Capstone Project-3

Coronavirus Tweet Sentiment Analysis

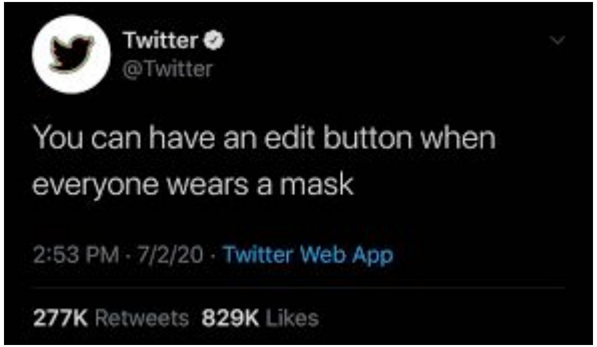
By

Harshavardhan M. Shete

AI

Twitter and Tweet Sentiment Analysis

- Twitter is microblogging platform
- Opinion platform used throughout the world
- Immense database of sentiments
- What is sentiment analysis?



Twitter
@Twitter

You can have an edit button when everyone wears a mask

2:53 PM · 7/2/20 · Twitter Web App

277K Retweets 829K Likes

Image Source: Google.com

Coronavirus Tweet Sentiment Analysis

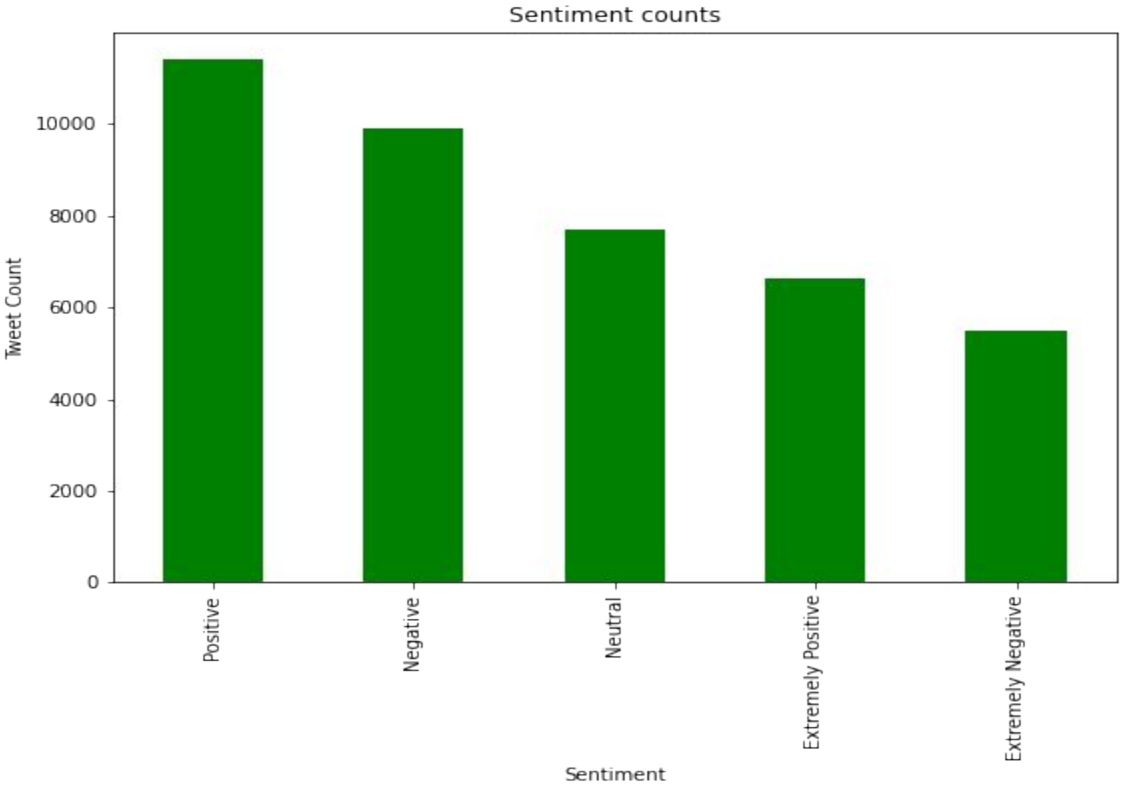
- The task is to build a classification model to predict the sentiment of Covid-19 related tweets
- The dataset contains following columns:
    - Location: The location at which the tweet was made
    - TweetAt: The date on which the tweet was made
    - OriginalTweet: This is the actual text of the 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

