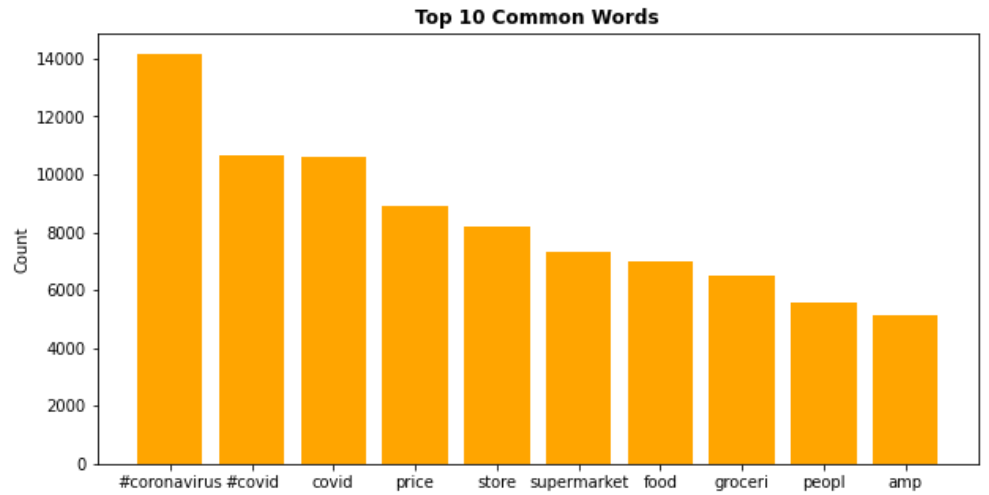
LOOKING INTO: ‘TweetAt’ COLUMN

- Here we have created countplot for the looking into the different tweeting dates given in the dataset.
- The tweeting date ranges from **16-03-2020** to **14-04-2020**, which clearly corresponds to one month (30 days).
- **20-03-2020** shows the maximum tweeting date.



