Detect sarcastic comments on reddit

Muskan Gupta, Sheenu Mittal, Shreya Bansal, Tanya Sharma, Sejal Banta, Neeraj Rani

Indira Gandhi Delhi Technical University for Women, Delhi, India

Abstract. Sarcasm refers to the use of satirical or ironic language to convey a message. It is usually used in social networks such as Reddit, Twitter etc. The identification of sarcasm can improve the efficiency of sentiment analysis. Sentiment analysis refers to analyzing the attitude of people towards a particular topic or scenario. Our proposed method uses a supervised approach to learn the sarcastic patterns for classification.So our aim is to built model which helps in detecting sarcastic comments on Reddit using Logistic Regression.We'll be using the dataset from the paper "A Large Self-Annotated Corpus for Sarcasm" with 1.3 million comments from Reddit, labeled as either sarcastic or not.

1 Introduction

Sarcasm is employed to convey a meaning different than the literal one, usually in satirical context. It is complex and a bit difficult to comprehend since the actual message in the text has to be interpreted by the user. Sarcasm is utilized for humor as well as criticism of ideas, people or events. Sarcasm actually conveys a different meaning than the actual text. The presence of sarcasm can alter the polarity of the sentiment analysis. Hence, the detection of sarcasm can enhance and refine the existing system for sentiment analysis. We are working on the reddit dataset. Reddit is an online discussion platform, where the community members can post information regarding news, politics, hobbies etc and any other areas of interest. The areas of interests are categorized as subreddits (/news, /politics etc) .Our test data is from the subreddit sarcasm where the members post sarcastic comments which are denoted by appending a /s to the text. To detect the sarcastic comments, the frequently occurring sentence patterns are detected along with its syntactic, semantic and sentiment based features.

Problem Statement. *Sarcasm is the unconventional way of conveying a message which conflicts the context. It is both positively funny and negatively nasty, which plays an important part in human social interaction. So our aim is to built model which helps in detecting sarcastic comments on Reddit.*

2 Related Work

Mikhail Khodak,Nikunj Saunshi,Kiran Vodrahalli [**?**] proposed Self-Annotated Reddit Corpus (SARC), a large corpus for sarcasm research and for training and

evaluating systems for sarcasm detection. The corpus has 1.3 million sarcastic statements – 10 times more than any previous dataset – and many times more instances of non-sarcastic statements, allowing for learning in both balanced and unbalanced label regimes.They evaluated the corpus for accuracy, construct benchmarks for sarcasm detection, and evaluated baseline methods.

Soujanya Poria,Roger Zimmermann, Erik Cambria [**?**] proposed CASCADE (a Contextual Sarcasm Detector) that adopts a hybrid approach of both content and context-driven modeling for sarcasm detection in online social media discussions. For the latter, CASCADE aims at extracting contextual information from the discourse of a discussion thread. CASCADE utilizes user embeddings that encode stylometric and personality features of the users. When used along with content-based feature extractors such as Convolutional Neural Networks (CNNs), they see a significant boost in the classification performance on a large Reddit corpus.

In our project we will be using logistic regression and naive bayes to detect sarcastic comments on reddit.

3 Methodology

Given that project is data-driven, the data set is explained as: This dataset contains 1.3 million labeled comments from the Internet commentary website Reddit. The dataset was generated by scraping comments from Reddit containing the (sarcasm) tag. This tag is often used by Redditors to indicate that their comment is in jest and not meant to be taken seriously, and is generally a reliable indicator of sarcastic comment content. Data has balanced and imbalanced (i.e true distribution) versions. (True ratio is about 1:100). The corpus has 1.3 million sarcastic statements, along with what they responded to as well as many non-sarcastic comments from the same source.

3.1 Dataset Description

We have used data set from Kaggle website which is open source website for data sets https://www.kaggle.com/danofer/sarcasm

Table 1 is an example describing the dataset through counts of some key entities involved in the dataset.

Details	**Count**
Number of unique parent comments	984286
Number of unique comments	962294
Number of authors	256561

Table 1. Details of the dataset.

Every dataset also comprises of data attributes. Table 2 describes attributes of data. This project aims to find a solution using logistic regression, a form of supervised learning, hence the *label* in this project is determining if the comment is sarcastic or not. If the comment is sarcastic, a target value of 1 is assigned, else taget value of 0 is assigned.

