Bulletin of Electrical Engineering and Informatics
Vol. 10, No. 5, October 2021, pp. 2845~2856
ISSN: 2302-9285, DOI: 10.11591/eei.v10i5.3078

Twitter sentimental analysis from time series facts: the implementation of enhanced support vector machine

Abhishek Kumar[1], Vishal Dutt[2], Vicente García-Díaz[3], Sushil Kumar Narang[4]
[1,4]Chitkara University School of Engineering and Technology, Chitkara University, Himachal Pradesh, India
[2]Department of Computer Science, Aryabhatta College, Ajmer, India
[3]Department of Computer Science, Universidad de Oviedo, Spain

Article Info

Article history:

Received Apr 18, 2021
Revised Jun 7, 2021
Accepted Aug 2, 2021

Keywords:

Bayes
Radial basis Kernel
Sentiment analysis
Senti WordNet
Support vector machine
Text data mining

ABSTRACT

Sentiment analysis through textual data mining is an indispensable system used to extract the contextual social information from the texts submitted by the intended users. Now days, world wide web is playing a vital source of textual content being shared in different communities by the people sharing their own sentiments through the websites or web blogs. Sentiment analysis has become a vital field of study since based on the extracted expressions, individuals or the businesses can access or update their reviews and take significant decisions. Sentimental mining is typically used to classify these reviews depending on its assessment as whether these reviews come out to be neutral, positive or negative. In our study, we have boosted feature selection technique with strong feature normalization for classifying the sentiments into negative, positive or neutral. Afterwards, support vector machine (SVM) classifier powered with radial basis kernel with adjusted hyper plane parameters, was employed to categorize reviews. Grid search with cross validation as well as logarithmic scale were employed for optimal values of hyper parameters. The classification results of this proposed system provides optimal results when compared to other state of art classification methods.

Corresponding Author:

Abhishek Kumar
Chitkara University Institute of Engineering and Technology
Chitkara University, Himachal Pradesh, India
Email: abhishek.kumar@chitkara.edu.in

1. INTRODUCTION

One of the most popular data analytics tools is sentiment analysis since it is frequently usedin different public as well as private sectors in the form of product/services survey or through social media like twitter data analysis. Usually the people or the communities place their opinions their websites forums,emails, blog forums,public blogs,and depict their opinions about the product, business processes or the decisions. Those opinions might be termed aspositive, negative or neutral about the indicative domains. Thesentiment analysis helps the targeted agencies to analyze people's reaction towards product usage and quality and future creations in the industries. Nowadays, world wide web is playing a vital role in the sentiment analytics because people generally like to share their own opinions through the websites or web blogs and allow their sentiments to be stored in the form of textual data. Sentiment expression analysis has become important since based on these expressions, individuals or the businesses can update or access their final verdict or reviews [1]. Sentiment analysis is mostly used to classify these textual reviews through some machine learning methodologies powered by natural language processing tools in the forms of probability scores of neutral, positive or negative classes.

Typically, sentiment analysis employs classifiers utilizing machine learning and lexicon related methodologies. Lexicon orientedmethodology is word reference and corpus based approach which calculates the polarity or orientation of words through bag of words representation. Both suervised and unsupervised machine learning methods are supposed to be higly reliable in strongly categorizing and forecasting the sentiments as either neutral or negative or positive sentiments. Supervised approach inputs the labeled dataset along with its corresponding assigned sentiment classes during training wheras unsupervised methods use datasets without any labels [2] and sentiments are not pre-classified with its own labelled data.

Supervised and unsupervised machine learning essentially square measure an outline of how during the training process we let the machines to analyze the provided labelled information set. In supervised learning approach, the system is provided with the outcome of the algorithm and all the system needs to do is to figure out the steps to reach to that outcome during the learning phase. In case of unsupervised learning the system is not made aware of the outcome of individual data items and due to this fact the input data mostly is not concrete and due to this unsupervised learning remains challenging.

The classification process of opinion mining may be staged at three different echelons which speak out to be at sentence level,document level, or at object-oriented level. This research effort is focused at document based classification of subjectivity of text extracted from overall text from the single text document. Here, initially unstructured reviews of a movie are converted into an organized arrangement so as to extract the features. Then a correspondingrank score based upon the extracted features is found outto labeled word arrangements. Then the rank score is fed to the support vector machine (SVM) classifier to predict the sentiment conveyed through the text as neural, negative or positive.

2. RELATED WORK

Multiple feature extraction methods like bigrams,unigrams, or combination of both, combinations of unigrams and POS labeling of POS, unigrams, and location are taken into account. The machine learning based supervised classification techniques like bayes, logical regression, and SVM algorithms are applied on thesepreprocessed data. Human prediction is not better compared to machine learning classification algorithms [3]. Fuzzy based classification followed by tokenization, term frequency-inverse document frequency (TF-IDF) [4], stop word removal and POS tagging appliedat preprocessing stage before this method, improved the performanceof the system trained for movie reviews dataset.

The study reports that machine learning related algorithms has provided good classification results with accuracy of above 85% when employed supervised training for emotion based datasets [5]. Film and Twitter surveys are classified utilizing WordNet with its POS by deriving words with the similar meaning in the same context and followed by assigning the corresponding polarity in SentiWordNet dictionary. The resultshows an increasedaccuracy by 7% using machine learning classifiers likeada boost, random forest, decision tree,WordNet synset [6]. They have proposed through a paper which is pointed towards usage of supervised learning methods which are more accurate and efficient than semantic orientation based techniques but at the same time, computation as well as time complexity is reportedly high [7]. Tweet textsentiment miningmodel consisted of the preprocessing, feature selection, and identification modules [8].