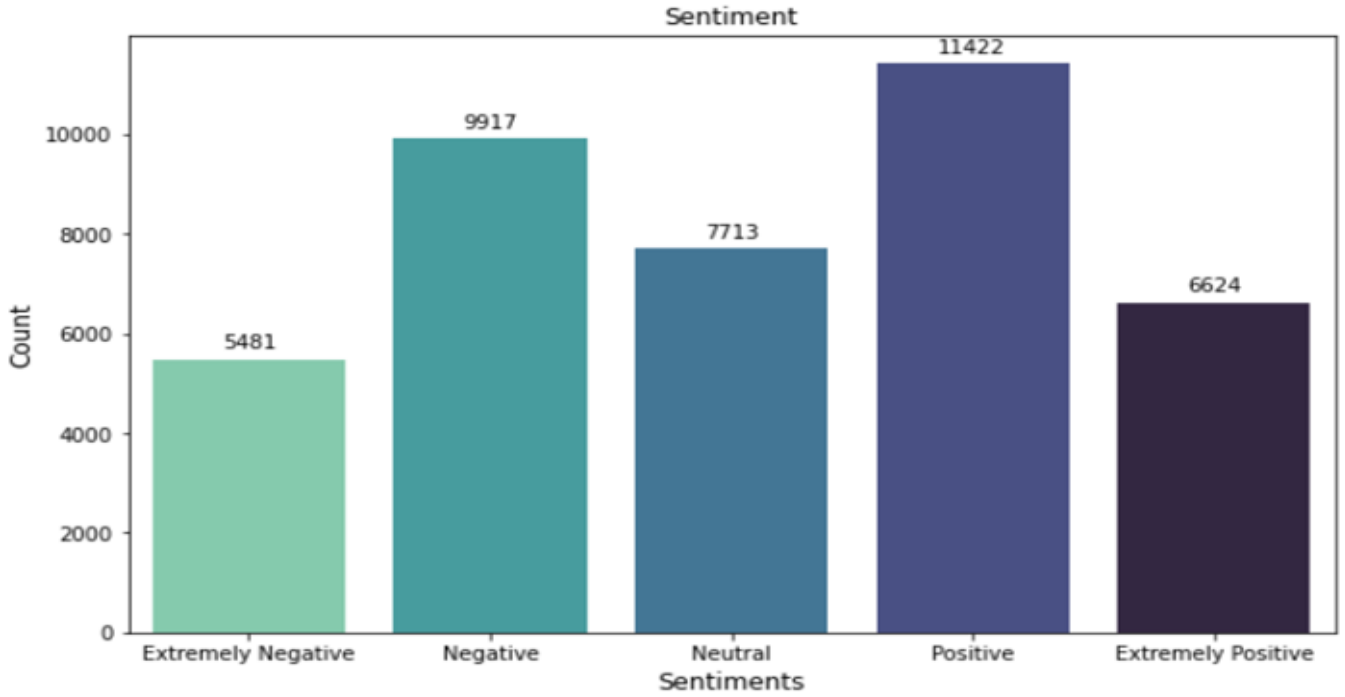
LOOKING INTO: ‘Original Tweet’ COLUMN

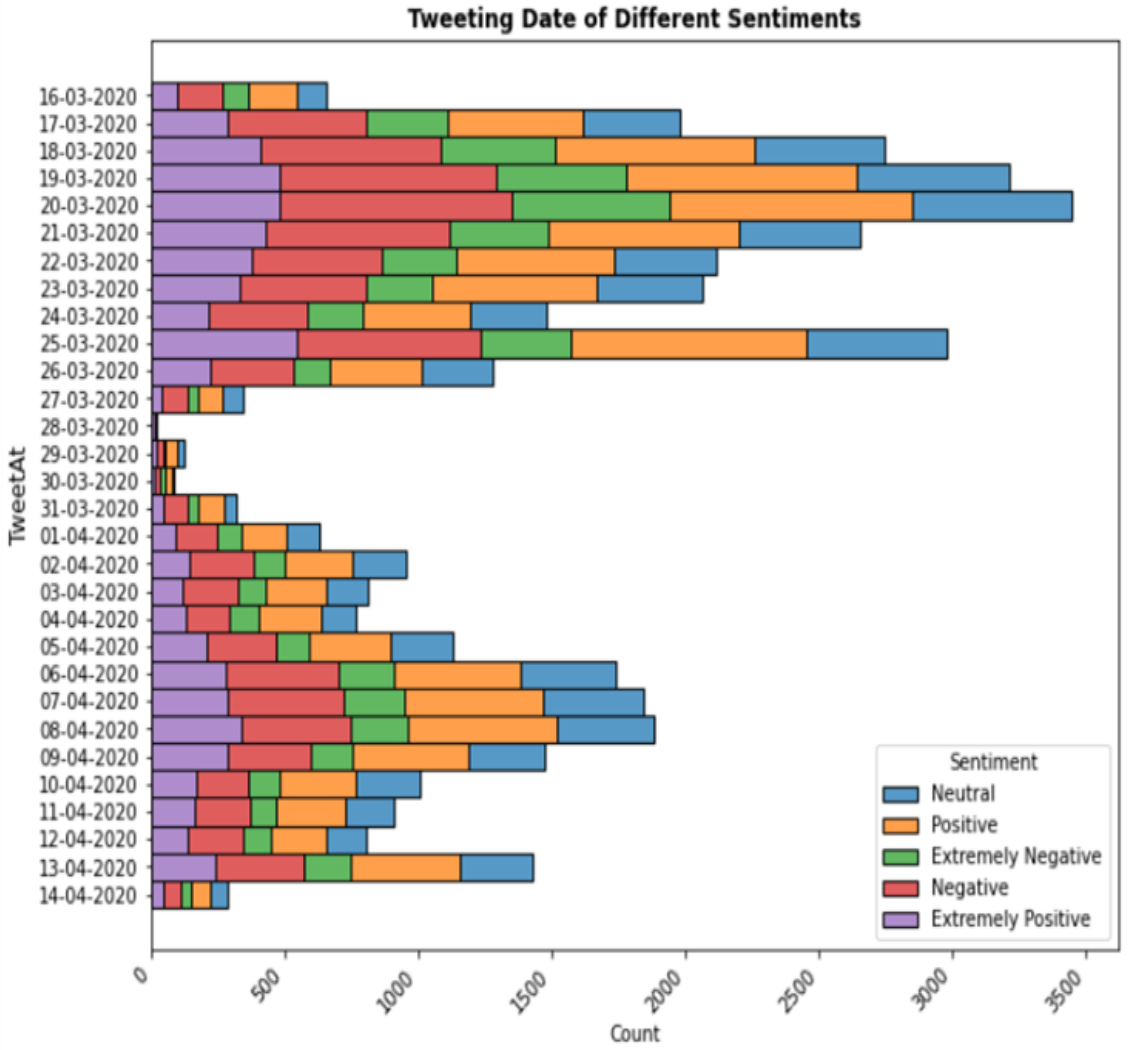
- There are some words like coronavirus, covid has the maximum frequency in the dataset from word cloud.
- realDonaldTrump, YouTube, Boris Johnson, etc were the top 10 trending mentions present in the tweets of the Original Tweet column.



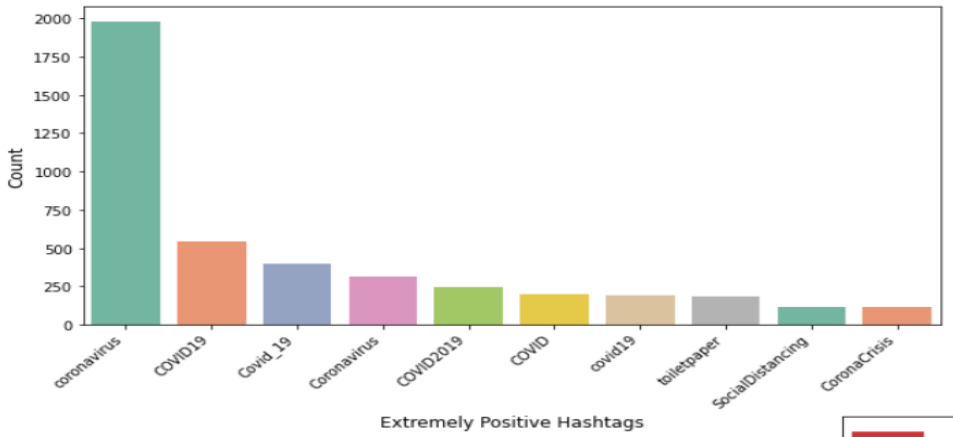
LOOKING INTO: ‘Sentiment’ COLUMN

- The count plot of sentiments shows the Positive sentiment has the highest value while extremely negative has the lowest among all 5 sentiments values.
- In relation to tweeting date, i.e. ‘TweetAt’ feature, the date of 20-03-2020 shows the highest number of tweets as well as the highest number of negative, extremely negative sentiment tweets.

