ISSN 2394-739X
IJRSET DECEMBER Volume 8 Issue 12
International Journal for Research in Science Engineering & Technology (IJRSET)
https://www.doi.org/10.5281/zenodo.5795058

ENSEMBLE BASED TWITTER SENTIMENT ANALYSIS ON BIG DATA USING APACHE SPARK

[1] Mrs. Sonal Devlekar, [2] Prof. Rashmi Thakur
[1] Student, [2] Assistant Professor,
[1, 2] Department of Computer Engineering,
[1, 2] TCET Mumbai, India.

ABSTRACT - The development of social media platforms like Twitter and Facebook has been a revolution in the history of mankind. Social media web applications played a great role in getting the world together. There are platforms like twitter where people can share their opinion about an issue which can be seen and heard by the entire world. Tweet count on twitter for a day is about 500 million tweets. By analysing these tweets on sentiment basis we can understand the public opinion regarding a particular matter. The system developed will perform a sentiment analysis on live twitter data using an ensemble algorithm to detect the nature of tweets (tweets are positive or negative). Implementing Sentiment analysis on twitter data is quite a difficult task as it involves operating on a large amount of diverse data. This thesis involves the implementation of the apache spark framework for working with a huge chunk of data that is big data. Sentiment analysis is a machine learning-based method. This thesis aims to examine various classification algorithm and their impact on live twitter data.

Keywords - [Naïve Bayes, Logistic regression, Random Forest , Decision Tree, Gradient Boosting, Confusion Matrix.]

1. INTRODUCTION

The development of the Internet has been a revolution in the sector of technology and the world. Some way or another internet has infiltrated the life of every human being. Many social media sites and applications have come into existence, using which people can express and share their views regarding an incident. One of the famous social media applications in the current age is Twitter. Twitter is an internet application where a different person can write their views about something for the whole world to see it. Various powerful and famous people use Twitter to express their views on a particular topic and the public can easily view their tweets. A common man can easily get influenced by the tweets of someone who is a famous person. A place where people can say anything and can be heard by all needs to be free of toxicity. A toxic comment on twitter by a powerful or famous person can create panic among the common people. It becomes very important to keep social media platforms like twitter free of toxic or hatred comments to keep society safe.

[1] Every sec on average about 6000 tweets are getting twitted on twitter. Tweet count for a day is about 500 million tweets. So it is not possible to view every tweet by a personal and delete the tweet if it is not appropriate. Hence, we require a smart system that automatically detects the tweet by using NLP that analyses each word in the tweet by performing operations of word bagging. By using the concept of Sentiment analysis a sentiment score for each word is generated for a review. If the overall polarity of sentiment score for a review is found to be greater than 0 than the review is termed as positive.

By using sentiment analysis for each comment the system figures out whether the comment is toxic or not. If the sentiment of a comment is found to be negative then it will be deleted automatically.

Operating on live streaming data is a huge challenge due to its size. Due to the continuous growth in data new frameworks came into existence like Apache Spark, Hadoop, Apache Storm, and distributed data storage like HBase and HDFS. These frameworks are designed to operate specially on a huge chunk of data. Libraries like Spark's MLlib came into existence that can perform machine learning operations on a large amount of data.

This system aims at developing a sentiment analysis based machine learning system using Spark's MLlib library. This system will operate on live twitter data, the streaming and handling of live data will be done with the help of Apache Spark framework. The classifier will classify the tweets as positive or negative. That is, whether a tweet is a hate speech or not. The form of learning used here is Supervised form of Learning.

Sentiment Analysis:

Sentiment analysis is the analysis of the input text and its classification of emotion like positive, negative, or neutral based on the words used in the sentence using the text analysing techniques. Sentiment analysis is used by a great number of companies for understanding the feedback for their service or products from the customers. It helps the company to gather useful data regarding their work.

There are various types of sentiment analysis. The type which this project will make use of is Fine-grained Sentiment Analysis. Which will classify a tweet into one of these categories

- Positive
- Negative

2. BIG DATA ANALYTICS:

For processing this large amount of data the system will

make use of Big data concepts. Big data generally refers to a large amount of data in a structured, unstructured, or semi-structured format. This large amount of data is mined and processed using big data concepts. After processing and cleaning of this chunk of data. The resulted data can be used as a dataset for Machine Learning algorithms. Doug Laney was first to identify these characteristics in his article in 2000. The categorization of big data is usually done with the help of 3V concepts. The 3V concept includes.

