International Journal of Engineering Technology, Management and Applied Sciences

IJETMAS

www.ijetmas.com August 2017, Volume 5, Issue 8, ISSN 2349-4476

Predictive Analysis of Tweets on Goods and Services Tax(GST) in India using Machine Learning

Karthik Ganesan

Ramaiah Institute of Technology

Akhilesh P Patil

Ramaiah Institute of Technology

Srinidhi Hiriyannaiah

Ramaiah Institute of Technology

ABSTRACT

The Goods and Services Tax has been a revolutionary change in the financial standards of India. This led to a widespread debate across all social media platforms on the severity of its effects on the common man of this country. The very fact that many reactions collected on the social media regarding this topic has brought about the need to bifurcate the reactions based on their sentiments. Our goal is to not only classify the reactions based on the sentiments, but also to predict whether the upcoming tweets on this issue is on a positive note, or a negative note.

This analysis can be done by classifying the dataset using various Machine Learning Algorithms. The main goal is to conclude, which of the models used is the most accurate in predicting the outcome.

Keywords: Classification Models: Linear Regression, Support Vector Machines, Naïve Bayes, Decision Tree, Random Forest, XGBoost, RNN-LSTM. Receiver Operating characteristic curves, Word Embeddings.

1.INTRODUCTION

1.1 Why Social Media?

Over the past decade, the social media has taken the internet by storm, rapidly increasing the number of users by a large fold. Social media has also become a major part of the digital market over the years. It is now a substantial part of the marketing budget and continues to grow.

This rapid increase in the users has served both as a boon as well as a bane for the society. People from across the globe take to the social media to express their thoughts regarding a trending topic. We witness various discussions regarding important topics by the people of the world across different social media platforms. One can easily express his or her thoughts by simply tweeting about it or posting about it on Facebook. Other social media users can express their regard or disregard on the very same topic by retweeting it or by commenting on it. This way, we get a clear idea about the number of users in favour of the topic and the number of users against the matter in discussion.

1.2 What is Social Media Analytics?

Let us consider a scenario where in we are required to determine whether the discussions about a trending global topic is on a positive side or a negative side. Suppose we consider many tweets on the topic of discussion, it becomes a difficult task to go through each tweet.

Hence, we utilise the different analytics tools to gather the valuable hidden insights from vast amounts of semi-structured and unstructured social media data to enable informed and insightful decision making. It involves systematically identifying, extracting and analysing social media data such as tweets, shares, and likes.

Data on social media can be broadly classified into textual data such as tweets and comments, network data such as Facebook friendship network or the twitter follow-following network, and actions such as likes, shares, views, and retweets. Gathering this data, pre-processing it, and arriving at reliable and convincing insights on the data becomes an integral part of social media analytics.

1.3 The significance of Social Media Analytics.

Let us start with the basic definition of social media analytics: social media analytics is the collection, aggregation and standardisation of data available on social media to convey certain trends and patterns that follow. Social analytics helps us simplify dozens of data on the network of millions of people. This can help us gather a deeper understanding of user behaviour and demographic data than google analytics could ever do.

Social media analytics find great importance in understanding the patterns in large amount social data on a brand. It is important to understand the who's who of social media. Analysing the reactions from more important people is more significant to us rather than gathering the reactions from lesser important persons. Through social media analytics, it is possible to know more about the followers a user has on twitter or Facebook. It is also useful to analyse the competition on social media. This must do with numbers such as thenumber of followers, number of shares or the number of likes.

1.4 Social media analysis on various topics in India.

India being a vast country with diverse traditions and cultures, many topics come into the discussion over the social media with regards to India alone. We witness almost every day tweets being exchanged about a topic in a debate on a news channel. It is important that the common man has a say in the matters of this country, and through the social media, he can accomplish this task. A few of them may be of the most trivial of cases, while a few others may be of great importance. Here are a few topics over the social media which were trending in the recent times:

1. There is a huge surge in the number of tweets during the annual cricketing extravaganza - the Indian Premier League, held for over a month. The number of tweets collected over the social media goes up to billions during this time.

2. One of the biggest change in the financial history of India- the demonetization of certain currency notes also saw a major uproar over the social media.

3. The general elections, deciding the prime ministerial candidate witnessed many reactions on the social media.

4. Another major change in the financial status of India - the GST had also witnessed and is still witnessing a lot of reactions over the social media.

These major topics and many more provide a lot of scope for big data analysis through machine learning for data scientists across India.

