Capstone Project – 3

Coronavirus Tweet Sentiment Analysis

by:-

Akshit Singh

Agenda

- ➔ Look at the problem statement
- ➔ Study the dataset
- ➔ Looking for null and duplicate values
- ➔ E.D.A and visualization
- ➔ Feature engineering
- ➔ Model comparison
- ➔ Conclusion

Approach

- **Understanding Business problem**
- **Exploratory Data Analysis**
  - Understanding features
  - EDA conclusion
- **Text Pre-processing**
  - Punctuations and stopwords removal
  - Stemming
- **Vectorization**
- **Modelling**
  - Train Test split
  - Fitting models to a Data
  - Hyperparameter Tuning
- **Model Performance & Evaluation**
  - Comparing model performance
  - Base models v/s Tuned models
- **Conclusion and Recommendations**

Problem Statement

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. We are given information like Location, Tweet At, Original Tweet, and Sentiment.

Data Summary

- There are 41,157 observations with various types of field in our dataset.
- List of columns:-

UserName
ScreenName
Location
TweetAt
OriginalTweet
Sentiment

Null and duplicate values

- Location column has around eight thousand five hundred null values.
- No duplicate rows are found in the dataset.

```
# checking if any null values are present in our dataset
count_of_null_values = tweet_df.isnull().sum()
count_of_null_values
```

```
UserName            0
ScreenName          0
Location         8590
TweetAt             0
OriginalTweet       0
Sentiment           0
dtype: int64
```

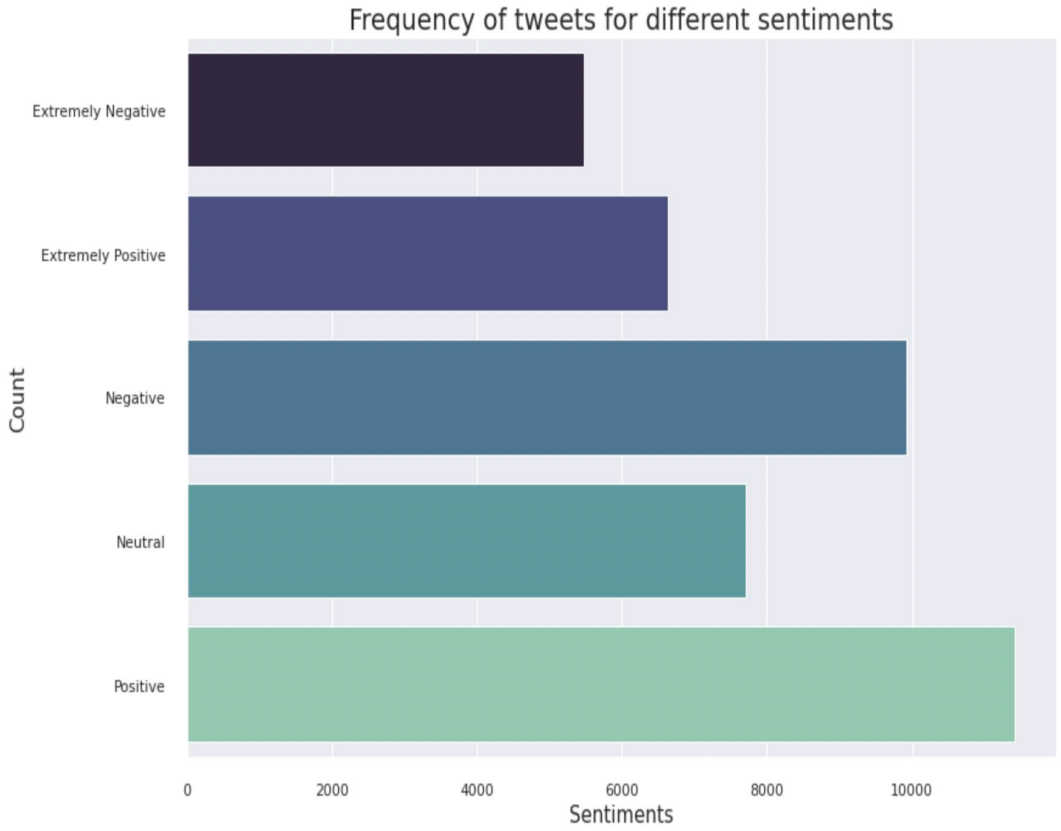
```
# checking duplicates in our dataset
value = len(tweet_df[tweet_df.duplicated()])
print("Total no. of duplicates = ", value)
```

```
Total no. of duplicates =  0
```

