International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. 10, No.1 (Jan-2021)

http://dx.doi.org/10.12785/ijcds/100119

Sentiment Prediction using Enhanced XGBoost and Tailored Random Forest

Supriya B N[1] and C.B. Akki[2]

[1]*Department of ISE, SJBIT, affiliated to VTU, Bangalore 560060, India*
[2]*Registrar, IIIT, Dharwad 580029, India*

Received 29 May 2020, Revised 22 Jul. 2020, Accepted 29 Jul. 2020, Published 1 Jan. 2021

Abstract: A large quantity of data is being generated in the form of blogs, tweets and updates of opinions on the topic of interest. People give their feelings and opinions on different topics such as movies, products, education, politics, news and so on. Analysis of such data is very useful to understand the views/opinions/sentiments of the society. Such analysis would also be more useful in decision making . The major challenge in analysis is the usage of jorgon words, spelling mistakes, hash tags, hyperlinks and irrelevant words. This research aims to know the opinion of people on particular topics considering their tweets. These can be evaluated as classification problem to analyse the tweets expressed in texts for hidden sentiments. For this purpose, we proposed and evaluated a tailored random forest and enhanced XGBoost algorithms. We achieved significantly better accuracy by enhancing XGBoost compared to tailored random forest and naive bayes for tweets classification.

Keywords: Twitter Sentiment Analysis (TSA), Machine Learning Techniques, Telecommunication Services, Feature Vector, Classification,Xgboost.

1. INTRODUCTION

Twitter being primary stage for online life where 500 million tweets are produced by almost 100 million users each day[1] .This tweets express their sentiments and perspective in the subject of interest, which are truly significant source of information that upgrades spaces for many domains like telecom, motion pictures, governmental issues, marketing and so on[2,3]. Analysts and tweet management practitioners have shown gigantic enthusiasm on tweets analysis, but still lacks on the classification performance which is below 70 %. These poor precision levels made tweets order issue more challenging. Tweets which are brief communication, constrain to 140 length of characters, consists variety of patterns like slang words, emojis, abbreviations etc. The conciseness of the tweet text offers considerably few words to evaluate with the lexicon sentiments that yields more sparsely dense tweet features. These structures will make performance of the tweet analysis to decrease and is also a challenging task. However, the tweets related to the brand are more frequently expressed by the consumers of the brand and these tend to be more required sentiments of interest [4]. From the brand owners' views, it is required to paramount the interest of consumers, so that they can improve continuously by themselves. In this perspective the tweeter sentiment analysis approaches are required to target the brand community. Consumers expressing strongly positive sentiments require no further intervention and those views of negative sentiments will entrenched in the evaluation of the product in the marketing unit. By considering these practical needs, tweeter sentiment analysis approaches are classifying the sentiment into positive, negative and neutral.

This paper is structured as follows: section 1, provides introduction to the tweets sentiment analysis and problems associated with it, section 2, reviews on existing techniques for tweets classification and its drawbacks, section 3, briefs about the proposed methodology, section 4 gives detailed description of random forest and parameter used for tailoring the random forest and enhanced XGBoost, section 5, provides the performance evaluation with accuracy results obtained and finally conclusion and future work is given in section 6.

2. RELATED WORK

Twitter Sentiment Analysis(TSA) is a more focused problem in the recent research areas of data science and computational linguistics. Many approaches for sentiment analysis have been identified and evaluated by the researchers to know the opinions expressed in text using

E-mail: bnsupriya@gmail.com , akki.channappa@gmail.com

automated methods. In recent trends twitter has been widely used to post the opinion on the particular issue of interest and it has become large communication platform for the people. For business and society, TSA makes an valuable source to understand the behavior, reaction of the people when the new brands are released in the market place. There are several constraints which makes the TSA problem more challenging these includes uncertainty in tweets , diverse information, informal language , slang words, evolving language and more imbalance in the opinion of the consumers. To find out the sentiment class from the tweet text is more complex due its unstructured nature. Apart from these challenges, other approaches to TSA that brings interest will be on communication platforms like new articles [5], reviews on the product [6] and forums in blogs [8].

Traditional techniques to TSA can be generally classified into main two types. The first type of approaches includes the utilization of lexicon based opinion extraction from the related terms in conjunction with sentiment scoring method to evaluate, this is termed as unsupervised learning application [10,11] where predefined class labels. These methods are widely used and less accuracy, the performance of the sentiment classification was limited and unable to identify the information context, vocabulary, opinion expressions and pronounced indicators. The second type of approaches, tries to derive relationship between the features in sentiment text and find the opinion by applying supervised learning mechanism by training the words in to the machine learning models. These models require large dataset to train the instances with complete opinion class labels. These features contains unnecessary, duplicate, infrequent appearance features in the representation and noises which will reduce classification performance. TSA evolved to find the opinion to address the unique problems from the literature. Many authors have focused on sanitization of tweet text to pre-processing to eliminate the slang words, misspellings, exaggerations, unwanted patters, hyperlinks, abbreviations and convert into the more readable format [12]. Other authors have found features like emoticons, user re-tweets, hashtags also give valuable information people express their feeling in terms of emoticons [13]