Data Attributes	**Brief Explanation**
Comment	Reply to the parent comment that need to be stated as sarcastic/non sarcastic.
Author	User Id of the writer of comment
Subreddit	subreddit for the current comment
Score	Count of number of upvotes minus number of downvote
Ups	Count of upvotes
Downs	Count of downvotes
Date	Contains month and year of creation of comment
Created-utc	Timestamp for the comment
Parent-comment	Parent comment for reply comment
Label	Determines if the comment is sarcastic(1) or non sarcastic (0)

Table 2. Details of Data Attributes.

3.2 Data Pre-processing

The dataset is first tested to remove any null values, if existing. The sarcasm dataset taken from kaggle had null values in the comment column. The comment

Data Attributes	**Number of null values**
label	0
comment	53
author	0
subreddit	0
score	0
ups	0
downs	0
date	0
created-utc	0
parent-comment	0

Table 3. Null values in Data Attributes

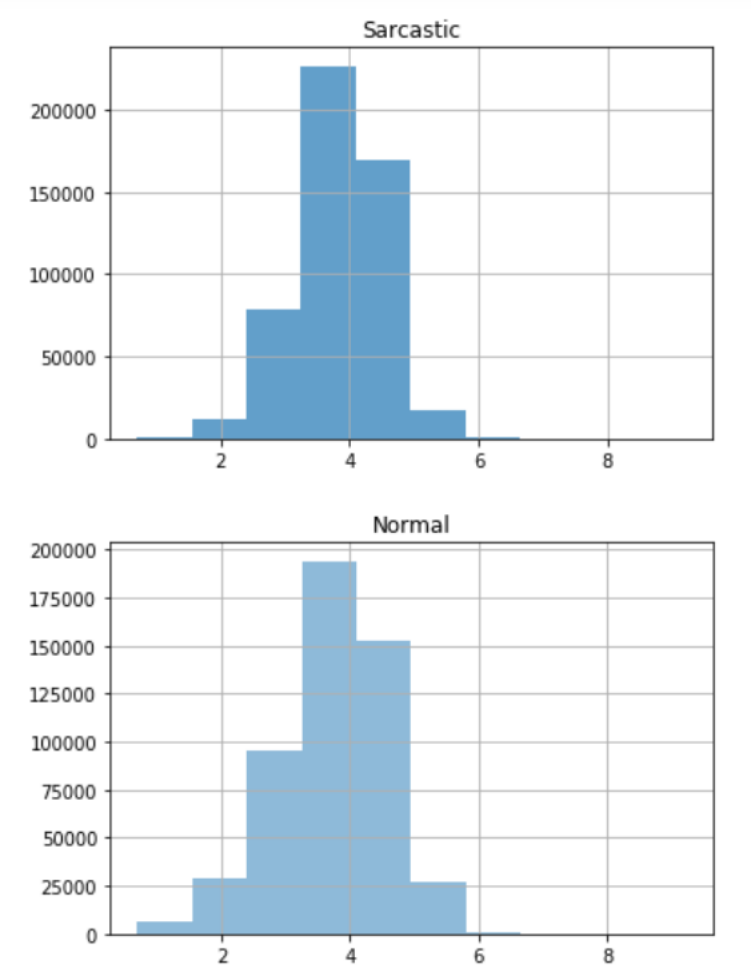
attribute had null values, which were removed from the dataset before processing the data. All the values are unique in the dataset. The dataset also contains outliers, which might be a consequence of some very popular or extremely unpopular comments.