Exploratory Data Analysis

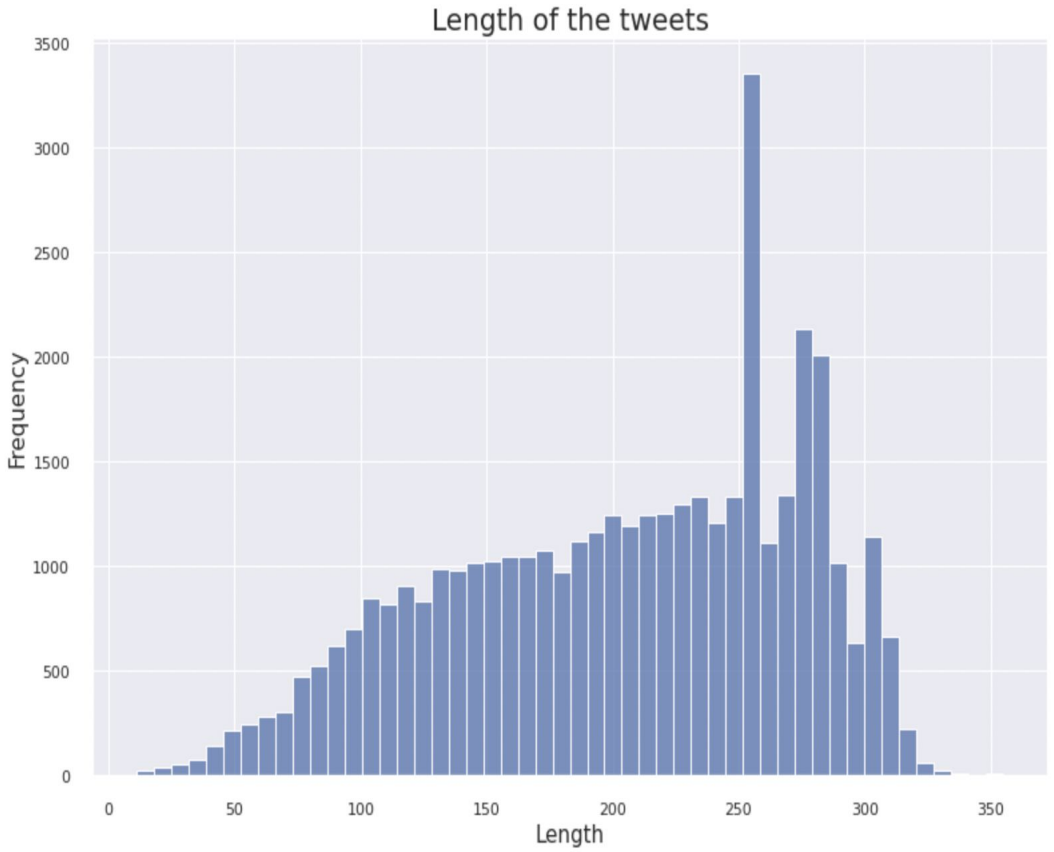
Frequency of tweets for different sentiments

- Most of the tweets are positive followed by negative tweets.
- Extremely negative tweets are the lowest.



