Classification of Sentiment on Business Data for Decision Making using Supervised Machine Learning Methods

Siji George C G, B. Sumathi

Abstract : Sentiment analysis is deals with the classification of sentiments expressed in a particular document. The analysis of user generated data by using sentiment analysis is very useful for knowing the opinion of a crowd. This paper is mainly aimed to tackle the problem of polarity categorization of sentiment analysis. A Detailed description of the sentiment analysis process is also given. Product review data set from UCI repository is used for analysis. This paper is giving a comparative analysis of four supervised machine learning algorithms namely Naive Bayes, Support Vector Machine, Decision Tree and Random Forest which are used for product review analysis. The result shows that, Random Forest classification algorithm provides better accuracy than other three algorithms.

Keywords : Decision Tree, Naive Bayes , Random Forest ,Sentiment Analysis, , Support Vector Machine(SVM).

I. INTRODUCTION

The popularity of internet changed the people's opinion methods. The people are sharing their thoughts, views and opinion through different social website such as blogs, facebook, twitter etc. Sentiment analysis is the process of figure out the feeling or opinion of people expressed in a document. The expressed opinion may be about a particular product or service .It is considered as one of the classification task. Sentiment analysis is mainly handled by machine learning approaches. In sentiment classification, a particular opinion/document is analyzed and classified as positive or negative.

The machine learning methods are classified in to two categories i.e, supervised machine learning and unsupervised machine learning. Among these, supervised machine learning methods give better performance. In this method, the machine is trained based on some labelled data and based on labelled data, it is expected to predict some unknown data in future. In this work, one of the supervised machine learning algorithms called Naive Bayes algorithm is used for product review analysis.

The product review can be analyzed to know the sentiment of a customer. The main objective of product review sentiment analysis is to analyze different classification algorithm to extract feature wise summary of a product. The companies can know about the customer expectations and limitations of their product and make modifications on it. It is also helpful for a new customer to purchase a particular product. The customers can perform comparative analysis to purchase a right product and right brand that suits their requirements.

Revised Manuscript Received on February 27, 2020.

Siji George C G, Ph.D. Scholar, CMS College of Science and Commerce, Coimbatore, Tamil Nadu, India. Email:-siji.gorg@gmail.com.

Dr. B. Sumathi, Associate Professor, CMS College of Science and Commerce, Coimbatore, Tamil Nadu, India. Email:-sumathithamizh@gmail.com

II. LITERATURE REVIEW

Huma Parveen and Prof.Shikha Pandey performed sentiment analysis on twitter reviews. The authors discussed pre-processing of data to remove noise from the data. The Hadoop framework helped the authors to analyze large number of tweets. The analysis performed on various perspectives like positive, negative and neutral. The proposed system also provides fast downloading for efficient twitter trend analysis.

D.MALI et al done sentiment analysis on product reviews in 2016 [2].The proposed model help the customers to gather information on product and also to take decisions. According to authors, NLTK with Naïve Bayes will give better performance for sentence level sentiment analysis.

Nurulhuda Zainuddin and Ali Selamat, [3] applied Support Vector Machine on benchmark datasets to train a classifier. They used N-grams and different weighting scheme to extract the most classical features. Authors explored Chi-square weight features to select informative features for classification. The result shows that the use of Chi-Square feature selection may provide high accuracy.

Sentiment analysis on Urdu Roman reviews are performed by Faiza Noor et al. in 2019[4].The authors collected 20.286 K reviews and are annotated in to three classes positive, negative and neutral. Bag of words model is applied for feature extraction and passed to Support vector Machine for classification. Different SVM Kernel is used for classification. Cubic Kernel provides high accuracy for classification.

Reviews on online transportation in Indonesia are analyzed by Rifkie Primartha et al. [5].The authors mainly concentrated on feature selection. Particle Swarm Optimization is applied on features, which are extracted by using TF-IDF. The authors performed different experiments based on different parameters of PSO such as population size, iteration size etc. The proposed method which combines both PSO and Decision Tree algorithm outperforms the original Decision Tree algorithm in terms of accuracy.

In "Sentiment analysis on product review data"[6], the authors used Amazon star scaled product reviews .Five level staring system is considered in this paper. A general process for sentiment polarity categorization is performed on dataset. They experienced both sentence-level categorization and review-level categorization. SVM and Naive Bayes classifiers provide better performance than Random Forest classifier.