An enhanced feature selection method is developed during the preprocessing step, unwanted word removal, stemming, and pos tagging are performed. Feature selection methods such as mutual information (MI), information gain (IG), chi-square (λ2), and TF-IDF are used to extract appropriate features [9]. Twitter sentiment analysis using binary cluster-based framework to map the related components in the tweet [10] helps the researcher to identify the twitter tweets, which includes implementation of ranking and the scoring of the features collected and reported as more accurate. The major advantage of this concept is to analyze the occurrence of a specific key-word pair. This process creates a model to understand the event handling in an effective manner. The process extracts and selects those feature variables which are required to train the model. The features being considered take the occurrence count and force the model to learn the combinations. This process is not supposed to work with the image based classification of the events and tweets.

Ensemble modelling [11] helpsin better implementation of the sentimental analysis. This process is based on fuzzy logic implementation where the tokenization of the keywords is considered withpart-of-speech (POS) tagging of the keywords in the feature. Every word is considered as the feature and the POS defines the analysis of the model. Ensemble modelling helps to learn the model with multiple occurrences. The occurrences with all the possible fuzzy logics are handled and accordingly processed. In a fuzzy set, all the possible mathematical operations can be performed with the probability of occurrence of the event. The time-series methodology can be implemented using fuzzy logic. The dataset consists of the time frequencies of the tweets on a specific topic and the model analyzes the frequency and the word count (specific key-

word) related to the domain of targeted tweet. If the tweet in unreadable or in concise format, then model may not identify the tweet and the prediction fails.

Statistical analysis [12] is the key concept in major implementations of sentimental analysis. Time series-based modelling ensures that frequency of occurrence of a tweet with specific key terms is measured and plotted. Time-series, M-model and autocorrection creates a path to predict some insights of the data using fuzzy logic. The statistical analysis makes users to understand the analysis of the event with the combinations on different standard based events. Statistical analysis takes the event in the form of time series with all possible data from the different time stamps. As mentioned Phan *et al*. in [11], here also, the tweets are calculated using the statistical modelling. The statistical modelling helps to analyze the intake of the tweets to the sever and in each server, the number of ways the analysis can be done, is predicted. The time-series models help to analyze the impact on intakes and predict the exact theory in the tweet. The same disadvantage as mentionedAhmad *et al*. in [12] is that it cannot be identified when there is unreadable format of tweet and also if the tweet is made from the virtual private network (VPN) based location.

A systematic special and temporal sentiment analysis [13] creates the model based on the user mental stability. For every tweet, an internal mental stability of the user is associated. Based on the internal mind stability, the tweets are posted. In this article, researchers have performed sentimental analysis on twitter data using mental stability of the users.

Geo-tag [14] based implementation have been helpful for the researchers to locate the user who is tweeting on a specific topic. Geo-tag reflects the motto of the tweet from a specific user. The location-based analysis helps the user to analyze the location-based tweets and the reason for happening of the tweets without the controversy. The tweet tag wars can be analyzed based on the location of tweets and this can help the stakeholders to monitor the issues on special events.

3. PROPOSED SYSTEM

The proposed method is explained is being as. The system consists of five major phases which arepreprocessing,feature extraction, feature normalization,feature selection, and classification. Here, we have used supervised learning approach. We have used two datasets for training the model, validating the classifier then finally the classifier classifies the input text based on training data. The radial basic $K(x, y)$ function kernel is used here and optimization is also performed to increase the performance of the system.

Step 1: Collection of online sentiment review dataset

In this paper, we have used a polarity based movie review dataset. Record of separate content is kept up for every survey. Moreover, Twitter and Gold dataset are additionally produced to show results of proposed technique on various datasets. Twitter application programming interface (API) is used for taking the Twitter dataset and amazon website is used to collect the gold dataset.

Step 2: Data-preprocessing of the obtained dataset

All the Reviews contents are not found to be completely informative or directly expressing the significance of the opinion because it contains some contamination therefore preprocessing is very much important to remove those impurities.

- Eliminate unwanted attentions: All attentions that genuinelycommunicate abundance are drained.
- Removal of stopping words: Usage of some words are quite common in any language known as stop words. These stop words should be removed as a step of cleaning the textual data. These words do not create a significant impact upon the contextual or subjective meaning of the whole sentence. Samples of stop words hold i, a, are, is, an and so on.
- Process of stemming: There are a lot of forms of a single word which are derivatively related and stemming is done to remove such affixes to the words to look like similar.
- Porter stemmer algorithm is employed for effectively completing the word during stemming. It limits the list of variant forms of the words and makes useful grouping of these words.
- During grammatical tagging, parts-of-speech (POS) of words may be used as a linguistics classification which is characterised by its syntactical or morphological conduct. Things, action words, modifiers, pronouns, relative words, combinations, and interposition area units fall under POS regular classifications.
- POS labeling is basically denoting each word with its appropriate POS during grammatical tagging. Here we have used stanford POS tagger for this tagging procecss.
- SentiWordNet is employed to provide sentiment scores to the tagged words which are used as an input to the SVM classifier to characterize reviews. The neutral, positive, and negative word scores are effectively characterised within the SentiWordNet lexicon.

Step 3: Classification using enhanced SVM

After completion of preprocessingphase, the preprocessed dataset is fed as input to the SVM and Bayes classifiers for prediction of sentiments. We have tweaked the hyper plane parameters for better classification. SVM classification powered with the radial basic function kernel allows all the data to be spread over and therebythe center is chosen based on nearest support vector enroute to classifying the input data. In SVM the Hyper Plane is defined is being as by the relation:

$$\sum_j Q_i K(x_i, x) \quad (1)$$

where: *Q*: denotes the affine subspace,

K (x_i, x)*:*is the Kernal function.