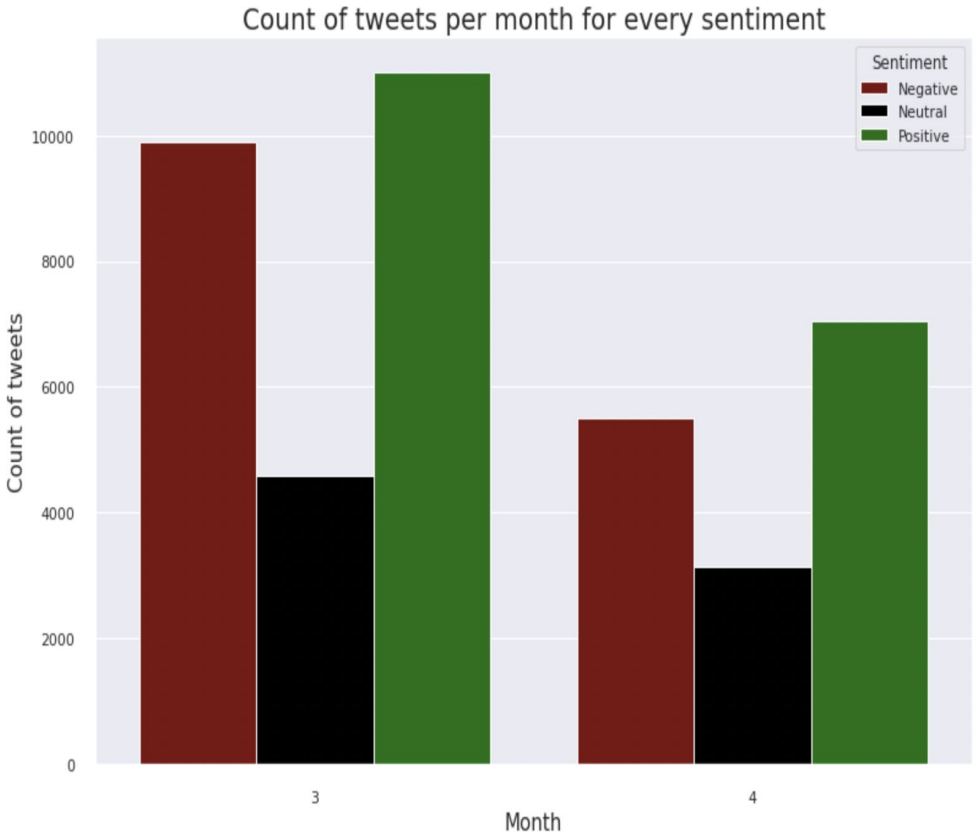
Length of the tweets

- From the histogram we deduce, the maximum number of tweets has a length of around 250.

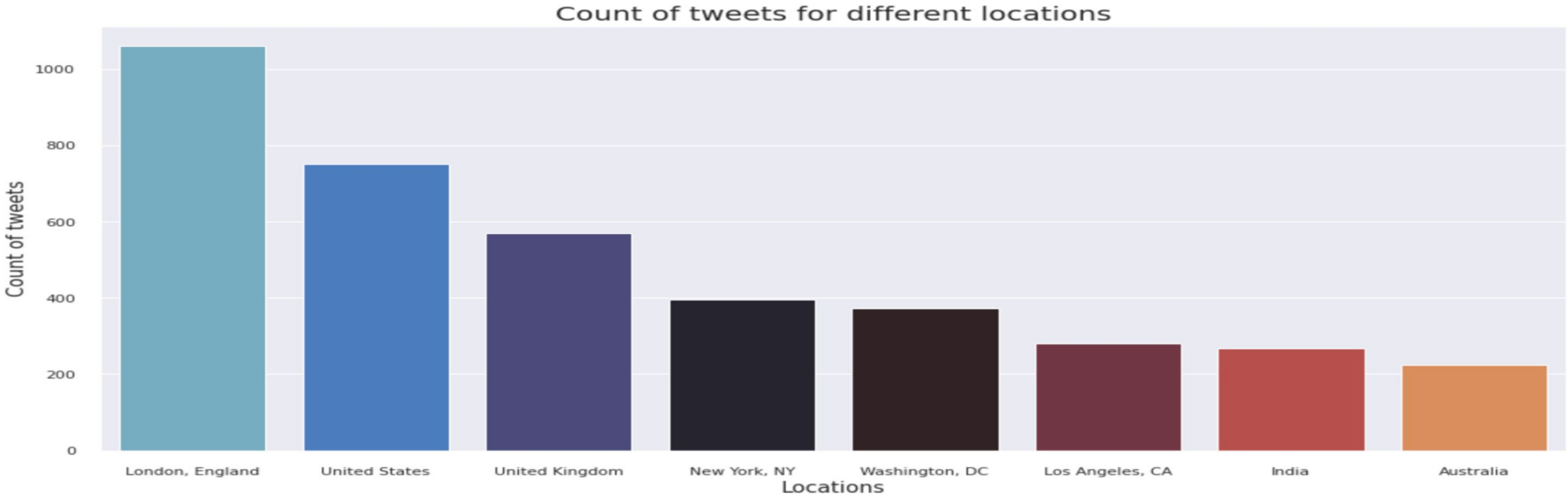


Tweets count per month

- People tweeted more in March than in April.
- During both months people tweeted positive tweets more than negative tweets although the difference is not that high.



