Taranpreet Report

This document aims to explain the processes that provide assistance in the completion of the coding and experimentation section of the research. The coding/experimentation section is divided into three subsections that handle the following:

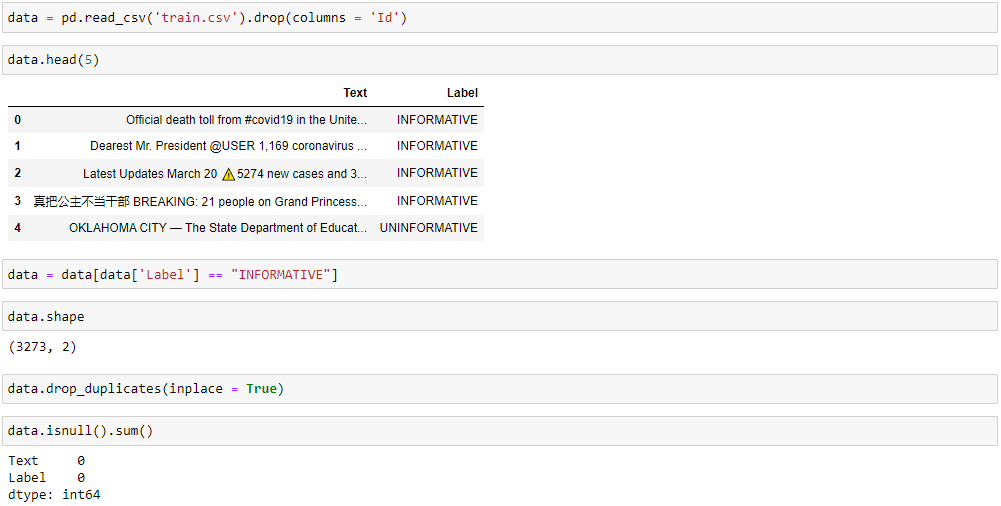
* Data Pre-processing
* Natural Language Processing
* Machine Learning Model

# Data Pre-processing

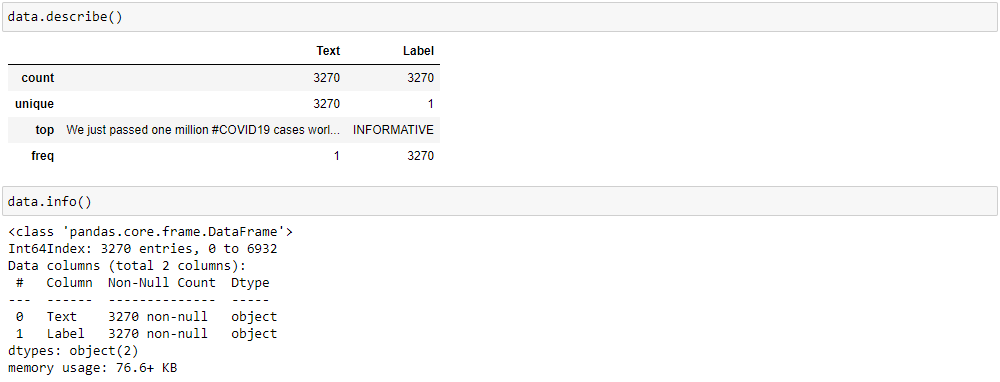
This section deals with the import of the dataset, studying it and adjusting it i.e. getting rid of the anomalies in it and preparing it for the training and experimentation. The features in the dataset are as follows:

Text: the tweet text content

Label: the output of the dataset indicating if the data row contains an informative tweet or not.



The above code does the data pre-processing section of the code. The code above imports the data into the code for usage using *pandas.* A sample of 5 instances from the dataset is displayed to understand the features in the dataset. The code also provides the shape of the dataset which is just the number of rows and columns in the dataset. The code then targets the duplicate and null instances in the dataset and removes them. The dataset has two labels i.e., “INFORMATIVE” and “UNINFORMATIVE”.