Gamon [7] proposed a technique for deriving TSA to expand into training instances by considering features emoticons and labeled from noisy text. Emoticon based sentiment classification work have been carried by using emoticons. No predefined class labels in the emoticons, need to determine the emoticons score and then machine learning models are trained for classification of the sentiment class [8]. Feature representation in the sentiment text is the main concern in the research of TSA so that number of training instances can be expanded to improve the overall accuracy of the sentiment classification. Montejo-Ráez[14] presents a novel way to deal with Sentiment Polarity Classification in Twitter posts, by removing a vector of weighted hubs from the diagram of WordNet. These loads are utilized in SentiWordNet to figure a last estimation of the extremity. In this way, the strategy proposes a non-regulated arrangement that is area autonomous. The assessment of a produced corpus of tweets shows that this procedure is promising. Ammar Hassan[15] designs a framework utilizes an expound bootstrapping group to subdue class irregularity, sparsity, and illustrative extravagance issues. Examination results uncover that the proposed approach is progressively precise and adjusted in its forecasts across slant classes, when contrasted with different correlation instruments and calculations.

Vasileios Athanasiou[17] explains that gradient boosting is a robust ensemble method that outperforms a family of methods used in sentiment analysis by handling sparsity in high dimensional data. Sung-Lin Chan[18], et al. presented result shows that the accuracy of different classification algorithms has implemented in their project. Authors tries to conclude that neural networks models are the best algorithm in kaggle competition. Babacar Gaye[19], concludes that if accuracy is the priority, a classifier like XGBoost can be used that uses high has the best accuracy. If processing and memory are small, then Naïve Bayes should be used due to its low memory and processing requirements. If less training time is available, but you have a powerful processing system and memory, then Random Forest can be considered. More focus is required in the data cleansing of raw data and transforming into cleaned data. Supervised learning needs more exploration to improve the accuracy of the TSA and enhance prediction of the new tweet text and determine the sentiment class automatically.
Sami Belkacem[21] in his paper proposed an approach that uses Random forest classifier as relevance prediction model for the news feed updates considering the four types of features :1. the relevance of the update content to the beneficiary's interests; 2. The social tie strength between the beneficiary and the update's author; 3. the author's authority; and 4 . the update quality. The results inferred that approach succeeded in relevance prediction and also concludes the critically for identifying the valuable updates in news feeds. Kiran Sangada[22] in their research work proposed the machine learning based classifier hyper parameter tuned random forest for classifying the Indian election related tweets. Through the tuned random forest the authors achieved the significant improvement in accuracy of 95.6%.This method has been compared with SVM algorithm.

Farideh Tavazoee[23] in her paper analysed the social media popularity of the candidates in the 2016 US Election. Besides some limitations they proposed a methodology utilizing combinations of the classifiers that extremely supple and works on variable data types. Random forest are specifically used in a sliding window fashion that accounts for trends popularity. Time dependencies were considered using ad-hoc weighing systems to estimate social media popularity in a recurrent manner.

3. PROPOSED METHODOLOGY

The work flow of our proposed methodology for twitter sentiment analysis is shown in Figure 1.

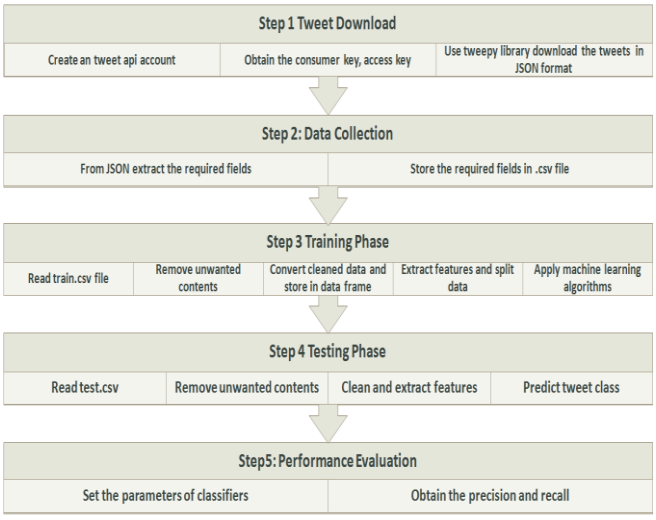

Figure 1. Work flow for twitter sentiment classification

A. Tweet Download

Twitter provides developers to create an account and API named as OAuthHandler, set_access_token, get_status and so on to access the tweets. These APIs offer low latency access to flows of twitter data. In our implementation we used OAuthHandler and set_access_token API to stream access to the tweets data and later required fields are filtered for further processing. In python tweetpy library can used to download the tweets from twitter web resource, tweet_curson API can be used to streaming the tweet data and downloaded into local system. We can mention the tweets range ans posted date to download the tweets and we also include search keywords in the parameters of the tweet_cursor API.