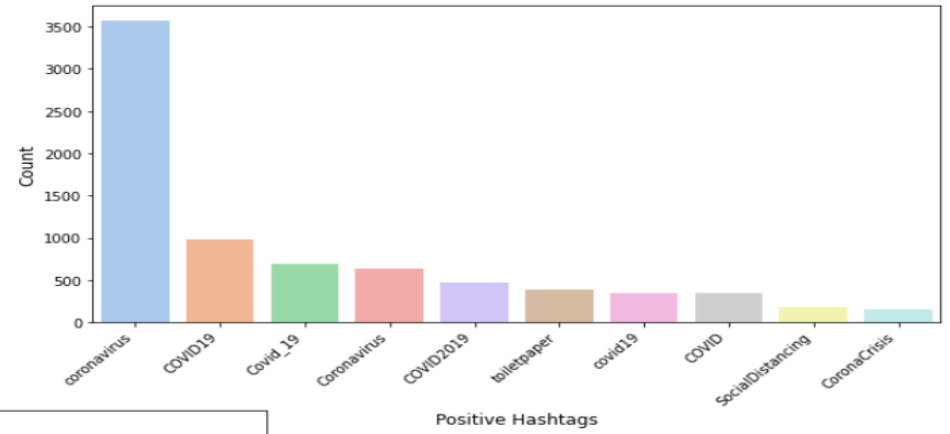




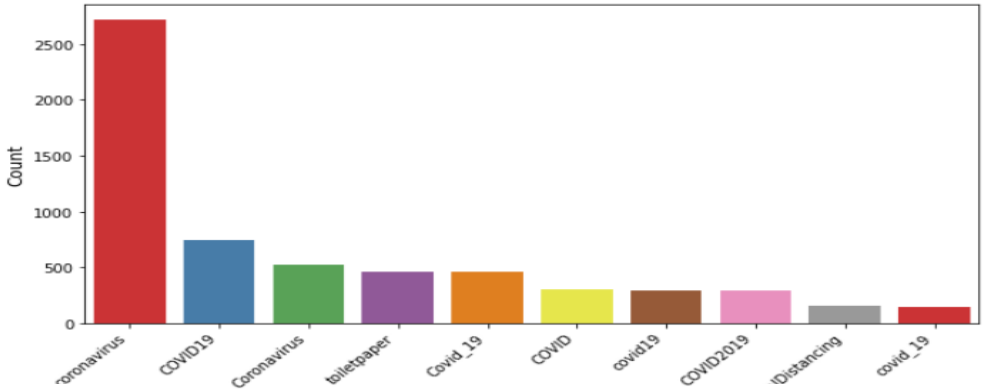
IMPACT OF HASHTAGS '#' ON TWEETS SENTIMENT



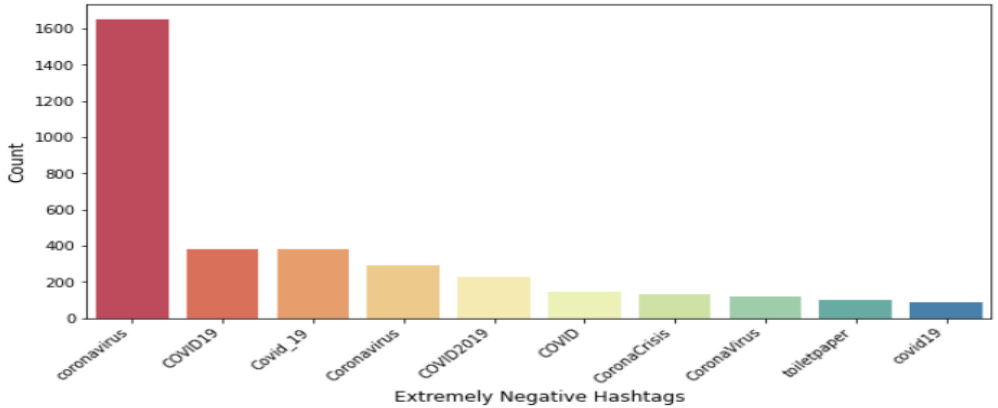
Extremely Positive Hashtags



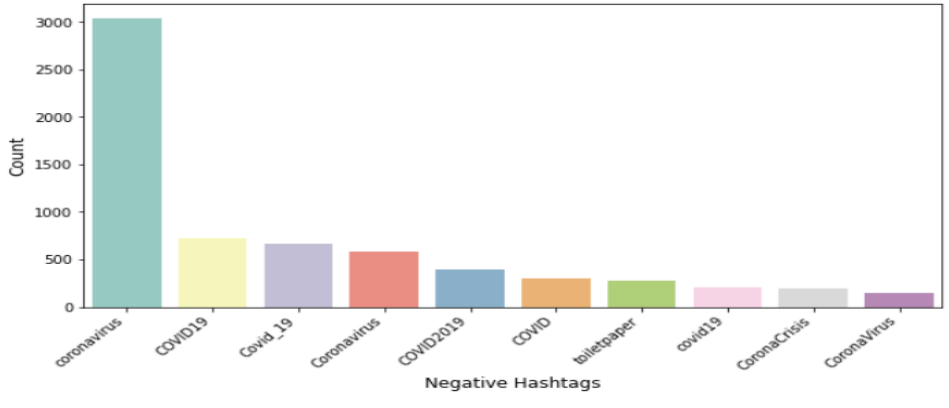
Positive Hashtags



Neutral Hashtags

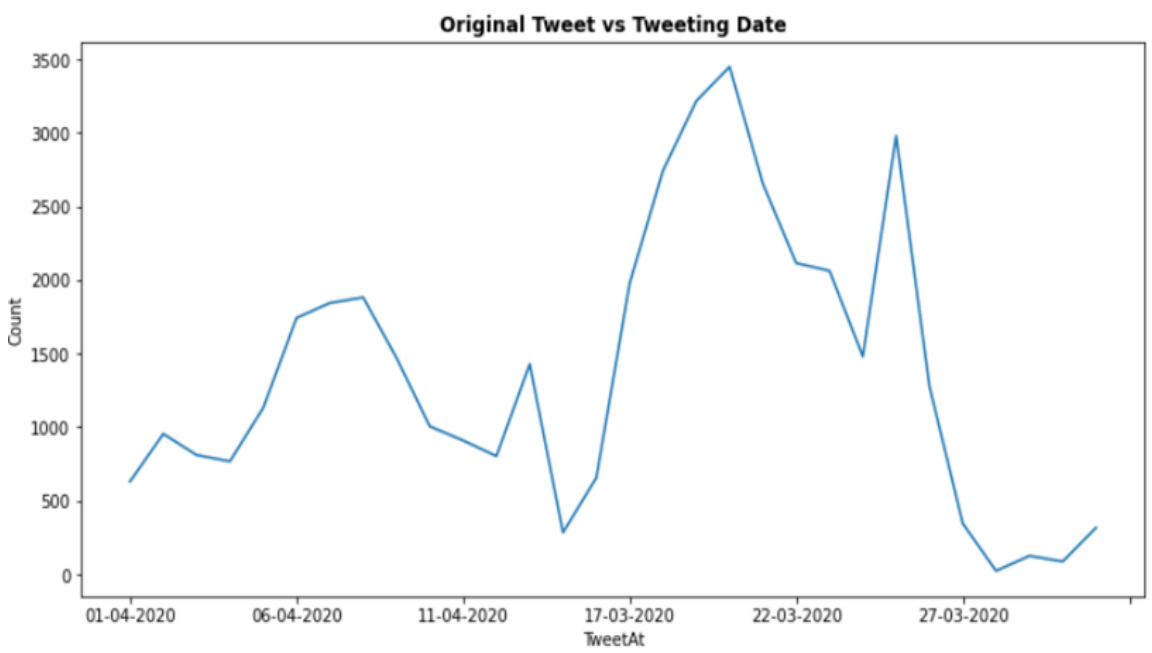


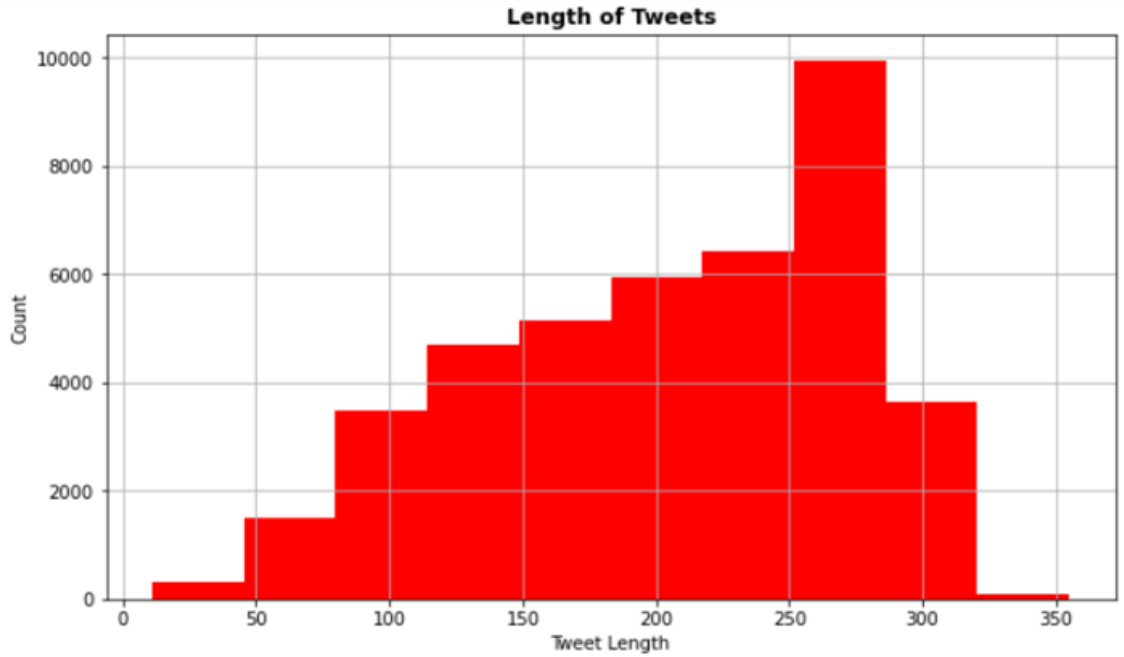
Extremely Negative Hashtags



Negative Hashtags

LOOKING INTO: ‘OriginalTweet’ COLUMN





- When line plot is plotted between original tweet and tweeting date columns, it can been seen that the maximum number of tweets were between the dates of 17-03-2020 & 22-03-2020.
- As well as a sudden drop can also been observed after 27-03-2020.
- In another plot, where tweet string length has been calculated. It shows the range of text length running between 10 to above 350 characters, with having maximum length between 250 to 300.

TEXT PREPROCESSING

- The text preprocessing of the text data is an essential step which makes raw text ready for mining.
- The objective of this step is to clean noise which are less relevant to find the sentiment of tweets such as user handle, punctuations, special characters, numbers and terms which doesn't carry much weightage in the context of the text.
- The cleaning for the tweets is done over **'OriginalTweet'** column.

Removing Tweets Handle (@user), url, http

- As we find lots of twitter handles (@user), different url with http links which are completely unnecessary for our further analysis.
- Hence, we need to clean all these and proceed to next step of text preprocessing.

