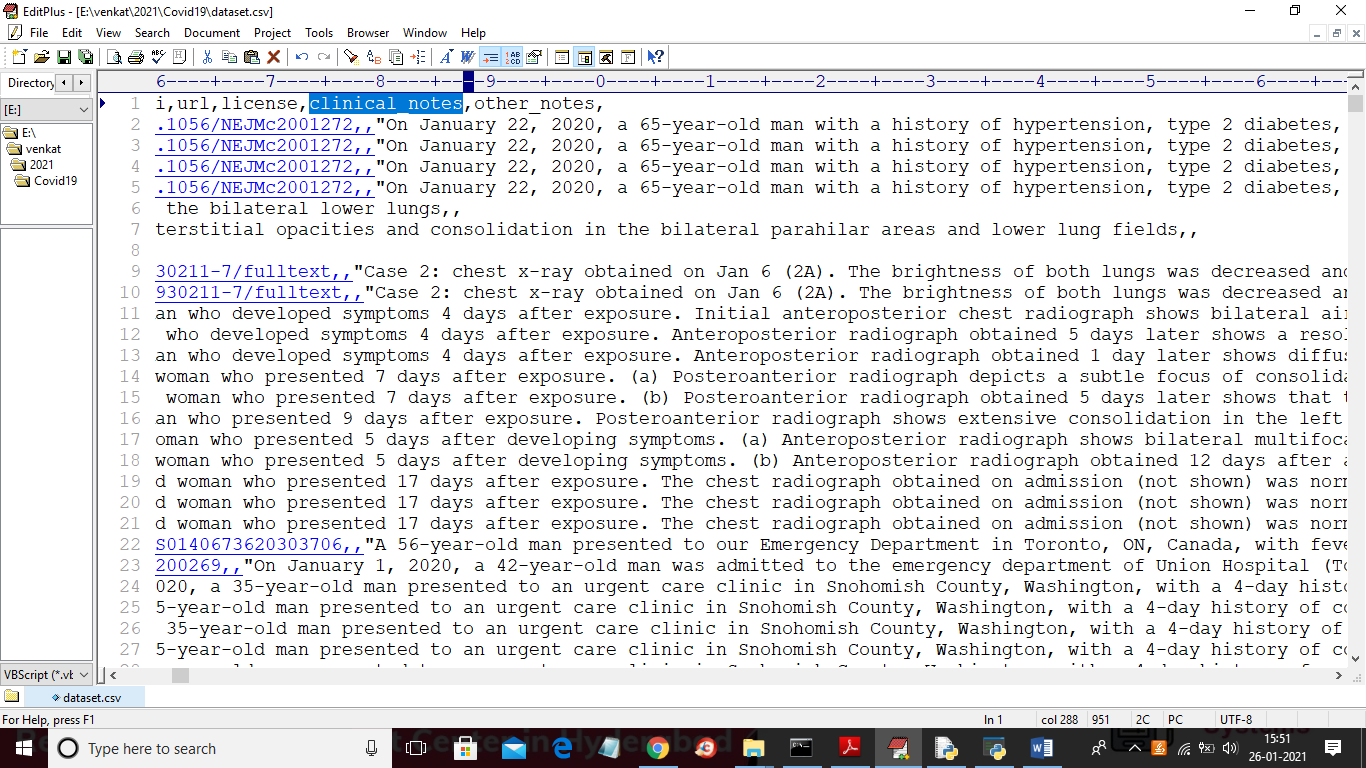
Machine Learning Based Approaches for Detecting COVID-19 using Clinical Text Data

In this project we are using traditional and classical machine learning algorithms to predict COVID-19 disease. In traditional algorithms we are using Logistic Regression, Naïve Bayes, SVM and Decision Tree and in classical algorithms we are using Bagging, AdaBoost, Random Forest and Stochastic Gradient Boosting classifier. In all algorithms Logistic Regression giving better performance.