1.5 Social Media Analysis through Machine Learning

As humans, it is possible for our brain to categorise content on social media, such as tweets into various categories. The question is, is it possible for computers also to differentiate these tweets into the very same categories as humans do? Yes, it is possible through Machine Learning to categorise the social media data.

Arthur Samuel, an early pioneer in the field of artificial intelligence, defined machine learning as **"the subfield of computer science that gives computers the ability to learn without being explicitly programmed."**

The idea about Machine Learning is to train the machine on a training set of data. Later using certain statistical models, we can test our results on a test set of data. What does this mean for social media analytics? Machine learning gives an analysis tool the ability to learn exactly what you're looking for in social media posts, and categorise posts based on that training. But why is this necessary? Why isn't there an algorithm that doesn't need to learn, it just already understands what I want it to find? The reason lies in the unstructured nature of social media data.

Structured data involves well-defined databases with rows and columns which make it easier to work with and hence easier to sort and categorise. Unstructured data is variable and complex, making it much more difficult to sort, categorise and analyse. Examples of unstructured data include emails, images, and any form of human language in a conversational format (like social media posts).

2. OBTAINING THE DATASET AND DATA PRE-PROCESSING.

The first step in obtaining data from Twitter is to create an application which interacts with the Twitter API and registering the application. Twitter provides REST APIs to interact with their services. Tweepy is the most interesting and straightforward to use. We have used Tweepy as a tool to access Twitter data with Python.

Further to start our analysis, we break the text into words, by tokenisation. The purpose of tokenisation is to split the stream of text into smaller units called tokens. But Twitter data pose some challenges because of the nature of the language. In every language, some words like conjunctions and some adverbs called stop words which do not affect the meaning of the sentence. We need to remove these stop-words. NLTK provides a simple list of English stop words.

3. CLASSIFICATION MODELS.

The various classification models used in the analysis are Logistic Regression, Naïve Bayes, Support Vector Machine, K-Nearest Neighbours, Random-Forest Classification, Decision Tree Classification, XGBoost and LSTM. Observing the ROC curves and the confusion matrices of the various models we will be able to conclude, which model is the best fitting model.

3.1RECEIEVER OPERATING CHARACTERISTIC CURVES.

In statistics, a receiver operating characteristic curve, i.e. ROC curve, is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied.

Let us consider a two-class prediction problem (binary classification), in which the outcomes are labelled either as positive (p) or negative (n). There are four possible outcomes from a binary classifier. If the outcome from a prediction is p and the actual value is also p, then it is called a true positive (TP); however, if the actual value is n then it is said to be a false positive (FP). Conversely, a true negative (TN) has occurred when both the prediction outcome and the actual value are n, and false negative (FN) is when the prediction outcome is n while the actual value is p.

3.2 Micro-Average and Macro-Average ROC Curves.

ROC Curves are generally used to study the output of a binary classifier. However, to extend the ROC curve area to multiclass or multi-label classification, it is necessary to binarize the output. One can draw a ROC curve by considering each element of the label indicator matrix as a binary prediction. This is micro-averaging.

On the other hand, macro-averaging gives equal weight to the classification of each label.

3.3 LOGISTIC REGRESSION

We know that the equation, $y = b0 + b1 * x$ fits a straight line, and by using a linear classifier we can predict whether the data point lies on the line or not. In the above equation y is the dependent variable which we are predicting based on x which is the independent variable.

Whereas, logistic regression uses the equation,

$p = \frac{1}{1+e^{-x}}$ where p is the probability of the event. Here p defines whether the sentiment is on a positive side or a negative side. On applying log on both sides of the equation we get,

$$\ln p = b0 + b1 * x$$

This gives a sigmoid curve between 0 and 1 (i.e. the positive and negative reactions).

3.3a Receiver Operating Characteristic Curve.

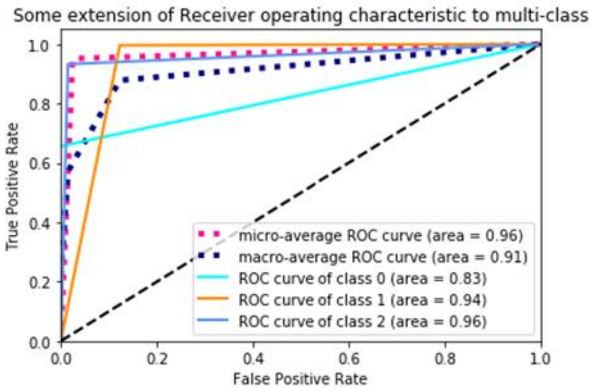


Fig 1: Taking the AUC for micro average and macro average we get around 94%.