	OriginalTweet	Sentiment	Clean_tweet
0	@MeNyrbie @Phil_Gahan @Chrisitv https://t.co/i...	Neutral	https://t.co/iFz9FAn2Pa and https://t.co/xX...
1	advice Talk to your neighbours family to excha...	Positive	advice Talk to your neighbours family to excha...
2	Coronavirus Australia: Woolworths to give elde...	Positive	Coronavirus Australia: Woolworths to give elde...
3	My food stock is not the only one which is emp...	Positive	My food stock is not the only one which is emp...
4	Me, ready to go at supermarket during the #COV...	Extremely Negative	Me, ready to go at supermarket during the #COV...

	OriginalTweet	Sentiment	Clean_tweet
0	@MeNyrbie @Phil_Gahan @Chrisitv https://t.co/i...	Neutral	
1	advice Talk to your neighbours family to excha...	Positive	advice Talk to your neighbours family to excha...
2	Coronavirus Australia: Woolworths to give elde...	Positive	Coronavirus Australia: Woolworths to give elde...
3	My food stock is not the only one which is emp...	Positive	My food stock is not the only one which is emp...
4	Me, ready to go at supermarket during the #COV...	Extremely Negative	Me, ready to go at supermarket during the #COV...

Removing Tweets Punctuations, Numbers & Special Characters

- In the next step we can see, the tweets also contain different punctuations, numbers and special characters, which are again cleaned.

OriginalTweet	Sentiment	Clean_tweet
@MeNyrbie @Phil_Gahan @Chrisitv https://t.co/i...	Neutral	
advice Talk to your neighbours family to excha...	Positive	advice Talk to your neighbours family to excha...
Coronavirus Australia: Woolworths to give elde...	Positive	Coronavirus Australia Woolworths to give elder...
My food stock is not the only one which is emp...	Positive	My food stock is not the only one which is emp...
Me, ready to go at supermarket during the #COV...	Extremely Negative	Me ready to go at supermarket during the #COVI...

Removing Tweets Stopwords

- In the next step we can see, the tweets contain different meaningless words which doesn't give much importance to sentence, which are again removed.

OriginalTweet	Sentiment	Clean_tweet
@MeNyrbie @Phil_Gahan @Chrisitv https://t.co/i...	Neutral	
advice Talk to your neighbours family to excha...	Positive	advice talk neighbours family exchange phone n...
Coronavirus Australia: Woolworths to give elde...	Positive	coronavirus australia woolworths give elderly ...
My food stock is not the only one which is emp...	Positive	food stock one empty please panic enough food ...
Me, ready to go at supermarket during the #COV...	Extremely Negative	ready go supermarket #covid outbreak paranoid ...

TOKENIZATION

- In tokenization, we convert the group of sentences into tokens. It is also called text segmentation or lexial analysis. It basically split the data into small chunk of words.
- Here tokenization has been performed via python NLTK library of tokenize().

OriginalTweet	Sentiment	Clean_tweet
@MeNyrbie @Phil_Gahan @Chrisitv https://t.co/i...	Neutral	
advice Talk to your neighbours family to excha...	Positive	advice talk your neighbours family exchange ph...
Coronavirus Australia: Woolworths to give elde...	Positive	coronavirus australia woolworths give elderly ...
My food stock is not the only one which is emp...	Positive	food stock not the only one which empty please...
Me, ready to go at supermarket during the #COV...	Extremely Negative	ready supermarket during the #covid outbreak n...

STEMMING

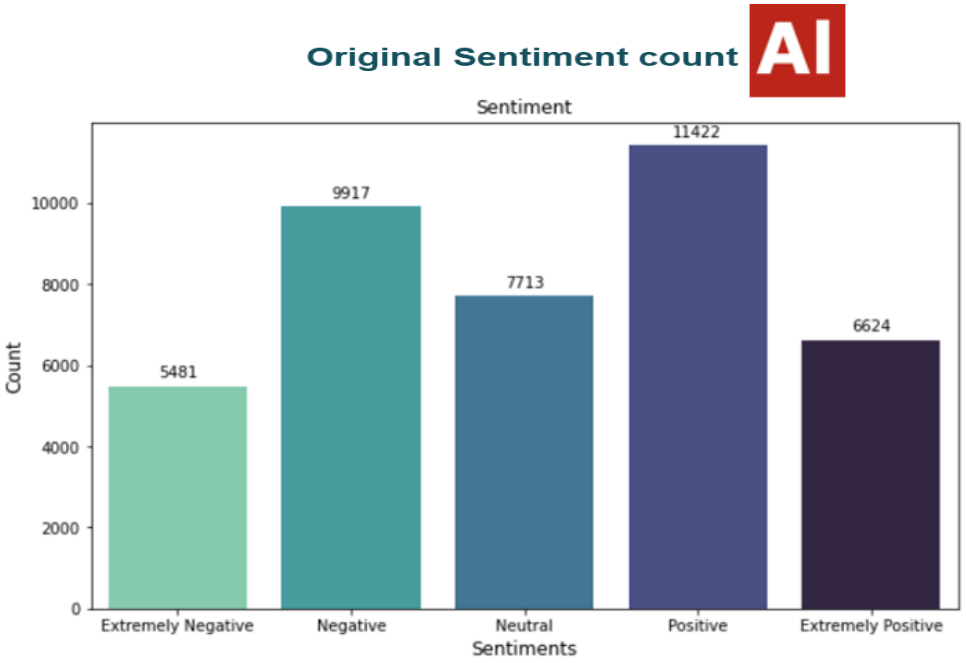
- Stemming is the rule based process for stripping the suffixes (“ing”, “ly”, etc.) from a sentence.
- It can be imported in python via NLTK library and here we have used **SnowballStemmer** and stemming is one of the important step of NLP.

