# INFORMATION RETRIEVAL TECHNIQUES

## CASE STUDY-SENTIMENT ANALYSIS

**Topic:** **Covid 19 Indian Sentiments on covid19 and lockdown**


**Submitted to:**

Dr.Keerthy A.S

(Assist Prof) Dept of Computer Science

RCSS

**Submitted by:**

Vishnu M

MSCCS (DataAnalytics)

Roll no:30

## Abstract

COVID-19 caused a significant public health crisis worldwide and triggered some other issues such as economic crisis, job cuts, mental anxiety, etc. This dataset contains cleaned tweets from India on topics like coronavirus, covid 19 and lockdown etc. The tweets have been collected between dates 23th march 2020 and 15th July 2020.hen the text have been labeled into four sentiment categories fear, sad, anger and joy.

This pandemic plies across the world and involves many people not only through the infection but also agitation, stress, fret, fear, repugnance, and poignancy. During this time, social media involvement and interaction increase dynamically and share one's viewpoint and aspects under those mentioned health crises. From user-generated content on social media, we can analyze the public's thoughts and sentiments on health status, concerns, panic, and awareness related to COVID-19, which can ultimately assist in developing health intervention strategies and design effective campaigns based on public perceptions. In this work, we scrutinize the users' sentiment in different time intervals to assist in trending topics in Twitter on the COVID-19 tweets dataset.

The approach of the study is that I used Bag of Words for extracting features from the Review Body. Sentiment classification aims to determine the overall intention of a written text which can be of fear, sad, anger and joy. This can be achieved by using machine learning algorithms such as Naïve Bayes, Random Forest, Decision Tree and Logistic Regression.

So that we can easily classify each comment into different groups.

Main file in this dataset is finalSentimentdata2.csv and the detailed descriptions are below

1. text- Cleaned twitter text
2. sentiment - Different sentiment label

## Introduction

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 Project will analyze various types of “Tweets” gathered during pandemic times. The study can be helpful for different stakeholders.

For example, Government can make use of this information in policymaking as they can able to know how people are reacting to this new strain, what all challenges they are facing such as food scarcity, panic attacks, etc. Various profit organizations can make a profit by analyzing various sentiments as one of the tweets telling us about the scarcity of masks and toilet papers. These organizations can able to start the production of essential items thereby can make profits. Various NGOs can decide their

strategy of how to rehabilitate people by using pertinent facts and information.

Sentiment analysis and classification is a computational study which attempts to address this problem by extracting subjective information from the given texts in natural language, such as opinions and sentiments. Different approaches have used to tackle this problem from natural language processing, text analysis, computational linguistics, and biometrics. In recent years, Machine learning methods have got popular in the semantic and review analysis for their simplicity and accuracy.

## 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.

In this project, we are going to Classify the Sentiments of COVID-19 tweets. The data gathered from the Tweeter and using Python environment to implement this project.

## Method of Analysis Performed

Following is the Standard Operating Procedure to tackle the Sentiment Analysis.

1. Exploratory Data Analysis.
2. Data Preprocessing.
3. Vectorization
4. Classification Models.

Machine learning algorithms work only with fixed-length vector of numbers rather than raw text, the input (in this case text data) need to be parsed. The method for transforming the texts into features is called the Bag of words model of text, which is a commonly used method of feature extraction. The approach works by creating different bags of words that occur in the training data set where each word is associated with a unique number. This number shows the occurrence of each word in the document. The model is called a bag of words because the position of the words is in the document discarded. Texts generated by humans in social media sites contain lots of noise that can significantly affect the results of the sentiment classification process. Moreover, depending on the features generation approach, every new term seems to add at least one new dimension to the feature space. That makes the feature space more sparse and high-dimensional. Consequently, the task of the classifier has become more complex. To prepare messages, such text preprocessing techniques as replacing URLs and usernames with keywords, removing punctuation marks and converting to lowercase were used in this program.

