Capstone Project

On

Coronavirus Tweet Sentiment Analysis

By

Roshan Tile

AlmaBetter

CONTENT

Following is the Standard Operating Procedure to tackle the Sentiment Analysis kind of project. We will be going through this procedure to predict what we supposed to predict.

- **Problem Statement**
- **Data Summary**
- **Exploratory Data Analysis (EDA)**
- **Text Pre-processing**
- **Classification Analysis**
- **Models Performance Metrics**
- **Conclusion**

Problem Statement

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. The given 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.

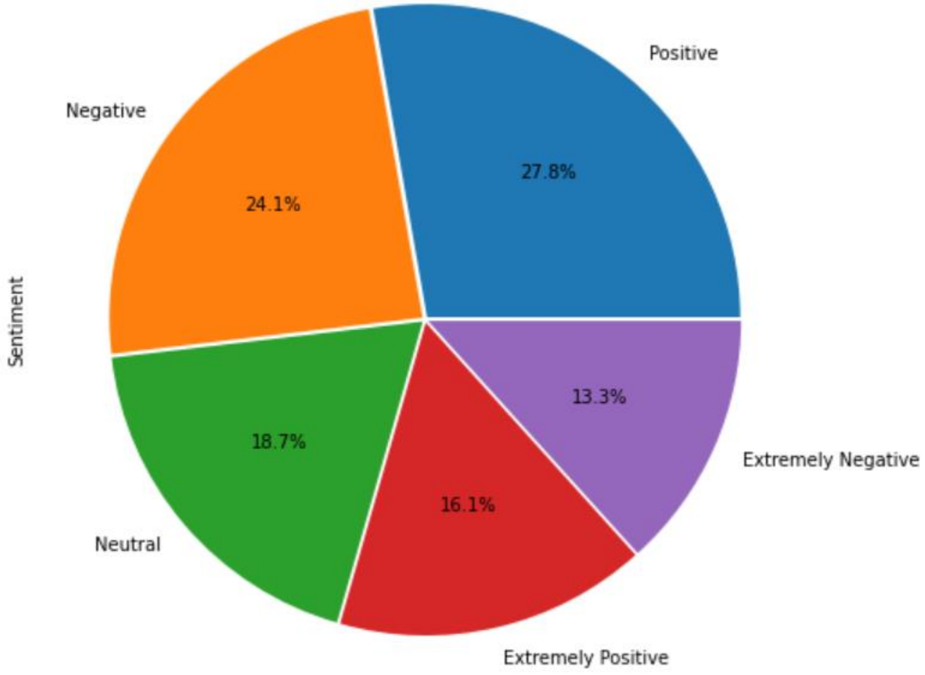
Data Summary

- The original dataset has 6 columns and 41157 rows. In order to analyze various sentiments, from this 6 feature 2 features are unusable so will ignore them
    1. Location = location (country) from where tweet is posted
    2. Tweet At = Date on which tweet is posted
    3. Original Tweet = Context of tweet
    4. Label = Type of sentiments

- We require just two columns named Original Tweet and Sentiment. There are five types of sentiments- Extremely Negative, Negative, Neutral, Positive, and Extremely Positive.

Exploratory Data Analysis (EDA)