Solutioning

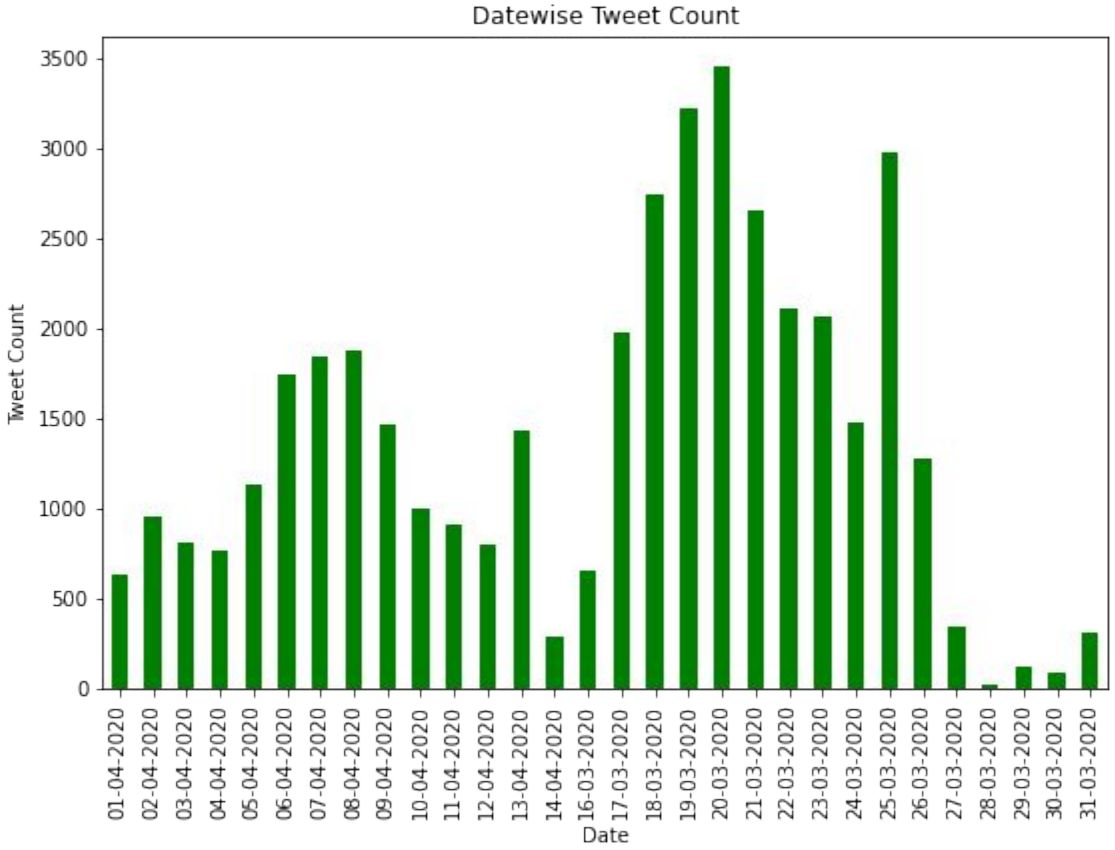
1. Exploratory Data Analysis
2. Data Pre-processing
3. Model Training and Performance Metrics