3.3 Data Visualization

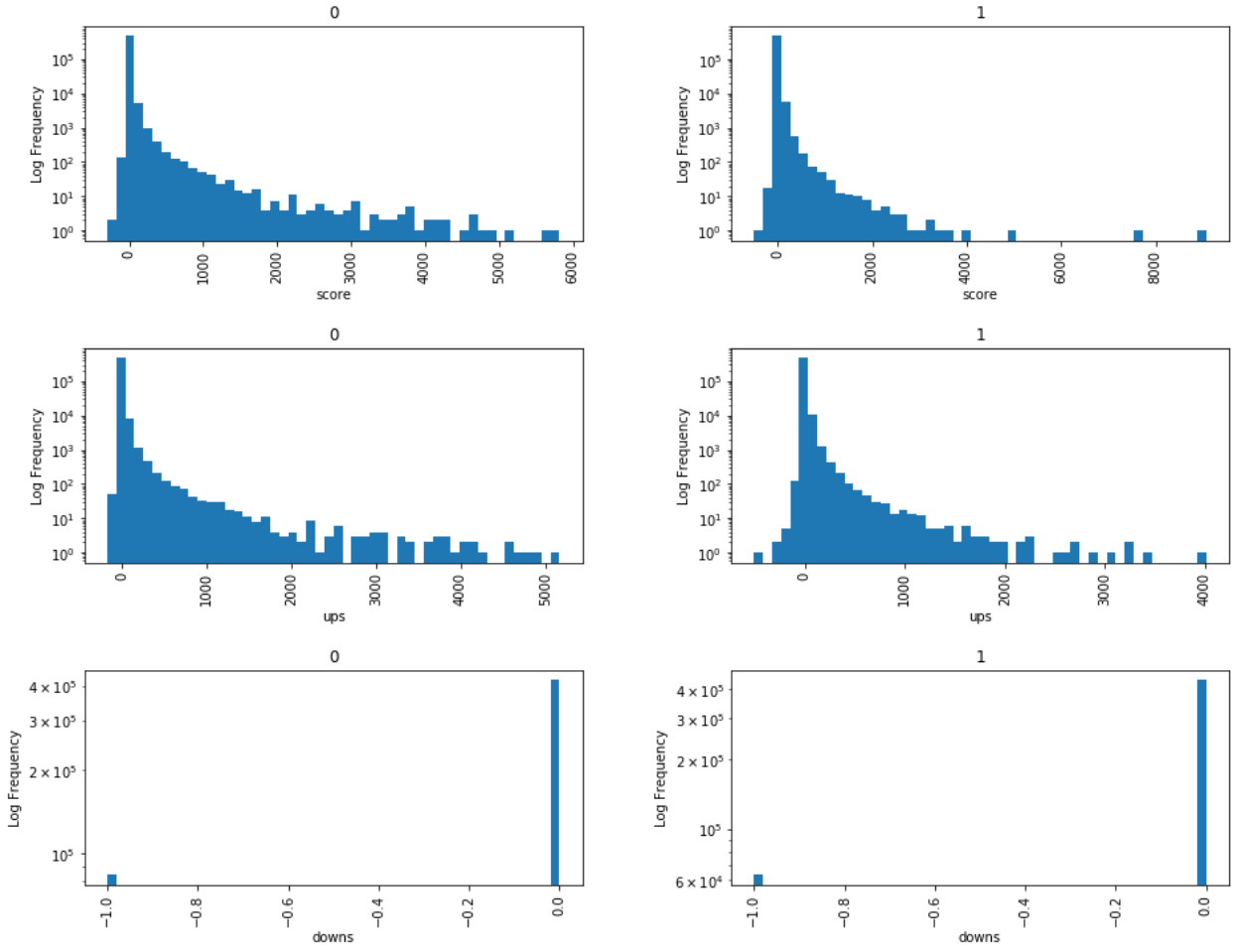
Step 1 First we observe the number of sarcastic and non-sarcastic comments.

Target Value	Number of rows
0	505405
1	505368

Step 2 We see the distribution of lengths for sarcastic and normal comments which is almost the same.