## 1. Exploratory Data Analysis.

This data contains 3046 unique rows were there are 3 attributes:

1.Unnamed

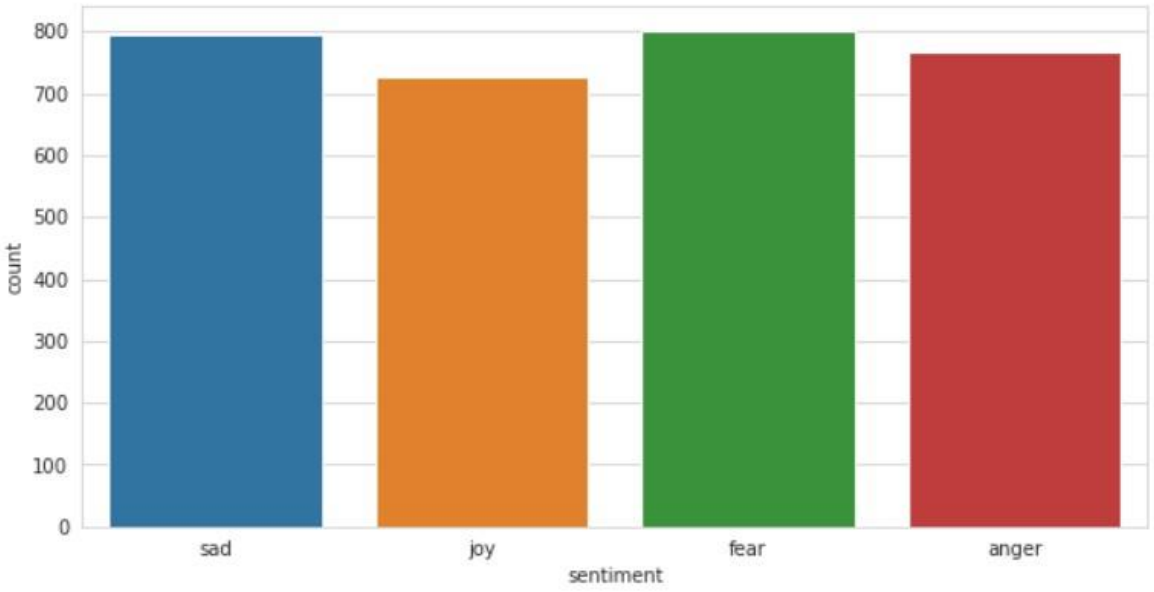
2.Text (comments)

3.Sentiments (fear, sad, anger and joy)

Here, we can see the no. of rows having each sentiment.

```
fear     801
sad      795
anger    767
joy      727
```

Bar Graph:



## Pre-processing

Here we delete the attribute which is not useful in our dataset and we check the missing values. As here we don't have any missing values.

## Vectorization

Word vectorization is a methodology in NLP to map words or phrases from vocabulary to a corresponding vector of real numbers which used to find word predictions, word similarities/semantics. The process of converting words into numbers are called Vectorization.

Hence, each sentiment are numbered as:

```
0    agree the poor in india are treated badly thei...
1    if only i could have spent the with this cutie...
2    will nature conservation remain a priority in ...
3    coronavirus disappearing in italy show this to...
4    uk records lowest daily virus death toll since...
```

*#1 for fear*
*#3 for sad*
*#0 for anger*
*#2 for joy*

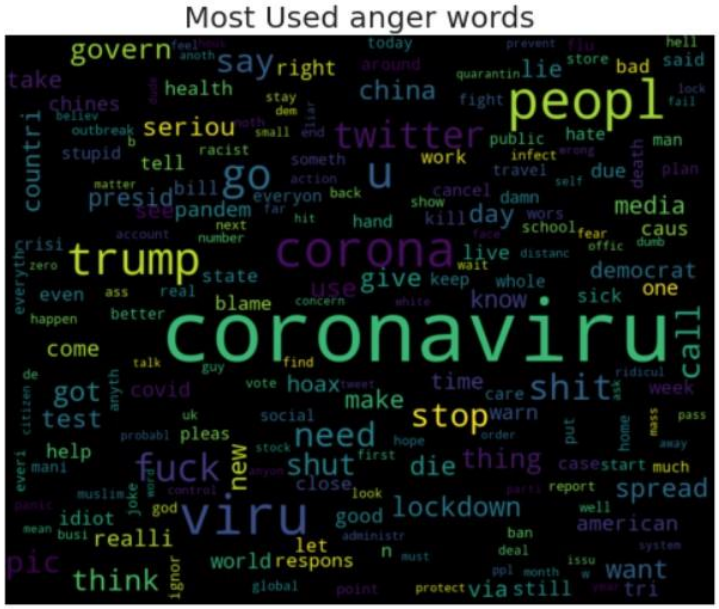
## Stemming and Cleaning the text.