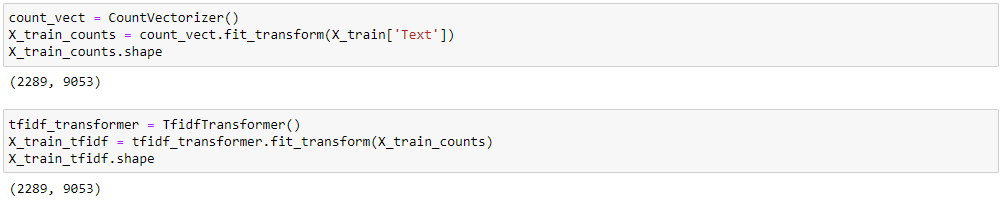
The dataset is also analyzed for its metadata i.e. data about its features like their data types and number of non-null features. The features are also analyzed statistically using “*data. Describe ()”.* The dataset is then divided into training and testing dataset in the 70:30 ratio for cross validation as follows:



This completes the data pre-processing section.

# Natural Language Processing

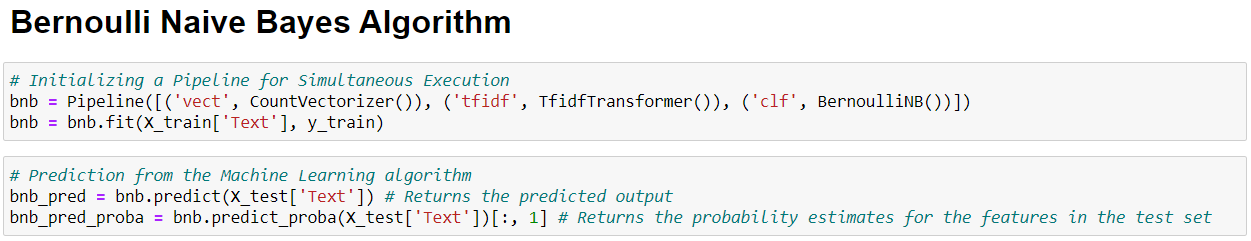
Since the paradigm followed for the code in this research is NLP based, the features have to be extracted from the dataset that satisfy the working of the NLP paradigm. The features here mean the keywords that will be extracted from the dataset to train the Machine Learning algorithm.

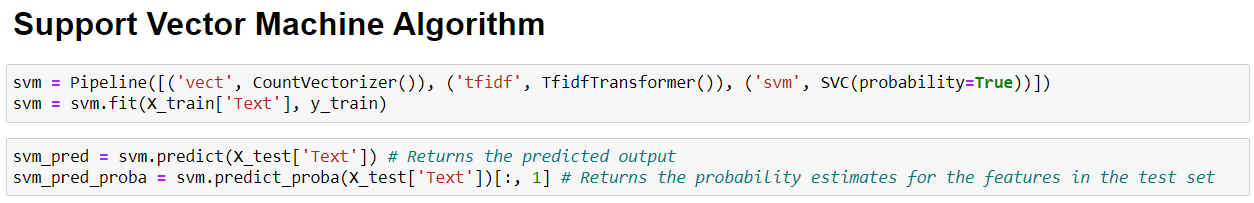


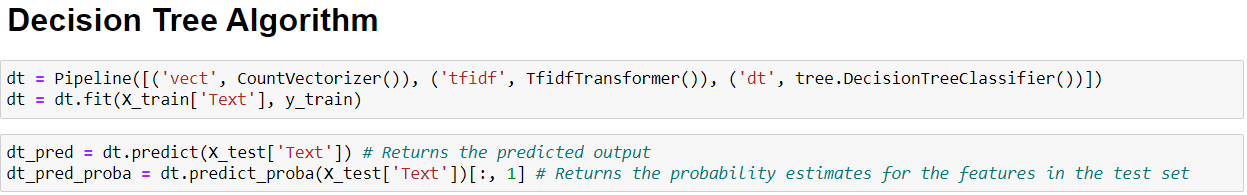
The code above performs the NLP processes for extracting keywords from the dataset to create a bag of words that will train the Machine Learning algorithm. The *CountVectorizer ( )* is used to extract the keywords from the dataset while the *TfIdfTransformer ( )* is used to transform the extracted bag of words into a td-idf representation. This completes the natural language processing section.