B. Data Collection

The downloaded data will be in JavaScript Object Notation (JSON) format, we require in the .CSV for further processing of data cleansing and transformation. Each attributes in JSON contains information and tags. Only the required fields are extracted from JSON object file and are stored in comma separated values (CSV) file. The .CSV file contains parameters like tweetID, tweet text, created date, retweeted count, status, hyper links, replyto and replytoSID and so on. We mainly interested in tweet text parameters which contains the user emotions.

C. Training Phase

In order to provide input to the machine learning algorithm the input data needs to be cleaned. Data cleaning not only improves the classification performance of the machine leaning models but also provide wide enhancement in the training phase. From the previous work [16], we have taken few steps for data cleaning. The unwanted content along with the action is tabulated in table 1.

Table 1. Unwanted content and action

Unwanted Content	**ACTION**
Punctuation (! ? , . ” : ;)	Eliminate
#hashtagas word	Eliminate#
@enduser	Substitute with “AT_USER”
Retweet (RT)	Eliminate
Tweet text in uppercase	Change to tweet text to lowercase
Hyper links and URL patterns	Substitute with “ “
Tweet	**Status**
Today @YouTubeGaming launches, with apps for iOS and Android devices in the US and UK, here is what you need to know http://t.co/Kf8DgnHX9b	Raw Data
YouTube Gaming Launches Tomorrow with iOS and Android Apps to Go HeadtoHead with Twitch ios game	After data cleaning
@jackstenhouse69 I really liked it, in my opinion it def is :)	Downloaded data in .CSV file
AT_USER really liked it, opinion def :)	After applying data cleaning
So pissed I just cracked my phone screen :(	Raw Data

After eliminating the irrelevant patters and unwanted symbols, stopwords are removed using Natural Language Processing (NLP). Then bag of words (BOW) are generated for each tweets through which features are extracted. To automatically classify the tweets we machine learning techniques such as navie bayes, random forest and XGBoost are applied.

Many decision trees of classification are structured by Random Forest. Each of these decision trees are trained by selecting the features randomly and based on the best split values and entropy the trees will be constructed. .The complex task of each decision tree is the high variation in the representation of the features that may affect the performance of the classifier. There are also the chances of co-relation problem between the decision trees that are the average produce Gaussian distribution. The average

value of each decision tree will be considered for classification.

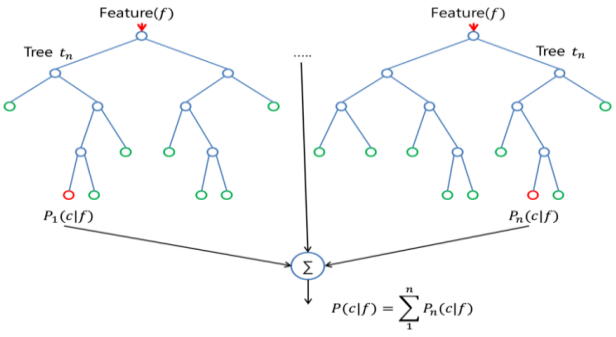

Figure 2. Random Forest Structure

The more values in decision, then the averaged value may become low in the variance. Reducing the overall variance will increase the classification performance in the machine learning models and reduce misclassification errors. To achieve better accuracy RF is modified and it is called as tailored random forest. Following sub section gives the description of each parameter.

Algorithm 1: Tailored Random Forest

Input: A training set S := (x1, y1), . . . ,(xn, yn), input features values F, and trees count in forest N, here {x1, x2, ……. Xn} are the BOWs and y1,y2, ……yn are the sentiment (positive, negative, neutral)} , M is an subtree.

1 procedure TailoredRandomForest(S , F)
2 M ← ∅
3 for i ∈ 1, . . . , N do
4 S (i) ← select the features with occurrences of terms >=4
5 K(i) ← build feature vector from S(i)
6 Initialize number of trees to a and seed value = 666
Where a= {200,250,300,350,403,450,500,600,700,800}
7 hi ← LearnRF(K (i), F)
8 M ← M ∪ {hi}
9 end for
10 return M
11 end procedure
12 procedure LearnRF(K , F)
13 Extract randomly b features from K to build a sub tree
Where b={20, 30,40,45,51,60,70,80,90,95}
14 At each node in sub tree:
15 f ← subset of F
16 based on best feature split in f by computing Entropy (E) and information gain G
17 return the Tree Learnt
18 end procedure

D. XGBoost

XGBoost is an ensemble machine learning built based on decision tree that uses gradient boosting [20]. Ensemble machine learning combines the predictive output of multiple learned models. The aggregated models can be either same algorithm learnt or different learning algorithms. Bagging and boosting are the most commonly used in ensemble learning techniques. In bagging technique many decision trees are computed in parallel from the initial learners. Data patters with replacement are provided to the learners during the training. The average output will be the final prediction from all the learners. In boosting technique, the built trees are consecutively aims at reducing the errors from the previous built trees. Every tree receives from its predecessors and residual errors are updated. Initial learning in boosting may be weak learners and the bias are high and power of predictive can be better than guessing randomly.

