PREDICTING SENTIMENT OF COVID 19 TWEETS

Capstone Project

AI AlmaBetter

Team Members:

- Kunal Gawande
- Chinma Rojatkar
- Deepali Mahajan
- Bipasha Zade
- Nikhil Aggarwal

CONTENT

1. Introduction.
2. Exploratory Data Analysis.
3. Data Pre-processing.
4. Vectorization.
5. Classification.
6. Evaluation.
7. Challenges.
8. Conclusion.
9. Q&A

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.

Presentations are column that can be used for analysis

- ✓ Location
- ✓ Tweet At
- ✓ Original Tweet
- ✓ Sentiment

Introduction

WE ARE ANALYZING DATA DURING PANDEMIC TIME TO GATHER CORRECT INFORMATION FOR MAKING POLICY FOR FURTHER USE

Sentiment Analysis is the process of computationally identifying and categorizing opinions expressed in a piece of text, especially in order to determine whether people attitude towards a particular topic is positive, negative, or neutral.
On 31st December, 2019 the Covid-19 outbreak was first reported in the Wuhan, Hubei Province, China and it started spreading rapidly all over the world. Finally, WHO announced Covid-19 outbreak as pandemic on 11th March, 2020, when the virus continues to spread.

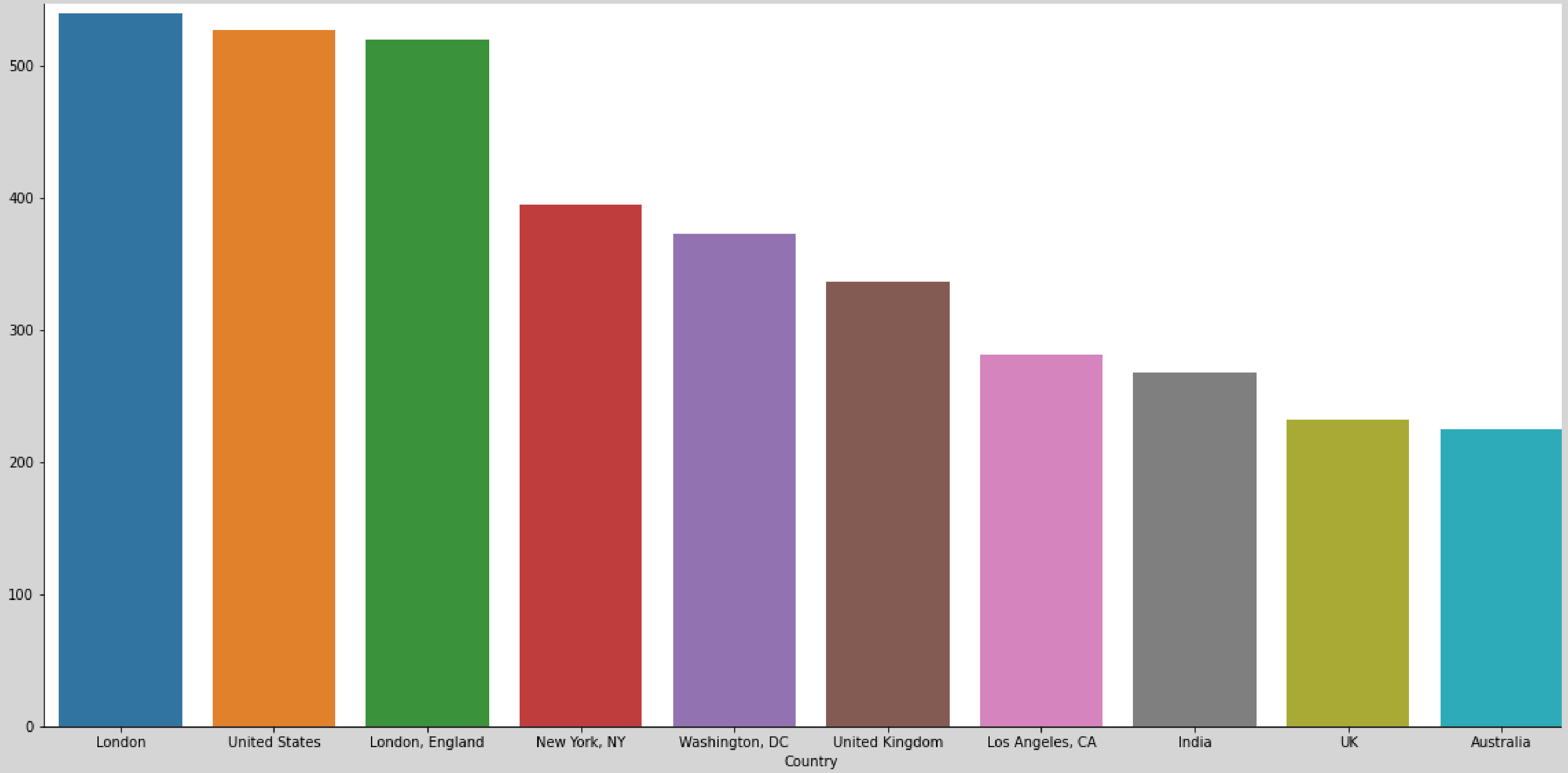
EXPLORATORY DATA ANALYSIS

- *LONDON , UNITED STATE AND ENGLAND HAS THE HIGEST TWEETS SENTIMENT*

- *AUSTRALIA AND UK HAS THE LOWEST TEWWTS SENTIMENTS*

- *According to graph we done the EDA OF TOP 10 COUNTRY TWEET WHOEVER A HIGHEST TWEET.*

- *Most of the tweets came from London followed by U.S*



COUNTRYWISE

Data Preprocessing

- Removing @user from Tweets
- Removed HTTP And URLS from Tweets
- Removing Punctuation, Numbers, and Special Characters
- Tokenization
- Removing Short Words

REMOVING STOPWORDS

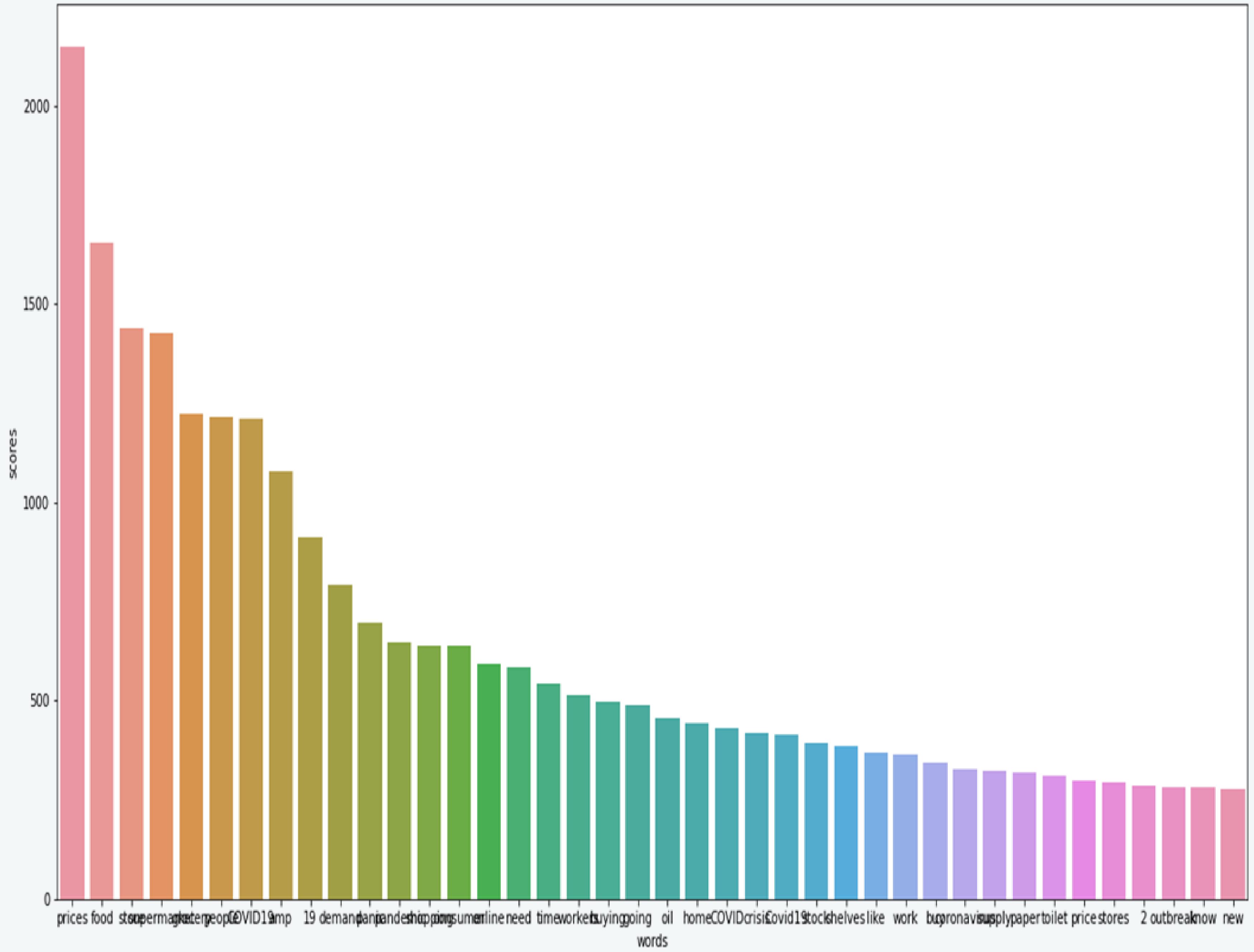
The words which are generally filtered out before processing a natural language are called stop words. These are actually the most common words in any language (like prepositions, pronouns, conjunctions, etc) and does not add much information to the text.