The following operation is performed to reduce the expression of the form by the soft margin of the SVM classifier:

$$\left[\frac{1}{n}\sum_{j=1}^{n} max\left(0,1 - y_j\left(w^T x_j - b\right)\right)\right] + \lambda||w||^2 \quad (2)$$

Step 4: Outcome

The confusion matrix characterizes the performance of the system and allows us to figure out the errors if any, incurred during the classification process. It provides us the total number of correct and incorrect predictions done with the test data along with the total number of counts in each class. The number of negative cases predicted correctly, the number of positive cases predicted correctly, the number of actual negative cases predicted positively, and the number of actual positive cases predicted negatives, which are called true negatives, true positves, false positives, and false negatives correspondingly, are provided in the matrix which are used to calculate the overall accuracy of classification. It is termed as the best operational tool to analyze the system performance.

Figure 1 depicts the complete process flow in the system. It is apparent that the textual data captured through tweets and reviews, is basically unstructured therefore we need to apply natural language text processing over the text as a part of preprocessing because it shall reduce the unwanted or noisy data and in turn make it homogeneous text and we shall be able to provide accurate data to the next level of processing. The input is preprocessed then the preprocessed data is passed to the feature extraction block then the feature normalizer performs the function of standardization and the best features are selected by the feature selector. Then this output is fed to the classifier in order to procure the decision of sentiment analysis. The Naïve Bayes and SVM classifiers are used to classify the input data. The same classification was performed with optimized SVM and the scores were compared.

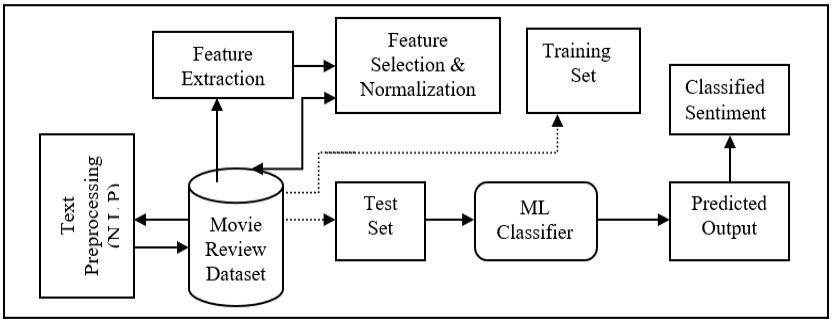

Figure 1. Proposed process flow

The optimized SVM provides better feature extraction which makes it helpful to extract the important information from the given data so that the classifier can differentiate the data between the classes. Feature normalization also helps in avoiding the over fitting and redundancy. We use feature normalization to reduce the data into double precision and the feature selection to reduce the dimensionality and select the best features for further training the model.

3.1. Feature normalisation

Min-max standardization is one amongst the foremost common ways to normalize data. For each feature, the minimum cost of that feature gets reworked into a zero, the maximum cost gets reworked into

one, and thereby each different cost gets reduced into a decimal between zero and one. When we do normalization, all the high and low feature values are reduced between zero and one. For example, if the minimum cost of a feature was twenty, and the maximum cost was forty, then thirty would be reworked to regarding 0.5 since it's halfway between twenty and forty.

Min-max social control has one fairly vital downside. It doesn't handle outliers alright, for instance, if you have got ninety nine values between zero and forty, and one cost is a hundred, then the ninety nine values can all be reduced to a worth between zero and one. That knowledge is simply as squished as before! Take a glance at the image below to visualize associate example of this. Figure 2 shows the min-max normalization of the word frequencies being normalized between 0 and 1.

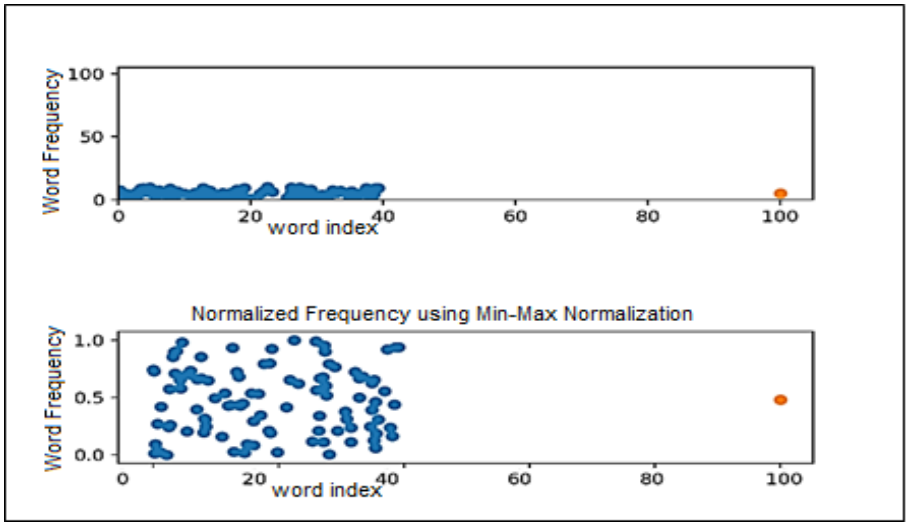

Figure 2. Min-max normalizaition

3.2. Feature selection

Feature selection is generally used to select the best features in order to assist the classifier to take better decision by reducing the number of training data values. The feature selection refines the features and provide them to the learning phase. This helps to make classifier more accurate and efficient as the presence of some of the features do not provide much information and in such cases the feature selector remove those features and provide worthier features to the model and helps us to receive increased system accuracy. We have employed wrapper method for feature selection.

In wrapper techniques, we tend to try and use a set of features and vicitimize the particularmodel under those selected features. Using Bi-directional elimination, we start with a null model and keep on adding a feature to it step wise using forward selection. Before adding a new feature, the significance of the existing feature is verified and if it is found insignificant, it is removed. The drawback of the method is that it is basically reduced to a hunt problem and sometimes becomes computationally terribly overpriced. Some common samples of wrapper ways include forward selection of features, elimination of features from backward, and algorithmic based elimination offeature.