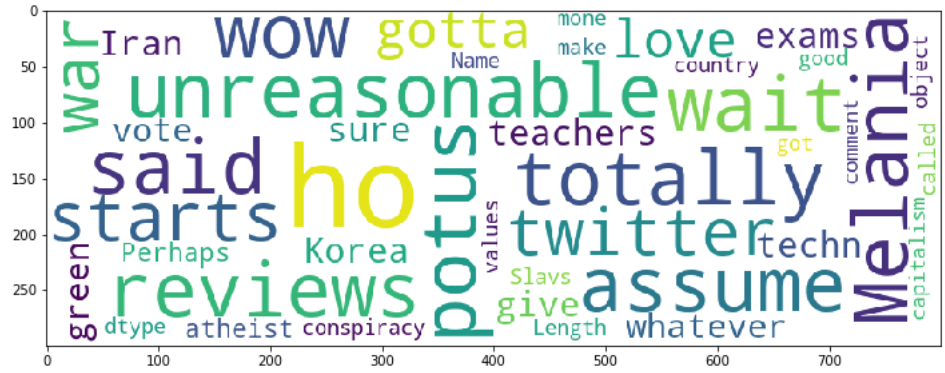


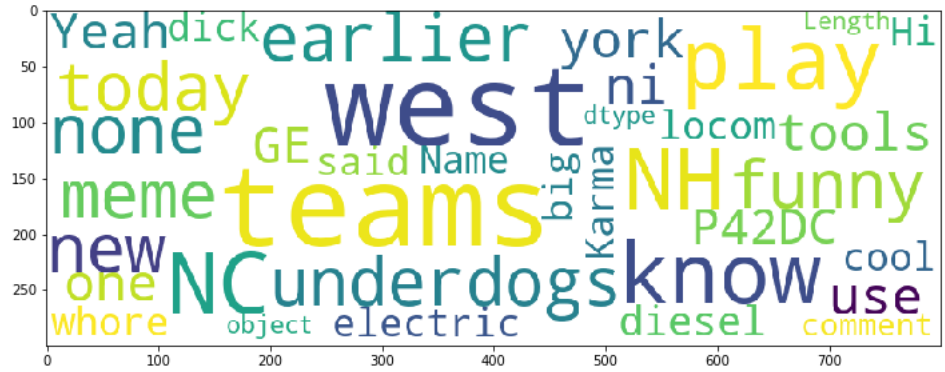
Step 3 Visualize data by plots that show frequencies of score, upvotes and downvotes for both the classes 0 and 1.



These graphs are frequency vs score, upvotes and downvotes.

Step 4 Word Cloud is a data visualization technique used for representing text data in which the size of each word indicates its frequency or importance. Significant textual data points can be highlighted using a word cloud. We now visualize the word clouds for both sarcastic and non-sarcastic words.





3.4 Proposed Approach

LOGISTIC REGRESSION

Logistic regression is a classification algorithm used to assign observations to a discrete set of classes. Some of the examples of classification problems are Email spam or not spam, Online transactions Fraud or not Fraud, Tumor Malignant or Benign. Logistic regression transforms its output using the logistic sigmoid function to return a probability value.

Types of logistic regression:

1. Binary (eg. Tumor Malignant or Benign)

2. Multi-linear functions failsClass (eg. Cats, dogs or Sheep's)

The hypothesis of logistic regression tends it to limit the cost function between 0 and 1. Therefore linear functions fail to represent it as it can have a

value greater than 1 or less than 0 which is not possible as per the hypothesis of logistic regression.

In order to map predicted values to probabilities, we use the Sigmoid function. The function maps any real value into another value between 0 and 1. In machine learning, we use sigmoid to map predictions to probabilities.

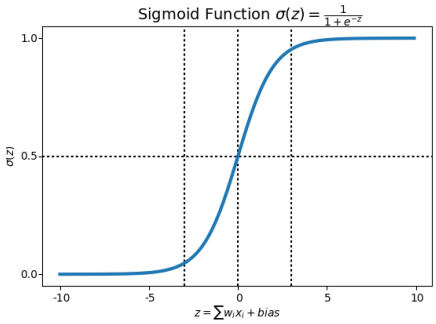


Fig. 1. Sigmoid function of graph

Example: We expect our classifier to give us a set of outputs or classes based on probability when we pass the inputs through a prediction function and returns a probability score between 0 and 1. For Example, We have 2 classes,

$$f(x) = \frac{1}{1+e^{-(x)}}$$

Fig. 2. Formula of a sigmoid function

let's take them like cats and dogs(1 — dog , 0 — cats). We basically decide with a threshold value above which we classify values into Class 1 and of the value goes below the threshold then we classify it in Class 2.

As shown in the graph below we have chosen the threshold as 0.5, if the prediction function returned a value of 0.7 then we would classify this observation as Class 1(DOG). If our prediction returned a value of 0.2 then we would classify the observation as Class 2(CAT).