3.4 K NEAREST NEIGHBOURS

In the implementation of this classification model the positive and the negative tweets are grouped together separately. Suppose we want to predict for a new data point we perform the following: we compute the Euclidean Distance from this data point to the K neighbours chosen using the formula for Euclidean distances. The distance closest to a group will be chosen for allocating the point to.

3.4a Receiver Operating Characteristic Curve

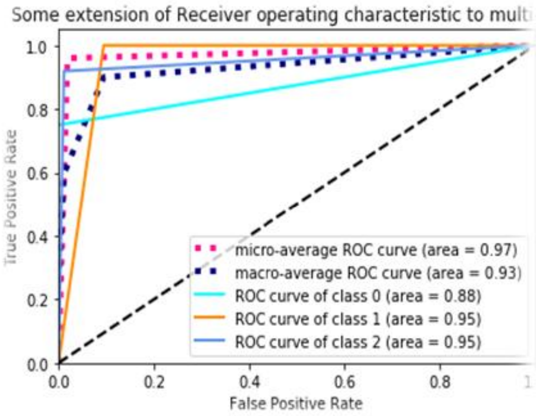


Fig 2: The average of the micro average and macro average AUC for the KNN model comes around 95%.

3.5 SUPPORT VECTOR MACHINES(SVMS)

In SVM's two data-points (support vectors) from each of the two categories of the data set are chosen, in such a way that the distance from that point to a line drawn separating the categories is maximum and equidistant. In this model, the machine tries to learn from the most likely positive or negative tweets. The least likely ones are chosen as support vectors.

While SVM is a linear classifier which uses a straight line to classify the two classes, the Kernel SVM is a non-linear type which uses characteristic curves and irregular boundaries to separate the classes.

3.5.1Receiver Operating Characteristic Curve.

(a) SVM

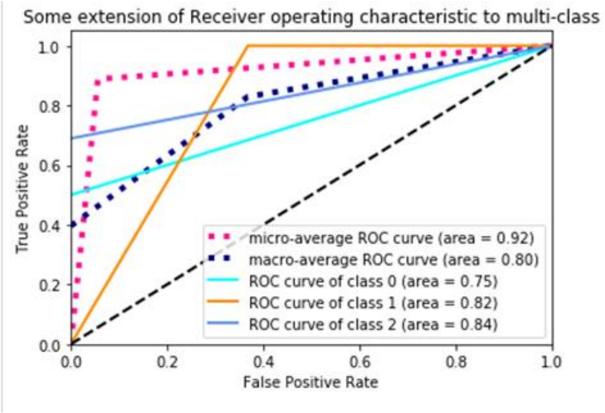


Fig 3: Calculating the average of the micro average and macro average AUC is around 86%.

(b) Kernel SVM

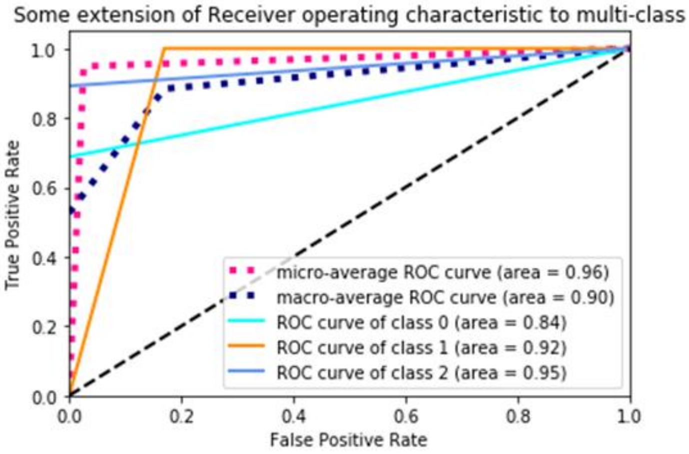


Fig 4: The average of the micro average and macro average AUC is 93%.

3.6 NAÏVE BAYES CLASSIFIER

We can use the Bayes Theorem given by:

$$P(A|X) = \frac{P(X|A) * P(A)}{P(X)}$$

Where P(A) is the Prior Probability, P(X) is the Marginal Likelihood, P(X|A) is the Likelihood, P(A|X) is Posterior Probability. Here X is the class taken into consideration. P(A|X) is the probability of occurrence of A given X and P(X|A) is vice versa.