- Selection in forward: Forward choice is a repetitious methodology during which we tend to begin with having no feature within the model. Eachand every iteration, we tend to retain accumulating the feature that increases the accuracy of the model till a tally of a fresh parameter variable does not provide any improvement in the performance of the system.
- Elimination from backward: In this backward exclusion context, we tend to begin with all options and eliminate the least volume of vital feature variable at each and every iteration that provides the best performance of the system model.
- Recursive feature exclusion: It is a dynamic improvement rule that aims to seek out the worst performing feature set. It frequently forms the new model and keeps aside the worst performing feature at every repetition. This constructs the successive model with the left options till all the options are exhausted. Then it ranks the choices maintained from the order of their elimination.

The easy and simple way for feature selection is the wrapper method as it provides the best or the optimal features with less computation time and finds the significanceof the each and every feature. Following steps are performed for its implementation.

a. First, it adds randomness to the given knowledge set by making shuffled copies of all options.
b. Then, it trains with the help ofrandom forest classifier on the comprehensive knowledge set and applies a feature significance to judge the significance level of every feature, wherever higher marks that feature vital.
c. At each iteration, it checks whether or not a true feature features a higher significance than the simplest of its shadow feature (i.e. whether or not the feature features a higher Z-score than the utmost Z-score of its shadow options) and perpetually removes features that the intended method deemed extremely unimportant.
d. At the end, the rule stops either once all selections get includedor excluded or it reaches a maximum bound of random forest runs.

4. CLASSIFICATION

SVM is the one of foremost classification tools which can be used for binary classification as well as multi class problems. Binary classification is used for classifying the data in to two classes while multi-class can be utilized to classify the data in to more than two classes. Here, we have used supervised SVM method and employed two datasets for the training. After the model is trained the classifier is tested for it performance when put under test data. A radial basis function kernal is used and optimization is also performed to increase the performance of the system.

SVM is a very useful system for grouping or classifying the labelled data. Before the classification task is intitiated, the whole information is divided into training and test data sets which comprise of a fixed percentage of data sets [15] respectively. Each case in the training set contains one objective output and a few characteristics. The objective of SVMis to deliver a model which predicts target estimation of data occurrences in the test set which comprises of unlabeled features only.

Classification process by SVMis recognized as supervised knowledge based system. The output labels help in the test data demonstrate whether the framework is acting in a correct manner or not. The aim of SVM classification is to discover a hyperplane which is possibly a line, 2 dimensional (2D) or a 3D plane depending upon the number of outcome classes.

SVM [16], [17] classifier after training finds the hyperplane, that sets the constraints α and b. This SVM has an alternative arrangement of parameters called hyper parameters [18], [19]: Gaussian radial basis kernel, the constant of soft marginal, C, and any constraints the centermay rely upon (widthor level of a kernels). In this paper, we show the eff ect of the hyper parameters on the boundary of a SVMutilizing two-dimensional models. For a huge estimation of C, a huge penalty is allocated to mistakes/edge blunders. This is found where the two nearest data points to the hyperplane aff ect its alignment, causing in a hyperplane which approaches a few other data values. The penalty parameter permits a specific level of misclassification, which is especially significant for non-detachable training sets. It gives a chance to control the exchange-off between permitting training blunders and compelling unbending edges. Expanding this worth additionally creates the expense of misclassifying targets and makes a progressive model that may not sum up well. A Threshold slider is utilized to show the degree of certainty that the nearest fragments of some random portion express a similar class as that section. Higher values mean more certainty, so just the closest portions are ordered.

The dimensions are transformed to higher order for nonlinear data, where multiplication of test inputwith each and every support vector is performed. So no need ofnonlinear mapping is generated. Further process is similar to that of linear data case. The data points which are closer to the hyperplane, are used to maximize the distance between classes so that the future data points are classified correctly. The SVM classifier, utilized to categorizereviews,uses radial basis functionkernel and is adjusted by its hyper planeparameters with marginal constant and Gamma. So theenhanced SVMgives better outcomesas compared tolinear/non linear SVM, logical regressionand naivebayes classifier. The output performance of this proposed system provides the optimal result compared to other state of art methods.

4.1. Classifiers

4.1.1. Logistic regression

Logistic regression, inspite of containing the term ‘Regression’ [20], is used for classification by employing a linear/non-linear regression curve to produce discrete outputs. It based on maximum probability estimation [21] andqualitative based modelselection. A threshold value is always which specifies the class to which a data case is espected to put into. Logistical regression can be used to construct the model for multi-classification problems too.

4.1.2. Naive Bayes

Bayes theorem for conditional probability is used for classification by assigning class labels to test inputs which are nothing but some feature sets. The naïve bayes classifier acts on the principle that value of a feature is independent of the value of all other features for a given class. Naive Bayes classifieremploys the principle of maximum likehood [22] for parameter estimation. The classifier expects the data set in the form of a frequency table which is utilized to generate a likelihood table after the probability of each feature is calculated. The the Bayes theorem is applied to calculate the posterior probability. Because our review dataset is multinomial distributed [23], we have implemented multinomial naïve bayes classifier.

4.1.3. Dataset

We have used SentiWordNet dataset here. Discrete document is conserved for each andevery single review. Twitter gold datasetis additionally taken to indicate result of projected methodology on completely dissimilar dataset. Twitter API [24] used for extracting the Twitter dataset and amazon website is used to collect the gold dataset. The following pseudocode will illustrate the procedure of all the steps involved in this implementation.In the following algorithms the dataset is taken from the twitter API and coverted into dataset.csv file. Then API() and Img_set() was executed on the dataset to get the accuracy and performance measure of the algorithm. Using SKLearn library of python, the accuracy matrix was plotted.

```
            Twitter_analysis(d1,d2):
                Def ta1():
                        #grab the datasets from twitter API
                API(self, SecretKey, AuthenticationKey)
            If(SecretKey == (Username, Password)):
                Pd.write("Dataset.csv","w")
            Else:
                Return 0
                Def ta2():
                        #grab the image datasets related to twitter analysis
            Img_set(Username, Password):
                Authenticate(username, password)
            Return 0
            Ob1.API()
            Ob2.Img_set()
            #Ob1
            #import SVM from SKLearn
            Plot Ob1.svm
            Plot Ob2.svm
            #import accuracy matrix from SKLearn
            Plot Accuracy matrix
            If( diff(y^,y)>= 0.5):
                Repeat SVM
            Else:
                Return 0
Exit
```

5. RESULTS AND DISCUSSION

Accuracy, sensitivity, and specificity [25] these are the three parameter we are consider for performance analysis.