- The large volume of the data present (Volume)
- The different variety in the present data. (Variety)
- The velocity or speed at which the data is produced and processed. (Velocity)

3. PROPOSED METHODOLOGY:

This section will consist of a brief discussion on the algorithms that will be used for the development of this system. The overall architecture of the system will also be discussed in this section.

- Naïve Bayes
- Logistic regression
- Random Forest
- Decision Tree
- Gradient Boosting

Naïve Bayes:
Naïve Bayes algorithm follows a simple assumption of all the available predictors of being independent of one another. Naïve Bayes considers that the result of a feature for a particular model does not affect the outcome of the other features for the given model. Every feature is considered independent of one another. Hence, the algorithm is termed as Naïve.

Logistic regression:
Logistic regression is very much like linear regression. In linear regression, a threshold is used for making a classification. While in the case of logistic regression a sigmoidal function is used. Logistic regression is used when the output required is categorical data.

Random forest
Random forest is a supervised learning algorithm. The "forest" it builds, is an ensemble of decision trees, usually trained with the “bagging” method. The general idea of the bagging method is that a combination of learning models increases the overall result

Decision Tree:
The decision tree algorithm falls under a supervised learning algorithm. It can perform both classification and regression. Decision tree as names suggest has a tree-based structure where the internal nodes represent features, decision rules are represented using branches and finally, the leaf node represents the result or the final prediction.

Gradient Boosting:
Gradient boosting is a machine learning technique that involves using multiple decision trees one after another to boost the accuracy of the system by minimizing the error rate during each iteration.
As said the main objective or concept behind gradient boosting is to reduce the loss function.

4. SYSTEM ARCHITECTURE:

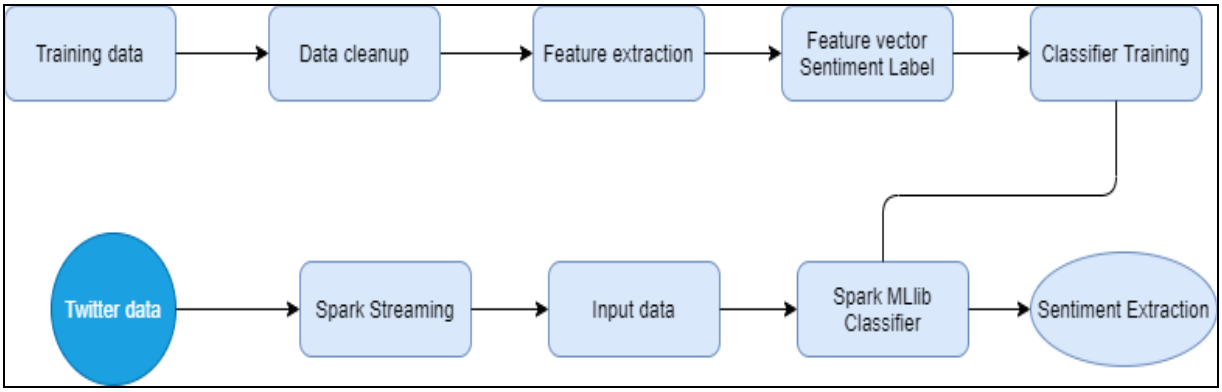

Figure. 1 System Architecture of Twitter Sentiment Analysis

Initially, a training dataset is created. This training dataset is in CSV file format. The dataset will consist of the input features and the output result.

Data Cleanup:
Here the input dataset is cleaned i.e the unwanted words which do not contribute towards calculating sentiments will be removed.

Feature extraction:
After the dataset is cleaned we will be left with only sentiment calculative relevant words. A sparse matrix will be created based on these words where the frequency of words for each sentence will be set.

Classifier Training:
This is the step where the algorithms will be applied on the given sparse matrix and a final model will be created which will predict input data (Tweets).

System specification:
The system will be developed on a computer with 4GB RAM, 10GB HDD, Intel 1.66 GHz Processor Pentium 4 processor, windows 7. Python 3.6.3 will be used as the

base language. VScode will be used as the text editor.