It is 80% of cleaned data now and it have to undergo **lemmatizing.**

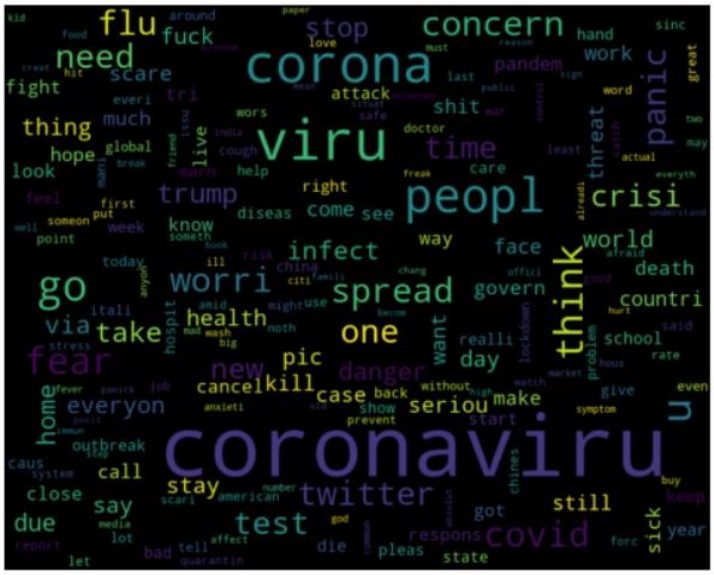
To achieve maximum accuracy of each comment.

From the cleaned data we can interpret each sentiment's word cloud.