We compare the probabilities for P(Positives|X) with the probabilities for P(Negatives|X) and classify the new tweet accordingly (i.e. whichever of the above two is greater).

3.6.1Receiver Operating Characteristic Curve.

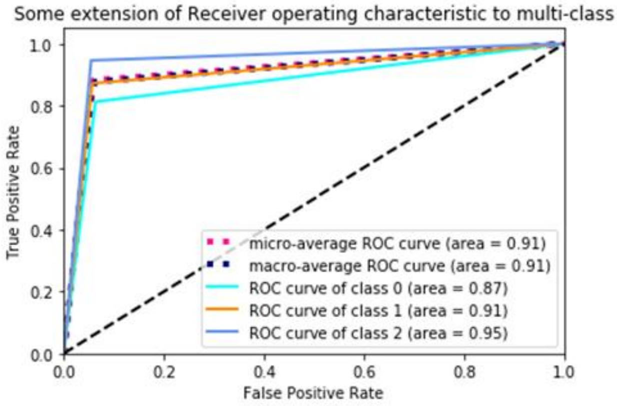


Fig 5: Taking the AUC for micro average and macro average we get around 91%.

3.7 DECESSION TREE AND RANDOM FOREST CLASSIFIERS.

Consider various groups of positive and negative data points on a 2-D graph with x1, x2 as the two classifiers on the horizontal and vertical axes respectively. The decision tree splits these groups and finally on making a few branches we can conclude, stating as to how the new tweet can get classified as.Random Forest is a collection of decision trees. The random forest algorithm takes the average of all the decision trees taken into considerations and gives a much more accurate result.

3.7.1Receiver Operating Characteristic Curve.

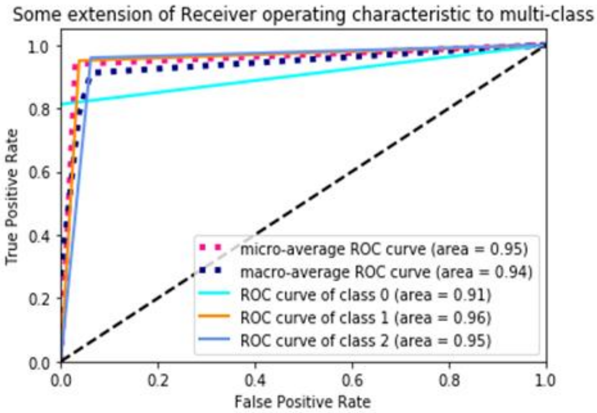


Fig 6 :The average of the micro average and macro average AUC for this model comes around 94.5%.

3.8 XGBOOST CLASSIFIER

XGBoost belongs to a family of boosting algorithms that convert week learners into strong learners. A week learner is one which is slightly better than random guessing. Boosting is a sequential process: i.e. trees are grown using the information from a previously grown tree one after another. This process slowly learns from the data and tries to improve its prediction in subsequent iterations.

XGBoost can be used to solve both classification as well as regression problems. To solve our problem, we use the **booster = gbtree**parameter, i.e. atree is grown one after other and attempts to reduce misclassification rate in subsequent iterations. Here the next tree is built by giving a higher weight to misclassified points by previous tree.

We push the statistics gi and hi to the leaves they belong to, sum the statistics together, and use the formula to calculate how good the tree is. This score is like the impurity measure in a decision tree used to build the structure of the tree. This is depicted in the below figure.

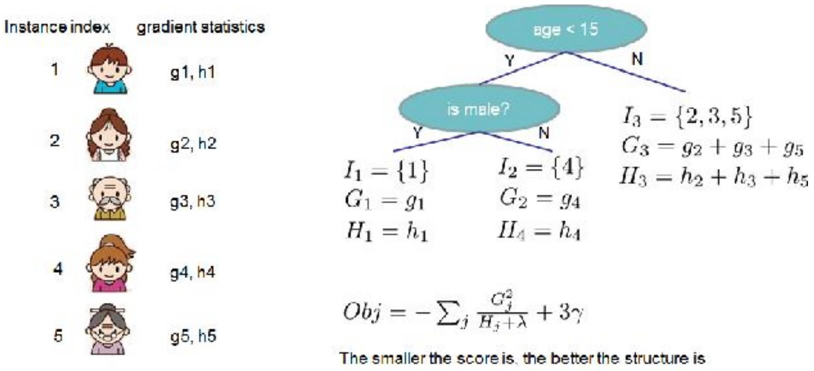


Fig 7: Classifier Score

3.8.1Receiver Operating Characteristic Curve.

