Capstone Project

SENTIMENT ANALYSIS FOR COVID 19

BY-

Hritik Sharma

Point of discussion

Introduction
Problem statement
Percentage of missing values in columns
Heating map of missing values
Bar plot for no. of unique values
Pie chart on top 10 Locations Of Tweets
Data preprocessing
Story Generation and Visualization from Tweets
Making frequency distribution on sentiments
Use Of several classifier For Multi Class Classification
All the multiclass models test accuracy in descending order
Converting our multiclass classification into binary classification
Evaluation of all binary classification models

POINT OF DISCUSSION

CONCLUSION

INTRODUCTION

Hi folks, I hope you are doing well in these difficult times! We all are going through the unprecedented time of the Corona Virus pandemic. Some people lost their lives, but many of us successfully defeated this new strain i.e. Covid-19. The virus was declared a pandemic by World Health Organization on 11th March 2020. **This article will analyze various types of "Tweets" gathered during pandemic times.** The study can be helpful for different stakeholders.

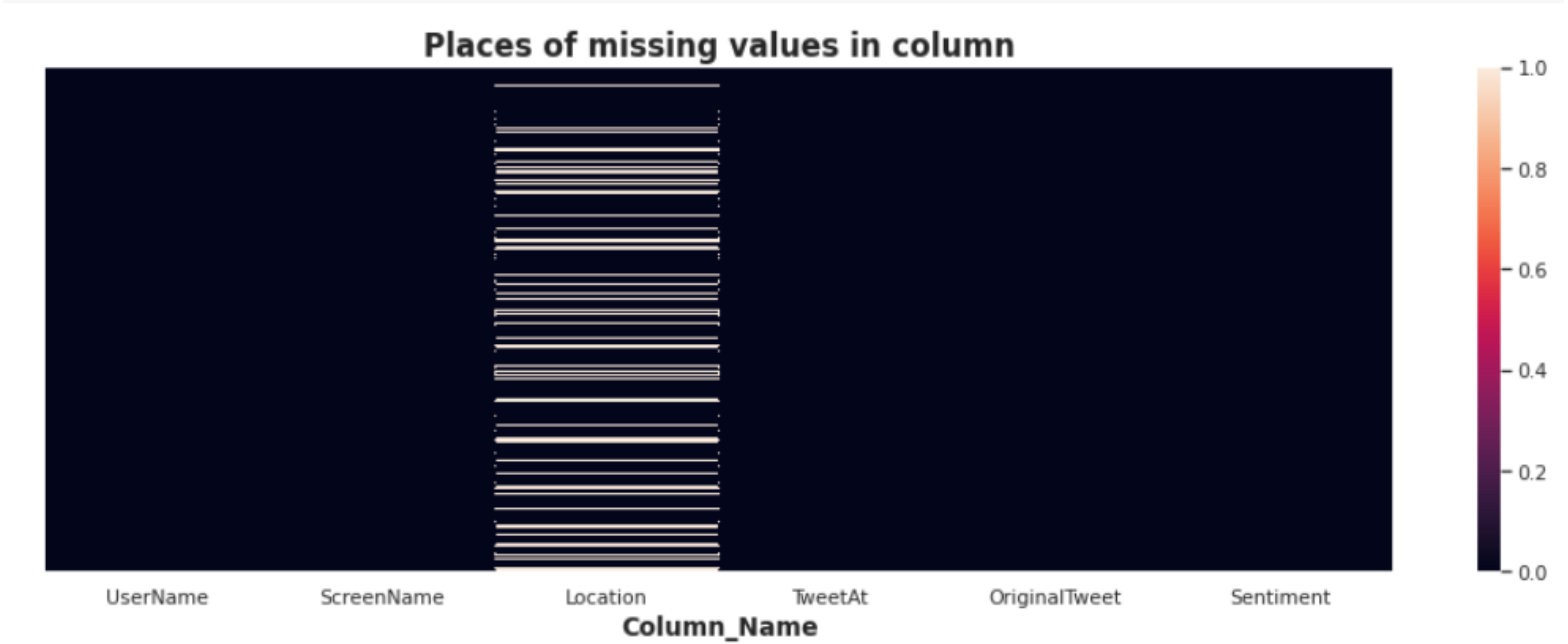
PROBLEM STATEMENT

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.