In 2019, Sirshendu Hore and Tanmay Bhattacharya analyzed social trends of girl child in India [7]. They collected reviews on "Beti Bachao Beti Padhao (BBBP)" government scheme, which aims women empowerment. Machine learning algorithms used to analyze review collected from Twitter. Among four used machine learning classifiers, Random Forest classifier provides better accuracy. The experiments show that many of the opinions are neither positive nor negative and some of them are out of contest.

The impact of feature extraction on sentiment analysis is analyzed by Ravinder Ahuja et al.[8].The impact of two feature TF-IDF word level and SS-Tweet on dataset is analysed by authors. The results shows that, TF-IDF provided 3-4% of performance higher than N-Gram. Six classifiers are used for comparison; among that Logic regression gave best predictions of sentiments.

Large scale Amazon product review analysis is carried out by Tanjim Ul Haque et al in 2018[9].The authors performed sentiment analysis using six machine learning algorithms on three different product review dataset and conducted cross validation methods also. TF-IDF and Bag-of-Words model gave better performance. The proposed supervised machine learning system with Support Vector Machine classifier provides high accuracy on large scale data.

Aashutosh Bhatt et al proposed a rule based extraction system [10] for customer reviews analysis for finding the sentiment of the reviews. The system was accurate enough to find the sentiment on iPhone 5 reviews on Amazon. The authors depicted the results of their analysis through different charts for easy evaluation.

III.RESEARCH DESIGN AND METHODOLOGY

Sentiment analysis architecture includes mainly three stages namely pre-processing, feature extraction and model development. It is depicted in Fig(1)

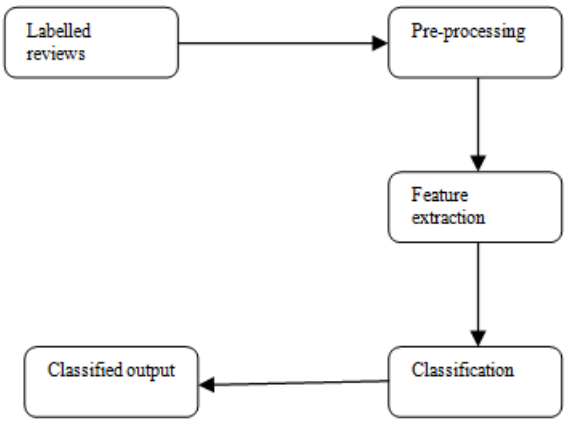


Fig. 1. Sentiment Analysis Architecture

A. Pre-processing

The pre-processing task will clean the dataset and prepare it for the classification task. The dataset may include number of noise and unwanted words which do not give any sentiment on it. Moreover, there are many words that do not have any impact on the general orientation of it and this will make the classification process difficult. This is the situation, in which pre-processing task has to be done. A proper pr-processing will improve the speed and performance of a classifier. The pre-processing task used in this work are

- Null fields removal
- Word tokenization
- Type conversion
- Stop words removal
- Stemming
- *Null fields' removal:*-The dataset may include records which do not have review text and score. All these records are removed in order to reduce the processing time. The records with score 3(neutral) is removed since the analysis mainly focus on binary classification i.e, positive and negative classes
- *Word tokenization:* - In tokenization, a sequence of string is breaking in to number of elements called tokens. Tokenization is vital to find out different pattern in a particular text. These tokens can be taken as the input for the classification process.
- *Type conversion:*-All upper case letters in the input record are converted in to lower case letters.
- *Stop words removal:* - The different stop words such as the, an, or, and etc which are not giving any sentiment is removed in pre-processing step.
- *Stemming:* - In stemming, each word is reduced in to their root word and it is known as stem. That is, it removes the suffixes from words in English. The number of words used for analysis can be reduced by using stemming.