Stop words are available in abundance in any human language. By removing these words, we remove the low-level information from our text in order to give more focus to the important information.

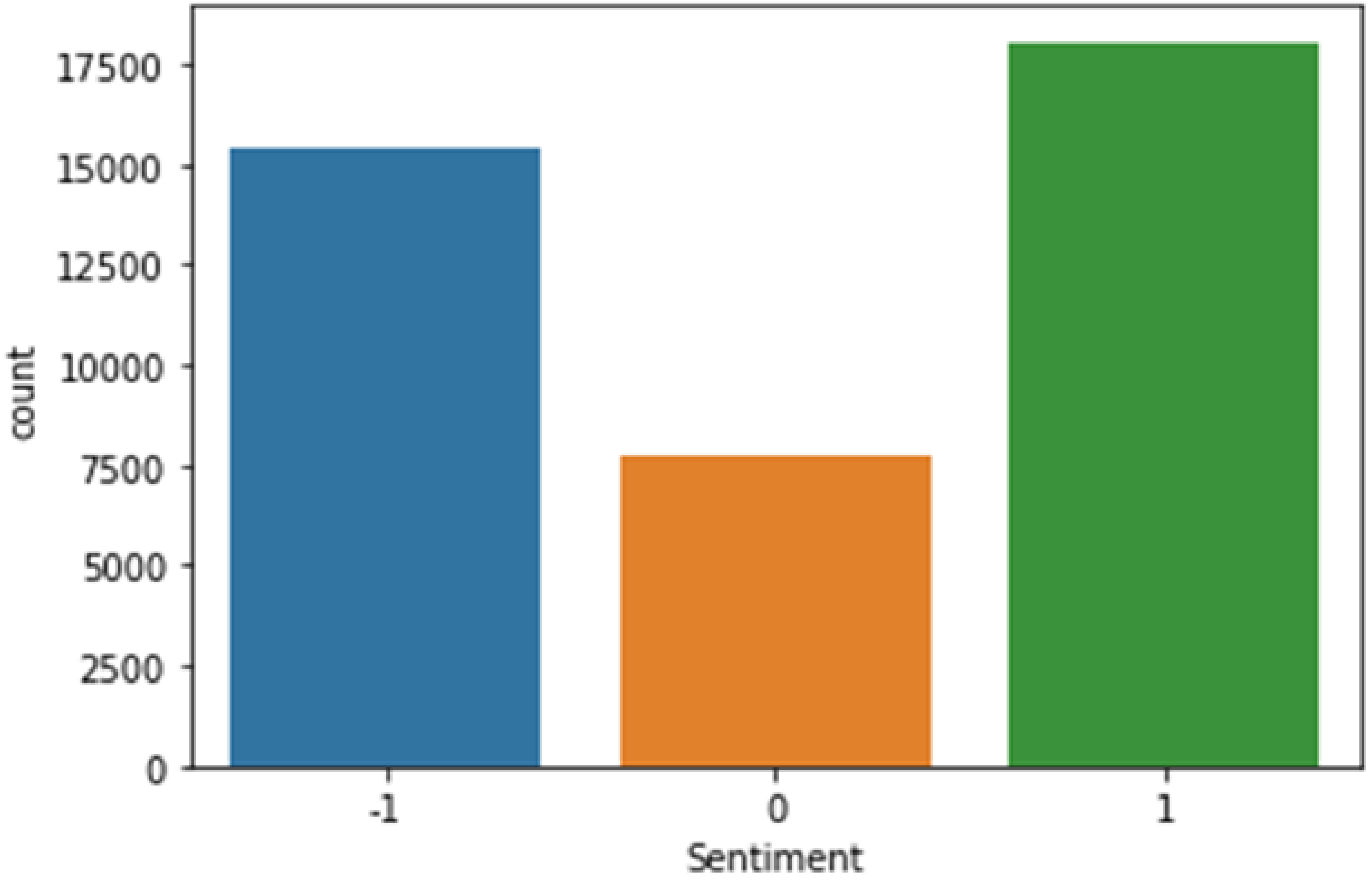
DATA SUMMARY

In order to analyses various sentiments, We require just two columns named Original Tweet and Sentiment
There are four types of sentiments- Extremely Negative, Negative, Neutral, Positive and Extremely Positive.

The original dataset has 6 columns and 41157 rows.