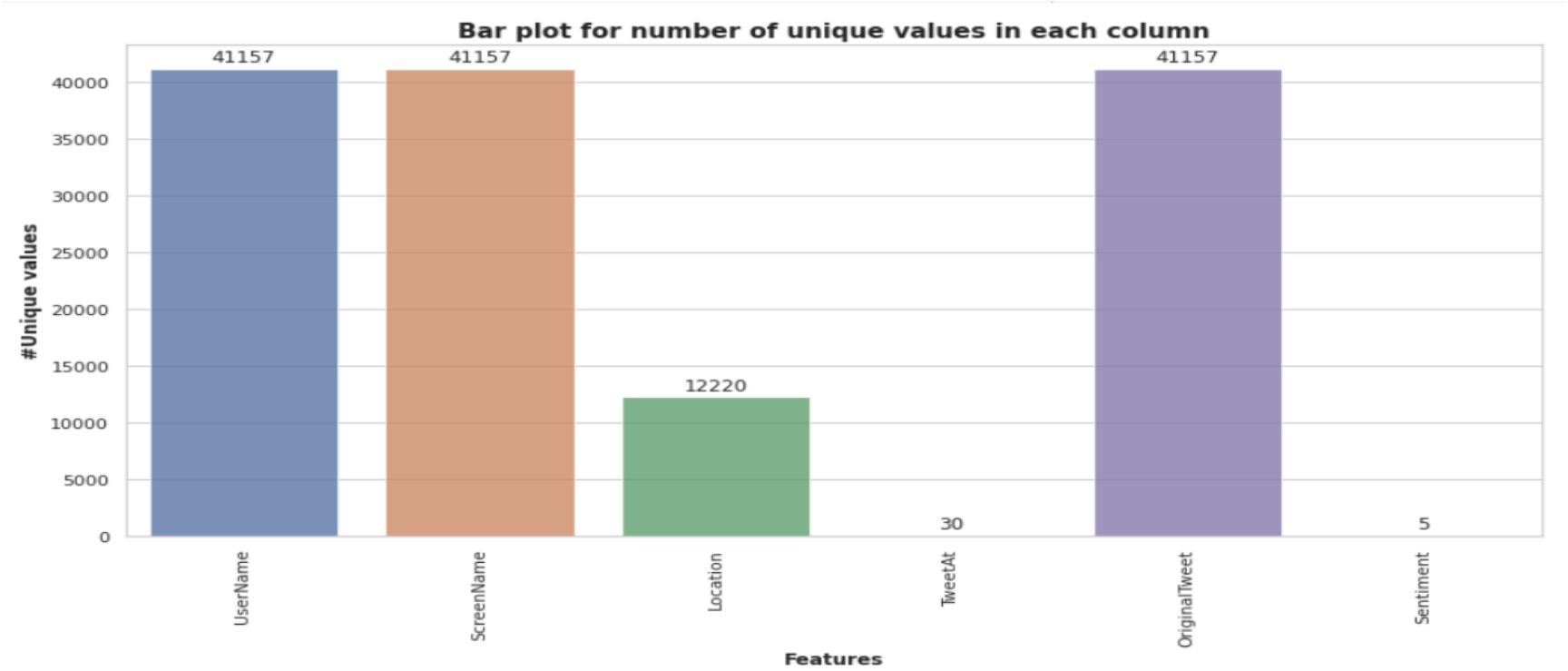
PERCENTAGE OF MISSING VALUES IN COLUMNS

After performing data cleaning on our dataset we found only location column which has 28% null values