To implement this project we are using Covid-19 text based dataset from below website

<https://github.com/Akibkhanday/Meta-data-of-Coronavirus>

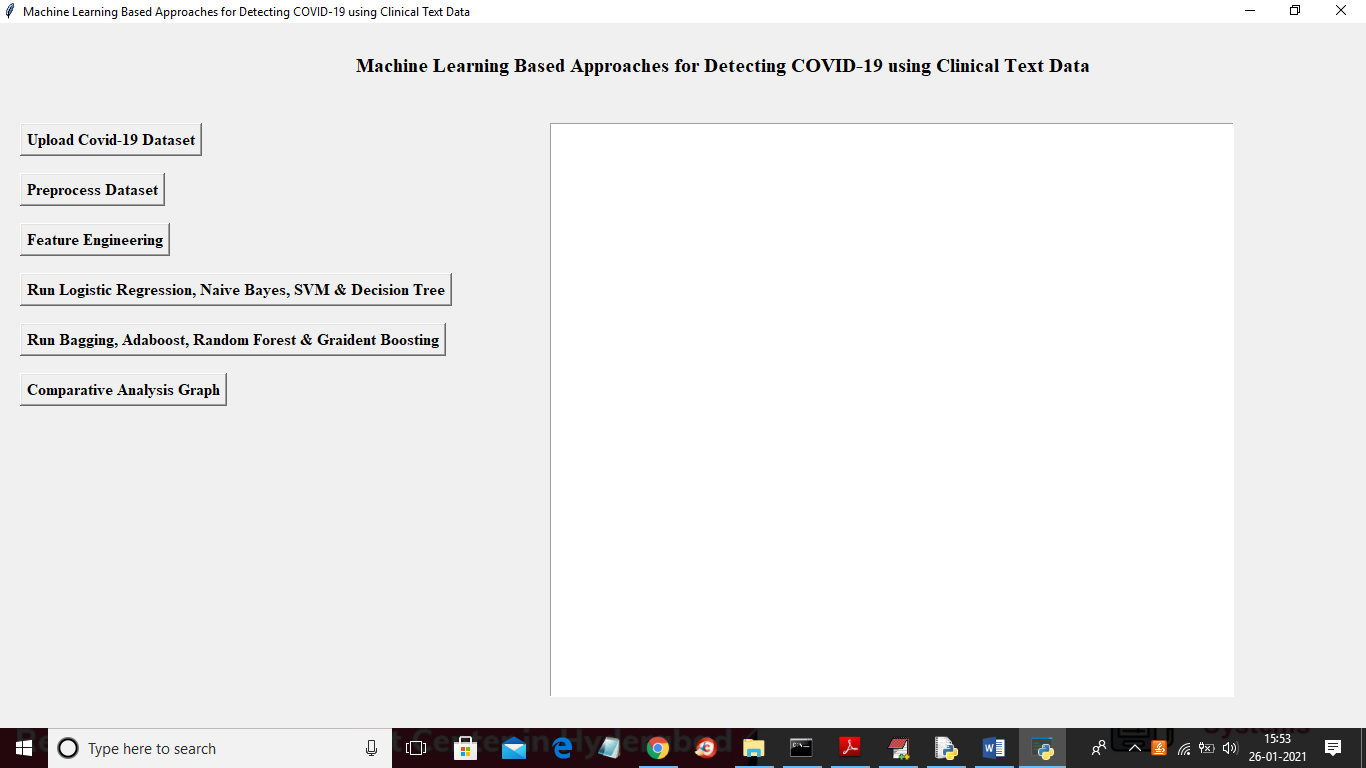
This dataset contains more than 30 columns but we are extracting two column values such as ‘clinical\_notes’ and ‘finding’. clinical\_notes column contains medical text data and this text data is preprocess using NLTK library to remove stop words, special symbols and then apply lemmatizer to remove ‘ing, tion etc.’ from text. After preprocess text we will apply TF-IDF to extract top 40 features from dataset. Below is the dataset screen shots and this dataset saved inside ‘dataset’ folder.