Accuracy: The accuracy can be calculated is being as:

$$Acc = \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad (3)$$

Sensitivity: The sensitivity can be calculated is being as:

$$Sen = \frac{TP}{(TP+FP)} \quad (4)$$

Specificity: The specificity can be calculated is being as:

$$\text{Spec} = \frac{\text{TN}}{(\text{TN}+\text{FP})} \quad (5)$$

F1 Score: The f1-score can be calculated is being as:

$$\text{Fsc}=2 \cdot \frac{(\text{Precision} * \text{Recall})}{(\text{Precision}+\text{Recall})} \quad (6)$$

Precision: The precision can be calculated is being as:

$$P = \frac{TP}{(TP+FP)} \quad (7)$$

Recall:The recall can be calculated is being as:

$$P = \frac{TP}{(TP+FN)} \quad (8)$$

The Table 1 shows performance of proposed method which far better than others. Figure 3 shows a comparative analysis of three classfiers employed for sentiment subjectivity analysis. Accuracy show the training accuracy of the three models and F1-score [26] being a better estimator of as compared to precision and recall, depicts the validation accuracy. The validation accuracy being less than training accuracy clearly points out that our model has not overfitted. Also a validation accuracy of 94% by SVM classifier is substantially ahead of 90% and 87% as shown by Naïve Bayes and logistic regression respectively. A high senstiivity and specificity ratio attained by SVM also suggests that the model is able to correctly classify true positives and true negatives.

Table 1. Performance with proposed method

Method	Accuracy	Sensitivity	Specificity	F1-Score	Precision	Recall
Bayes	87	85	89	87	82	83
Logistic Regression	91	89	93	90	90.65	92.45
Proposed Methodology	97.5	96	97	94	96	93.01

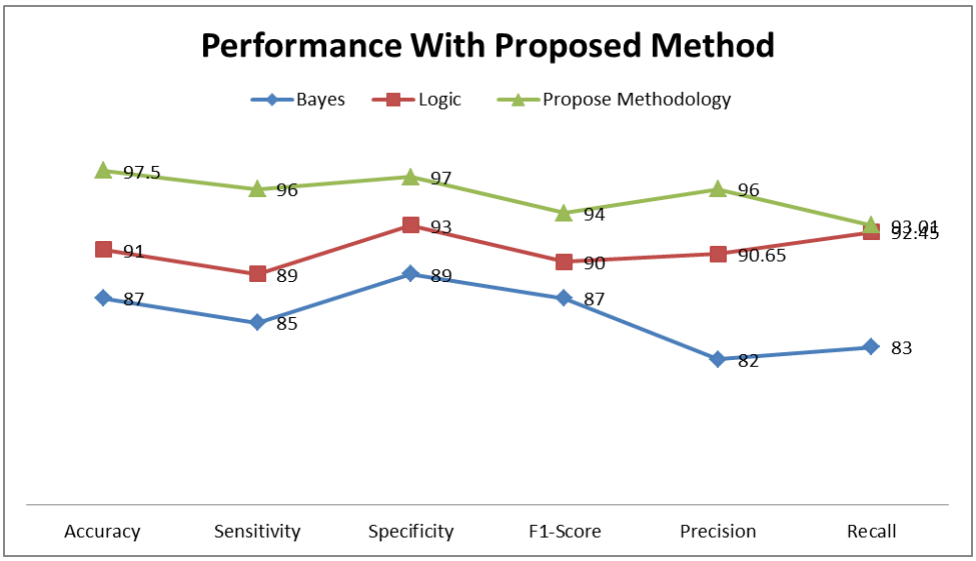

Figure 3. The performance visualization

Figure 4 demonstrates the receiver operator characteristic (ROC) curve [27] plotted for true positive rate versus false positive rate and it indicates the exact trade off between the sensitivity (TPR) and specificity (1-FPR). All the classifiers tend to be closer to the left corner specified by 1 TPR and SVM being the closest to the corner shows its improvement over others. The performance of the region of convergence provides the better AUC for proposed system which is better compared to other conventional methods. The Table 2 shows performance accuracy, sensitivity, specificity of proposed method which is far better than others with 70-30 training and testing partition.

Figure 5 shows the performance of the classifiers after a cross validation is performed with a split of 70% train data and 30% test data. Figure 6 exhibits true positive rates, true negative rates, false positive rates and false negative rates from the confusion matrix by proposed classifer with that of other classifier. The table shows the true positive true negative, false positive and false negatives performance of proposed method s better

than others. The Table 3 shows the accuracy sensitivity and and specificity of proposed method higher than the others. Figure 7 shows the graphical representation of the classification summary parameters of the classifiers when modelled against some different real world reviews data from a different source.

