

PREDICTING SENTIMENT OF COVID 19 TWEETS

Capstone Project

AI AlmaBetter



Team Members:

- Kunal Gawande
- Chinmay 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

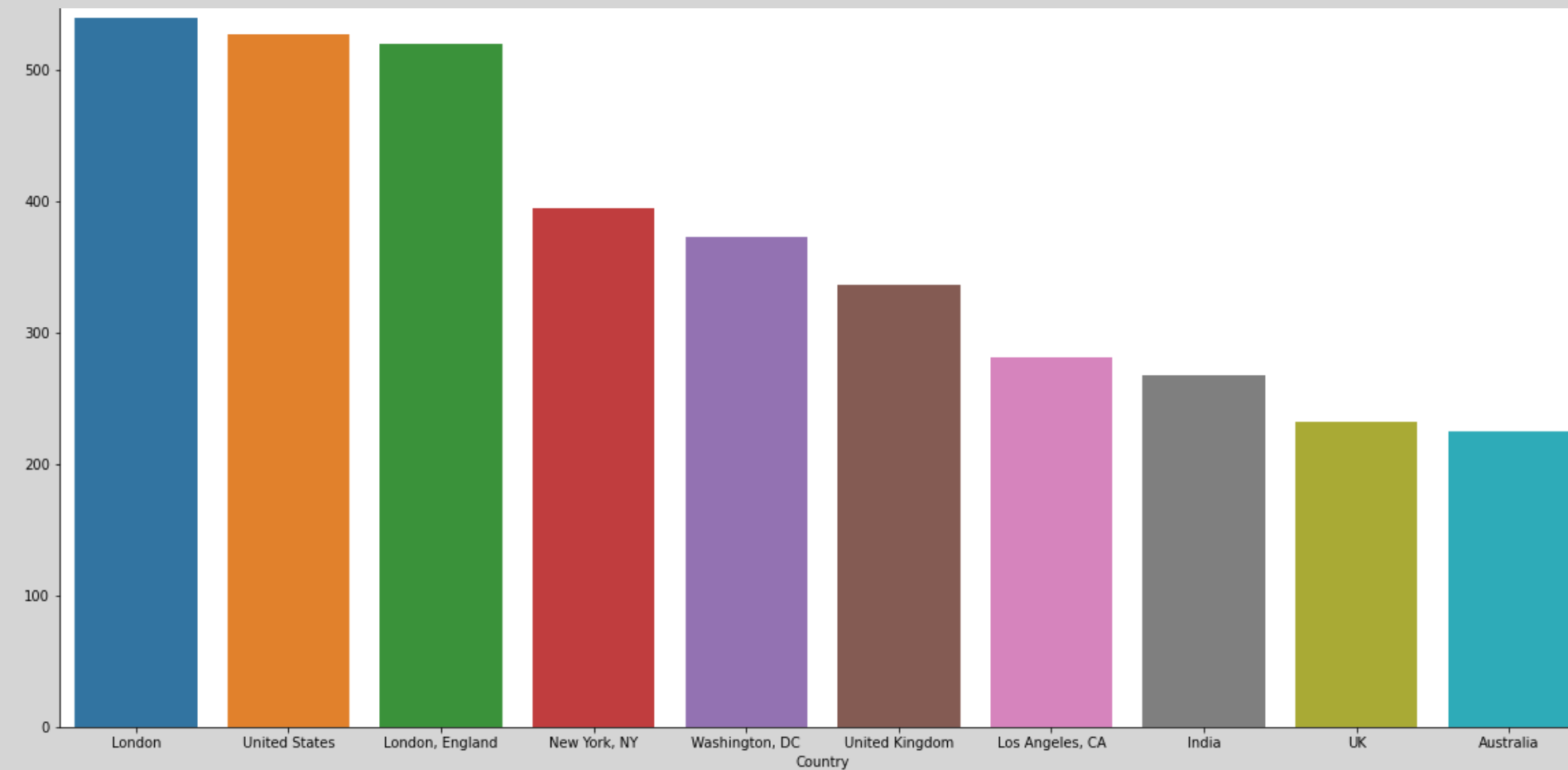
WE ARE ANALYZING DATA DURING PANDEMIC TIME TO GATHER CORRECT INFORMATION FOR MAKING POLICY FOR FURTHER USE