Apache Spark Framework:
Apache spark is a framework very similar to the Hadoop but an advanced version of it. It can perform processing of data on a large scale. It can distribute data on multiple systems for parallel execution.

5. PROPOSED EVALUATION:

For each algorithm, the result will consist of a confusion matrix, AUC-ROC Curve, precision, recall, and f1-score value. All these measures will help to gain a proper analysis of results obtained by each algorithm in detail.

Confusion Matrix:
A confusion matrix is a table that summarizes the result or the output obtained from the training or testing data set on the machine learning model. The table displays the count of all the predictions made. That is, a count is stored for all the true positives, False Positives, True negatives, false negatives. This is the key concept of the confusion matrix, it displays the actual state our model is in after working on the dataset.

	Class 1 Predicted	Class 2 Predicted
Class 1 Actual	TP	FN
Class 2 Actual	FP	TN

Figure 2. Confusion Matrix

AUC-ROC Curve:
AUC - ROC curve is a performance measurement for the classification problems at various threshold settings. ROC is a probability curve and AUC represents the degree or measure of separability. It tells how much the model is capable of distinguishing between classes. Higher the AUC, the better the model is at predicting 0 classes as 0 and 1 classes as 1. The ROC curve is plotted with TPR against the FPR where TPR is on the y-axis and FPR is on the x-axis..

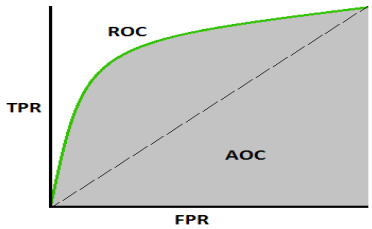

Figure 3. AUC-ROC Curve

Precision:
It is the measure of the value of correctly classified samples from all the value which are classified positive by the algorithm.

$$\text{Precision} = \frac{\text{True Positive}}{(\text{True Positive} + \text{False Positive})}$$

Precision

Figure 4. Formula for Precision

Recall:
It is the measure of the value of positive classified samples from all the samples which are present in the dataset.

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

$$= \frac{True\ Positive}{Total\ Actual\ Positive}$$

Figure 5. Formula for Recall

F1-score:
F1-score is a harmonic mean of precision and recall. It displays a balance between precision and recall.

$$\text{F1-score} = \left(\frac{\text{Recall}^{-1} + \text{Precision}^{-1}}{2}\right)^{-1} = 2 * \frac{(\text{Precision} * \text{Recall})}{(\text{Precision} + \text{Recall})}$$

Figure 6. Formula for F1-score

All these values will help to perform a comparative study of algorithms on the dataset. These values will help to analyse which algorithm performs well in which department.

6. Results and Discussion

The experimental results of different algorithms used for sentiment analysis provide us insights on which method is best suited for real-time analysis and gives accurate results. The outputs of these techniques have been presented and analysed in the form of graphs and a close match have been found between the different results obtained. We have performed sentiment analysis on static data. For static analysis, data is collected beforehand and stored in a file. Basic counting methods and machine learning algorithms are applied to this stored data to identify the sentiments. This has allowed us to process a large number of tweets and obtain accurate trends. For the experimental analysis, we have used a sample dataset having 600 static tweets.