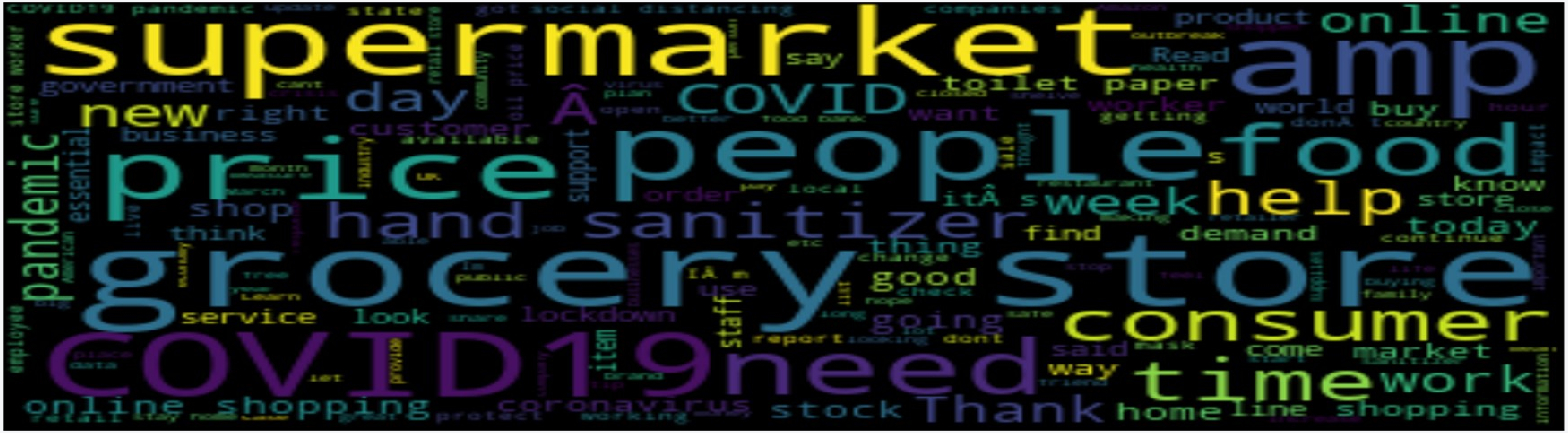
ONE HOT ENCODING



We have classified the tweets on the basis

the compound sentiments into two different classes, i.e. Positive is (1), Negative (-1).

Then we have assigned the sentiment rating for each tweet based on the algorithm presented also shows in Graph.



EDA On “Original Tweet” Column.

- There are some words like ‘coronavirus’,’super market’, having the maximum frequency in our dataset.