In contrast to RF, where tress are grown to maximum length, boosting technique helps to use less splits in trees. Small trees can be highly interpretable because of not very deep in generating. Parameters such as iteration, number of trees, depth of trees and learning rate of gradient boosting can be optimised by validating through k-fold cross validation. Obtaining more number of tress may lead to overfitting. So, there is required to properly selecting the termination criteria in boosting. Boosting technique consists of three steps:

Initial built model P0 is determined to predict target parameter‘t’. This model will be correlated with an residual (t- P0)

An new generated model m1 is fitted with residual in previous step.

Now, P0 and m1 gives the P1 , the mean square error of P1 will be lesser than P0.

These steps can be made in ‘n’ iterations until the residual errors are minimized as shown in below equation.

$$P_n(x) < P_{n-1}(x) + m_n(x)$$

For gradient boosting following steps are followed.

P0 (x) with initial model are determined and function to minimise the Mean Square error in this case is:

$$P_0(x) = \arg\min_{\phi} \sum_{i=1}^{n} S(\phi_i - \phi)^2$$

The loss function f_{in} in gradient are determined iteratively, where δ is an rate of learning :

$$f_{in} = -\delta \left[\frac{\partial (S(\phi_i, P(x_i))}{\partial P(x_i)} \right]_{P(x)=P_{n-1}(x)}$$

To derive the best solution, we have divided the proposed methodology in to two parts namely training and

testing phase. The training and testing phase is discussed below:

Proposed Algorithm: Training Phase
Given input: A set of training sample tweets

Output: Find the sentiment from the consumer reviews Method:
1. Read the sample.csv file for training phase
2. Remove the unwanted content (Punctuation (! ? , . ” : ;) , #word, @user, Emoticons (:), :D, :(, ;), :-)), URLs and web links and prepare clean data
3. Store the clean data into python data frame
4. From python data frame extract the features (bag of words)
5. Split the cleaned data for training (Ratio: 80:20, 90:10)
6. Create a data model for analysing tweet class label positive, negative, neutral
7. Apply random forest, naive bayes, tailored random forest, enhanced XGBoost to train and build the model.

Algorithm: Training Phase Ends

E. Testing Phase

To test the classifiers, test.csv file is created where no class labels are specified. Based on the input features the TRF and XGBoost need to predict the class label to negative, positive or neutral. To train the proposed ML models the pickle file is created. In python, the trained ML model can be stored on disk in pickle file so that it can used at anytime. To test the classifier the tweets were split into different ratios such as 80:20 and 90:10. After the split of tweets, input features are given for navie bayes, random forest, tailored random forest, enhanced XGBoost classifiers.

Algorithm: Testing Phase

Input: i. A Tweet data model created in train phase
ii. New tweets (NT).
Output: Class labels are predicted (positive, negative, neutral)
Method:
1. Read new tweets from test.csv file for testing new tweets
2. Remove the unwanted content (Punctuation (! ? , . ” : ;) , #word, @user, Emoticons (:), :D, :(, ;), :-)), URLs and web links and prepare clean data
3. Store the clean data into python data frame
4. From the cleaned data extract the features (bag of words)
5. Using data model, predict tweet class (positive, negative, neutral) using random forest, naive bayes, tailored random forest and XGBoost.

Algorithm: Training Phase Ends

4. IMPLEMENTATION

Implementation of the work is done using python 3.7 version in anaconda version 3. Libraries such as numpy, sklearn, pandas and plotly are used in the development. Numpy library in the python packages provides an scientific computing functionality for numerical analysis. Scikit learn is an open source software ML package in the python programming language. It consist of algorithms for solving classification, regression and clustering problems. It can designed and incorporate with python language for scientific application. Another package pandas, which is an open source library for implementing data structures and data analysis tools in python language. For experimenting on tweets data set that consists of positive tweets, negative tweets and neutral tweets are considered, figure 3 shows sentiment type distribution in training data set.