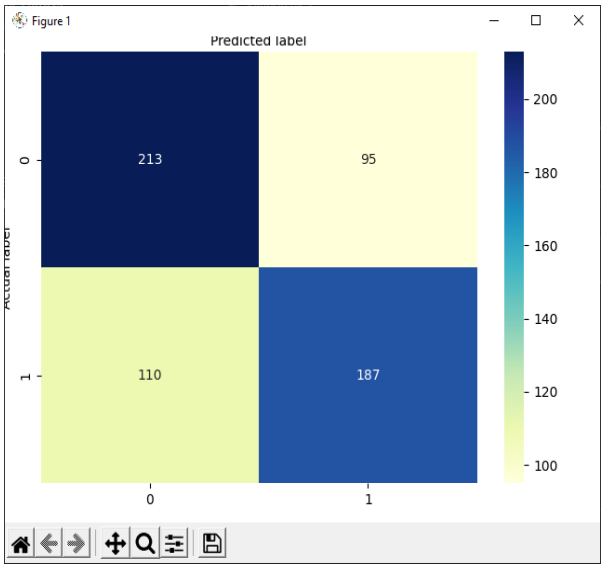

Figure. 7 Confusion Matrix using Naïve Bayes algorithm

ISSN 2394-739X
IJRSET DECEMBER Volume 8 Issue 12

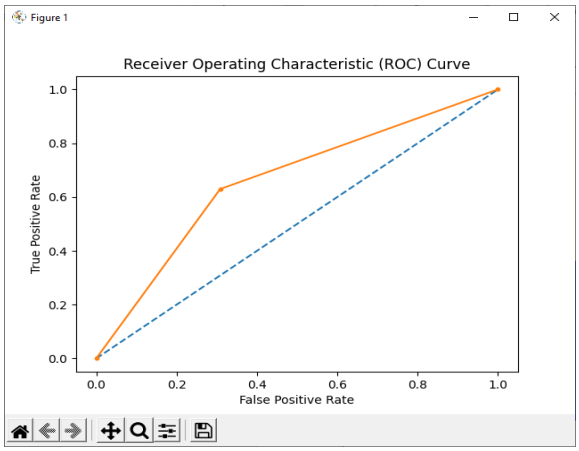

Figure 8 AUC-ROC Curve using Naïve Bayes algorithm

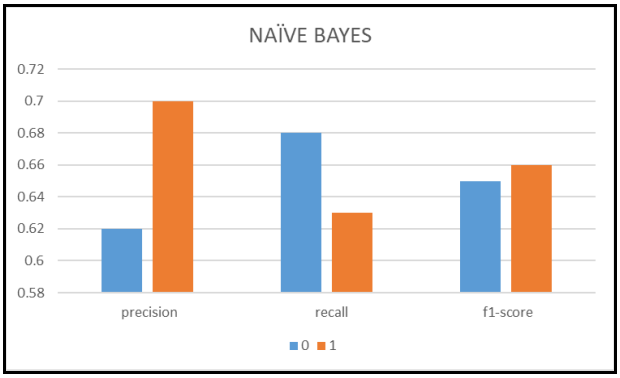

Figure 9 Precision, recall and F1 score using Naïve Bayes algorithm

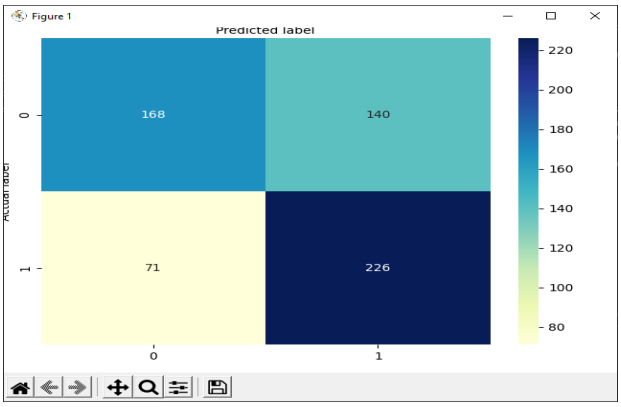

Figure 10 Confusion Matrix using Logistic Regression algorithm

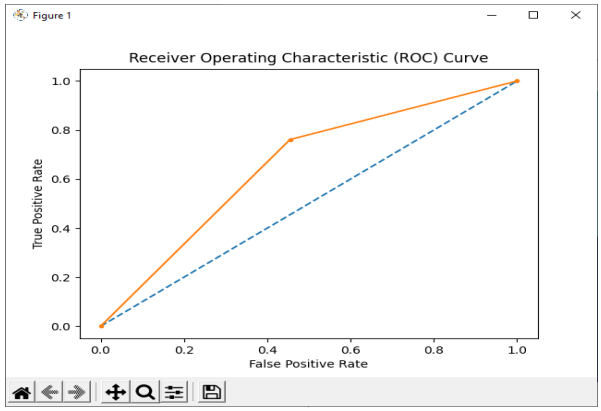

Figure 11 AUC-ROC Curve using Logistic Regression algorithm

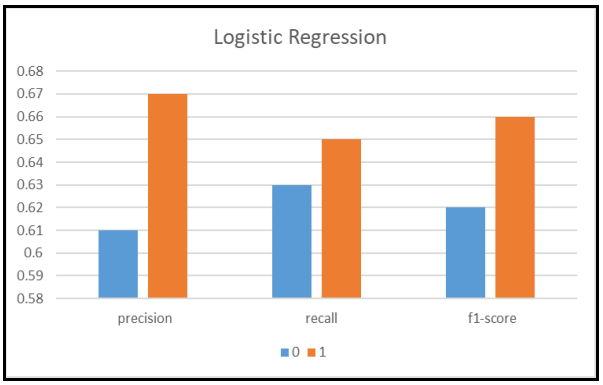

Figure 12 Precision, recall and F1 score using Logistic Regression algorithm

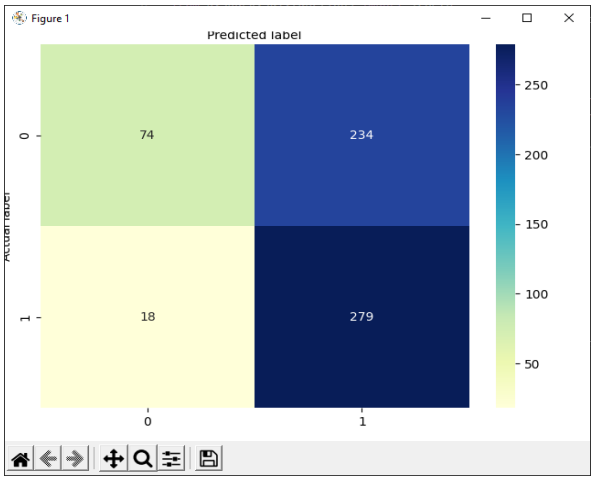

Figure 13 Confusion Matrix using Random Forest algorithm

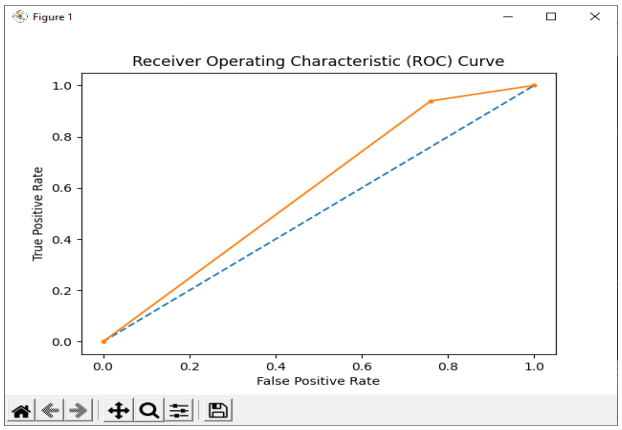

Figure 14 AUC-ROC Curve using Random Forest algorithm

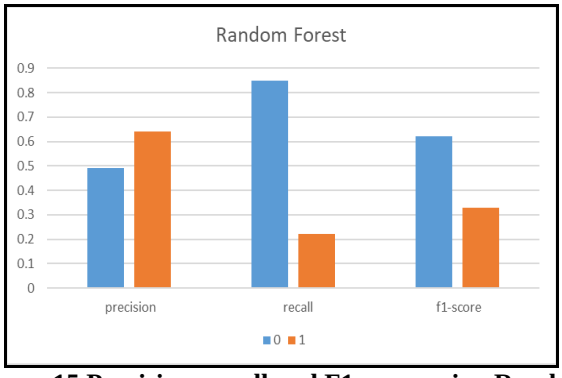

Figure 15 Precision, recall and F1 score using Random Forest algorithm

ISSN 2394-739X
IJRSET DECEMBER Volume 8 Issue 12

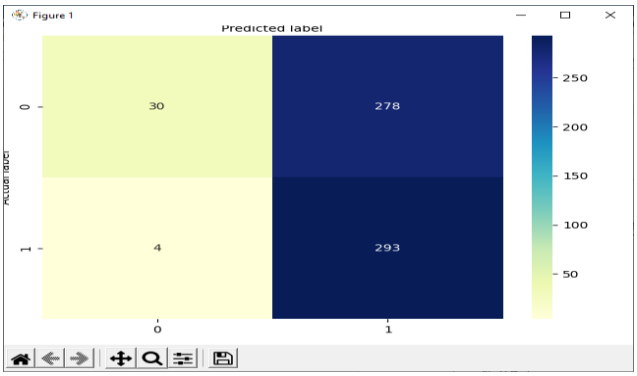

Figure 16 Confusion Matrix using Decision Tree algorithm

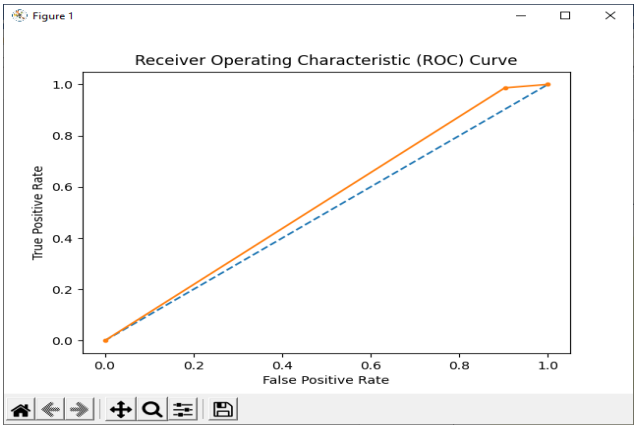

Figure 17 AUC-ROC curve using Decision Tree algorithm

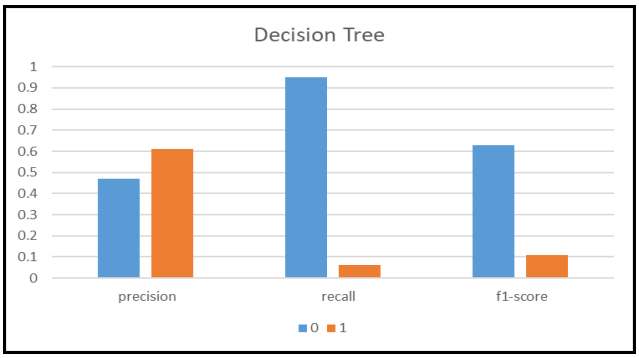