- There are various #hashtags in tweets column. But they are almost same

TOKENIZATION

In tokenization we convert group of sentence into token . It is also called text segmentation

It is basically splitting data into small chunk of words.

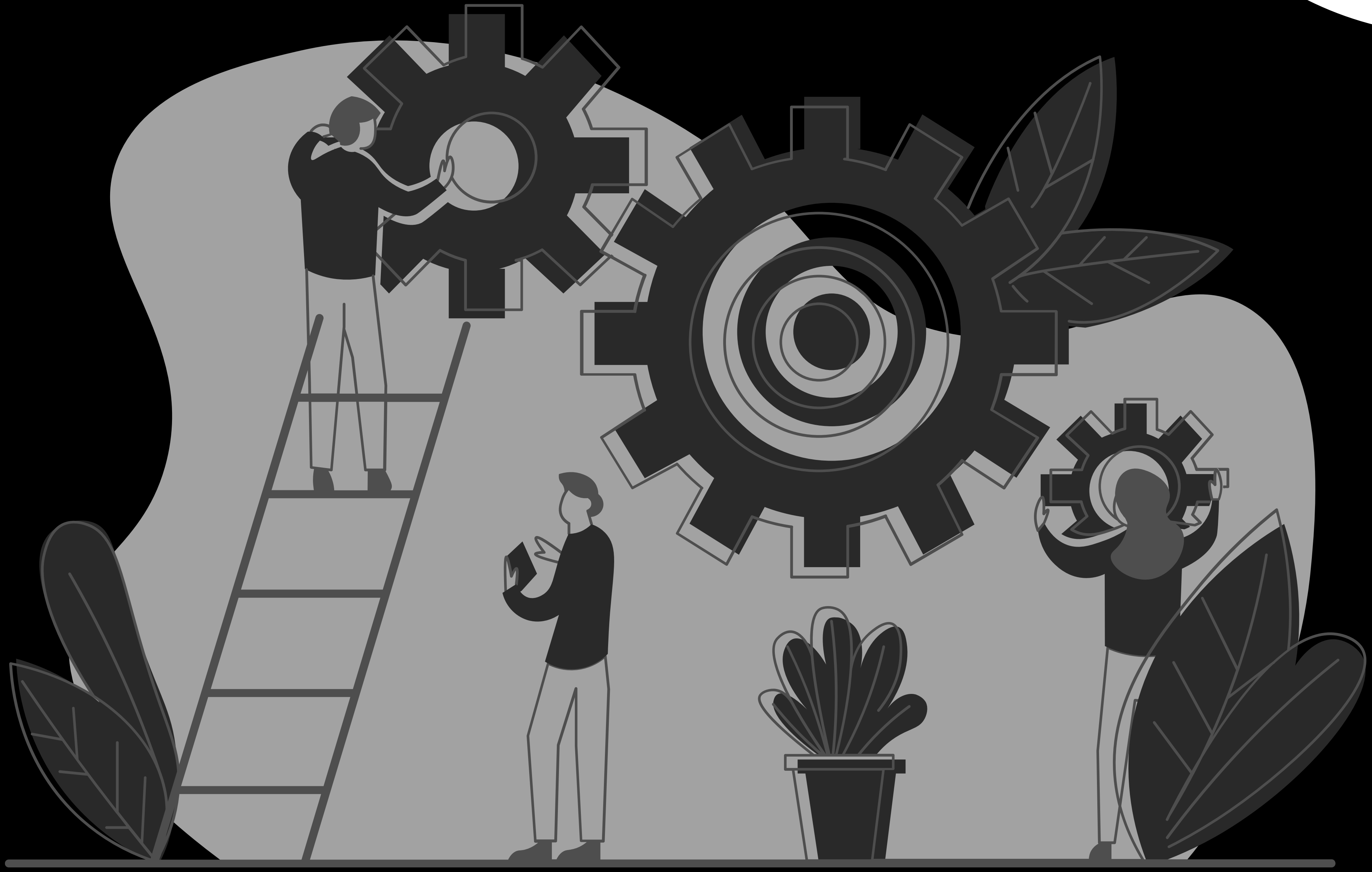
Tokenization in python can be done by python NLTK library's word tokenize() function.