B. Feature Extraction

The machine can not directly deals with the collected input. So, it should be converted in to machine readable format. Bag-of-Words model is used for feature extraction. It is mainly used to extract features from text document. The model will create a vocabulary of all the unique words appear in the training dataset. It will consider only frequency of words in the given document and disregard its order and their relationship with words. In this the occurrence of words are used as features for classification. The two Bag-of-Words feature extraction model used are

- CountVectorizer
- TF-IDF(Term Frequency and Inverse Document Frequency) Vectorizer

- *CountVectorizer*:-It is the simplest feature extraction method. Number of occurrence of each word or token in the given document will consider as count. CountVectorizer is used to convert collection of text in to number of counts.
- *Term Frequency and Inverse Document Frequency*: The word count by CountVectorizer is basic. The significance of each word for a given document can be finding by using Term Frequency and Inverse Document Frequency. It find both word count (Term Frequency) in the given document and importance of a word to each document(Inverse Document Frequency).Both CountVectorizer and TF-IDF Vectorizer are used for extracting features from given set of customer product reviews.

C. Classification

Classification is one of the major data mining function and it is used to assign records/items to different category or class. The main objective of this work is to perform binary classification on product review data set. Four machine learning algorithms are used for this purpose and they are

- Naive Bayes (NB)
- Support Vector Machine (SVM)
- Decision Tree
- Random Forest

Naive Bayes:- It is the simplest machine learning algorithm which is used for classifying text document. High dimensional training data set is involved in this algorithm. The Bayes theorem is the basis of Naive Bayes algorithm. Since the occurrence of one feature is not dependant on other's occurrence, the Bayes theorem is called Naive. For each class, the Bayes theorem will predict the membership probabilities. The most likely class will be the class with highest probability. The Naive Bayes is given by the equation

$$P(c|x) = P(x|c)P(c)/p(x) \quad (1)$$

Where P(c|x) is the probability of class x, P(c) is the probability of class c,P(X) is the probability of class x and P(x|c) is the probability of class c given class x.

In this work, Naive Bayes classifier is used to estimate whether the probability of a product review document is to be positive or to be negative. The procedure for Naive Bayes classifier used in this work is as follows

NBProc( )

- Consider product review data set D which include positive (P) and negative (N) reviews.
- Find the prior probability for each class P and N
  - Class P=number of object of class P/total number of objects
  - Class N=number of object of class N/total number of objects
- Calculate the total number of word frequencies n_p (for class P) and n_n(for class N)
- Calculate the conditional probability of keyword occurrence of each class
  - $P(word_i/Class_i)$=wordcount/n_i(P)
  - $P(word_i/Class_i)$=wordcount/n_i(N)
  - where i=1 to n
- Predict the class of a new product review document M by calculating probability for each class P and N, P(M/W)
  - Find P(P/W)=P(P)*P(word1/classp)*P(word2/classp).....*P(wordn/classp)
  - Find P(N/W)=P(N)*P(word1/classn)*P(word2/classn).....*P(wordn/classn)
- Class with higher probability will be assigned to the document M.

Support Vector Machine(SVM):-The main objective of Support Vector Machine classifier is to draw a hyper plane that separate or different classes of data points. In SVM, the hyper plane with maximum margin will be selected. Maximum margin will improve the performance of the classifier. The mathematical representation of the hyper plane is

$$\beta o + \beta 1X1 + \beta 2X2 + \beta 3X3 + \cdots \beta pXp = 0 \quad (2)$$

If X satisfies this equation, then the point will be on the plane, otherwise, it will be on one side of the plane. The performance of SVM is relies on particular kernel function .The selection of SVM kernel is very important for the performance of the classifier.The SVM procedure for this work is as follows

SVMProc( )

- Consider the product review data set D of n couple of elements (x_i, y_i)
- Each x_i is associated with a value y_i indicating if the elements belongs to the class (+1) or not (-1)
- Select the two hyper planes separating the data with no points between them by using the formula

$$w.x + b = 0$$

- For each vector x_i either

$$w.x_i + b \geq 1 \text{ for } x_i \text{ having the class 1}$$

or

$$w.x_i + b \leq 1 \text{ for } x_i \text{ having the class -1}$$

- Maximize the distance (margin) between the two hyper planes by using the formula

$$m = \frac{2}{||w||}$$

Decision Tree Classifier:-Decision Tree classifier repeatedly divides the working space in to number of sub parts. It builds a classification model like a tree structure. Both discrete and continuous variables can be used for Decision tree classifier. A decision tree is a tree, in which every internal node is labelled with an input feature. Data comes in the form of

$$(X, Y)=(x_1, x_2, x_3, x_4,.....x_k, Y) \quad (3)$$

Where Y is the dependant variable that is going to be classified and X is the feature set used for classification.