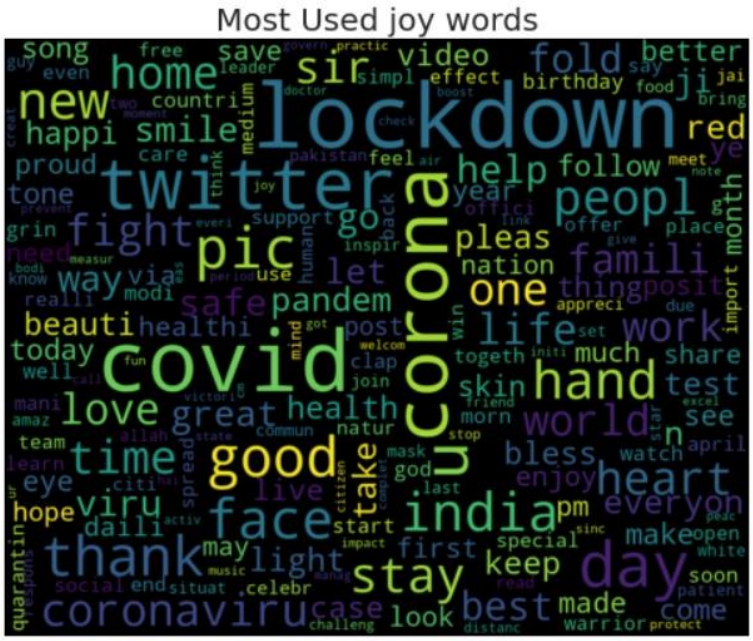
## Most commonly used Anger words



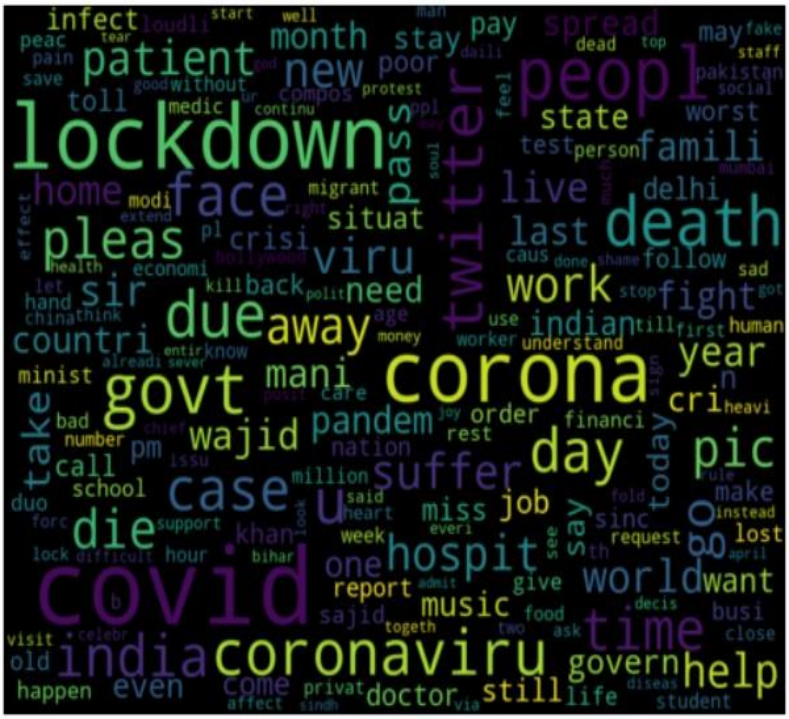

## Most commonly used Fear words



## Most commonly used Joy words

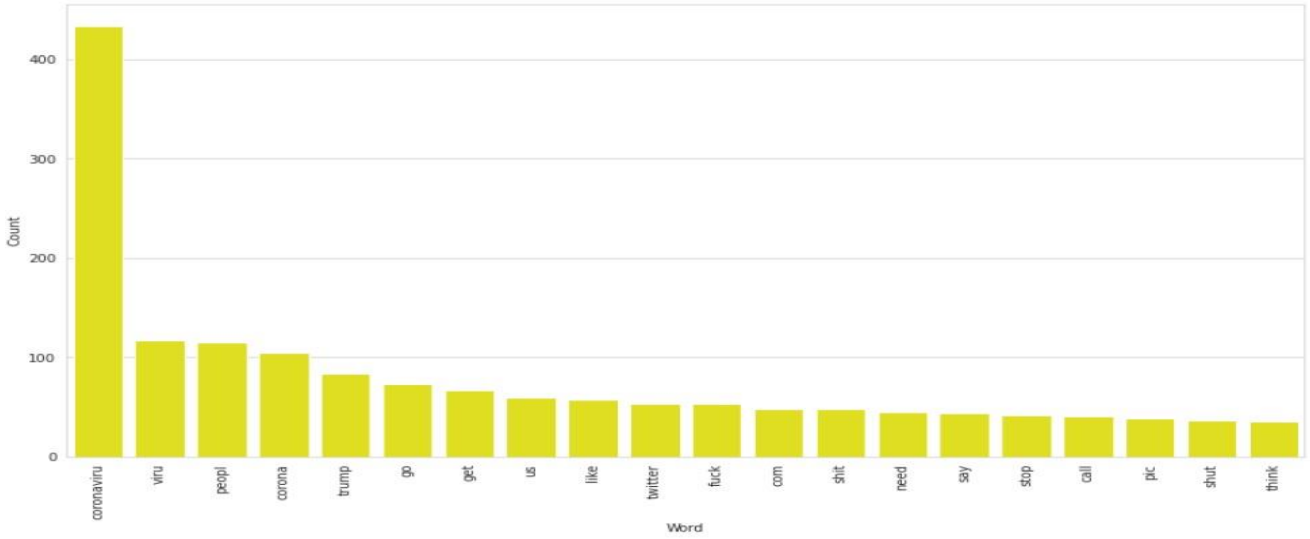