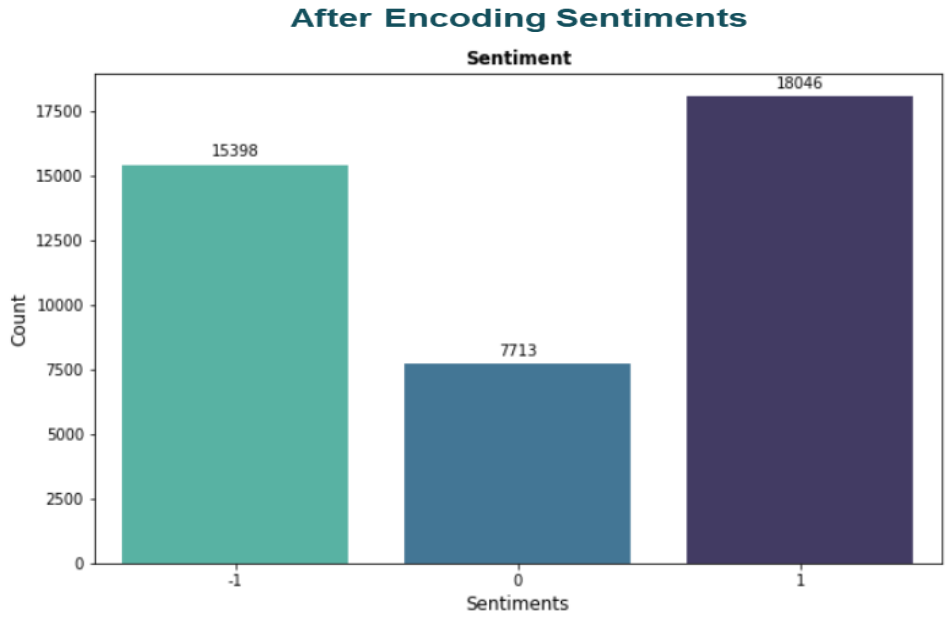
OriginalTweet	Sentiment	Clean_tweet
@MeNyrbie @Phil_Gahan @Chrisitv https://t.co/i...	Neutral	
advice Talk to your neighbours family to excha...	Positive	advic talk neighbour famili exchang phone numb...
Coronavirus Australia: Woolworths to give elde...	Positive	coronavirus australia woolworth give elder dis...
My food stock is not the only one which is emp...	Positive	food stock one empti pleas panic enough food e...
Me, ready to go at supermarket during the #COV...	Extremely Negative	readi go supermarket #covid outbreak paranoid ...

FEATURE ENGINEERING

Encoding The Sentiment

- There are 5 different sentiments are present in dataset. neutral , positive, extremely positive, negative & extremely negative.
- To make problem simpler we encoded the sentiments into 3 different values where we merged Extremely positive and Positive into one and Extremely Negative and Negative into another.
- Now we have Sentiments are – Positive: '**1**', Neutral: '**0**', Negative: '**-1**'





VECTORIZATION

- In the next step, we choose **CountVectorizer()** as our vectorizer.
- It will create a sparse matrix of all the words and the number of times they are present in the document.
- By default, Countvectorizer converts the text to lowercase and uses word-level tokenization.
- Here we have used CountVectorizer with <`decode_error = 'replace',stop_words = stop`>
- CountVectorizer() can be imported in python via sklearn library.

MODEL TRAINING

Models used for Classification:

- Multinomial Naive Bayes
- Logistic Regression
- Support Vector Machine Classifier
- Random Forest Classifier
- Stochastic Gradient Descent Classifier
- CatBoost Classifier

Why Multinomial Naive Bayes ?

- Good accuracy for classification if the feature independence condition holds.
- Space and time effective.
- Can handle high dimensional data pretty well.
- A good baseline model.

Why Logistic Regression ?

- Unlike Naive Bayes it makes no assumption about the feature independence.
- Logistic Regression with L1 regularization is well known for feature reduction.
- Fast to train.

Why Support Vector Machine ?

- It is well known to handle high dimensional data.
- It allows misclassification as well with soft margins.

Why Random Forest Classifier?

- Random Forest takes random samples and features to make train the model.
- Time taking, but Decision tree like model with less chance to overfit.

Why Stochastic Gradient Descent?

- It is neural network based.
- It converges comparatively faster for large datasets.
- It fits one sample at a time.
- Computationally Fast.