# Machine Learning

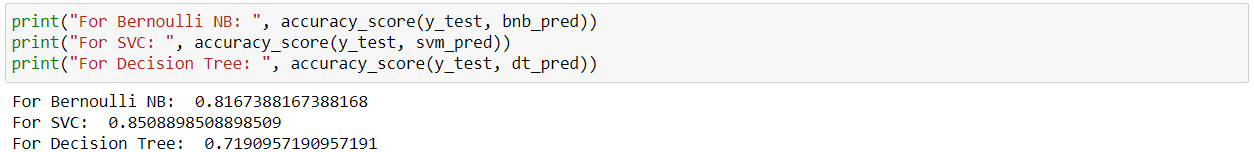
The algorithms used in this research are the Bernoulli Naïve Bayes Algorithm, The Support Vector Machine and the Decision Tree since the dataset has binary output and these algorithms specialize in classification. The algorithms are trained on the bag of words extracted above and the tested using the test dataset. The performance is measured using the mean method.

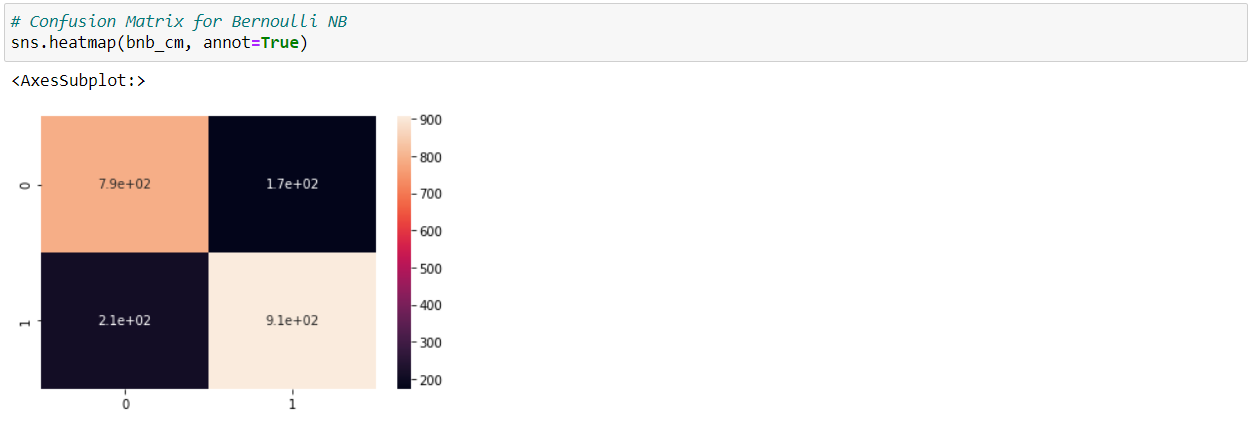


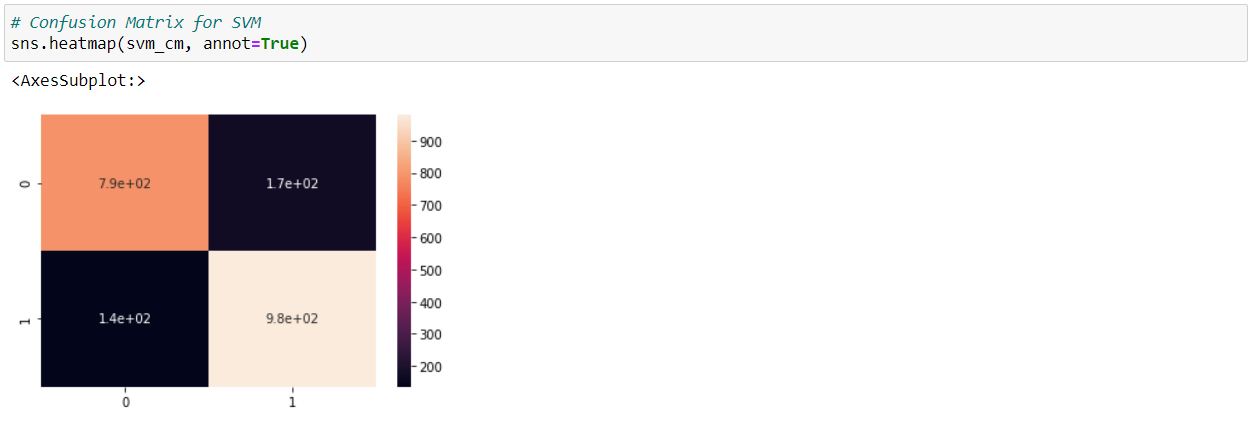


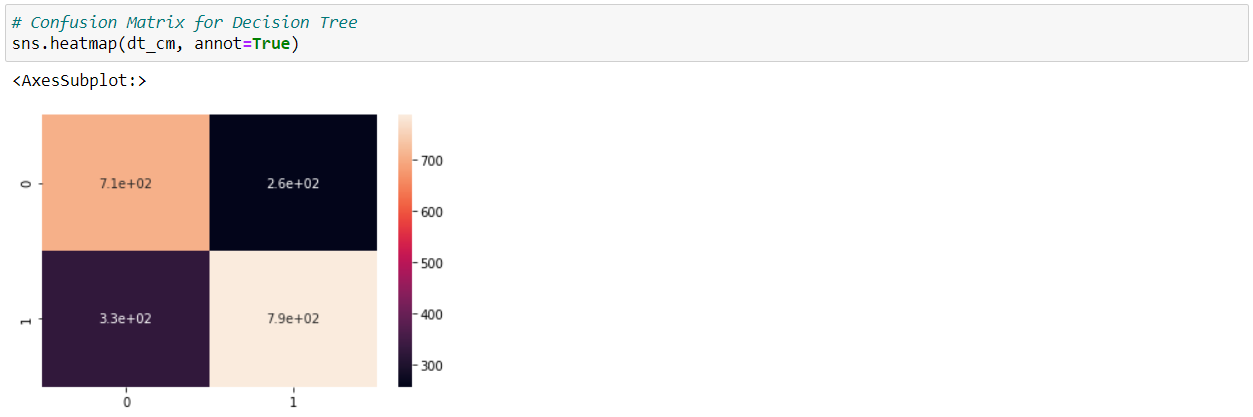


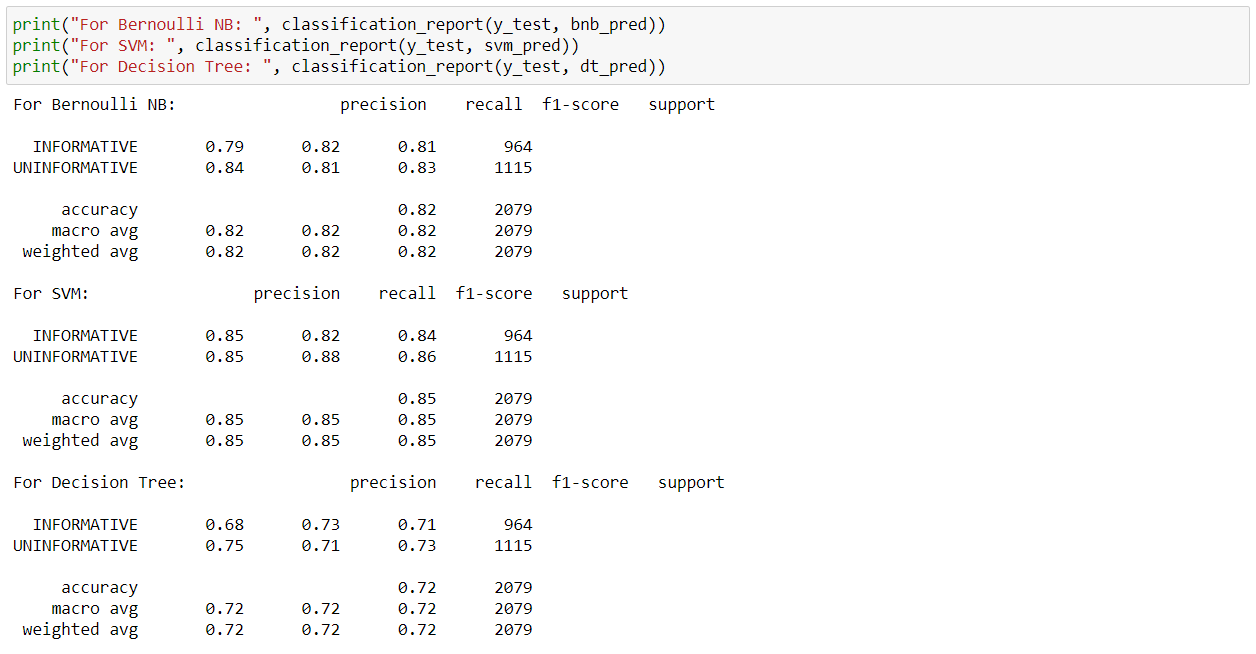