## Most commonly used Sad words

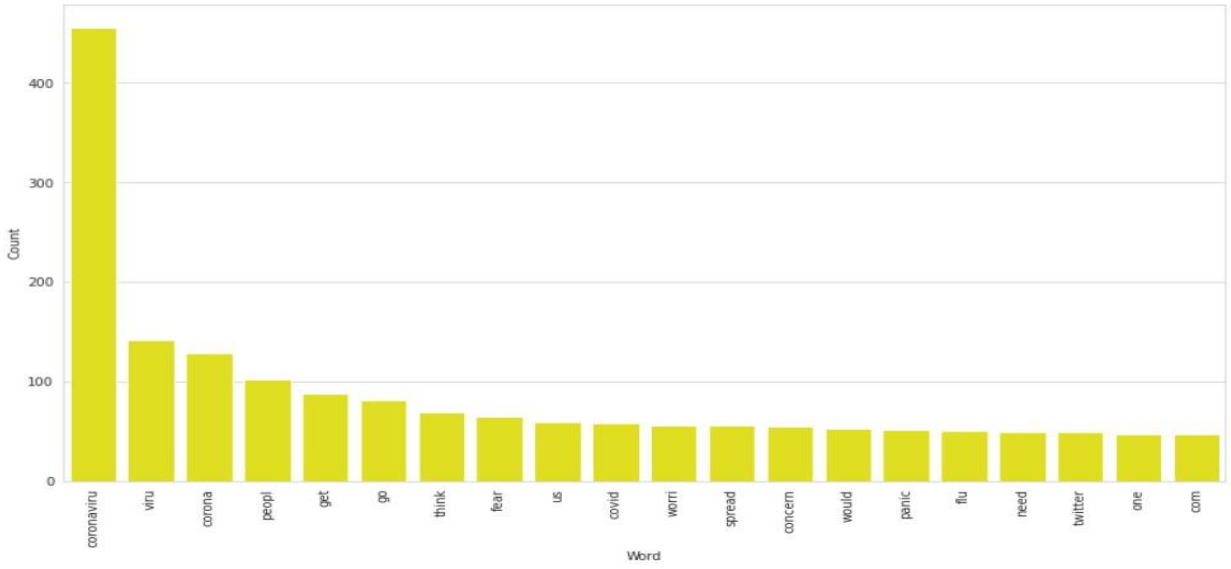


These are the visualization of most occurred words for each sentiment. Here we can plot bar graph by the frequency.