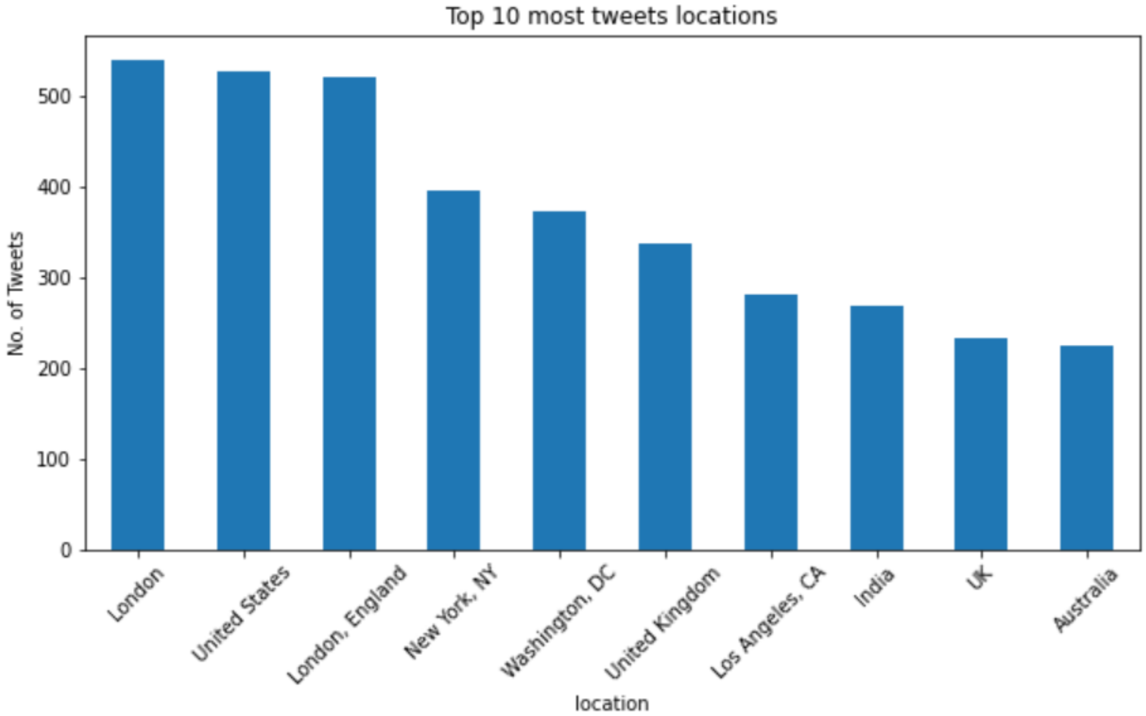
Percentage wise sentiments

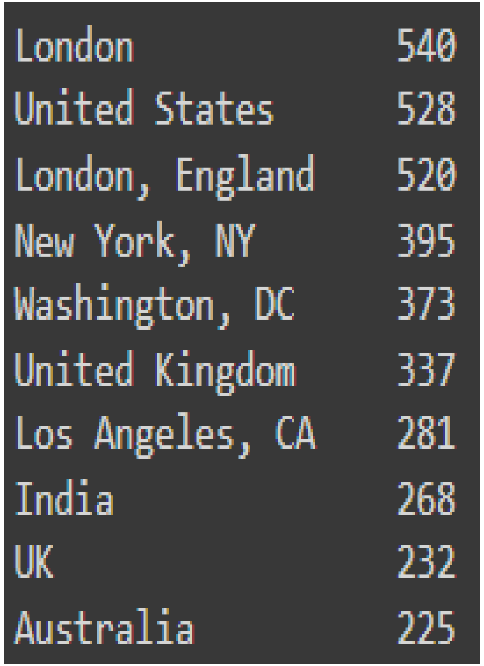


When we try to explore the 'Sentiment' pie chart, we came to know that:

- ✓ Most of the peoples about 44% are having positive sentiments about various issues shows us their optimism during pandemic times.
- ✓ Very few people about 37.4% are having negatives thoughts about Covid-19.
- ✓ While 18.7% people have neutral opinion.