In this work, ID3 decision tree algorithm is used. Entropy and Information Gain plays major role in ID3 algorithm. Homogeneity of a sample can be found by using Entropy. The entropy is reduced by calculating information gain. It compares the entropy of the given dataset before and after transformation. The Decision Tree classifier procedure used for in this work is as follows

DecisionTreeProc( )

- Given a set of product reviews D, their attribute values and S set of positive and negative examples.
- The entropy of set S relate to this binary classification is calculated as
  - $E(S) = -p(P)log2\,p(P) - p(N)log2\,p(N)$
- The data set is split on different attribute
- Find entropy for each attribute
- Find the total entropy for the split
- The resulting entropy is subtracted from the entropy before split

- $Gain(T,X) = E(T) - E(T,X)$
- The attribute with highest information gain is set as the decision node ,repeat the same process on every branch
- A branch with entropy 0 is set as leaf node
- A branch with entropy>0 need further splitting
- The steps repeated on non-leaf node until all data is `classified.

Random Forest:-It is a kind of ensemble algorithm. From training set, the Random Forest classifier generates a group of decision trees. It then aggregate results form decision tree to decide the final class The result of this algorithm is a combined output of each tree in the ensemble. It is a flexible algorithm. Forest consists of lot of trees. When the number of tree increases, the robustness of the classifier will also increase. The Random Forest procedure used for product review classification is as follows

RandomForestProc( )

- Consider the product review document with N training cases and M variables
- m number of input variable is used to determine the decision at a node of the tree, m should be much less than M
- Choose a training set by selecting n times with replacement from N available training cases.
- For each node of the tree, randomly choose m variable on which to base the decision at that node.
- Based on these m variables, calculate the best split
- Each node is fully grown and not pruned.
- For prediction, a new sample is pushed down the tree and it is assigned the label of the training sample in the terminal node.
- This procedure is iterated overall tree in the ensemble
- The average vote of all tree is taken as random forest prediction

IV.RESULT AND ANALYSIS

The aim of the work is to analyze the product reviews and thereby support the business organization to know the popularity of a particular product. Four machine learning algorithms are used for finding the sentiment on product reviews. Product review is represented in the form of positive or negative sentiment of users. Sentiment analysis on product review will help the business organization to know the market place of their products. The proposed system used supervised machine learning methods to train a classifier and to predict the sentiment on unknown product review in future.1500 product reviews are collected from UCI Repository. The reviews are the opinion of the customers regarding different product purchased from Amazon and it is in .csv file format. A star-scaled rating system is used for all these product reviews. Reviews with 5 stars and 4 stars are considered as positive reviews and reviews with 2 stars and 1 star is considered as negative review.

Table- I: Star scale rating

Star Level	General Meaning
☆☆☆☆☆	I Love it.
☆☆☆☆	I Like it.
☆☆☆	It's Okey.
☆☆	I Don't Like it.
☆	I Hate it.

There are different attributes in each review record such as id, brand, reviewdate, score, reviewstext but reviewtext has been considered for this proposed research. The main aim is to perform analysis on these reviews and conclude the reviews which are positive and negative.

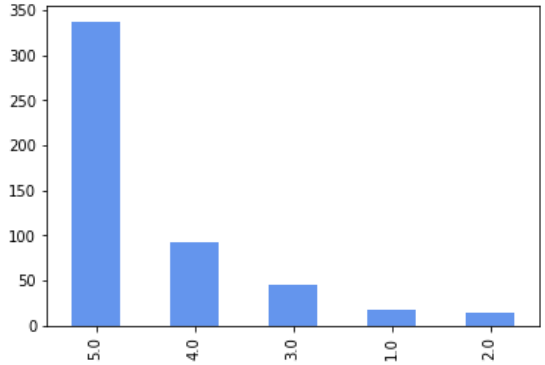


Fig.2. Data based on score levels

The purpose of the system is to perform binary classification with two classes of sentiments: positive and negative. The distribution of sentiments across all the product reviews by the proposed system is given in pie chart