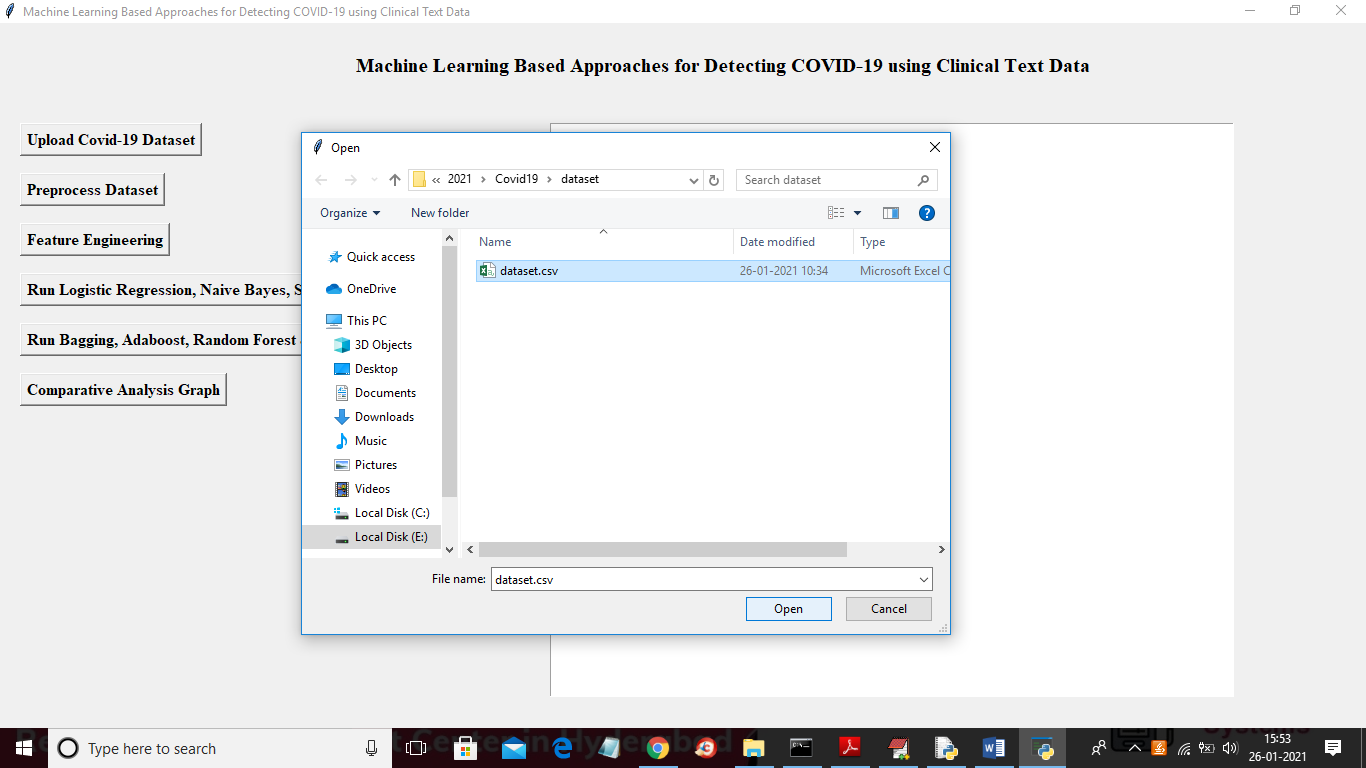
In above dataset we extracted clinical notes and then build features for machine learning algorithms and then apply various ML algorithms to calculate accuracy, precision, recall and FScore.