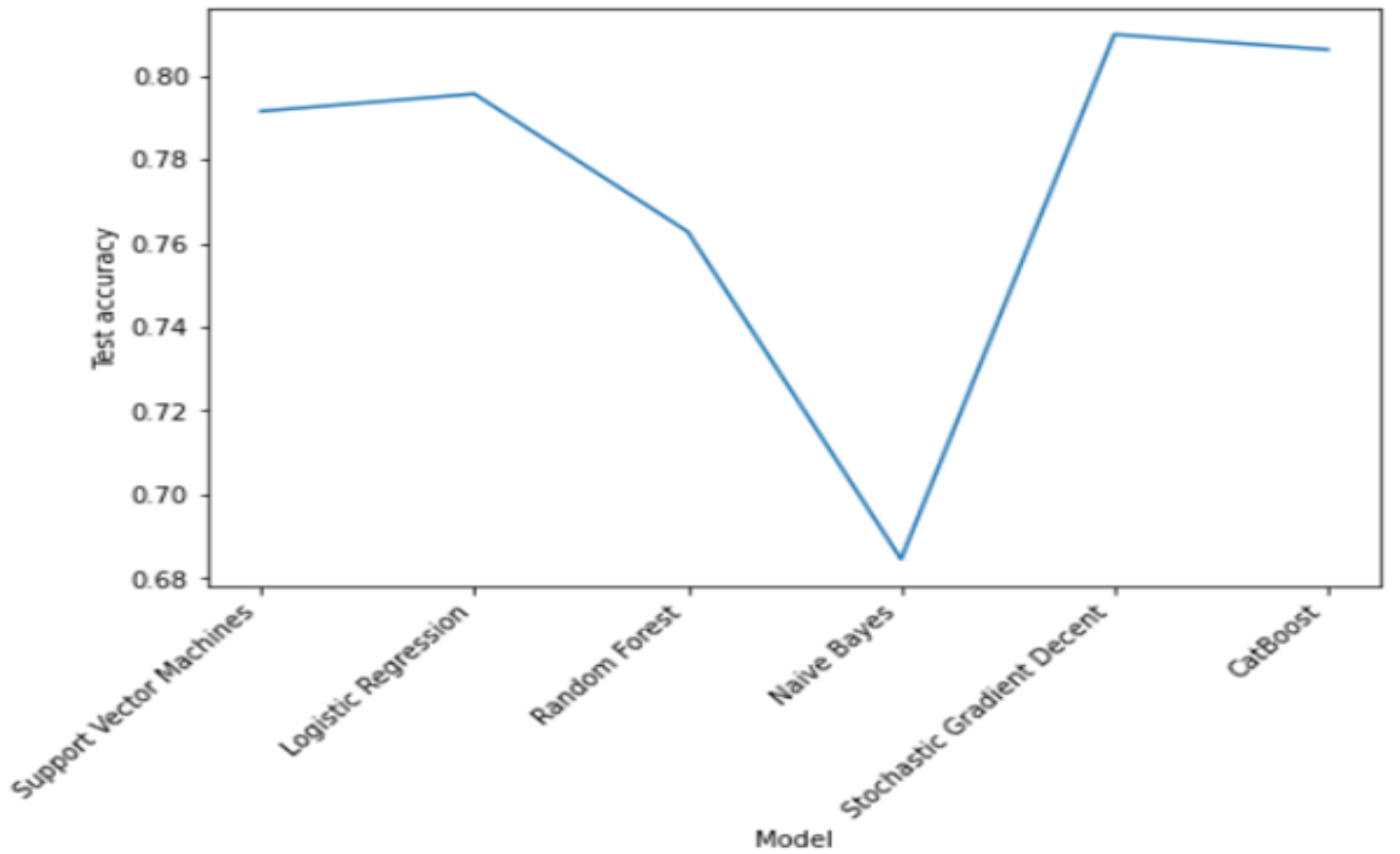
Why CatBoost Classifier?

- CatBoost is a high-performance open source library for gradient boosting on decision trees.
- Uses symmetric trees, which result in a Fast Inference.
- It is good in handling sophisticated categorical features.

MODEL EVALUATION

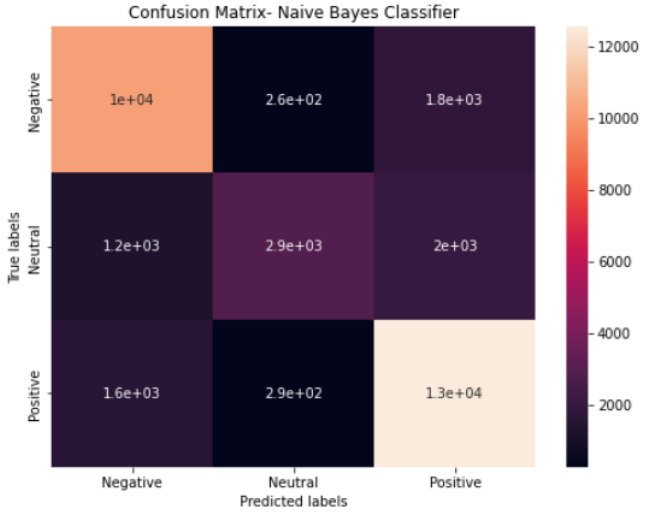
Model	Test accuracy	Recall	Precision	F1-Score
Support Vector Machines	0.791545	0.791545	0.791029	0.791169
Logistic Regression	0.795675	0.795675	0.795269	0.795460
Random Forest	0.762755	0.762755	0.764281	0.762980
Naive Bayes	0.684524	0.684524	0.747144	0.705568
Stochastic Gradient Decent	0.809888	0.809888	0.809558	0.809685
CatBoost	0.806244	0.806244	0.805694	0.805287