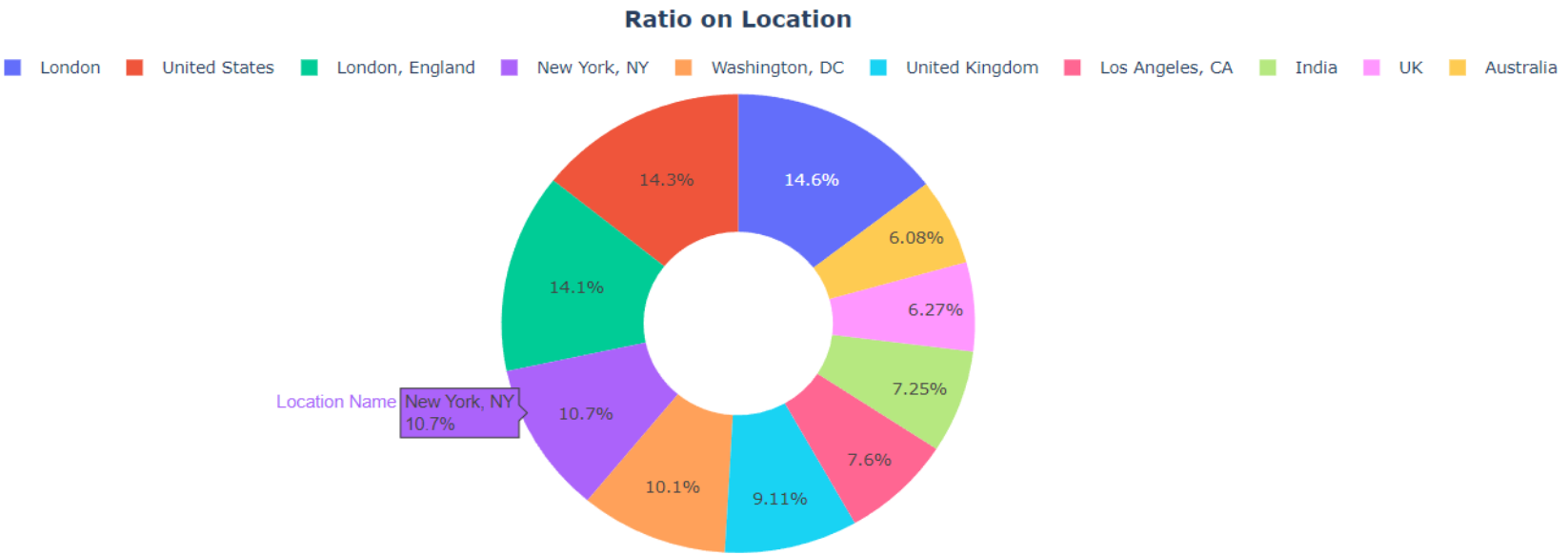
HEAT MAP OF MISSING VALUES



BAR PLOT FOR NO. OF UNIQUE VALUES



PIE CHART ON TOP 10 LOCATION OF TWEETS



DATA PREPROCESSING

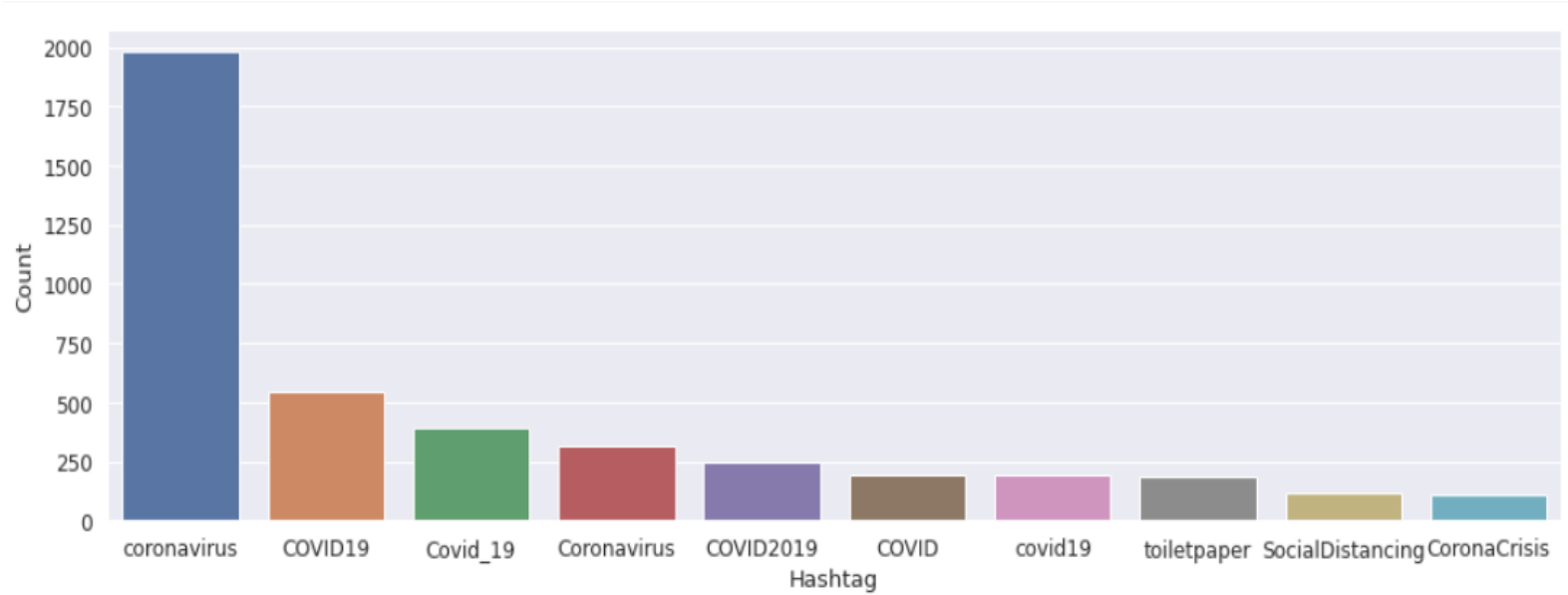
- Removing @user
- Removed http and urls from tweets
- Removing Punctuations, Numbers, and Special Characters
- Removing Short Words
- Tokenization
- Stemming