SCREEN SHOTS

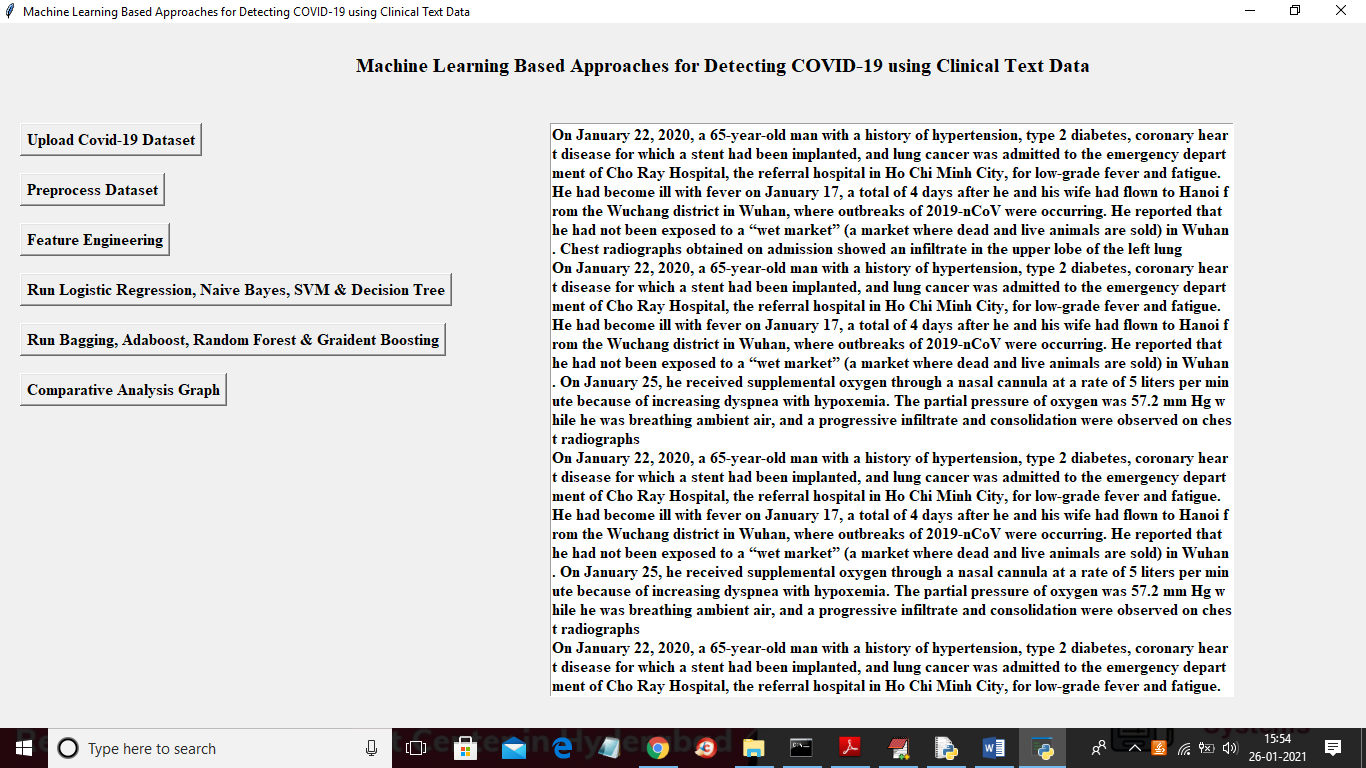
To run project double click on ‘run.bat’ file to get below screen