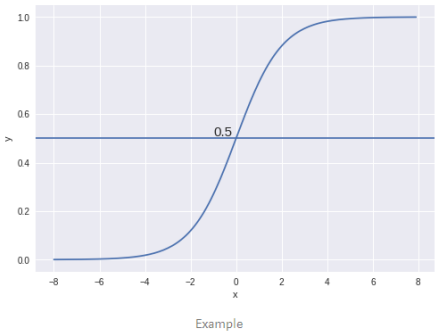


Fig. 3.

In our proposed project logistic regression is used to classify the comments as sarcastic or non sarcastic.It is selected as sentiment analysis is a binary classification. With the help of this classifier,massive datasets can be executed.In order to train classifier,a manually produce training set is used.An X:Y relation is provided in training set.The variable X represent the reddit comments whereas variable Y represents whether a comment is sarcastic(1) or non- saracastic(0).

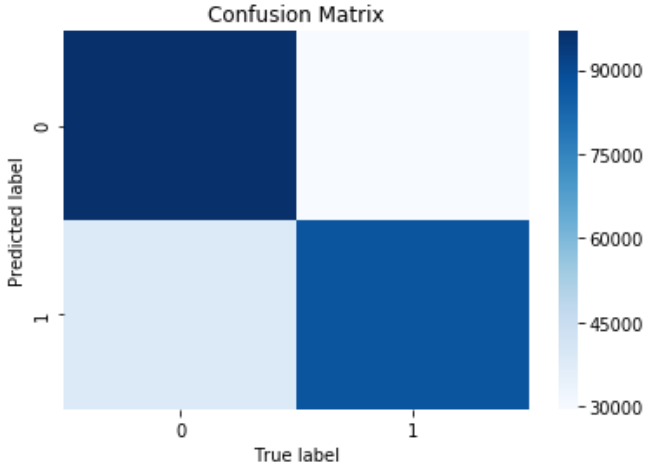


Fig. 4. Confusion Matrix

Accuracy of the test data is : 67.85

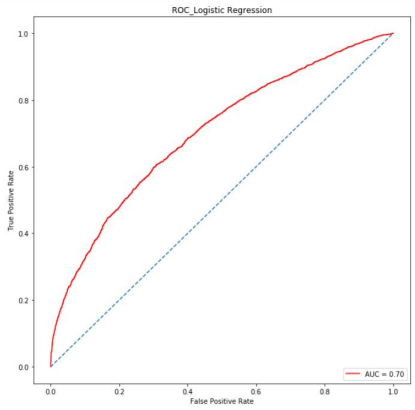


Fig. 5. ROC Curve

NAIVE BAYES

Naive Bayes is a simple but surprisingly powerful algorithm for predictive modeling. It is a statistical classification technique based on Bayes Theorem which provides a way that we can calculate the probability of a hypothesis given our prior knowledge. Naive Bayes classifier is the fast, accurate and reliable algorithm. Naive Bayes classifiers have high accuracy and speed on large datasets.

It is a classification algorithm for binary (two-class) and multi-class classification problems. The technique is easiest to understand when described using binary or categorical input values. Naive Bayes classifier assumes that the effect of a particular feature in a class is independent of other features.

Naive Bayes can be extended to real-valued attributes, most commonly by assuming a Gaussian distribution. This extension of naive Bayes is called Gaussian Naive Bayes.

Other functions can be used to estimate the distribution of the data, but the Gaussian (or Normal distribution) is the easiest to work with because you only need to estimate the mean and the standard deviation from your training data.