Exploratory Data Analysis

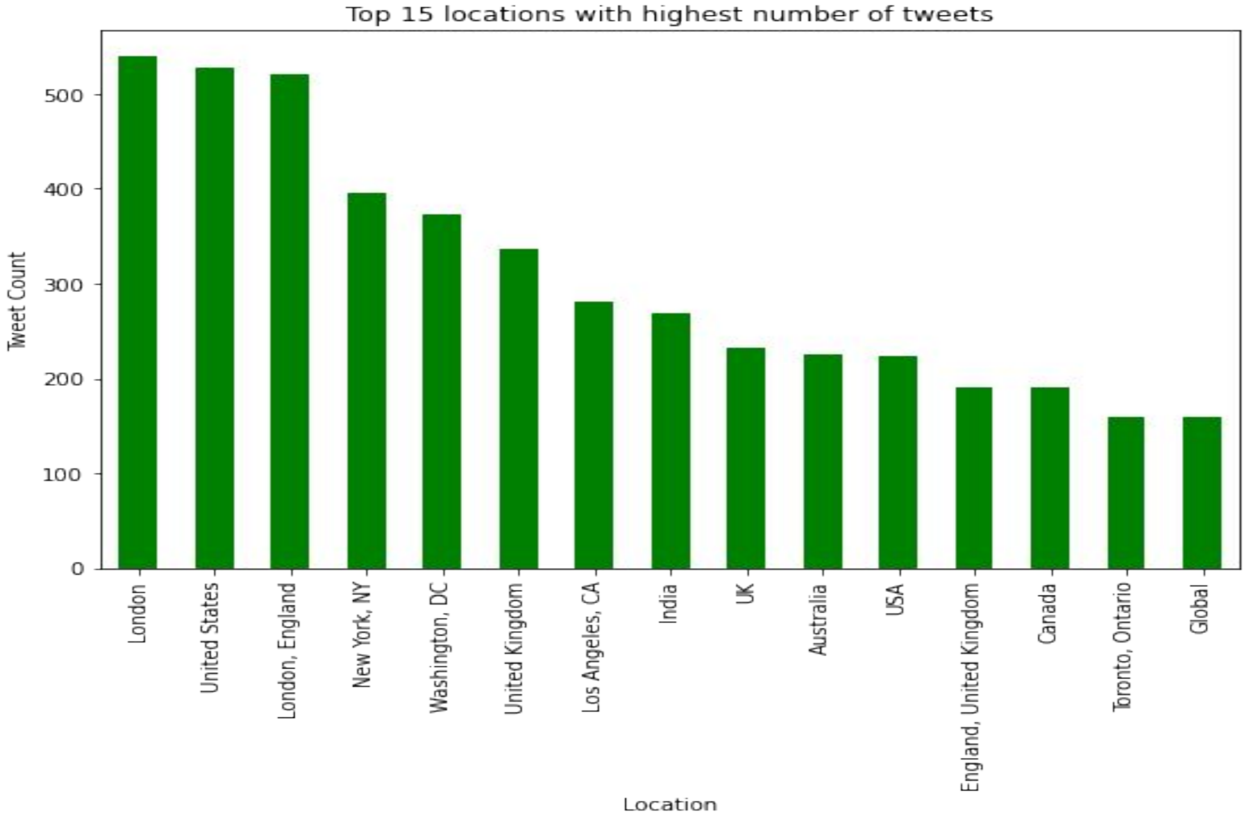
Sentiment Count



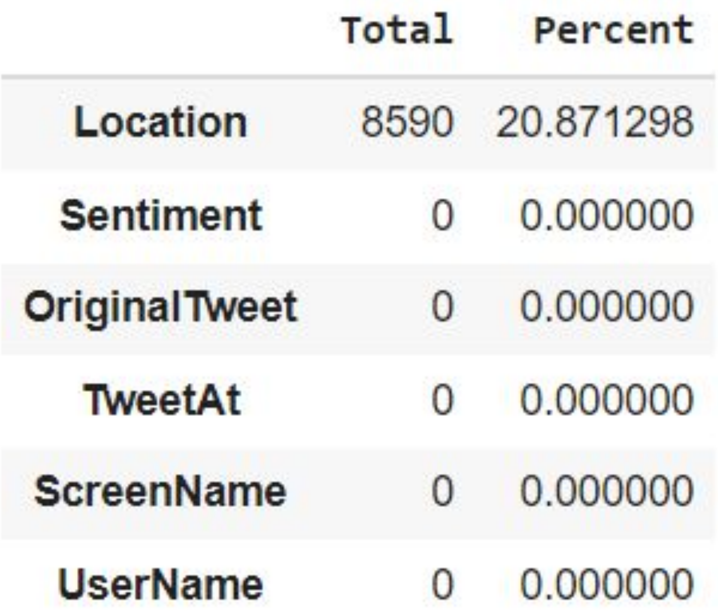
Date-wise Tweet Count



Top 15 locations with highest number of tweets



Handling Null Values



	Total	Percent
Location	8590	20.871298
Sentiment	0	0.000000
OriginalTweet	0	0.000000
TweetAt	0	0.000000
ScreenName	0	0.000000
UserName	0	0.000000

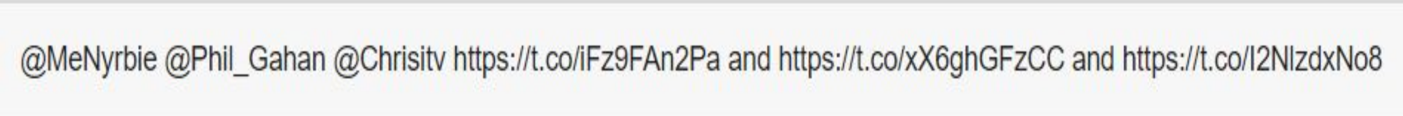
Data Pre-processing

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

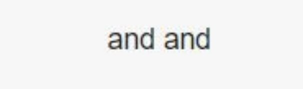
Few examples of before and after tweets

- Original Tweet



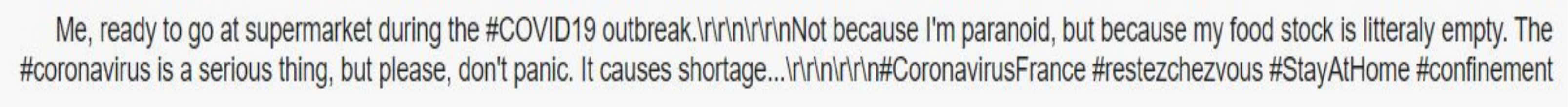
@MeNyrbie @Phil_Gahan @Chrisitv https://t.co/iFz9FAn2Pa and https://t.co/xX6ghGFzCC and https://t.co/I2NlzdxNo8