# Performance Metrics

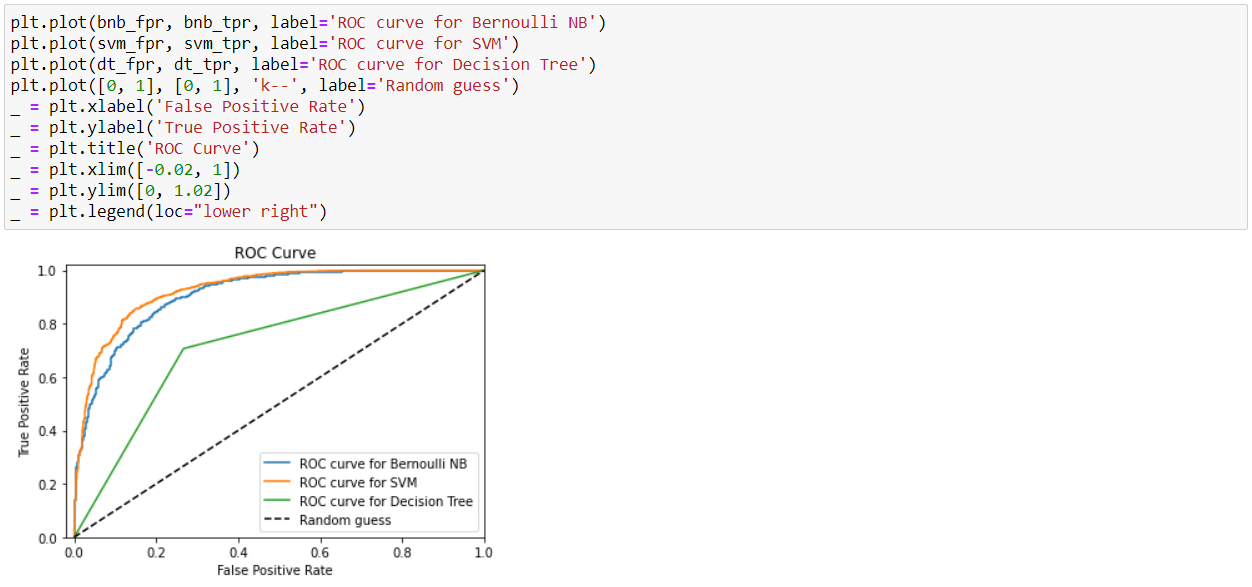


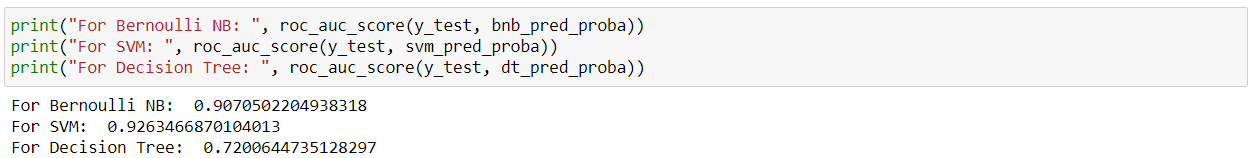












As the results show, out of the three algorithms, SVM performs the best giving 85% accuracy over the Bernoulli NB that gives a close 82% accuracy. The decision tree comes in last with an accuracy of 72%.