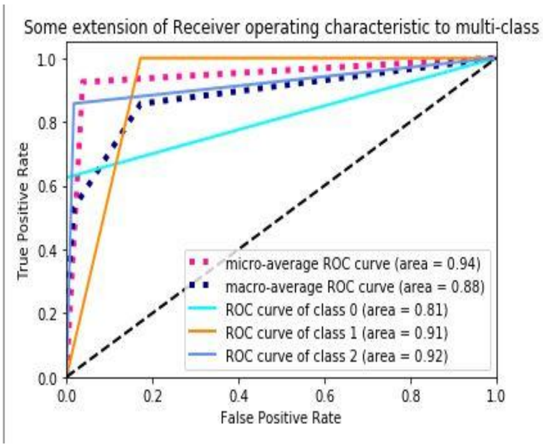


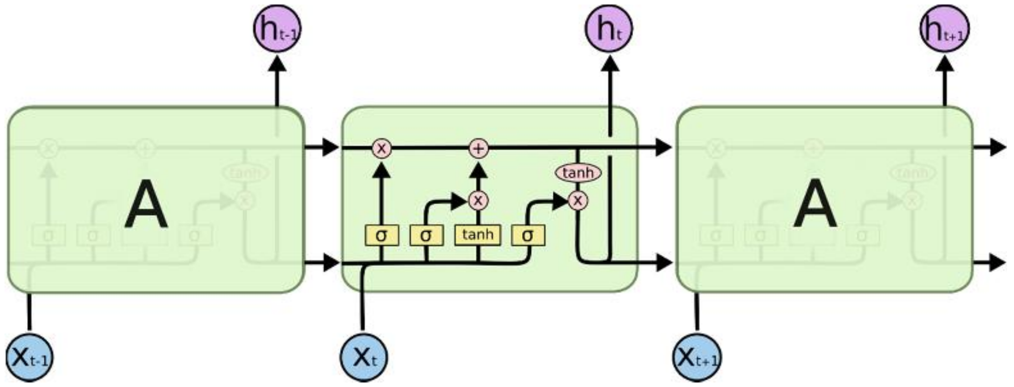
Fig 7:The average of the micro average and macro average AUC is 90.5%.

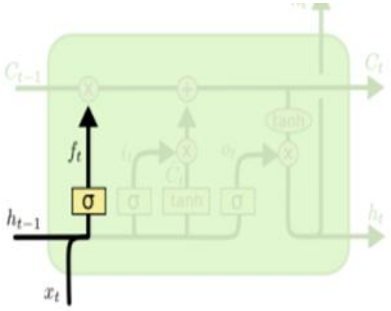
3.9 LSTM NETWORKS

Long Short-Term Memory networks are a special kind of Recurrent Neural Networks, capable of learning long-term dependencies. They are explicitly designed to avoid long-term dependency problem.

In our problem where we predict the sentiment of the tweets, the whole idea here is that the tweets are made of a sequence of words and order of words encode a lot of information about the sentiments.

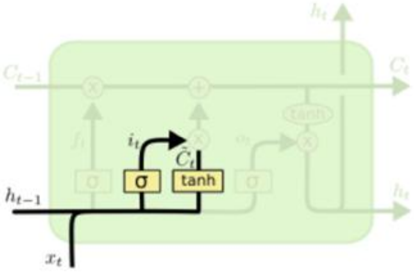
Step1 is to map the word to word embeddings.Step2 is the RNN that obtains the vectors as input and considers the order of vectors to generate predictions. From the embedding layer, the new representations will be passed to LSTM cells. These will add recurrent connections to the network so we can include information about the sequence of words in the tweets collected. Finally, the cells of the LSTM will go to sigmoid output layer. We use a sigmoid because we are trying to predict if the tweet is negative or positive. LSTMS have chain like structure, the repeating module has a different structure. Instead of having a single neural network layer, there are four, interacting in a very special way. This is represented in the below diagram.





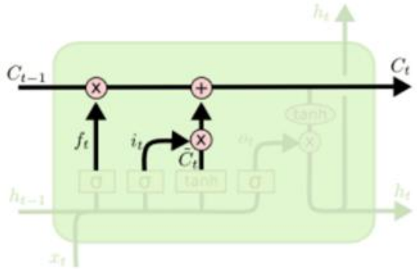
$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] \; + \; b_f\right)$$

In the first step, the LSTM model decides by a sigmoid layer, looks at h(t-1) and x(t) and outputs a number between 0 and 1 for each number in state C(t-1).