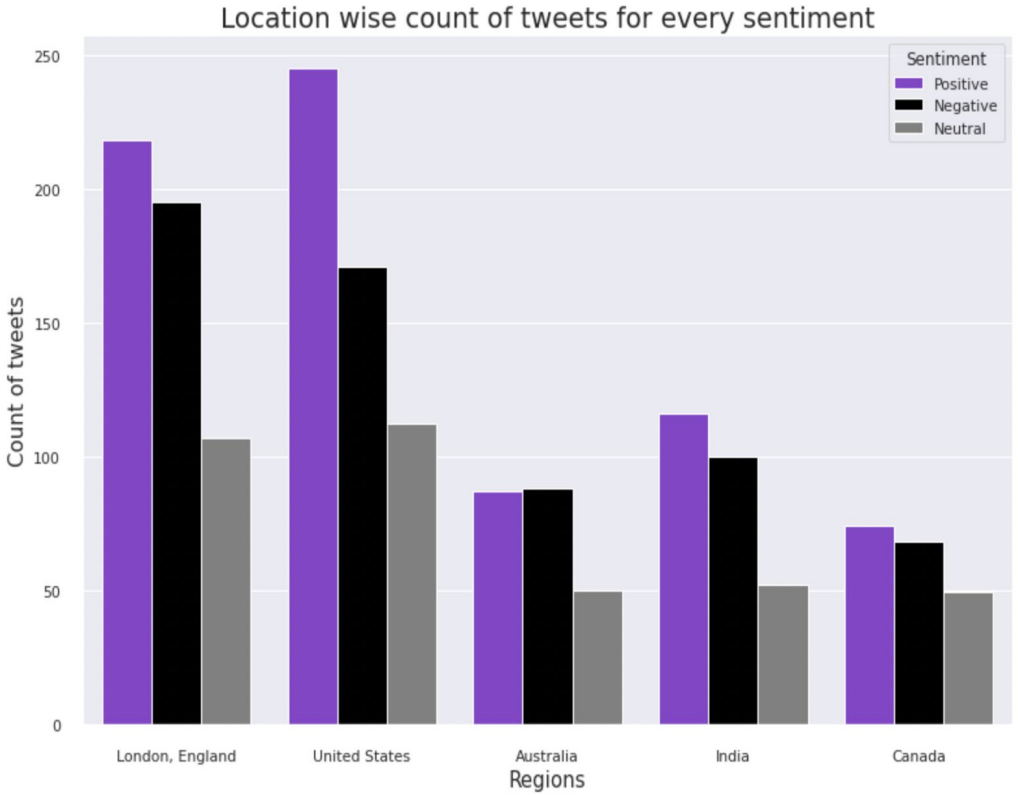
Locations with maximum tweets



- ➔ The maximum number of tweets are from London (England), followed by the United States.
- ➔ Tweets from Australia are the lowest among the above locations.

Location-wise count of tweets for every sentiment

- For all the countries except Australia, positive tweets are more in comparison to negative tweets.
- For Australia, negative and positive tweets are almost the same.
- Neutral tweets are the lowest for all the countries.