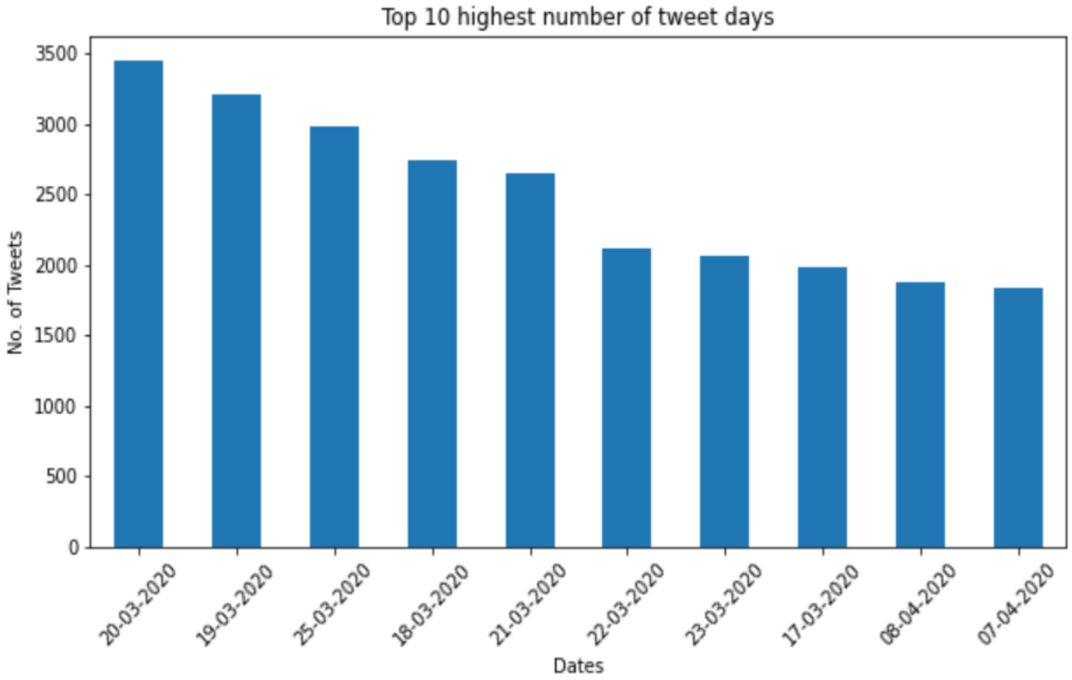
Top 10 most tweet’s locations

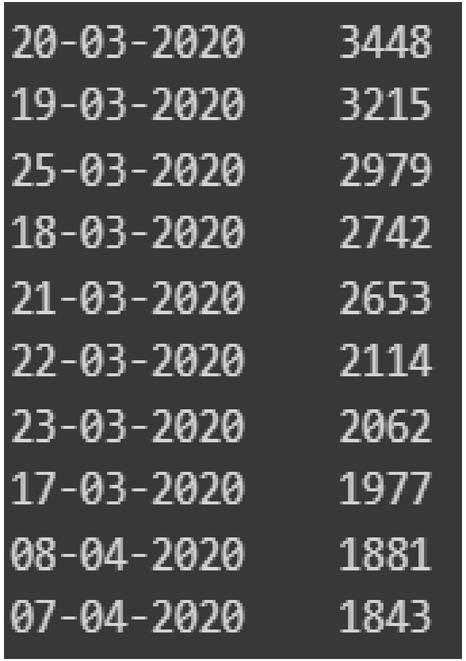




```
London             540
United States      528
London, England    520
New York, NY       395
Washington, DC     373
United Kingdom     337
Los Angeles, CA    281
India              268
UK                 232
Australia          225
```

Top 10 highest number of tweet days





```
20-03-2020    3448
19-03-2020    3215
25-03-2020    2979
18-03-2020    2742
21-03-2020    2653
22-03-2020    2114
23-03-2020    2062
17-03-2020    1977
08-04-2020    1881
07-04-2020    1843
```

Text Pre-Processing

Text pre-processing of the text data is an essential step as it makes the raw text ready for mining and making it suitable for a machine learning model. The objective of this step is to clean noise those are less relevant to find the sentiment of tweets such as :

- ✓ Url links (HTTPS: / HTTP:)
- ✓ Username/tweeter handle (@Xyz)
- ✓ Punctuation (.,?,” etc.),
- ✓ Special characters (@,%,&,$, etc.),
- ✓ Numbers (1,2,3, etc.)

Other Essential Steps are:

- ✓ Stop words
- ✓ Positive Negative Word Count
- ✓ Stemming
- ✓ Tokenization
- ✓ Encode the Sentiments

Vectorization

- vectorization is a methodology in NLP to map words or phrases from vocabulary to a corresponding vector of real numbers which used to find word predictions, word similarities/semantic. i.e., Process of converting text into numerical representation.

- Techniques:
  - ✓ One hot encoding
  - ✓ Bag Of Words
  - ✓ Ngrams
  - ✓ TFIDF
  - ✓ Word2Vec
  - ✓ CountVectorizer