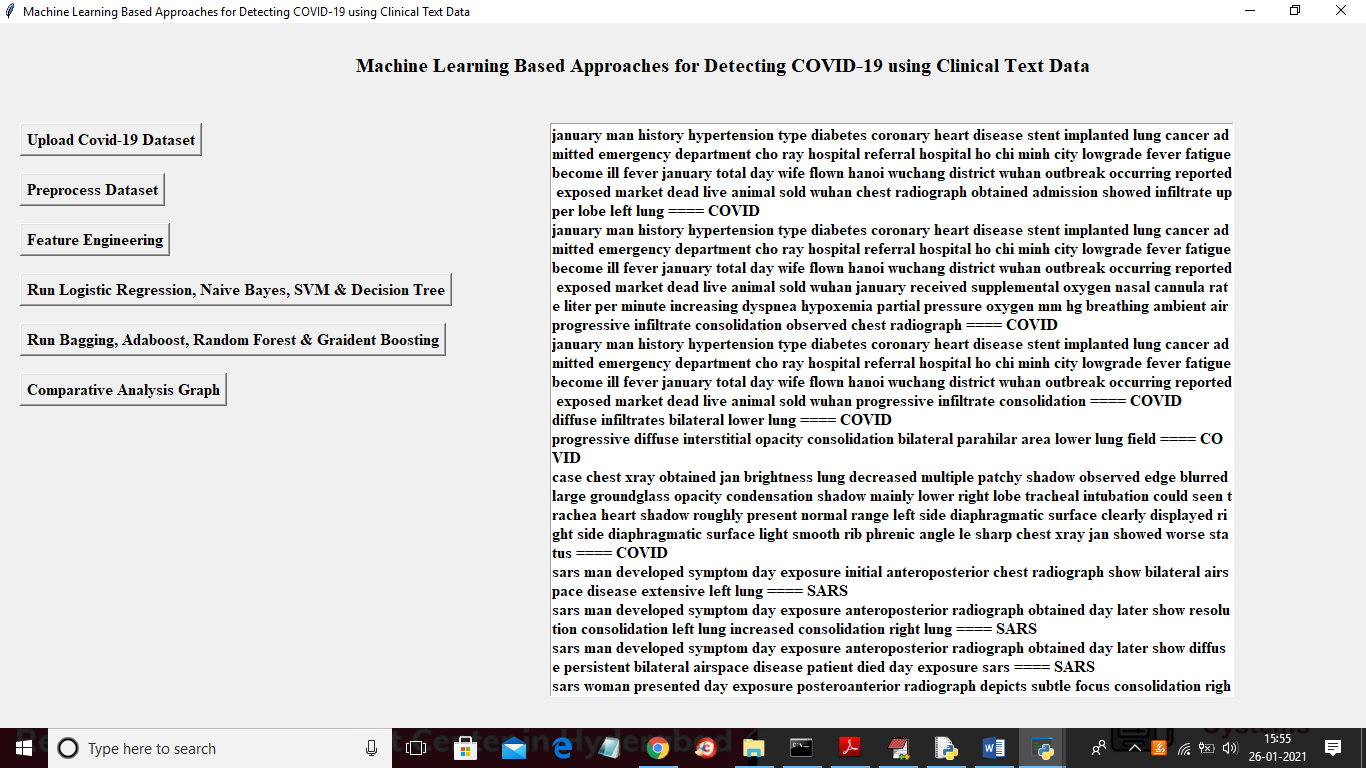
In above screen click on ‘Upload Covid-19 Dataset’ button and then upload dataset



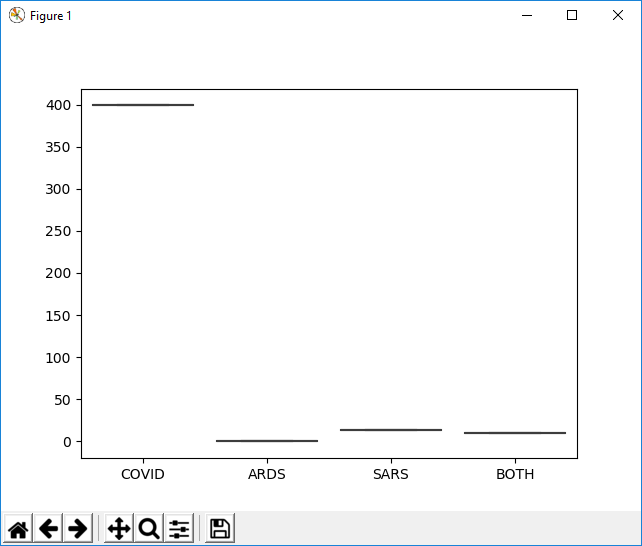
In above screen selecting and uploading ‘dataset.csv’ file and then click on ‘Open’ button to load dataset and to get below screen