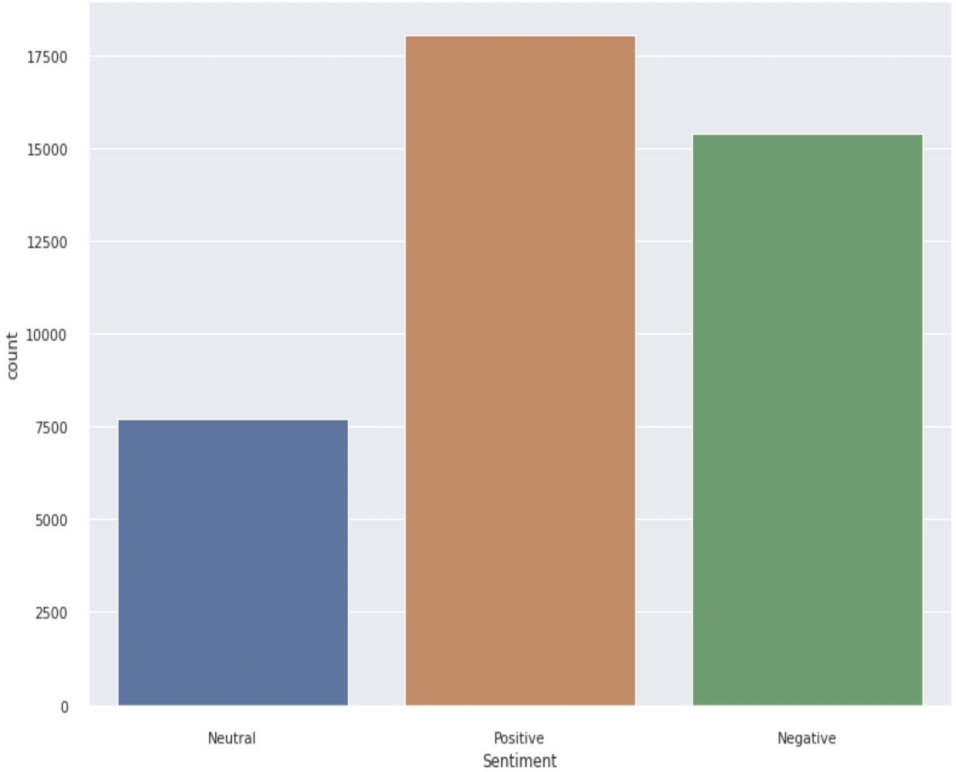


Feature Engineering

Analysis of Sentiments

- In the dataset, sentiment column contains 5 different types of sentiments which are:
  - Extremely Negative
  - Negative
  - Neutral
  - Positive
  - Extremely Positive

- In order to make it simpler and improve the performance of the models, the classes Extremely Positive and Extremely Negative were merged into Positive and Negative classes respectively.



Text pre-processing

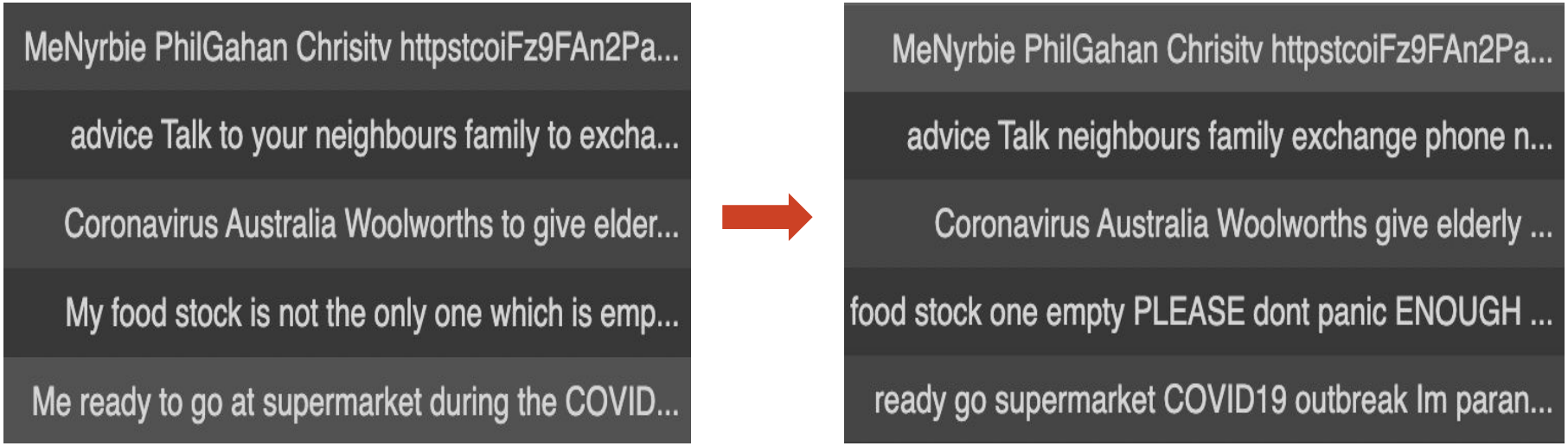
The text preprocessing is an important step where the words that are insignificant for the machine learning model are removed such as punctuations, special characters, stopwords etc. The objective is to remove words that carry less weightage in context to the text.

