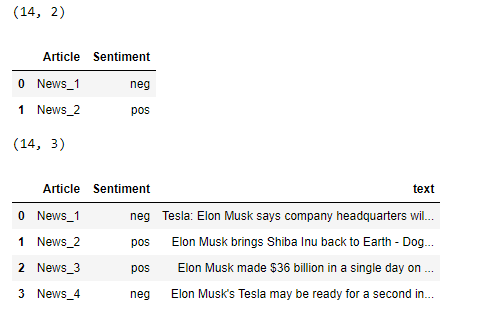
**Report**

**Part 1:**

In the part we need to find the last three days sentiment using the data of the last 11 days sentiment using the machine learning algorithm for this we have done the following steps. Firstly, we loaded and preprocess all the txt file number wise. Then we merge all files to create a new data frame.



Now we separate the data for training and testing through indexing. After splitting data apply the count vectorizer to transform data from textual to numeric form so the model can understand the data. Applying a Naive-Bayes Algorithm to Find the Predictions.



**Part 2:**

In part, two we create a list of positive and negative word files by using their text file through for loop and storing the results in a list. After this apply Frequency Distribution a function of count vectorizer which is used to count the number of each word.

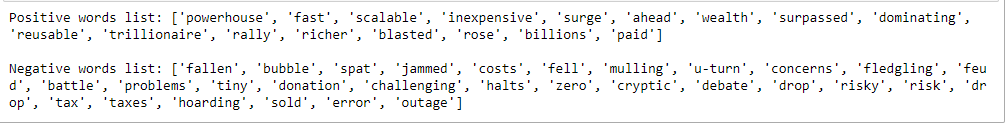


Figure : postive and negative word list

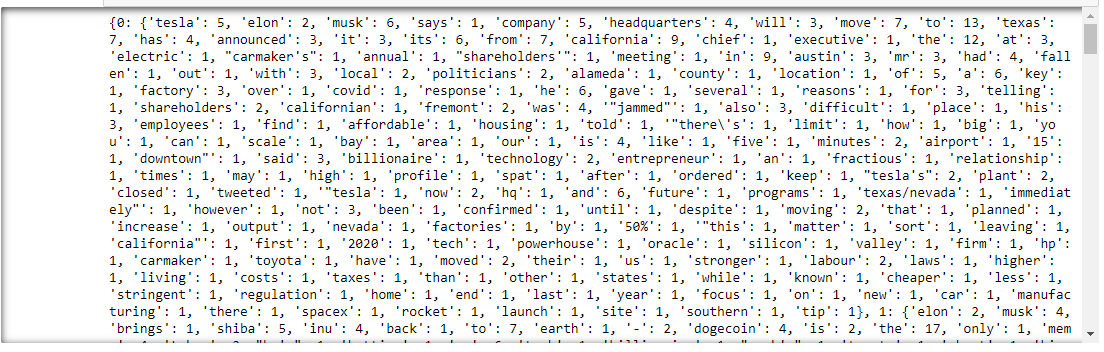


Figure : Word Frequency

Apply a proportion formula to find positive words for each article and negative words for each article and store the values in the list.

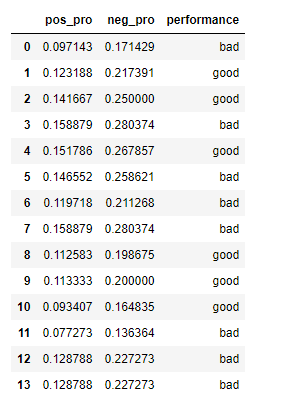


Figure : negative and Positive proportion dataframe

Import a CSV file through the panda's library in which the values of Return of Equity (ROE) and Net Profit Margin (NPM) are stored. Calculate a company's performance based on ROE and NMP. Create a data frame with features and labels. Separate a feature and labels in training and testing data through indexing.



Figure : Train and test data shape

Apply the first Logistic Regression Model to a data set to find the predictions. We use a Scikit-learn function feature importance to find an important feature of Logistic Regression. The second Model is the Random Forests Model, which is also used to find the predictions. We also use a Scikit-learn Regression feature importance to find an important feature of Random Forests.

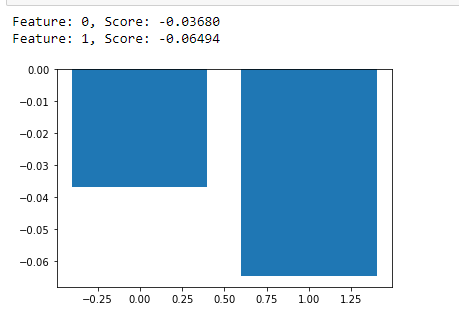


Figure : Feature Importance of Logistics Regression

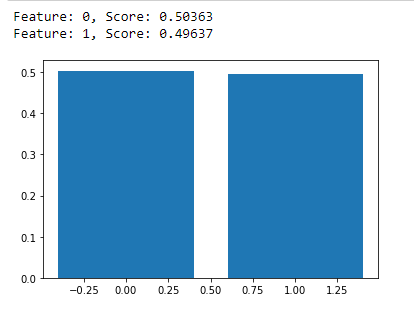


Figure : Feature Importance of Random Forest

Here from this project we have observed that approaches both are useful for the problem to get the sentiment of the text here from the count vectorizer. Here it predicts the unseen values as the negative values but by using the negative and positive proportion formula the calculated value have predicted the target value of the last three records as the positive value. It may due to the rule base or model base approach, It may also base upon the importance of the feature that is shown above and calculated by the feature importance function of the both Logistics regression and Random forest function. These function shows that the first feature have more importance then the second feature that is shown in the above graph.