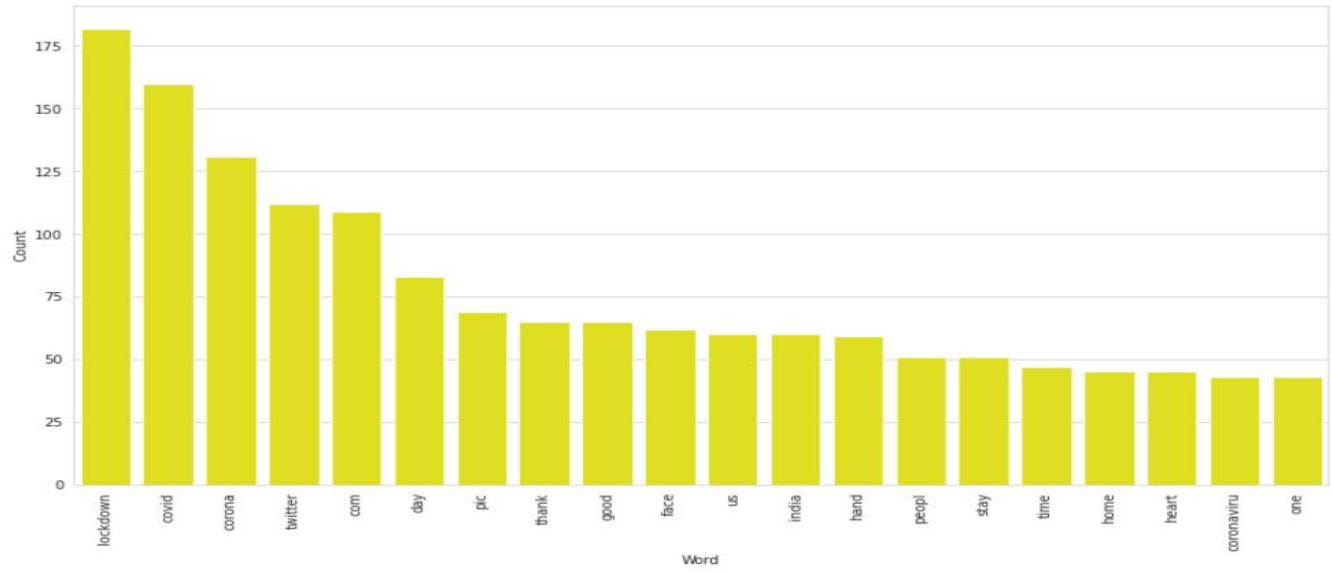
*frequency of most anger words*



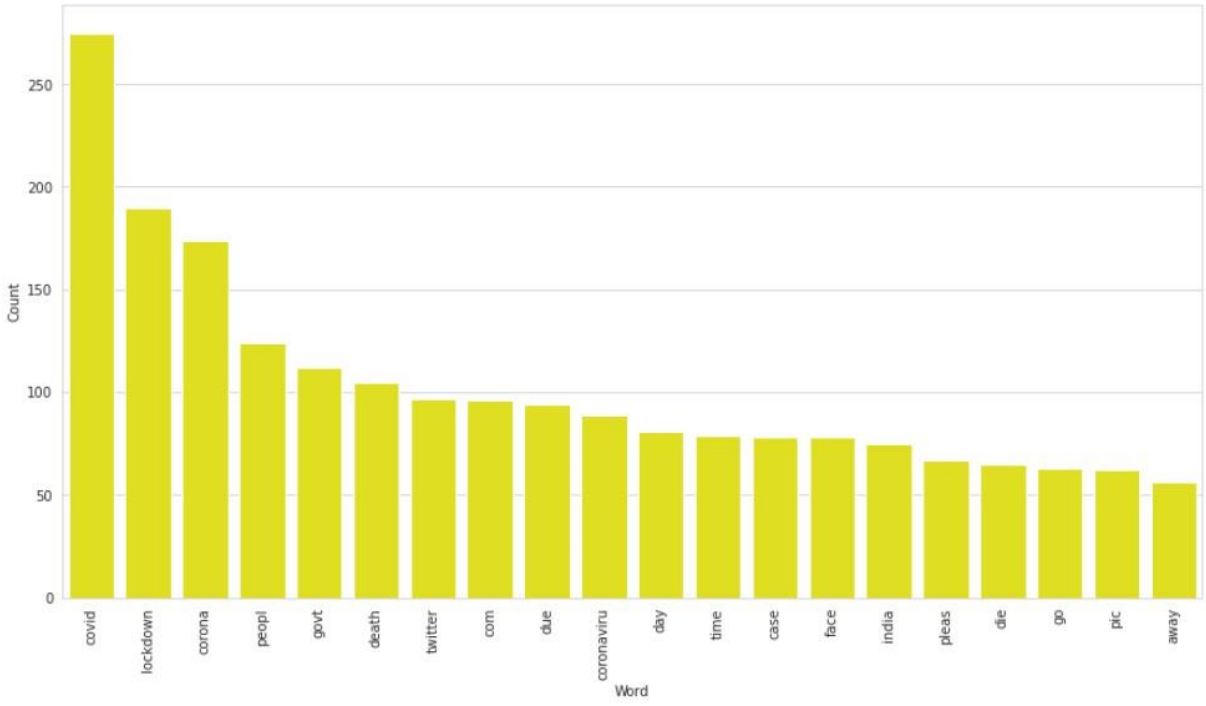

*frequency of most fear words*




*frequency of most joy words*




*frequency of most sad words*



## Classification Models

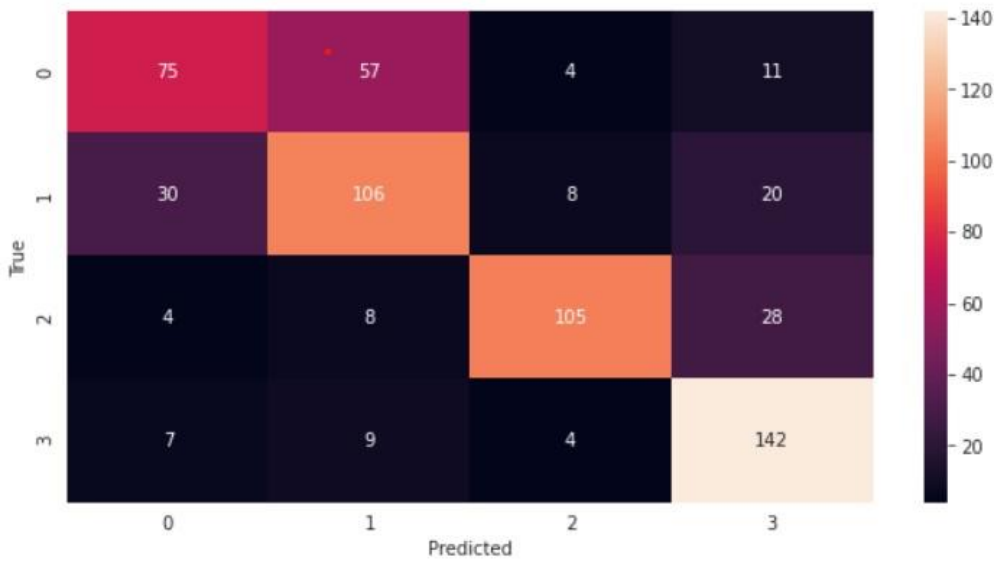
- Using Naïve Bayes method

Naive Bayes classifiers are a collection of classification algorithms based on Bayes' Theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other. When used for textual data analysis, such as Natural Language Processing, the Naive Bayes classification yields good results. Simple Bayes or independent Bayes models are other names for nave Bayes models. All of these terms refer to the classifier's decision rule using Bayes' theorem. In practice, the Bayes theorem is applied by the Naive Bayes classifier. The power of Bayes' theorem is brought to machine learning with this classifier. Multinomial Naive Bayes classification algorithm tends to be a baseline solution for sentiment analysis task. The basic idea of Naive Bayes technique is to find the probabilities of classes assigned to texts by using the joint probabilities of words and classes

## Result Analysis using Naïve Bayes

To carry out the experiments, each classifier algorithm needs to be trained before being tested. In order to train and use the classifiers, the data was divided into two data sets as training and testing data sets. After using SentimentIntensityAnalyzer()

Accuracy=65%




The accuracy after applying hyperparameter in MultinomialNB when alpha=0.6 its gives the maximum accuracy of 0.6747572815533981

```
Alpha: 0.0, Score : 0.6245954692556634
Alpha: 0.1, Score : 0.6650485436893204
Alpha: 0.2, Score : 0.6682847896440129
Alpha: 0.30000000000000004, Score : 0.6715210355987055
Alpha: 0.4, Score : 0.686084142394822
Alpha: 0.5, Score : 0.6893203883495146
Alpha: 0.6000000000000001, Score : 0.686084142394822
Alpha: 0.7000000000000001, Score : 0.6925566343042071
Alpha: 0.8, Score : 0.6909385113268608
Alpha: 0.9, Score : 0.6909385113268608
```

Checking with other classifier algorithm like Decision Tree Classifier, Random Forest Classifier, Logistic Regression by using GridSearchCV and cross validation.