- After removing Twitter Handles and URL links



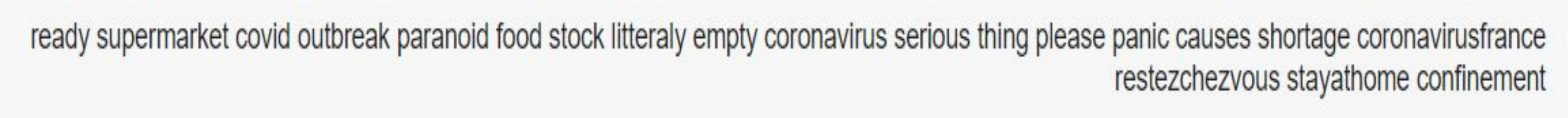
and and

- Original Tweet



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

- After removing # symbols, punctuations, special characters and stopwords



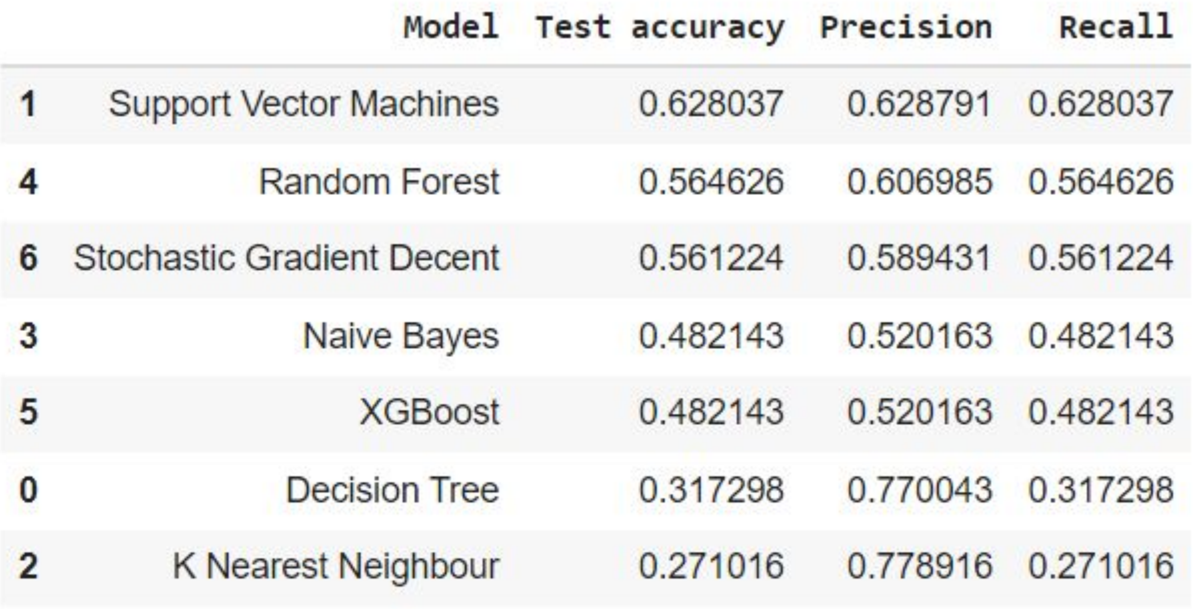
ready supermarket covid outbreak paranoid food stock litteraly empty coronavirus serious thing please panic causes shortage coronavirusfrance restezchezvous stayathome confinement

Model Training and Performance Metrics

Models used

- Decision Tree Classifier
- Support Vector Machine Classifier
- K Nearest Neighbour Classifier
- Naive Bayes classifier
- Random Forest Classifier
- XGBoost Classifier
- Stochastic Gradient Descent- SGD Classifier

Performance Comparison



	Model	Test accuracy	Precision	Recall
1	Support Vector Machines	0.628037	0.628791	0.628037
4	Random Forest	0.564626	0.606985	0.564626
6	Stochastic Gradient Decent	0.561224	0.589431	0.561224
3	Naive Bayes	0.482143	0.520163	0.482143
5	XGBoost	0.482143	0.520163	0.482143
0	Decision Tree	0.317298	0.770043	0.317298
2	K Nearest Neighbour	0.271016	0.778916	0.271016

Challenges

- Sarcastic Tweets
- Huge number of location, difficult to draw any conclusion
- Longer execution time

Conclusion

For the given problem of multiclass tweet sentiment classification the ‘Support Vector Machine’ classifier has performed the best with 62.8% accuracy and equally fair performance in terms of precision and recall.