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*





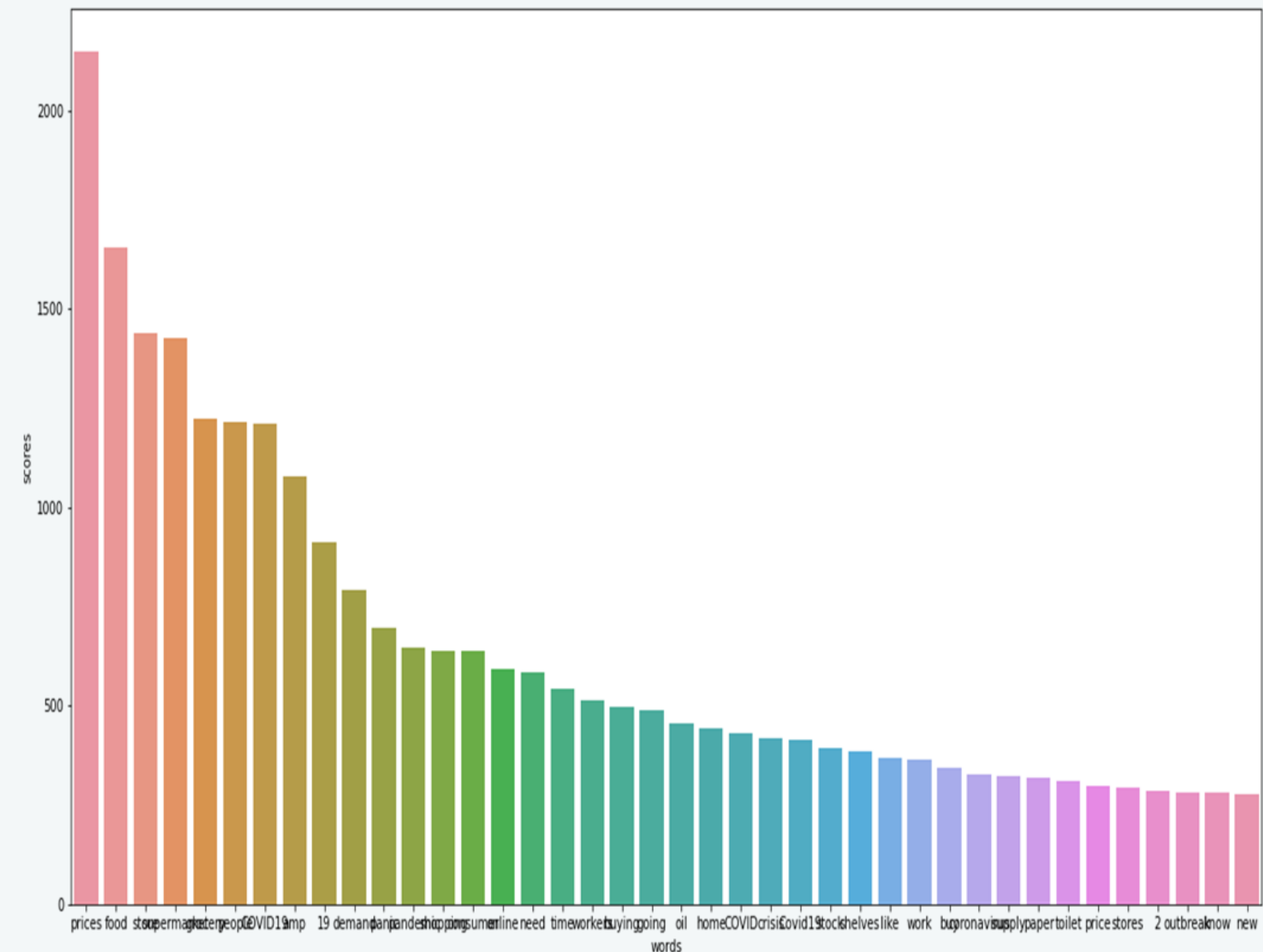
DATA SUMMARY

In order to analyse various sentiments, We require just two columns named Original Tweet and Sentiment