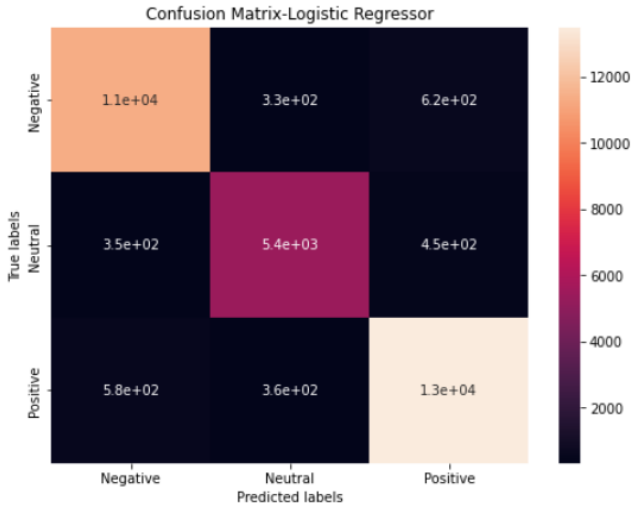
Model Test Accuracy

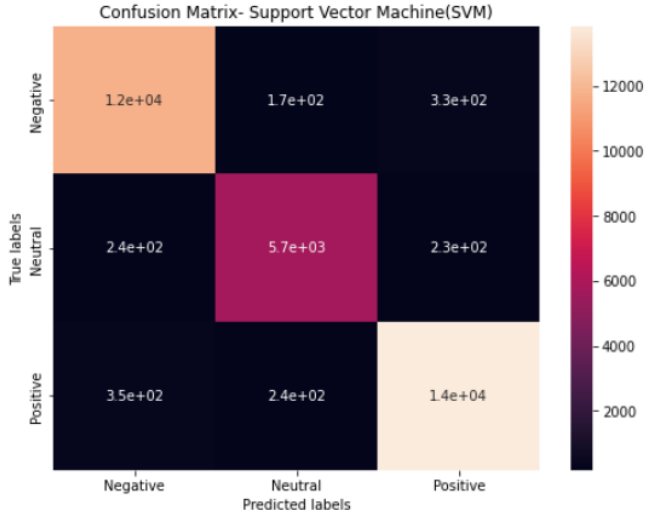


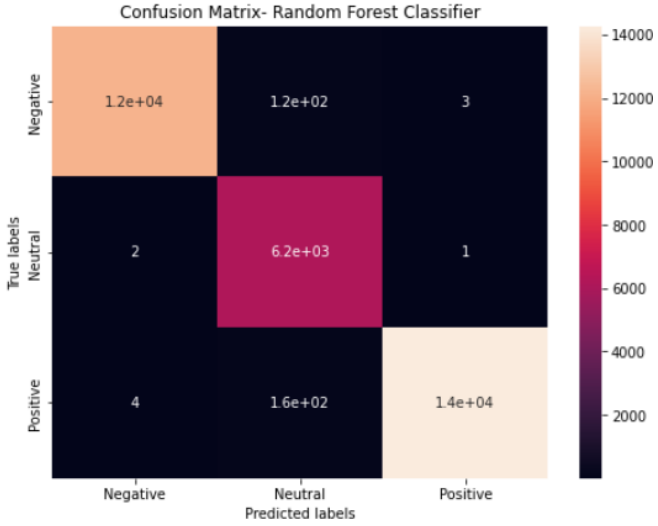
- Overall the best test accuracy is for **SGD Classifier** with accuracy of **80.98%**.

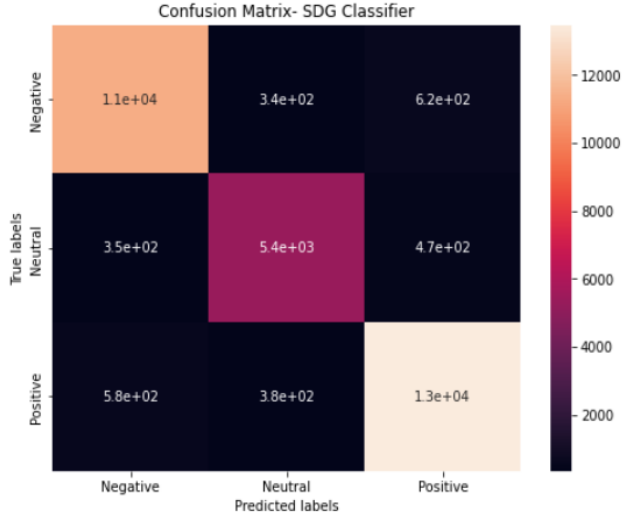
Confusion Matrix of all the models in order...

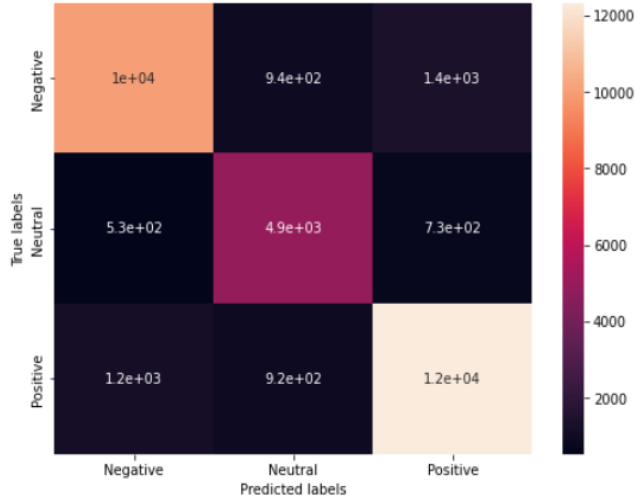












CHALLENGES

- Locations being too many/unformatted/irrelevant
- Sarcastic tweets
- Computation time/crashes
- Selection of best model
- Cleaning of tweets

CONCLUIOSN

- ‘Location’ column contains approx. **20.87%** of Null values.
- The columns such as “UserName” and “ScreenName” does not give any meaningful insights for our analysis.
- There are **five types** of **sentiments**- Extremely Negative, Negative, Neutral, Positive and Extremely Positive. So, we merged Extremely Positive with positive and Extremely Negative with Negative. And use encoding with value **‘-1’** for **negative**, **‘0’** for **neutral** and **‘1’** for **positive**.
- All tweets data collected between months of **March** and **April 2020** and of around 30 days.
- Most of the tweets came from **London** followed by U.S.
- Among top 10 mentions in tweets real Donald Trump was the top mentioned name and "#coronavirus" was the most trendiest hashtag.
- After evaluation of different model, best test accuracy of **80.98%** delivered by **Stochastic Gradient Descent Classifier (SGD)** for multi-class classification.


thank
you