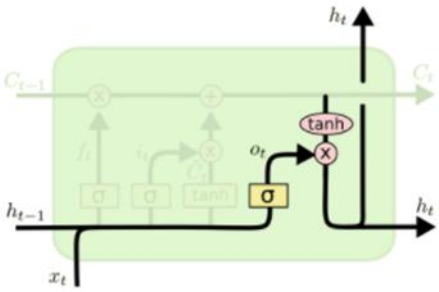
$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] \; + \; b_i\right)$$
$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] \; + \; b_C)$$

The next step is to decide what new information we are going to store in the cell. In the example of our predicting model we’d want to add the sentiment of the new tweet to the cell state, to replace the old one we are forgetting.



$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

In the next step, we update the old cell state C(t-1) to new state C(t). In the case of our model this is where we drop the information of the old tweet and add new information, as decided earlier.



$$o_t = \sigma\left(W_o\left[h_{t-1}, x_t\right] + b_o\right)$$
$$h_t = o_t * \tanh\left(C_t\right)$$

Finally, we need to decide what we are going to output. For the predictive model, it might want to output information relevant to a sentiment, so that we know what kind of sentiment should be conjugated into if that's what follows.

However, the amount of training data on the neutral sentiment for the tweets collected were not large enough for the RNN LSTM to give a high rate of accuracy. The accuracy obtained through RNN-LSTM was around 75 percent.

4. FUTURE WORK AND SCOPE FOR IMPROVEMENT.

In our models which have been used above, using word embeddings can immensely contribute to the practicality and accuracy of each of the models. Word embedding is the collective name for a set of language modelling and feature learning techniques in natural language processing(NLP), where words and phrases are from the vocabulary are mapped to vectors of real numbers. A word is encoded as a vector from 0 to N associated with a word in our vocabulary. Consider an example where we are processing reviews for a detergent called 'Tide'. Our machine must know the difference between the sentences- 'Time and Tide wait for none' and 'Tide is a very good detergent'. By vectorising each of the words we can make our machine to do so.

Doc2Vec is one such algorithm which we can use, to generate vectors from sentences, paragraphs and documents. Unlike sequence models like RNN, where word sequence is captured in generated sentence vectors, doc2vec sentence vectors are word order independent.

However, for input corpus with a lot of spelling errors like in that of tweets, this algorithm may not be the correct choice. One may be better off generating word vectors constructed from character n grams using Fast Text.

5. RESULTS

Observing each of the classification models, we were able to come up with the confusion matrix for each of them. From the confusion matrices precession, recall, F1 Score and accuracy can be calculated. The ROC curves also can be considered as means to give an idea about the accuracy of each of the models. Observing the mean of the micro-average and macro-average ROC curve areas, we can tell as to which is the best performing model.

6. CONCLUSION

For each of the classifiers we have the F1 scores, accuracy and the ROC curves. In the ROC curve areas, class 0, class 1, class 2 is the ROC curve areas for negative, neutral and positive tweets respectively. Computing the mean of each of the micro-average and macro-average ROC curves we get the following-93.5% for Logistic Regression, 95% for K Nearest Neighbours, 86% for SVM, 93% for Kernel SVM, 91% for Naïve Bayes, 94.5% for Random Forest and 91% for XGBoost. The classification models classify a large number of tweets as a neutral class. This can be observed by the ROC curves, where in the curve area depicting the class 1 is on a higher side for all of the classification models.

From the above results, we can conclude that the K Nearest Neighbours as the suitable model for our classification. On observing the ROC curves for the majority of the models we can conclude that most of the tweets are classified as neutral by the models.

7. REFERENCES

[1] Nehal Mamgain, Ekta Mehta, Ankush Mittal and Gaurav Bhatt.2016 Sentiment Analysis of Top Colleges in India using Twitter Data.

[2] Hwi-Gang Kim, Seongjoo Lee and Sunghyon Kyeong.2013 Discovering Hot Topics using Twitter Streaming Data.

[3] Colah’s Blog.2015 Understanding LSTM networks.

[4] Bo Pang, Lillian Lee and Shivakumar Vaithyanathan.2002 Thumps up? Sentiment Analysis using Machine Learning Techniques.

[5] Neethu M S and Rajashree R. Sentiment Analysis in Twitter using Machine Learning Techniques.