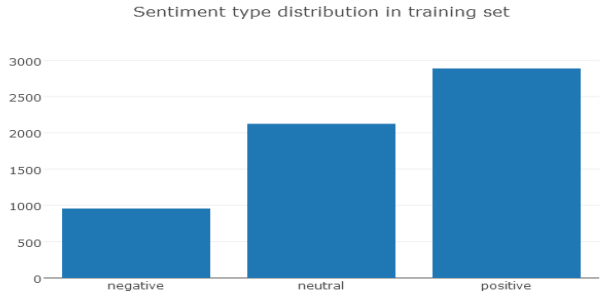

Figure 3. Sentiment type distribution in training data set.

Tweets are tokenised and words are separated to filter the stop words. The most common stop words like not, n’t are filtered out for analysis as they can influence the sentiments greatly. Having this in mind, this word will be whitelisted. Once the stop words are filtered the word list is created. Top most words generated in wordlist is as shown in Figure 4. Figure 5 shows the most common words in sentiment word list. After pre-processing the training set data, the feature vector is obtained. Unigrams at the end of pre-processing end up with 2586 features and each of the features have equal weights. These 2586 features are stored in data frame in python.

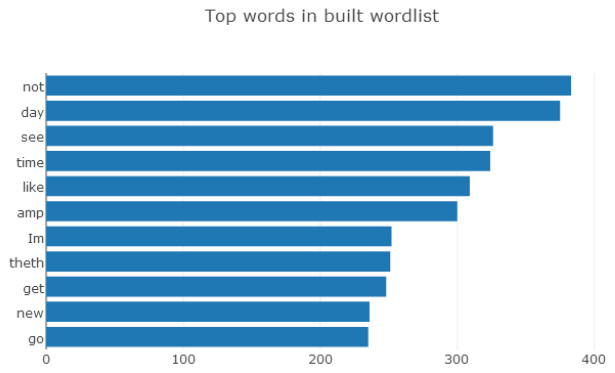

Figure 4. Top most words generated in wordlist.

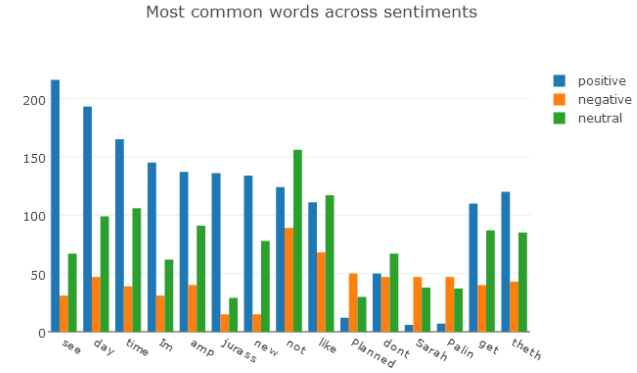

Figure 5. Most common words in sentiment word list

To train the classifier the tweets were split into different ration such as 80:20 and 90:10. After the split of tweets, input features are given for naive bayes, random forest, tailored random forest, Enhanced XGBoost classifiers. Naive Bayes (NB) classifier will classify the sentiment based on the probability of the specified input parameters. NB works based on bayes theorem.

5. Performance Evaluation

To evaluate the performance of the proposed T-RF and XGBoost first, we compare the sentiment classification accuracy of normal RF and T-RF and then we compare the XGBoost and enhanced XGBoost using k-fold cross validation. Some of the parameters in RF will either increase the predictive accuracy of the ML model or it may become easy for the training the model. The following parameters are tailored to improve the accuracy level of tweets classification.

max_features: when more features are considered the classification accuracy improved at the node in each level of the tree.

n_estimators: this provides the number of decision tree need to builds for increasing the voting averages for prediction accuracy. More number of decision trees will improve the overall performance.

min_sample_leaf: here we are reducing the leaf length to minimize the noise in the training data. Leaf nodes are the last nodes in the decision trees.

n_jobs: while building an machine learning model we need to tell the in how may multiple processors we need to allowed to use. A numeric value of "-1" indicates there is no constraint where as "1" means it can run on only one processor.

random_state : this parameter make solution simple to replicate. A proper value for random state value will make machine learning model to produce same results with same features if the given training data given.

The cross-validation technique is selected to increase the level of the classification accuracy and also to improve the training and testing of dataset. The split of training and testing datasets commonly used is into 70% training and 30% test data. We evaluated the classifiers in two trails:
Trail 1: 80% training and 20% test data
Trail 2: 90% training and 10% test data

We evaluated this trails with classification accuracy, it is defined as a ratio of total number of correctly predicted to the total number of given sample data. The obtained classification accuracy result after applying normal RF and T-RF is shown in Figure 6.

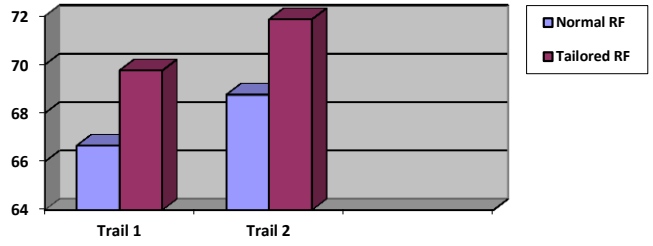

Figure 6. Accuracy comparison graph of Normal RF and TRF classifiers in tweet classification

To enhance the XGBoost, parameters such as number of estimators, learning rate and depth of trees are varied with following values as shown in Table 2.