Accuracy of the test data is : 49

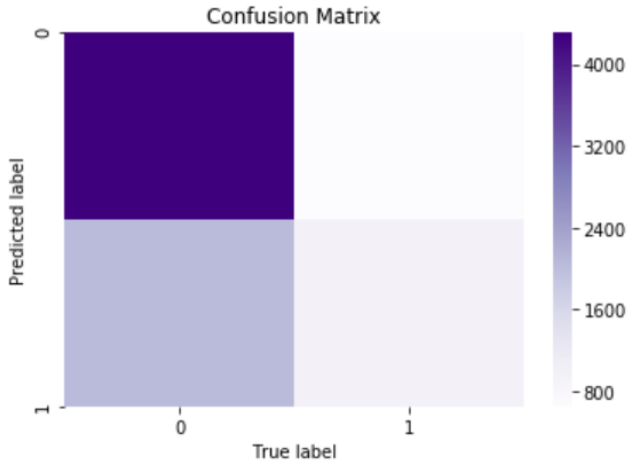


Fig. 6. Confusion Matrix

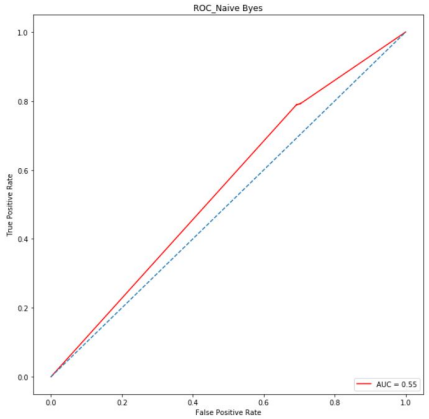


Fig. 7. ROC Curve

LINEAR SUPPORT VECTOR CLASSIFIER

Support vector machine algorithm isa supervised learning algorithm which is used to find a hyperplane in an N-dimensional space(N — the number of features) that distinctly classifies the data points.

Support vectors are data points that are closer to the hyperplane and influence the position and orientation of the hyperplane. Using these support vectors, we maximize the margin of the classifier. Deleting the support vectors will change the position of the hyperplane. These are the points that help us build our SVM.

In SVM, we take the output of the linear function and if that output is greater than 1, we identify it with one class and if the output is -1, we identify is with another class. Since the threshold values are changed to 1 and -1 in SVM, we obtain this reinforcement range of values([-1,1]) which acts as margin.

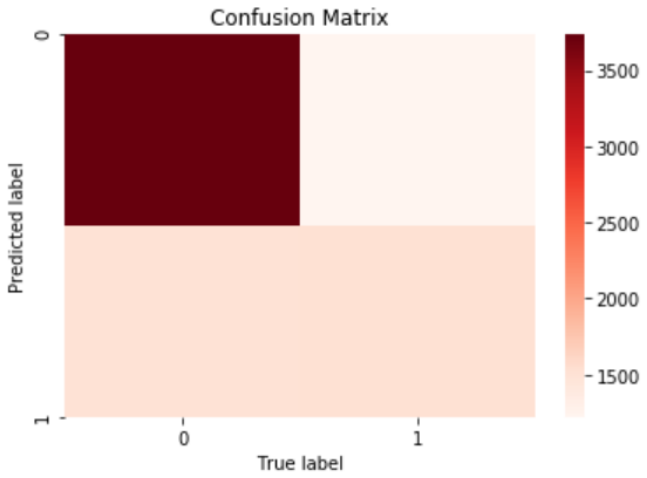


Fig. 8. Confusion Matrix

Accuracy of the test data is : 66.91

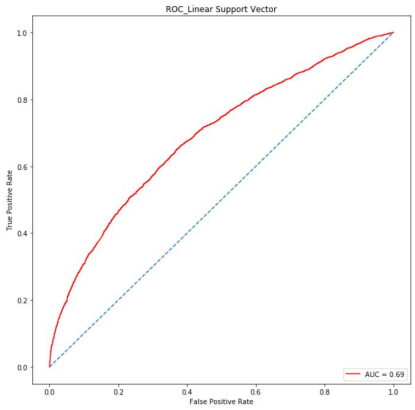


Fig. 9. ROC Curve

RANDOM FOREST CLASSIFIER

Random Forest is a supervised learning algorithm. It creates a forest and makes it somehow random. The forest it builds', is an ensemble of Decision Trees. Random Forest adds additional randomness to the model, while growing the trees. Instead of searching for the most important feature while splitting a node, it searches for the best feature among a random subset of features. This results in a wide diversity that generally results in a better model.