Spitting Our Dataset into Training And Testing Dataset

We need to split a dataset into train and test sets to evaluate how well our machine learning model performs. The train set is used to fit the model, and the statistics of the train set are known. The second set is called the test data set.

VECTORIZATION

- We chose Count Vectorizer as our Vectorizer with minimum document frequency =10.
- It will create a sparse matrix of all words and the number of times they are present in a document
- Countvectorizerto transform a given text into a vector based on the frequency (count) of each word that occurs in the entire text.



Building Classification Models

THERE ARE FIVE TYPES OF SENTIMENTS SO WE HAVE TO TRAIN OUR MODELS SO THAT THEY CAN GIVE US THE CORRECT LABEL FOR THE TEST DATASET. I AM GOING TO BUILT DIFFERENT MODELS LIKE NAIVE BAYES, LOGISTIC REGRESSION, RANDOM FOREST, XGBOOST, SUPPORT VECTOR MACHINES AND STOCHASTIC GRADIENT DESCENT.

Classification

ML MODELS

1

Naive Bayes Classifier for MULTICLASS Classification

2

RANDOM FOREST CLASSIFIER FOR BINARY CLASSIFICATION

3

LOGISTIC REGRESSION BINARY CLASSIFICATION)

4

XG BOOST BINARY CLASSIFICATION

5

SUPPORT VECTOR MACHINE(BINARY CLASSIFICATION)

5

Stochastic Gradient Descent-SGD Classifier BINARY CLASSIFICATION

NAIVE BAYES CLASSIFIER FOR MULTICLASS CLASSIFICATION

Training Accuracy	81.11%
Validation accuracy	67.37%

RANDOM FOREST CLASSIFIER FOR BINARY CLASSIFICATION

Training Accuracy 99.9%

Validation accuracy 78.30%

LOGISTIC REGRESSION BINARY CLASSIFICATION

Training Accuracy	97.14%
Validation accuracy	81.82%

XG BOOST BINARY CLASSIFICATION

Training Accuracy	61.11%
Validation accuracy	78.30%

SUPPORT VECTOR MACHINE(BINARY CLASSIFICATION)

Training Accuracy	94.85%
Validation accuracy	78.74%

WINNING MODEL

STOCHASTIC GRADIENT DESCENT-SGD CLASSIFIER(BINARY CLASSIFICATION)

Training Accuracy 95.16%

Validation accuracy 83.62%

Concluson

We focused on sentiment analysis for sentence labelling. We described the preprocessing steps, and pipeline steps within which text normalization and model cross-validation is included, and performance has been measured using balanced accuracy, f1 score etc. We used “Stemming” instead of Lemmatization to reduce dimensions, for the same reason we haven’t tried tf-idf or term frequency vectorizer. We concentrated on feeding our model with word count information. We assume, in the case of binary classification, we can further improve this score.

THANK YOU!