TABLE 2.PARAMETERS SETTINGS FOR ENHANCING XGBOOST

Parameters of XGBoost model	**Various values experimented**
Number of estimators	100, 200, 300, 400, 500, 600,700 , 800,900,1000
Rate of Learning	0.05,0.075,0.1,0.25,0.5, 0.75,1,1.05,1.075,1.01
Max_depth	1, 6, 2, 11,

Iteratively, k-fold validation is applied using stratified k-fold to find the best learning rate, number of trees, maximum depth. Then enhanced XGBoost function is created for predicting the tweets sentiment. The obtained classification accuracy result after applying normal XGBoost and enhanced XGBoost is shown in Figure 7.

Figure 7. Accuracy comparison graph of XGBoost and enhanced XGBoost classifiers in tweet classification

Table 3 shows the accuracy obtained by applying our proposed T-RF and enhanced XGBoost classifiers for tweets classification. From result it shows that better accuracy is obtained by enhanced XGBoost compared to tailored random forest when executed. Figure 6 gives the accuracy comparison graph of NB, T-RF and enhanced XGBoost classifiers on tweet classification

TABLE 3. SHOWS THE PERFORMANCE EVALUATION OF NB, T-RF AND ENHANCED XGBOOST CLASSIFIERS

Classifiers	Trail 1 (80:20) Accuracy	Trail 2 (90:10) Accuracy
Navie Bayes	67.8205	65.3427
Tailored RF	69.7638	71.8701
Enhanced XGBoost	**72.2623**	**72.5496**

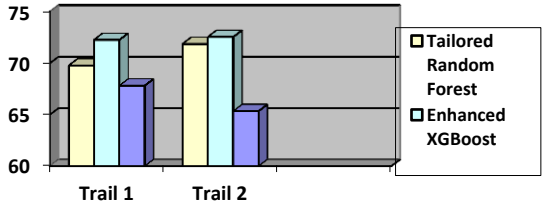

Figure 6. Accuracy comparison graph of different classifiers in tweet classification

We also evaluate the performance of the classifiers with the help of the metrics like F1, Precision, and recall. This has been tabulated in the tables at the end of this paper. The table 4 depicts the value of the metrics for the classifier Tailored Random forest and table 5 for the classifier Enhanced XGBoost.

6. CONCLUSION

The research work carried out in this paper provides an sentiment class prediction using NB, RF, T-RF and enhanced XGBoost. There is need of more data cleansing algorithms for eliminating the unwanted data present in the raw tweets because of complex unstructured natures of tweet text. Data cleansing method is implemented for eliminating the unwanted patters and cleaning the data. Major problems for tweet sentiment analysis are crafting the input raw text machine understandable and data skewness. Along with bag of words addition features need to considered. We observed that random forest classifier accuracy for tweets is low, this is due to data skewness. In order to enhance the accuracy of tweets classification RF parameters such as max_features, n_estimators, min_sample_leaf, n_jobs and random_state have been tailored and In XGBoost parameter such as n_estimators, max_depth and learning rates are enhanced to find the best fit for obtained better accuracy. From the results obtained it shows that enhanced XGBoost gives better accuracy compared to navie bayes and random forest and tailored random forest. The enhanced XGBoost will select the features from BOW with best split based on more information gain and constructs the tree which gives better accuracy.

ACKNOWLEDGMENT

This research was supported by Visvesvaraya Technological University, Jnana Sangama, Belagavi-590018 for Grant of financial assistance.