Figure 18 Precision, recall and F1 score using Decision Tree algorithm

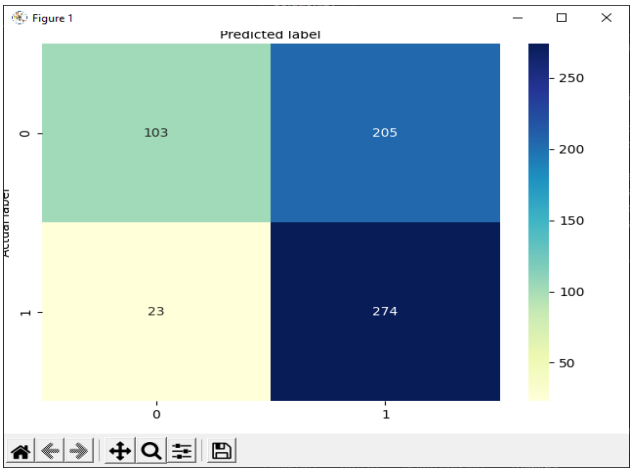

Figure 19 Confusion Matrix using Gradient Boosting algorithm

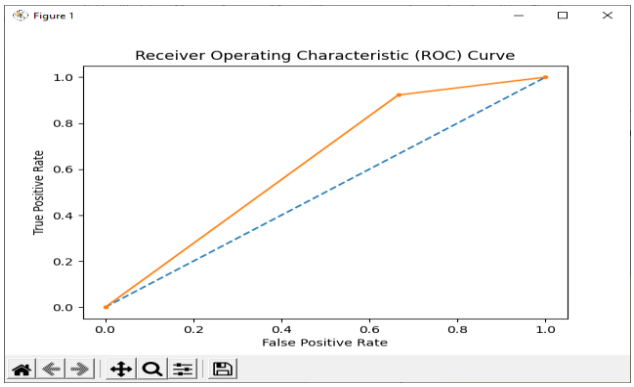

Figure 20 AUC-ROC Curve using Gradient Boosting algorithm

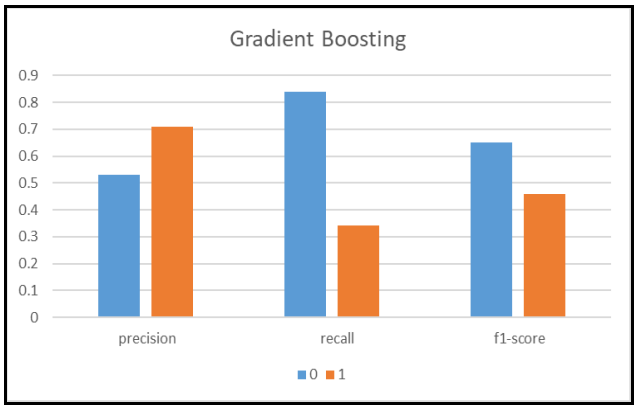

Figure 21 Precision, recall and F1 score using Gradient Boosting algorithm

CONCLUSIONS

Twitter is one of the major platforms with a large number of users worldwide. People, their interest, their opinion, likes, dislikes, events, sports tournaments, politics, movies, and the music everything are part of it. Analysing such a rich data content platform and observing trends in it definitely will be beneficial. Analysing Twitter trends helps to know what people are more interested in and thus helps business organizations or brands to improve their sales, political parties to understand people's emotions and needs, movie industries to get valid feedback for their performances, and much more.

Considering the nature of the data generated, we use the Spark framework to stream the data in realtime and perform the sentiment analysis model on it. Apache Spark was preferred over Hadoop for Big Data processing, considering its performance in regards to time and scalability.