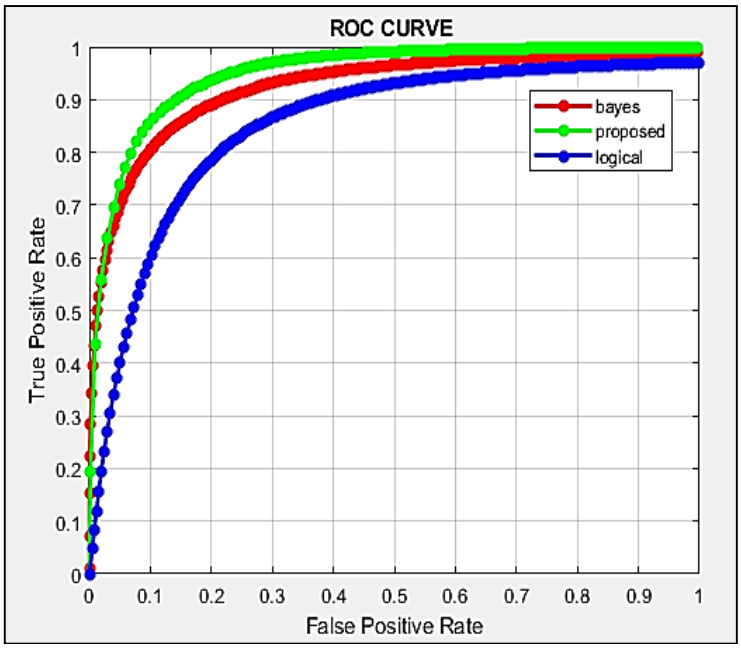

Figure 4. ROC performance

Table 2. performance with proposed method with 70-30 partition

Method	Accuracy	Sensitivity	Specificity	F1-Score	Precision	Recall
Bayes	81	83	82	81	79.5	80.2
Logistic Regression	89	87	91	81.36	82.12	80.98
Proposed Methodology	96.5	94	96	95	94.5	96.9

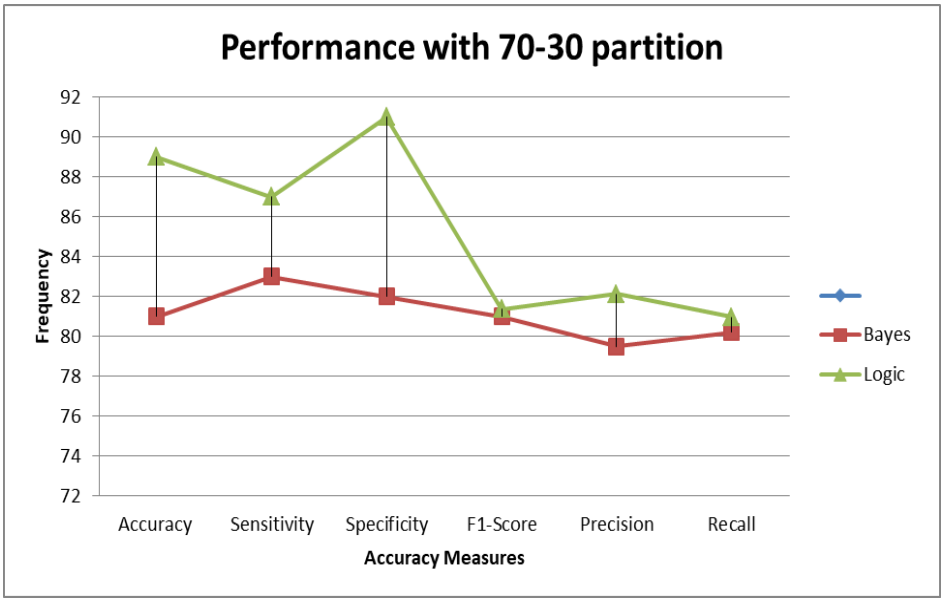

Figure 5. Performance visualization with 70-30 partitions

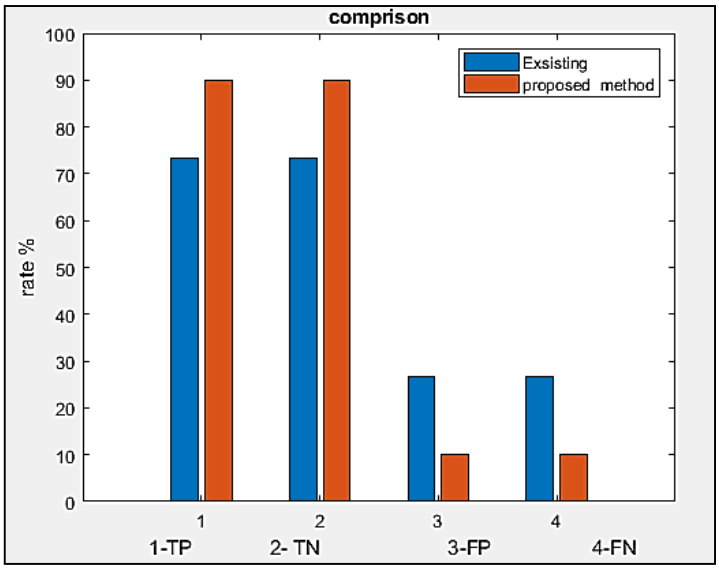

Figure 6. Feature SEMI vs FULL (hybrid)

Table 3. Performance with proposed method with real word data

Method	Accuracy	Sensitivity	Specificity	F1-Score	Precision	Recall
Bayes	86	84	87	85	84	82
Logistic Regression	91	88	92	94	91	91
Proposed Methodology	96.5	95	96.5	95	94.5	96.9