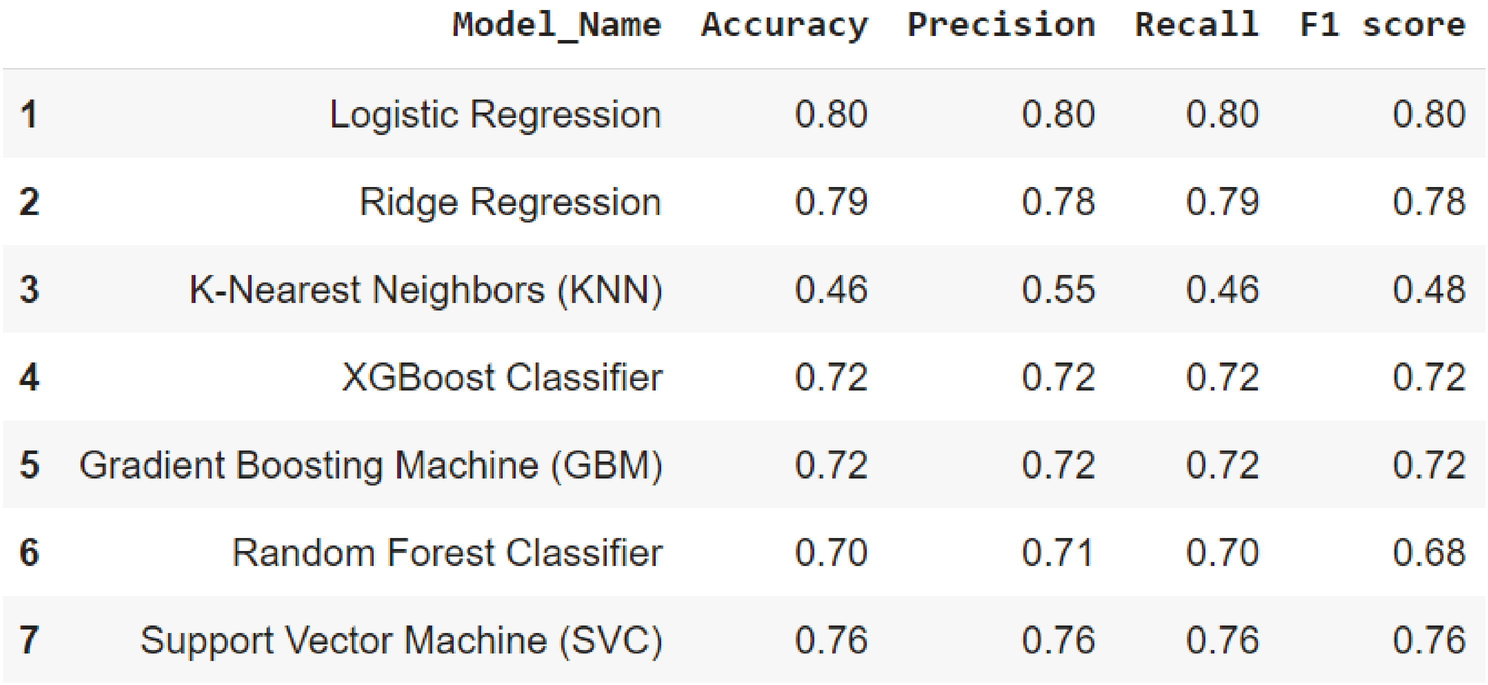
Classification Analysis

Building Classification Models

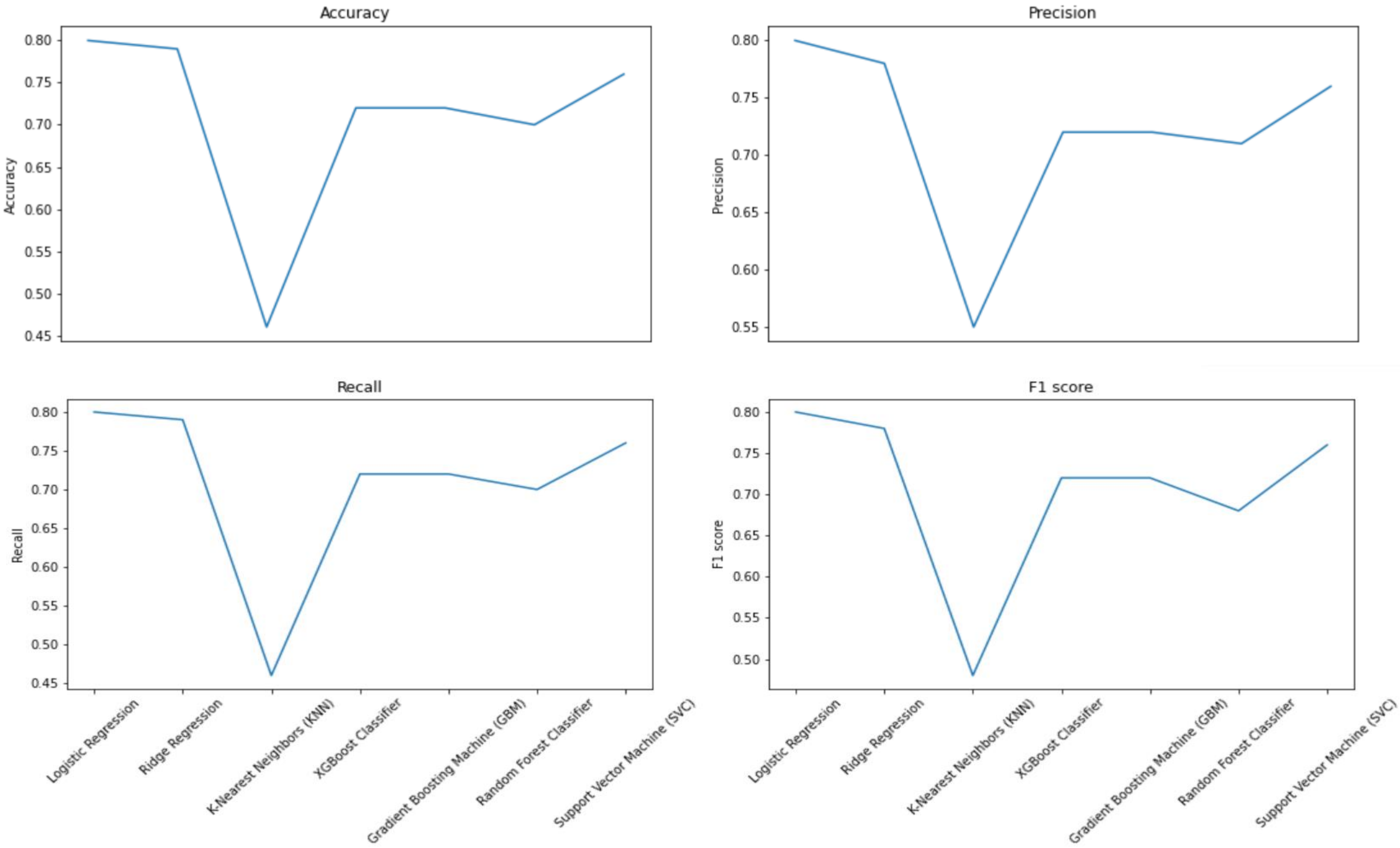
The given problem is Ordinal Multiclass classification. We had five types of sentiments and we converted them into three type, We have trained our models on different classification models are:

1. **Logistic Regression**
2. **Ridge Classifier**
3. **K-Nearest Neighbors (KNN)**
4. **XGBoost Classifier**
5. **Gradient Boosting Classifier (GBM)**
6. **Random Forest Classifier**
7. **Support Vector Machine (SVC)**

Models Performance Metrics



	Model_Name	Accuracy	Precision	Recall	F1 score
1	Logistic Regression	0.80	0.80	0.80	0.80
2	Ridge Regression	0.79	0.78	0.79	0.78
3	K-Nearest Neighbors (KNN)	0.46	0.55	0.46	0.48
4	XGBoost Classifier	0.72	0.72	0.72	0.72
5	Gradient Boosting Machine (GBM)	0.72	0.72	0.72	0.72
6	Random Forest Classifier	0.70	0.71	0.70	0.68
7	Support Vector Machine (SVC)	0.76	0.76	0.76	0.76


Accuracy

Precision

Recall

F1 score

Conclusion

- ✓ K-Nearest Neighbors (KNN) doesn't work well with a large dataset and with a high number of dimensions. It didn't classify the sentiments efficiently and gives worse results than all the other implemented models.

- ✓ The Gradient Boosting classifier (GBM) and XGBoost classifier gave almost identical results of 0.72 F1-score.

- ✓ Gradient Boosting classifier (GBM) , XGBoost classifier and Random Forrest take a lot of time to run.

- ✓ Logistic regression gives the highest result of about 0.80 F1-score of all the implemented models , Followed by the Ridge Regression (0.79 F1-score) and Support Vector Machine(SVC) (0.76 F1-score) .

- ✓ While selecting a model, it should need to have good explainability and less complexibility. As per the result, We have all three models with higher accuracy and less error. Therefore, we will select Logistic Regression.

Thank You