This is done in three steps:

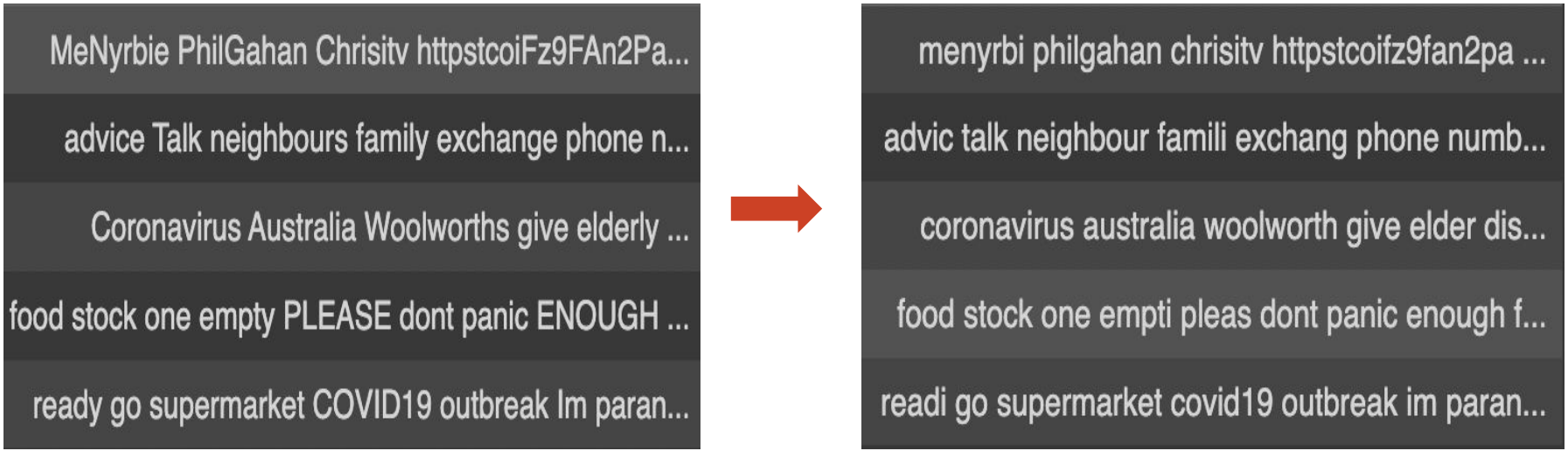
1. Remove punctuations
2. Remove stopwords
3. Apply stemming

Step 1- Remove punctuations

Step 2 - Remove stopwords



Step 3 - Stemming



Word Cloud

We can see some of the most frequently used words in tweets like COVID19, coronavirus, store, grocery, shopping, people, supermarket, price etc.

Model Comparison

Model comparison

S.No	Algorithm	Accuracy	Precision	Recall	F1 score
1.	Logistic Regression	0.80	0.79	0.80	0.79
2.	Random Forest Classifier	0.74	0.74	0.74	0.73
3.	XGBoost Classifier	0.70	0.70	0.70	0.69
4.	Multinomial Naive Bayes	0.63	0.70	0.63	0.57
5.	Support Vector Classifier	0.78	0.78	0.78	0.78

Conclusion

- Around the globe, people tweeted positive tweets more in comparison to negative and neutral tweets.
- Most of the tweets' length was around 250 which means there was a curiosity among the people related to COVID-19.
- People tweeted more in March than in April as many countries imposed lockdown during this period.
- The maximum number of tweets were from London (England) followed by the United States.
- During the pandemic, we saw mixed reactions from Australia as the count of positive and negative tweets were almost the same.
- Most of the tweets contain words like COVID19, grocery, supermarket, store, price etc. which shows during the pandemic, people were mainly concerned about food supplies and their prices.

Contd.

- Logistic regression has performed the best as it has got the highest accuracy, precision and recall values with a decent f1-score.
- Support Vector Classifier also performed well with not much of a difference in performance from logistic regression.
- Multinomial Naive Bayes is the worst performer which means the assumption of independence among the variables does not hold for this dataset.