TABLE 4. SHOWS THE PERFORMANCE METRICS OF T -RF CLASSIFIER

Estimator, Learning Rate	Accuracy		F1						Precision						Recall					
	80:20	90:10	80:20			90:10			80:20			90:10			80:20			90:10		
			neg	Neu	pos	neg	Neu	pos	neg	Neu	pos	neg	Neu	pos	Neg	Neu	pos	neg	Neu	pos
100,0.05	66.8	65.4	64.7	60.2	67.3	61.9	60.1	65.6	89.8	51.5	66.4	88.2	49.4	69.2	50.5	72.3	68.2	47.7	76.7	62.3
200,0.075	69.4	70.3	69.9	60.9	69.6	70.5	62.9	69.8	85.1	55.04	67.5	87.2	55.1	70.2	59.3	68.1	71.8	59.1	73.3	69.4
300,0.1	69.4	72.0	71.5	60.1	68.7	70.8	61.7	69.4	82.5	56.3	65.8	82.4	56.1	68.8	63.2	64.4	71.8	62.1	68.8	70
400, 0.25	69.4	65.8	72.7	59.3	68.2	71.8	60.3	69.0	78.3	56.1	67.8	77.9	56.8	68.6	67.8	62.9	68.6	66.7	64.2	69.4
500,0.5	72.2	72.5	72.9	59.9	67.9	71.9	60.4	69.7	79.7	56.4	67.2	77.4	54.9	69.8	66.7	61.3	69.4	61.9	65.9	70.9
600,0.75	68.8	69.1	71.0	58.2	68.9	70.5	59.3	69.6	74.9	55.7	68.8	76.5	55.1	70.4	67.4	61.0	69.0	65.5	64.2	68.8
700,1	68.6	68.2	70.7	57.4	69.6	68.8	59.5	65.8	75.3	54	70.5	77.1	52.1	71.2	66.7	61.4	68.8	62.0	69.3	61.1
800,1.05	69.0	65.4	69.7	58.3	70.9	67.9	58.5	65.4	73.7	54.7	72.3	75	51.2	71.8	66.1	62.5	69.4	62.0	68.1	60
900,1.075	69.3	69.6	67.8	59.7	71.1	65.8	58.6	67.8	75.7	56.8	72.9	76.3	53.9	72.9	67.6	66.7	70.1	61.9	67.9	70.1
1000,1.01	70.5	71.5	68.3	59.9	71.2	67.1	59.7	68.9	76.1	57.3	71.8	77.1	54.9	72.8	68.3	65.9	71.3	63.7	68.1	70.9

TABLE 5. METRICS OF ENHANCED XGBOOST CLASSIFIER

Estimator, Max-Feature	Accuracy		F1						Precision						Recall					
	80:20	90:10	80:20			90:10			80:20			90:10			80:20			90:10		
			neg	Neu	pos	neg	Neu	pos	neg	Neu	pos	neg	Neu	pos	neg	Neu	pos	neg	Neu	pos
200,20	69.7	71.8	74.2	62.6	69.4	73.6	63.5	69.8	87.1	57.7	67.2	89.3	55.47	70.92	64.8	68.6	71.8	62.7	73.7	68.9
250,30	68.5	68.5	72.2	59.9	68.6	71.6	60.8	68.7	81.6	56.3	66.3	82.7	54.5	69.7	64.7	64.0	71.0	63.2	68.7	67.5
300,40	68.7	68.5	71.9	61.2	68.4	70.4	61.3	69.2	84.8	56.6	65.7	83.4	54.9	69.0	62.4	66.7	71.4	60.9	69.3	69.4
350,45	68.9	65.7	71.1	62.2	68.9	64.9	57.9	68.3	85.5	54.7	71.1	81.0	49.2	72.3	60.9	72.1	67.6	54.0	70.5	64.7
403,51	69.5	68.3	71.3	61.2	69.7	70.5	61.04	69.02	83.1	56.6	67.4	84.7	54.2	69.2	62.4	66.7	72.1	60.3	69.9	68.9
450,60	69.1	65.0	71.1	62.2	68.8	64.8	57.9	68.3	85.4	54.7	70.1	81.0	49.2	72.3	60.9	72.1	67.6	54.0	70.4	64.7
500, 70	69.3	69.3	72.2	61.1	69.6	70.3	62	70.9	84.1	57.3	66.3	86.6	55.4	69.5	63.2	65.5	73.4	59.1	70.4	72.4
600,80	69.2	69.2	72.4	61.3	69.5	71.2	62.3	69.9	84.6	57.8	69.1	86.9	54.6	71.25	62.1	66.3	72.9	60.1	71.1	72.5
700,90	68.1	69.9	72.5	61,5	69.1	71.1	62.5	70.1	85.4	57.9	68.3	85.9	53.9	71.45	62.3	67.9	71.3	60.9	71.3	72.9

REFERENCES

[1] , Inc. (2013). IPO Prospectus. Downloaded on February 2,2014,*(http://www.sec.gov/Archives/edgar/data/1418091/000119312513390321/d564001ds1.htm).

[2] Jansen, B., Zhang, M., Sobel, K., and Chowdury A, " Twitter Power: Tweets as Electronic Word of Mouth". Journal of the American Society for Information Science and Technology, 2009,60.

[3] Abbasi, A., Hassan, A., and Dhar, M , "Benchmarking Twitter Sentiment Analysis Tools", Proc of LREC Conf, 2014.

[4] Ghiassi, M., Skinner, J., and Zimbra, D, " Twitter Brand Sentiment Analysis: A Hybrid System using Ngram Analysis and Dynamic Artificial Neural Network" Expert Systems with Applications,2013, 40.

[5] Tetlock, P, " Giving Content to Investor Sentiment: The Role of Media in the Stock Market", The Journal of Finance, 2007,62.