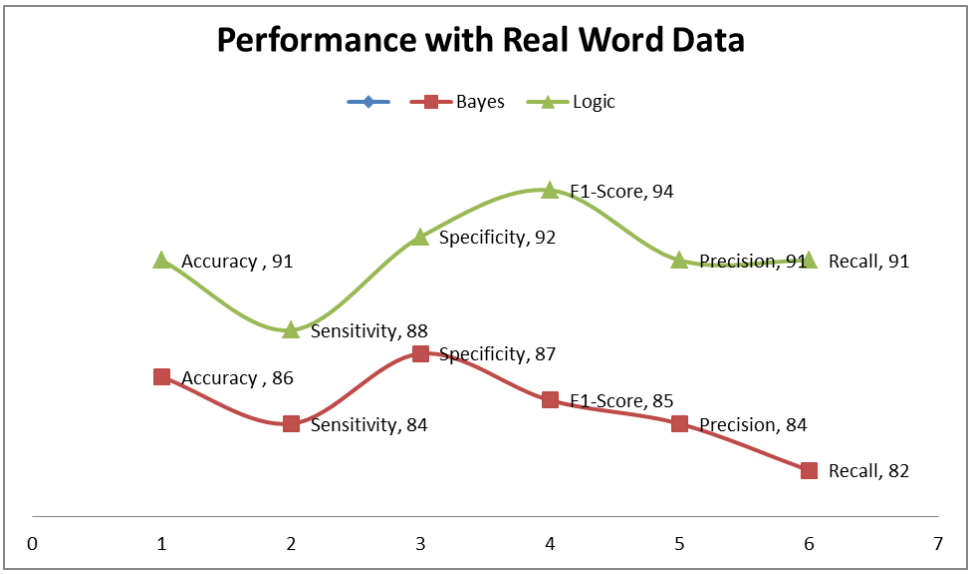

Figure 7. Performance visualization with real world data

6. CONCLUSION

Sentiment mining is an important analysis to categorize the user or human opinions for the future predictions and valuable outcomes. Here we have designed a novel sentiment mining system to build a better performing system. Also we have developed an enhanced SVM algorithm for better classification by changing the hyper parameter values exploitingthe feature selection and feature normalization processes. Both feature normalization and feature selection prove to be very helpful for better classification of classifiers.We have used featurenormalization to reduce the data into double precision and the feature selection to select the optimal features for further processing.The output performance of this proposed system

provide the optimal result as compared to other state of art methods. So this method can be used to analyse the sentimental data in the real time environments since it provides the better accuracy compared to other conventional methods.

REFERENCES

[1] B. Liu, ‘‘Sentiment analysis and opinion mining,’’*Synthesis lectures on human language technologies*, vol. 5, no. 1, pp.1-167, May 2012, doi: 10.2200/S00416ED1V01Y201204HLT016.

[2] Y. Singh, P. K. Bhatia, and O. Sangwan, "A review of studies on machine learning techniques," *International Journal of Computer Science and Security*, vol. 1, no. 1, pp. 70-84, 2007.

[3] B. Pang, L. Lee, and S. Vaithyanathan, ‘‘Thumbs up?: sentiment classification using machine learning techniques,’’in *Proceedings of the ACL-02 conference on Empirical methods in natural language processing*, *Association for Computational Linguistics*, vol. 10, pp. 79-86,July 2002, doi: 10.3115/1118693.1118704

[4] K. Mouthami, K. N. Devi, and V. M. Bhaskaran, ‘‘Sentiment analysis and classification based on textual reviews,’’ In *2013 international conference on Information communication and embedded systems (ICICES), IEEE*, February 2013, pp.271-276.

[5] A. Go, R. Bhayani, and L. Huang, ‘‘Twitter sentiment classification using distant supervision,’’*CS224N Project Report, Stanford*, vol. 1, p.12, 2009

[6] M. Kanakaraj and R. M. R. Guddeti, ‘‘Performance analysis of Ensemble methods on Twitter sentiment analysis using NLP techniques,’’in*2015 IEEE International Conference on Semantic Computing (ICSC)*, February2015pp. 169-170, doi: 10.1109/ICOSC.2015.7050801.

[7] P. Chaovalit and L. Zhou, ‘‘Movie review mining: A comparison between supervised and unsupervised classification approaches,’’in *Proceedings of the 38th Hawaii International Conference on System Sciences*, January 2005, pp. 112c- 112c,doi: 10.1109/HICSS.2005.445.

[8] X. Zhou, X. Tao, J. Yong, and Z. Yang, ‘‘Sentiment analysis on tweets for social events,’’in *Computer Supported Cooperative Work in Design (CSCWD), 2013 IEEE 17th International Conference on*, 2013, pp. 557-562, doi: 10.1109/CSCWD.2013.6581022.

[9] P. H. Shahana and B. Omman, ‘‘Evaluation of Features on Sentimental Analysis,’’*Procedia Computer Science*, vol. 46, pp.1585-1592, 2015, doi: 10.1016/j.procs.2015.02.088.

[10] M Bibi,W. Aziz, M. Almaraashi, I. H. Khan, M. S. A. Nadeem, and N. Habib, "A Cooperative Binary-Clustering Framwork based on majority voting for twitter sentimental analysis", *IEEE Access*, vol. 8, pp. 68580-68592, 2020, doi: 10.1109/ACCESS.2020.2983859.

[11] H. T. Phan,V. C. Tran, N. T. Nguyen, and D. Hwang, "Improving the Performance of Sentiment Analysis of Tweets Containing Fuzzy Sentiment Using the Feature Ensemble Model," *IEEE Access*, vol. 8, pp. 14630-14641 , 2020, doi: 10.1109/ACCESS.2019.2963702.

[12] M. Ahmad, S. Aftab, and S. S. Muhammad, "Machine Learning Techniques for Sentiment Analysis: A Review," *International Journal of Multidisciplinary Sciences and Engineering*, vol. 8, no. 3, p. 27, 2017.

[13] T. Hu, B. She, L. Duan. H. Yue, and J. Clunis, "A Systematic Spatial and Temporal Sentiment Analysis on Geo-Tweets," *IEEE Access*, vol. 8, pp. 8658-8667, 2019.

[14] W. L. Lim, C. C. Ho, and Choo-Yee Ting, "Sentiment Analysis by Fusing Text and Location Features of Geo-Tagged Tweets,"*IEEE Access*, vol. 8, pp.181014-181027, 2020.

[15] J.P.Lewis, ‘‘Tutorial on SVM, CGIT Lab, USC,’’ 2004, pp. 14-20.

[16] A. Tripathy, A. Agrawal, and S. K. Rath, ‘‘Classification of Sentimental Reviews Using Machine Learning Techniques,’’ Procedia Computer Science, vol. 57, pp.821-829, 2015, doi: 10.1016/j.procs.2015.07.523.