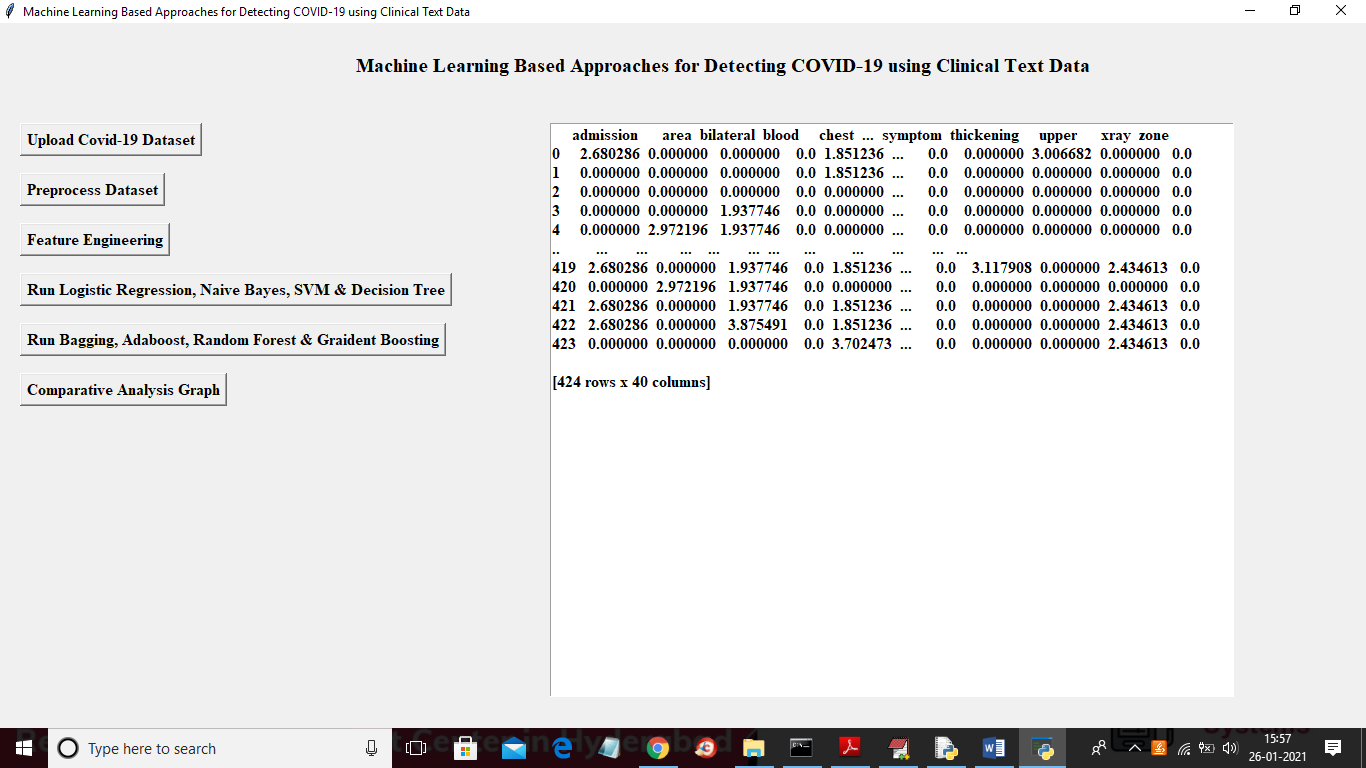
In above screen we extract all text data from dataset and now in above screen text in first sentence we have ‘on’ stop words and many number of numerical values and to remove those stop words and to clean data then click on ‘Preprocess Dataset’ button