[6] Pang, B., Lee, L., and Vaithyanathan, S.," Thumbs Up?: Sentiment Classification using Machine Learning Techniques", Proc of the ACL Conf on EMNLP, 2002.

[7] Gamon, M." Sentiment Classification on Customer Feedback Data: Noisy Data, Large Feature Vectors, and the Role of Linguistic Analysis", Proc of Conf on Computational Linguistics, 2004.

[8] Abbasi, A., Chen, H., and Salem. A," Sentiment Analysis in Multiple Languages: Feature Selection for Opinion Classification in Web Forums", ACM Transactions on Information Systems, 2008, 26(3).

[9] Kim, S. and Hovy. E, "Determining the Sentiment of Opinions", Proc of the Intl Conf on Computational Linguistics, 2004.

[10] Sentiment140 (2015). www.sentiment140.com.

[11] Go, A., Bhayani, R., and Huang, L, "Twitter Sentiment Classification using Distant Supervision. Technical Report", Stanford Digital Library Technologies Project, 2009.

[12] Pak, A. and Paroubek, P," Twitter as a Corpus for Sentiment Analysis and Opinion Mining.", Proc of LREC Conf, 2010.

[13] Barbosa, L. and Feng, J,"Robust Sentiment Detection on Twitter from Biased and Noisy Data", Proc of COLING Conf, 2010.

[14] Montejo-Ráez, A., Martínez-Cámara, E., Martín- Valdivia, M. T., and Ureña-López, L. A," Ranked WordNet Graph for Sentiment Polarity Classification in Twitter", Computer Speech and Language, 2014, 28.

[15] Hassan, A., Abbasi, A., and Zeng, D, "Twitter Sentiment Analysis: A Bootstrap Ensemble Framework", Proc of the ASE/IEEE Conf on Social Computing, 2013.

[16] B.N. Supriya, Vish Kallimani, S.Prakash, C.B. Akki(2016). Twitter Sentiment Analysis using Binary Classification Technique, Proc of ICTCC Conf.

[17] Sung-Lin Chan, Xiangzhe Meng, S¨uha Kagan K¨ose, “EPFL Machine Learning CS- 433 - Project 2 Twitter Sentiment Analysis “, Article · January 2018.

[18] Babacar Gaye1, Aziguli Wulamu, “ Sentimental Analysis for Online Reviews using Machine learning Algorithms”, International Research Journal of Engineering and Technology (IRJET), Volume: 06 Issue: 08 | Aug 2019, e-ISSN: 2395-0056, p-ISSN: 2395-0072.

[19] Carol Anne Hargreaves,” Analysis of Hotel Guest Satisfaction Ratings and Reviews: An Application in Singapore”, American Journal of Marketing Research Vol. 1, No. 4,2015, pp. 208-214.

[20] Chen, Tianqi, and Carlos Guestrin. "Xgboost: A scalable tree boosting system." In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pp. 785-794. ACM, 2016.

[21] Sami Belkacem, Kamel Boukhalfa, Omar Boussaid,” Expertise-aware news feed updates recommendation: a random forest Approach”, Springer Science+Business Media, LLC, part of Springer Nature 2019, Cluster Computinghttps://doi.org/10.1007/s10586-019-03009-w.

[22] Kiran Sangada, Jitendrakumar Dhobi,” Sentiment Analysis of Tweets for Indian Election Using Random Forest Classifier”, International Journal of Technical Innovation in Modern Engineering & Science,2019 e-ISSN: 2455-2585 Volume 5, Issue 04, April-2019,page no1030-1033.

[23] Farideh Tavazoee, Claudio Conversano, Francesco Mola” Recurrentrandomforestfortheassessmentofpopularity insocialmedia 2016USelectionasacasestud”, KnowledgeandInformationSystems, Springer-VerlagLondonLtd.,partofSpringerNature2019, https://doi.org/10.1007/s10115-019-01410.

Supriya B N, received the Bachelor’s Degree in Computer Science and Engineering from East Point College of Engineering and Technology, Bengaluru, India in 2005; Completed Master’s Degree in Computer Science and Engineering from Bangalore Institute of Technology , Bengaluru India in 2012. Currently an Assistant Professor at SJB Institute of Technology, Bengaluru, India. Have academic experience. Special interests includes Social Network Analysis and Big Data . Published research papers in reputed conferences.

Dr. C.B.Akki, he received the Bachelor’s Degree in Electrical Engineering from University Vishvesvaraiah College of Engineering, Bengaluru, India in 1982, He received his Master’s Degree and Ph.D in Computer Science and Technology from University of Roorkee (IIT), India in 1990 and 1997 respectively. He is currently a professor at IIIT, Dharwad, India. He has both academic and Industrial experience in India and abroad. His Special interests are Wireless Communication, Mobile Computing and Computer Networks. He is a member for several research bodies and published numerous research papers in reputed journals.