A big data of tweets will be classified as positive or negative using different machine learning algorithms from the Apache Spark's Machine Learning Library, entitled MLlib. From the system proposed and evaluation metrics used we can conclude that the ensemble algorithm should have an increase in the accuracy of the system designed. The spark framework should increase the overall performance of the system, due to its caching abilities after each processing step.

REFERENCES

[1]. Anon., n.d. Decision Tree Classification Algorithm. [Online] Available at: https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm

[2]. Anon., n.d. Machine Learning-Based Sentiment Analysis for Twitter Accounts.
[3]. Baltas, A., Kanavos, A. & Tsakalidis, A. K., 2017. An Apache Spark Implementation for Sentiment Analysis on Twitter Data. Springer.
[4]. Kamal, R., Shah, M. A., Hanif, A. & Ahmad, J., 2017. Real-time Opinion Mining of Twitter Data using Spring XD and Hadoop. IEEE.
[5]. Khan, M. & Malviya, A., 2020. Big data approach for sentiment analysis of twitter data using Hadoop framework and deep learning. IEEE.
[6]. Kolchyna, O., Souza, T. & Treleav, . P., 2015. Twitter Sentiment Analysis: Lexicon Method, Machine Learning Method and Their Combination. ARXIV.
[7]. Pang, B. & Lee, L., 2008. Opinion Mining and Sentimental Analysis Foundations and Trends in Information Retrieva. IEEE.