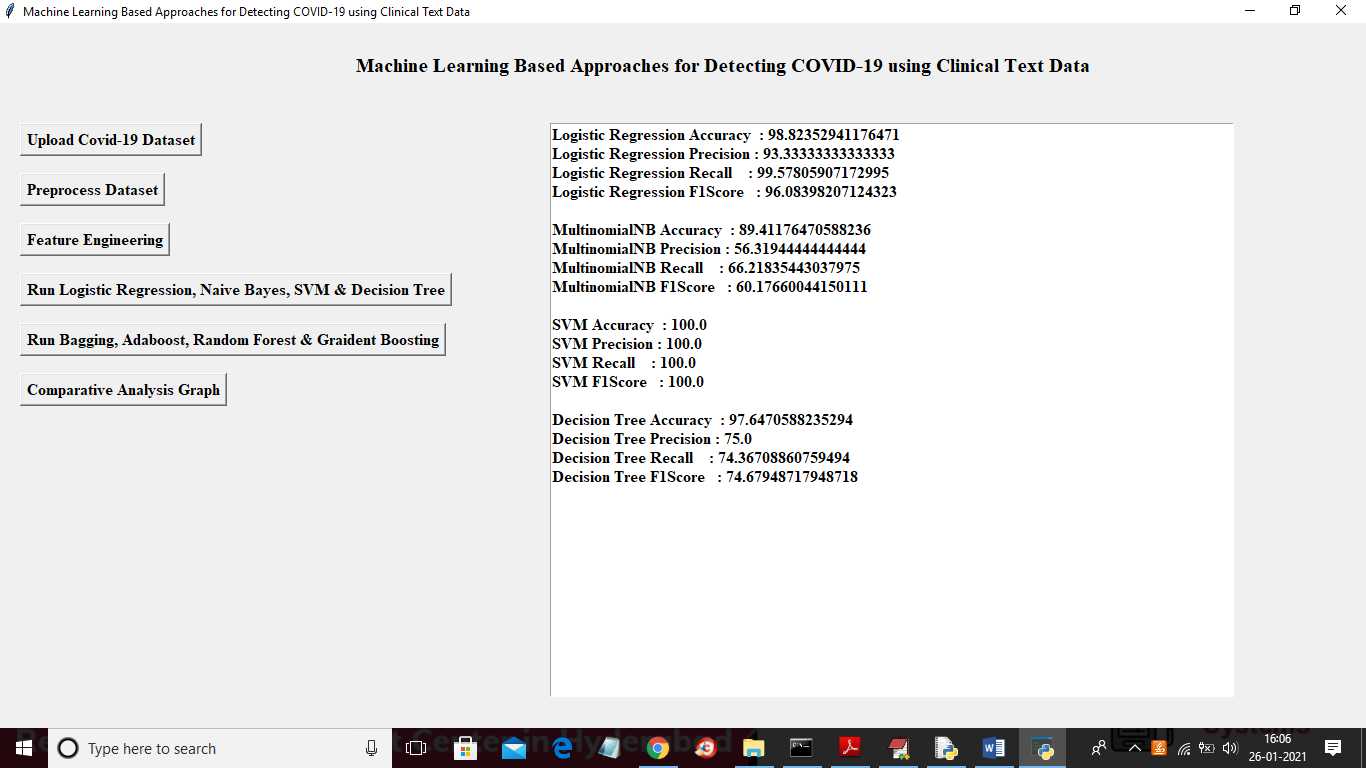
In above screen after preprocess all stop words removed out and in above ‘on’ stop word removed out and then will get below graph