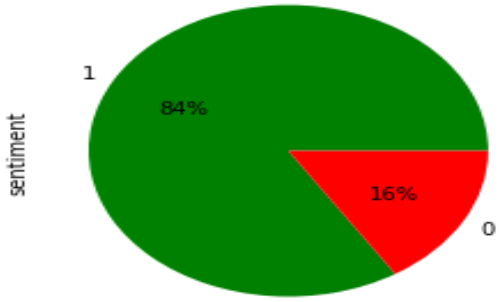
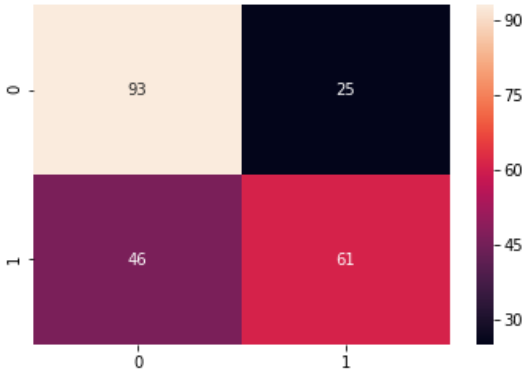
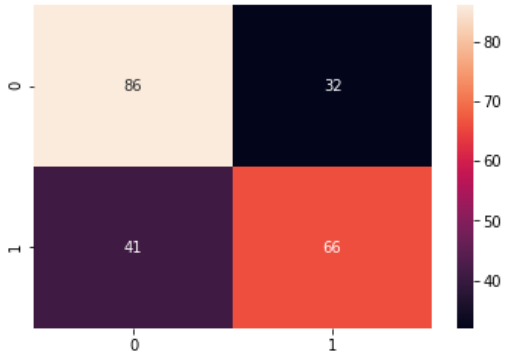


Fig.3. Sentiment distribution

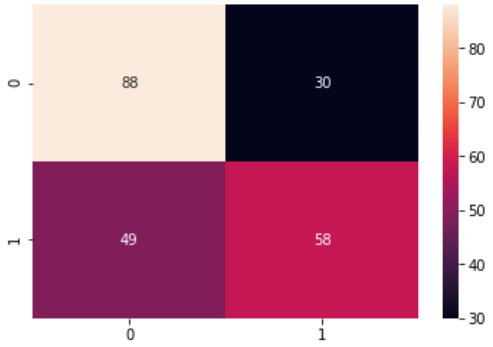
In this paper,four supervised machine learning algorithms(Naive Bayes,SVM,Random Forest,Decision Tree) are used for product review analysis and a detailed comparison is done.Heatmap for the four classifier is given in figure(1).It is the colourful 2D graphical representation of the confusion matrix.



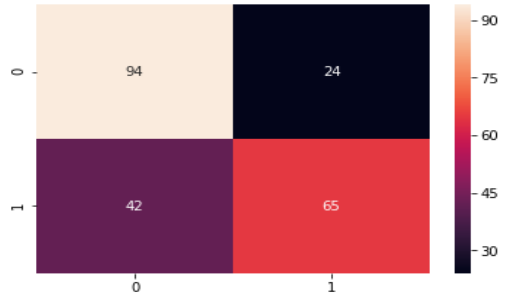
Naive Bayes



SVM



Decision Tree



Random Forest

Fig. 4. Heatmap for classifiers

The performance evaluation for the four used algorithms are given in Table(II)

Table –II: Evaluation parameters

Classifier	Evaluation Parameters			
	Precision	Recall	F-measure	Accuracy
Naive Bayes	69	70	84	70
SVM	68	73	70	67
Decision Tree	62	71	66	63
Random Forest	67	79	62	73

The ROC curve for the four used classifier is given in the Fig(5). ROC is a graphical plot that illustrates the ability of a binary classifier. It is created by using true positive rate and false positive rate.