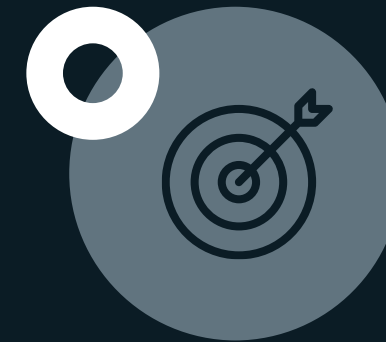
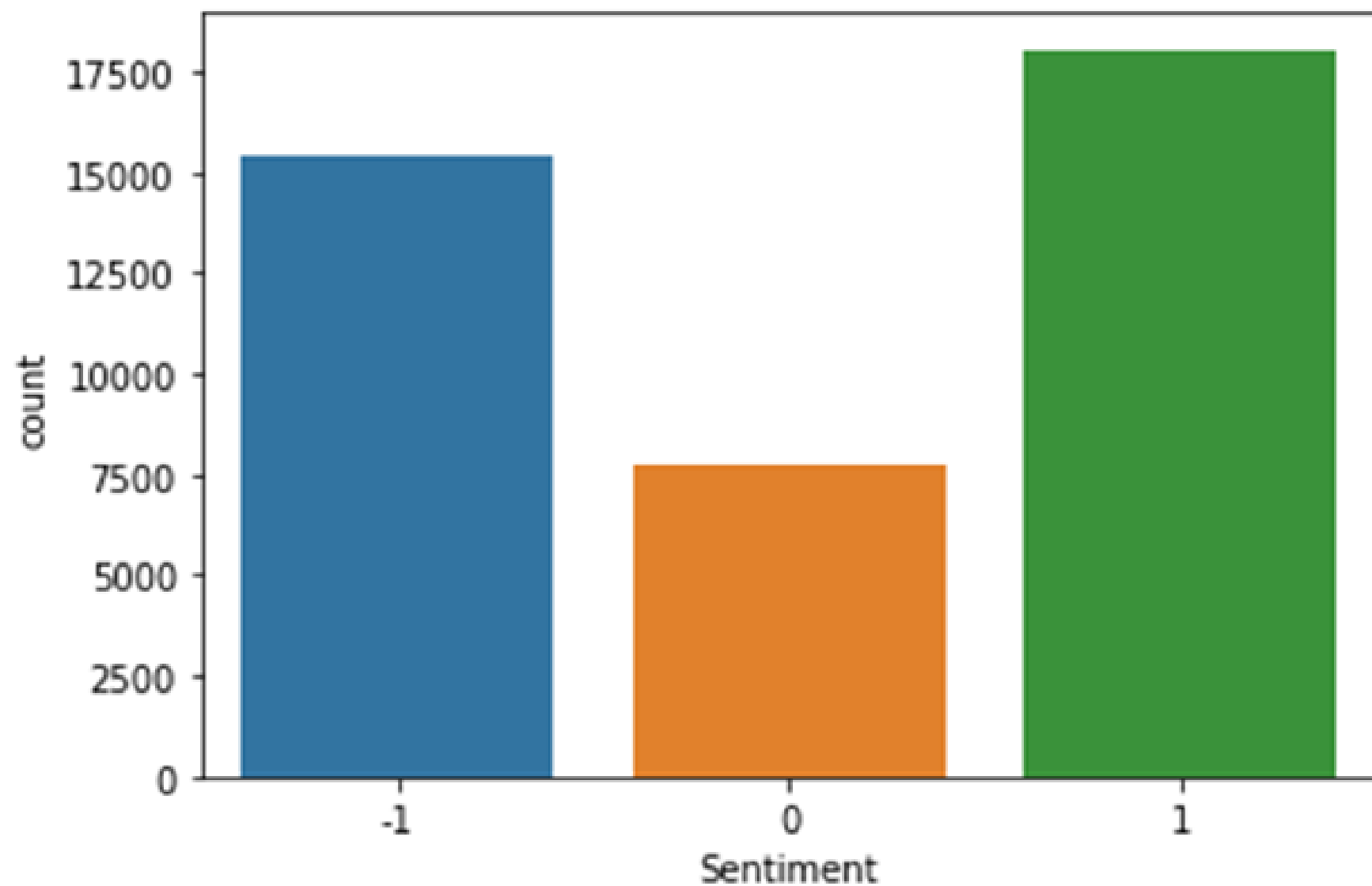
There are four types of sentiments- Extremely Negative, Negative, Neutral, Positive and Extremely Positive.



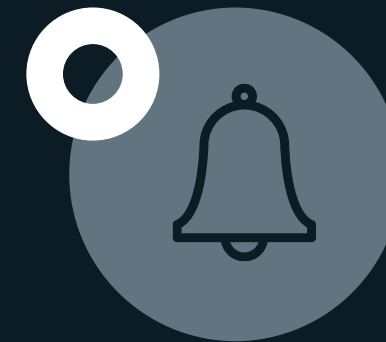
The original dataset has 6 columns and 41157 rows.



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 polarity rating for each tweet based on the algorithm presented also shows in Graph.



EDA On “Original Tweet” Column.

- There are some words like 'coronavirus', 'grocery store', having the maximum frequency in our dataset.
- There are various #hashtags in tweets column. But they are almost same in

TOKENIZATION

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



It is basically splitting data into small chunk of words.



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



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,
CATBOOST, 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

Training Accuracy

0.8111%

*NAIVE BAYES
CLASSIFIER FOR
MULTICLASS
CLASSIFICATION*

Validation accuracy

0.6737%

*RANDOM FOREST
CLASSIFIER FOR
BINARY
CLASSIFICATION*

Training Accuracy

0.999%

Validation accuracy

0.7830%

*LOGISTIC
REGRESSION
BINARY
CLASSIFICATION*

Training Accuracy

0.9714%

Validation accuracy

0.8182%

*XG BOOST BINARY
CLASSIFICATION*

Training Accuracy

0.6111%

Validation accuracy

0.7830%

*SUPPORT VECTOR
MACHINE(BINARY
CLASSIFICATION)*

Training Accuracy

0.9448%

Validation accuracy

0.7874%

WINNING MODEL

Training Accuracy

0.9516%

*STOCHASTIC
GRADIENT
DESCENT-SGD
CLASSIFIER(
BINARY
CLASSIFICATION)*

Validation accuracy

0.8362%



Conclusion

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.