STORY GENERATION AND VISUALISATION FROM TWEETS

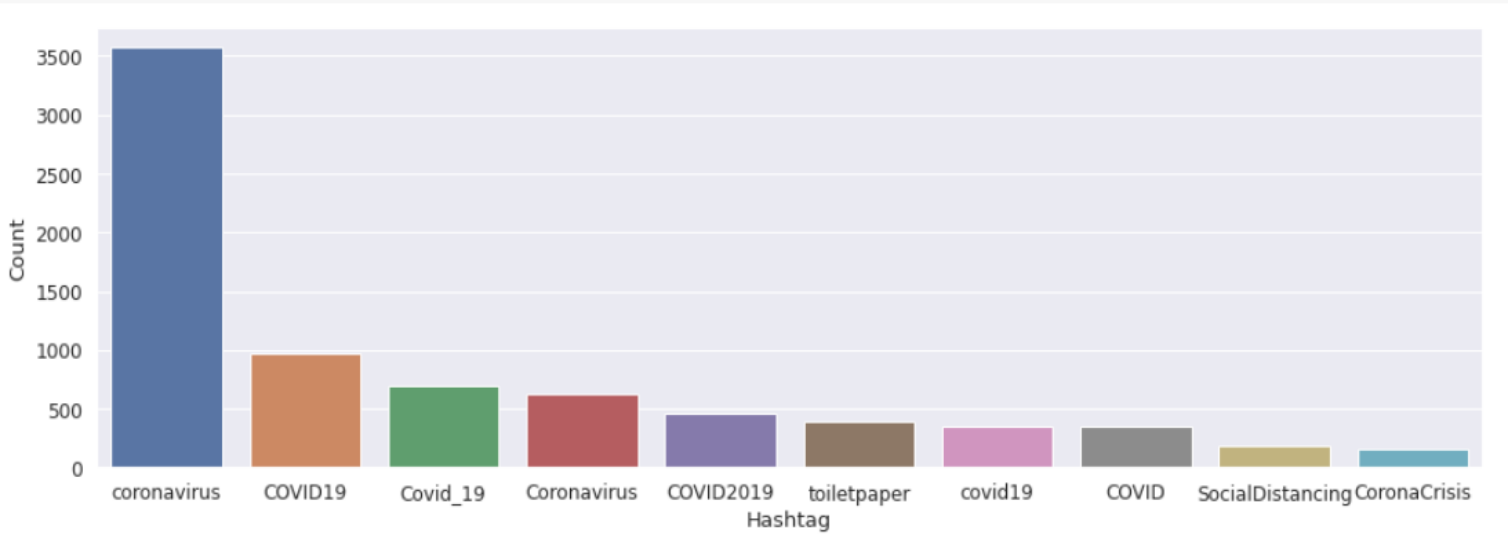
- What are the most common words in the entire dataset?
- What are the most common words in the dataset for negative and positive tweets, respectively?
- How many hashtags are there in a tweet?
- Which trends are associated with my dataset?
- Which trends are associated with either of the sentiments? Are they compatible with the sentiments?

MAKING FREQUENCY DISTRIBUTION ON SENTIMENTS

- Making frequency distribution top 10 Extremely Positive hashtags



Making frequency distribution top 10 Positive hashtags



USE OF SEVERAL CLASSIFIER FOR MULTI CLASS CLASSIFIER

- Counter vectorizer for multiclass classification
- Naïve bayes classifer for multiclass classification
- Stochastic gradient descent sgd classifier
- Random forest classifier
- Extreme gradient boosting classification
- Support vector machine
- Logistic regression

All the multiclass models test accuracy in descending order

	Model	Test accuracy
1	Logistic Regression	0.617954
0	Support Vector Machines	0.607264
4	Stochastic Gradient Decent	0.572643
2	Random Forest	0.566448
5	XGBoost	0.486880
3	Naive Bayes	0.479470

Evaluation of all binary classification models

Winner model – Stochastic gradient descent

	Model	Test accuracy
4	Stochastic Gradient Decent	0.862488
1	Logistic Regression	0.859451
6	CatBoost	0.850705
0	Support Vector Machines	0.845603
2	Random Forest	0.832483
3	Naive Bayes	0.791667
5	XGBoost	0.739553

CONCLUSION