## DecisionTreeClassifier

A decision tree a tree like structure whereby an internal node represents an attribute, a branch represents a decision rule, and the leaf nodes represent an outcome. This works by splitting the data into separate partitions according to an attribute selection measure, which in this case is the Gini index (although we can change this to information gain if we wanted). This essentially means that we each split aims toreduce Gini impurity which measures how impure a node is according to incorrectly classified results.

## Random Forest Classifier

Random forest classifier **creates a set of decision trees from randomly selected subset of training set**. It then aggregates the votes from different decision trees to decide the final class of the test object.

## Logistic Regression

Logistic regression is a statistical analysis method used to predict a data value based on prior observations of a data set. Logistic regression has become an important tool in the discipline of machine learning. The approach allows an algorithm being used in a machine learning application to classify incoming data based on historical data. As more relevant data comes in, the algorithm should get better at predicting classifications within data sets.

| | model | best_score | best_params |
|---|---|---|---|
| 0 | svm | 0.660200 | {'C': 1, 'kernel': 'linear'} |
| 1 | random_forest | 0.616924 | {'n_estimators': 25} |
| 2 | logistic_regression | 0.707929 | {'C': 1} |
| 3 | decision_tree | 0.551388 | {'criterion': 'gini'} |

## Conclusion

In this paper, we proposed machine learning NLP techniques for classification of restaurant reviews. We removed stop words from the given dataset and apply stemming for efficiency.

The data was divided into two data sets as training and testing data sets. We have fitted many classifier to the training set and predicted the test results. Then we have plotted a confusion matrix for the test. From which we **can conclude that Logistic Regression is the best classifier with the accuracy of 69% precision of 0.75%.**

References

https://www.analyticsvidhya.com/blog/2021/02/sentiment-analysis-predicting-sentiment-of-covid-19-tweets/

https://www.kaggle.com/poulamibakshi/covid-19sentiment-analysis