[8]. RAY, S., 2017. 6 Easy Steps to Learn Naive Bayes Algorithm with codes in Python and R. [Online] Available at: https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/
[9]. Ristevski, B., 2018. Big Data Analytics in Medicine and Healthcare. ResearchGate.
[10]. Rodrigues, A. P., Robnik, M. & Chiplunkar, N. N., 2014. Real-time Twitter data analysis using Hadoop ecosystem. IEEE.
[11]. Shen, Y. & Wang, J., 2008. An Improved Algebraic Criterion for Global Exponential Stability of Recurrent Neural Networks With Time-Varying Delays. IEEE.
[12]. Socher, R. et al., 2013. Recursive deep models for semantic compositionality over a sentiment treebank. IEEE.
[13]. S, V., L. & P., 2002. Thumbs up? Sentiment classification usingmachine learning techniques.
[14]. Taboada, M., 2016. Sentiment Analysis: An overview from Linguistics. ResearchGate.
[15]. T, W. & Hoffman, J., 2005. Recognizing contextual polarity in phraselevel sentiment analysis.
[16]. Vaithyanathan, S., Lee, L. & Pang, B., 2010. Thumbs up sentiment classification using machine learning. IEEE.
[17]. Vinodhini , G. & Chandrasekaran, R., 2012. Sentiment analysis and opinion mining: A survey. IEEE.
[18]. Woldemariam , Y., 2016. Sentiment Analysis in A Cross-Media Analysis Framework. IEEE.