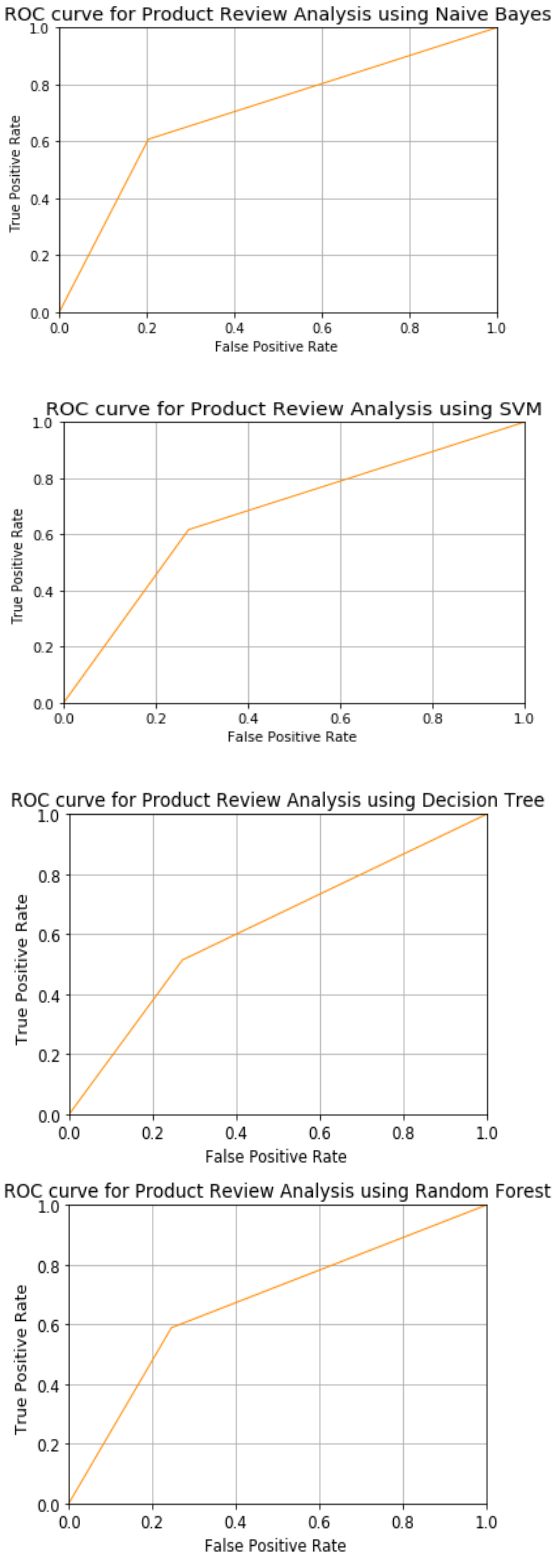


Fig. 5. ROC curve for classifiers

V. CONCLUSION

Sentiment analysis researches use machine learning approach for processing. Product review dataset is used to train a classification model using Naïve Bayes, SVM, and Decision Tree and Random Forest algorithm. The collected data is converted into metrics form using two feature extraction techniques, TF-IDF and CountVectorizer. A detailed comparison of four algorithms is done on the basis of evaluation parameters. It is noticed that Random Forest classifier provides better accuracy for product review analysis and it is 73%. The developed model helps the business organization in decision making.