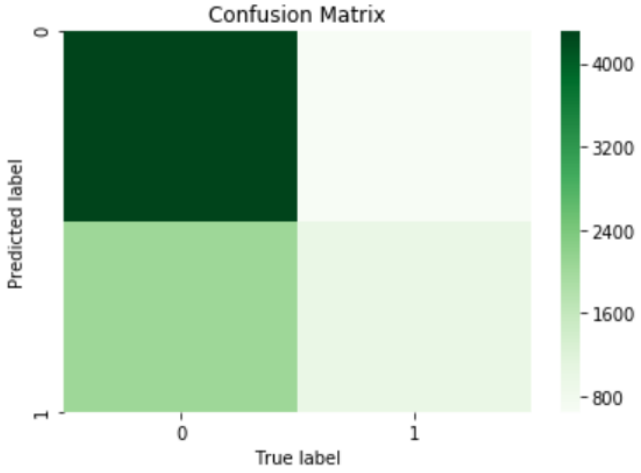


Fig. 10. Confusion Matrix

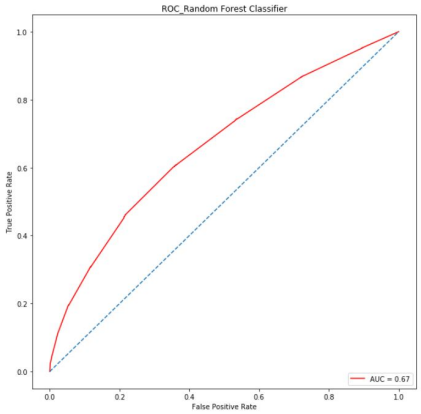


Fig. 11. ROC Curve

Accuracy of the test data is : 65.63

4 Result

Comparison of all the models with respect to their accuracy, precision and recall is shown as graphs below.

Accuracy Comparison Graph

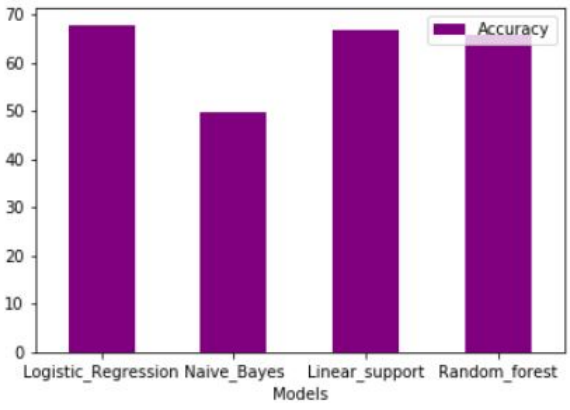


Fig. 12. Accuracy Graph

Precision Comparison Graph

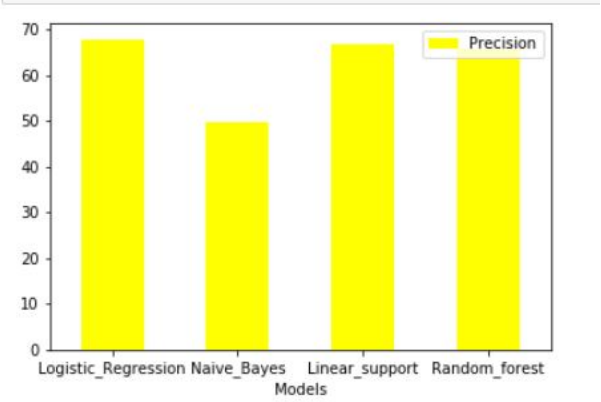


Fig. 13. Precision Graph

Recall Comparison Graph

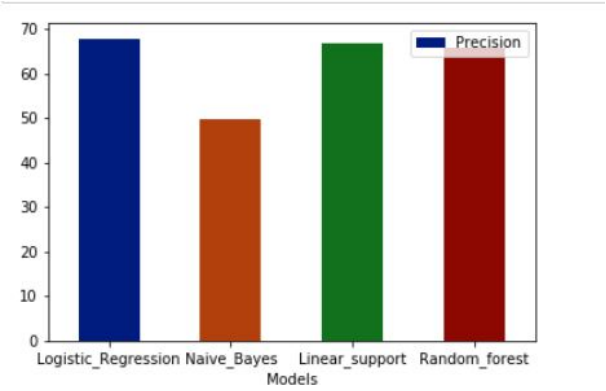


Fig. 14. Recall Graph

On comparing the four models, Logistic Regression has the best performance in terms of accuracy followed by linear support vector, random forest and naive bayes.

5 Future Work and Conclusion

Some future work that will be done later is:

- Detecting trending subreddits
- Evaluating author activeness
- Predicting trending hashtags