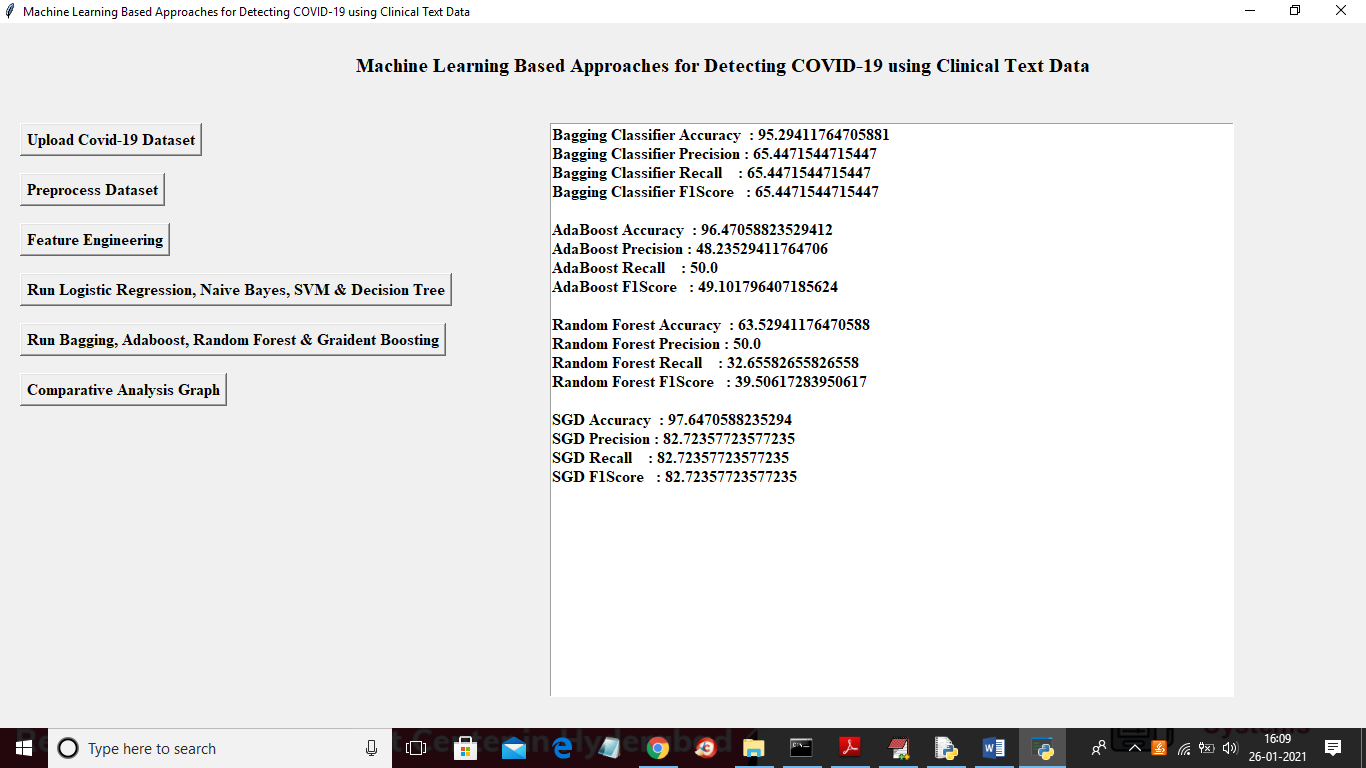
In above graph showing count/finding of each label and now close above graph and then click on ‘Feature Engineering’ button to apply TF-IDF on above text data and to get below features



In above screen all text data converted to above TF-IDF features and now click on ‘Run Logistic Regression, Naive Bayes, SVM & Decision Tree’ to run all traditional algorithms on features data and to calculate accuracy



In above screen displaying accuracy, precision, recall and FScore for each algorithm and now click on ‘Run Bagging, Adaboost, Random Forest & Gradient Boosting’ button to calculate accuracy of classical algorithms



Above screen showing classical algorithms accuracy and other metrics values and now click on ‘Comparative Analysis Graph’ button to get below graph



In above screen we can see accuracy, precision, recall and fscore for each algorithm in group bar chart and in above graph x-axis represents algorithm name y-axis represents values