REFERENCE

1. Huma Parveen and Prof.Shikha Pandey,"Sentiment Analysis on Twitter Data-set using Naive Bayes Algorithm", 2nd International Conference on Applied and Theoretical Computing and Communication Technology,IEEE 2016.
2. D.Mali, M.Abhyankar, P.Bhavarthi, K.Gaidhar, M.Bangare-"Sentiment Analysis of Product Reviews for Ecommerce Recommendation", International Journal of Management and Applied Science, Volume-2, Issue-1, Jan.-2016.
3. Nurulhuda Zainuddin and Ali Selamat," Sentiment Analysis Using Support Vector Machine", 2014 IEEE 2014 International Conference on Computer, Communication, and Control Technology (I4CT 2014), September 2 - 4, 2014 - Langkawi, Kedah, Malaysia.
4. Faiza Noor, Maheen Bakhtyar(B) , and Junaid Baber," Sentiment Analysis in E-commerce Using SVM on Roman Urdu Text", ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering,Springer,2019.
5. Rifkie Primarthaa, Bayu Adhi Tamab?, Azhary Arliansyaha, Kanda Januar Miraswana," Decision tree combined with PSO-based feature selection for sentiment analysis", IOP Conf,2018.
6. Xing Fang and Justin Zhan, " Sentiment analysis using product review data", Journal of Big Data,Springer,2015.
7. Sirshendu Hore and Tanmay Bhattacharya," Analyzing Social Trend towards Girl Child in India: A Machine Intelligence-Based Approach", Recent Developments in Machine Learning and Data Analytics, Advances in Intelligent Systems and Computing, Springer, 2019.
8. Ravinder Ahuja,Aakarsha Chug,Shruti Kohli,Shaurya Gupta,Prathyush Ahuja,"The Impact of Feature Extraction on the Sentiment Analysis",Sciencedirect,2019.
9. Tanjim Ul Haque, Nudrat Nawal Saber and Faisal Muhammad Shah-" Sentiment Analysis on Large Scale Amazon Product Reviews", IEEE International Conference on Innovative Research and Development, May 2018, Bangkok.
10. Aashutosh Bhatt, Ankit Patel, Harsh Chheda, Kiran Gawande," Amazon Review Classification and Sentiment Analysis", International Journal of Computer Science and Information Technologies, Vol. 6 (6) , 2015.
11. Milica Ciric , Aleksandar Stanimirovic , Nikola Petrovic , Leonid Stoimenov," Comparison of Different Algorithms for Sentiment Classification", IEEE 2013.
12. Indhra om Prabha M, G. Umarani Srikanth," Survey of Sentiment Analysis Using Deep Learning Techniques", ICIICT, IEEE, April 2019.
13. Adyan Marendra Ramadhani,Hong Soon Goo, "Twitter Sentiment Analysis using deep Learning",IEEE,2017.
14. Biswarup Nandi, Mousumi GhantiSouvik Paul, "Text Based Sentiment Analysis",IEEE,2017.
15. Harpreet Kaur, Veenu Mangat,Nidhi, "A survey of Sentiment Analysis", IEEE, International Conference 2017.
16. Shahid Shayaa,Noor Ismawathi Jaffar,Shamshul Bahri, "Sentiment Analysis of Big Data:Methods,Applications,and Open Challenges", IEEE, June 2018.
17. Yong Chen, Bin Zhou, Weina Zhang, "Sentiment Analysis Based on Deep Learning and Its Application in Screening for Perinatal Depression", IEEE Third International Conference on Data Science in Cyberspace,2018.
18. Yuling Chen, Zhi Zhang, "Research on Text Sentiment Analysis Based on CNNs and SVM",IEEE ,13th International Conference on Industrial Electronics and Applications,2018.
19. Peng Yang, Yunfang Chen, "A Survey on Sentiment Analysis by using Machine Learning Methods",IEEE,2017.
20. Sahar A.El Rahman, Feddah , Wejdan, "Sentiment Analysis on Twitter Data",IEEE,2019.
21. Priyanka Thakur, Dr.Rajiv Shrivastava, "Sentiment Analysis of Tourist Review using Supervised Long Short Term Memory Deep Learning Approach", IJIRCCE, Volume 7,Issue 2,2019.
22. Abdulaziz M.Alayaba, Vasile Palade, Matthew England, " Arabic language sentiment analysis on health services", IEEE,2017.
23. Rushlene Kaur Bkakshi,Navneet Kaur, Ravneet Kaur, "Opinion Mining and Sentiment Analysis", IEEE,2016.
24. Mika V.Mantyla,Daniel Graziotin, "The Evalution of Sentiment Analysis –A Review of Research Topics,Venues and Top Cited Papers", ELSEVIER, Computer Science Review, Volume 17,February 2018.
25. Pushpak Bhattacharya, "Sentiment Analysis", IEEE, ICETACS,2013
26. Saufian Jebbara, Phillip Cimiano, " Aspect-Based Sentiment Analysis Using a Two-Step Neural Network Architecture", Semantic Web Challenges, pp-153-167, Springer,2016.
27. Jorge Carrillo de Albornoz, Laura Plaza, Pablo Gervás, Alberto Díaz, "A Joint Model of Feature Mining and Sentiment Analysis for Product Review Rating", Advances in Information Retrieval,pp-55-66,Springer,2011.

AUTHORS PROFILE

Siji George C G, Ph.D. Scholar, CMS College of Science and Commerce, Coimbatore, Tamil Nadu, India. Email:-siji.gorg@gmail.com.

Dr. B. Sumathi, Associate Professor, CMS College of Science and Commerce, Coimbatore, Tamil Nadu, India. Email:-sumathithamizh@gmail.com