[17] G. A. Miller, ‘‘WordNet: a lexical database for English,’’ Communications of the ACM, vol. 38, no. 11, pp.39-41, November 1995, doi: 10.1145/219717.219748.

[18] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," in *Data Mining and Knowledge Discovery*", vol. 2, no. 2, pp. 121-167, 1998, doi: 10.1023/A:1009715923555.

[19] V. Vapnik, S. E. Golowich, and A. Smola,‘‘Support vector method for function approximation, regression estimation, and signal processing, *Advances in neural information processing systems*, 1997, pp. 281-287.

[20] E. Theodorosand P. Massimilliano, ‘‘Statistical Learning Theory: a Primer,’’ 1998.

[21] O. Bousquet, S. Boucheron, and G. Lugosi, "Introduction to Statistical Learning Theory", *Summer School on Machine Learning*, *Springer*, Berlin, Heidelberg, February 2003, pp. 169-207, doi: 10.1007/978-3-540-28650-9_8.

[22] N. Cristianini and J. Shawe-Taylor, "An Introduction to Support Vector Machines and Other Kernel-based Learning Methods," *Cambridge University Press*, 2000.

[23] D. M. Skapura, ‘‘Building Neural Networks,’’*ACM press*, 1996., ISBN: 0201539217, 9780201539219

[24] M. Rodríguez-Ibáñez,Francisco-Javier Gimeno-Blanes, P. M. Cuenca-Jimenez, S. Munoz-Romero, C. Soguero, and J. L. Rojo-Alvarez, "On the Statistical and Temporal Dynamics of Sentiment Analysis," *IEEE Access*, vol. 8, pp. 87994-88013, 2020, doi: 10.1109/ACCESS.2020.2987207.

[25] M. Hu and B. Liu, "Mining and summarizing customer reviews," in Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '04, 2004, p. 168.

[26] X. Ding, B. Liu, B. Liu, and P. S. Yu, "A holistic lexicon-based approach to opinion mining," *WSDM '08: Proceedings of the 2008 International Conference on Web Search and Data Mining*, pp. 231–240, 2008, doi: 10.1145/1341531.1341561.

[27] M. Bibi, W. Aziz, M. Almaraashi, I. H. Khan, M. S. A. Nadeem and N. Habib, "A Cooperative Binary-Clustering Framework Based on Majority Voting for Twitter Sentiment Analysis," in *IEEE Access*, vol. 8, pp. 68580-68592, 2020, doi: 10.1109/ACCESS.2020.2983859.

BIOGRAPHIES OF AUTHORS

Abhishek Kumar He is doctorate in computer science from University of Madras and done M. Tech in Computer Science & Engineering from Government engineering college Ajmer, Rajasthan Technical University, Kota India. He has total Academic teaching experience of more than 8 years with more than 80 publications in reputed, peer reviewed National and International Journals, books &Conferences.he has edited 18 books and authored 6 text books. His research area includes Artificial intelligence; Image processing, Computer Vision, Data Mining and Machine Learning. He has been in International Conference Committee of many International conferences. He has been the reviewer/editor of various peer-reviewed journals.

Vishal Dutt He is doctorate in computer science from University of Madras and done MCA (Gold Medallist) from MDS University, Ajmer, Rajasthan, India. He has been working as the Assistant Professor of Computer Science at Aryabhatta College, Ajmer, and also visiting faculty in Maharshi Dayanand Saraswati University (State Govt. University) Ajmer. He has total Academic teaching experience of more than 4 years. His research includes Data Science, Data Mining, Machine Learning and Deep Learning.He hasmore than 22+ publications in reputed, peer reviewed National and International, Scopus Journals, IEEE EXPLORE. He also has data analytics experience in Rapid Miner, Tableau, and WEKA. He hasbeen working as a freelancer for more than 6 years in the field of data analytics, Freelance Writer, Java, Assembly Programmer, Desktop Designer, and Android Developer.

Vicente García-Díaz He is an Associate Professor in the Department of Computer Science at the University ofOviedo. He has a PhD in Computer Science from theUniversity of Oviedo and a Diploma in Advanced Studies, as well as Degrees in Computer Engineering and TechnicalSystems Computer Engineering. In addition, he possessesa Degree in Occupational Risk Prevention. He is part of theeditorial and advisory board of several international journals. He has supervised 90+ academic projects and published 90+ research papers in journals, conferences, and books. His teaching areas are algorithm design techniques, anddesign and development of Domain-Specific languages. His research interestsalso include decision support systems and the use of technologies in teachingand learning.

Sushil Kumar Narang He is an Associate Professor in the Department of Computer Science and Engineeringat Chitkara University, Rajpura, Punjab (India) since 2019. From 2006-2019, He was head of IT Department at SAS Institute of IT & Research, Mohali, Punjab (India). From 1996-2006, He was AssistantProfessor at Department of Computer Science & Applications, MLN College, Yamunanagar, and Haryana (India). He Completed his Ph.D. at Panjab University, Chandigarh (India). His Research on “Feature Extractionand Neural Network Classifiers for Optical Character Recognition for Good quality handwritten Gurmukhiand Devnagari Characters” focused on various image processing, machine as well as deep learningalgorithms. His research interests lie in the area of programming languages, ranging from theory todesign to implementation, image processing, data analytics and machine learning. He has collaboratedactively with researchers in several other disciplines of computer science, particularly machine learningon real world use cases. He is a certified deep learning Engineer from Edureka. He possesses expertise in Object-Oriented Analysis; Design and Development using Java and Python programming using OpenCV in ImageProcessing and Neural Network construction. He has strong knowledge of C++ and Java withexperience in component architecture of product interface. With Solid training and management skills, He has demonstrated proficiency in leading and mentoring individuals to maximize levels ofproductivity, while forming cohesive team environments.