International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 10 Issue: 02 | Feb 2023 www.irjet.net p-ISSN: 2395-0072

Sentiment Analysis on Product Reviews Using Supervised Learning Techniques

Siddhes Tripathy [a], Ankit Pradhan [a], Dibya DasMohapatra[a], Narayan Sharma[a], Mr. Ramesh Jana[a]

[a] *Department of computer science and engineering, Balasore college of engineering and technology, Odisha, India*
**Corresponding Author: Siddhes Tripathy*

Abstract: This paper represents a comparison between three Machine learning approaches for analysing the sentiment of the Customer's reviews on various products. Eventually, reviews of a Product help the customers to understand the product quality. It involves computational study of behaviour of an individual in terms of his buying interest and then mining his opinions about company's Business entity. In this paper, dataset has taken from Amazon.com, Flipkart.com which contains reviews of Alexa's Mobile Phones. After pre-processing, we applied Machine Learning Algorithms to classify the reviews or opinions that are positive or negative Behaviour.

Keywords: Sentiment Analysis, Natural Language Processing, Product reviews, Machine Learning, Classifiers.

I. INTRODUCTION

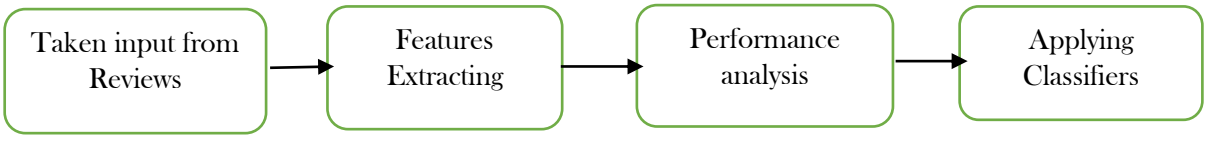
Nowadays, people are mostly interested in trading products on e-commerce websites instead of the offline market because of time efficiency and conventional property. In electronic commerce, product reviews are used on shopping sites to give customers an opportunity to rate and comment on products they have purchased, right on the product page. A common situation is to go through a review of the product before buying. So, the reviews of the product have inclined the consumer positively or negatively. On the other hand, reading thousands of reviews is an inhuman accomplishment. In this era of growing machine learning-based algorithms, it is rather time-consuming to go through thousands of remarks to discern a product where a review of a specific category can be polarized to know its popularity among consumers worldwide.

Sentiment analysis is the part of text mining that attempts to define the opinions, felling, and attitudes present in a text or set of texts. Online reviews are so important to businesses because they ultimately increase sales by giving the consumers the information, they need to make the decision to purchase the product. One very important factor in elevating the reputation, standard, and evaluation of an e-commerce store is product review. Product review is the most valuable resource available for Customer Feedback.

This paper aims to categorize customers' positive and negative feedback on distinct products and create a supervised learning model to polarize a wide range of reviews. Last year, research on amazon disclosed more than 88 percent **[1]** of internet shoppers trust reviews as much as a private recommendation. However, any online product with a big number of favourable reviews offers the item's legitimacy with a power remark. On the other hand, without reviews, electronic gadgets or any other item online puts prospective potential in a state of distrust.

Sentiment analysis helps to analyse these opinioned data and extract some important insights which will help to other user to make decision.

In this research paper, pre-processing is performed to reduce the dimensionality of the features applying the Multi-Domain Sentiment Dataset **[2].** Later, all frequent words beyond threshold value are considered as features. After Complete extraction, three machine learning algorithms have been applied to find the best potential based on statistical analysis.



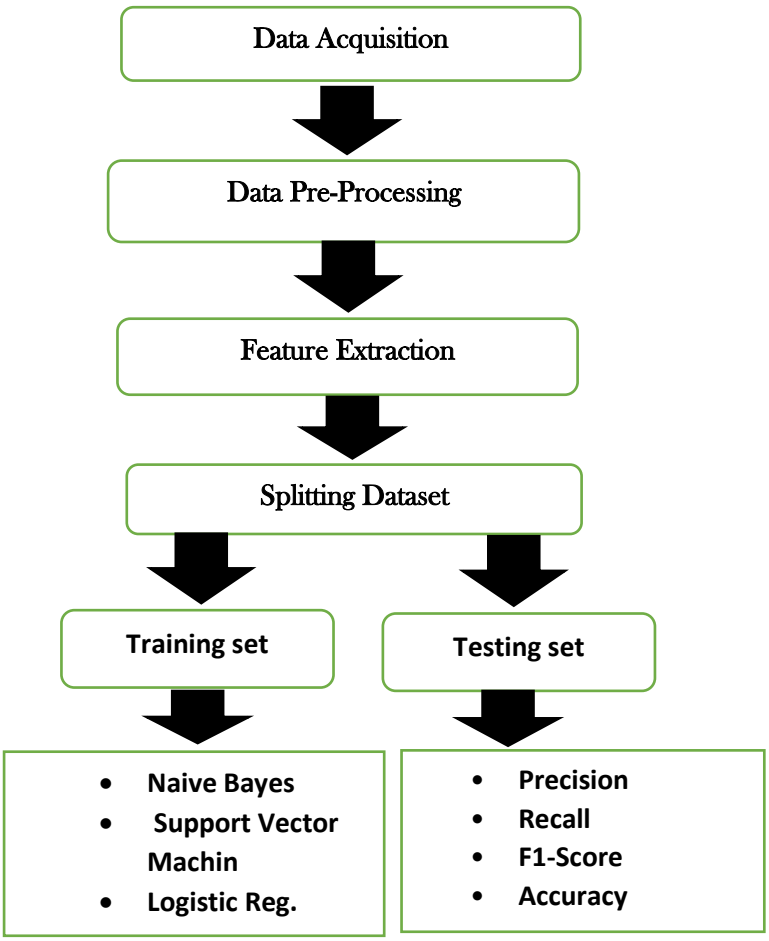
II.LITERATURE REVIEW

Sr.no.	Paper Name	Authors	Year of publication	Merits	Demerits	Methods/tools
1	Research on Aspect-Level Sentiment Analysis Based on Text Comments	Jing Tian, Wushour Slamu, Miaomiao Xu, Chunbo Xu and Xue Wang[3]	2022	It is found that, there are affective regions that evoke human sentiment in an image.	In the problem of comparative level sentences, where it is difficult to judge polarity of the comment.	NLP, combines capsule network and BERT and CapsNet-BERT.
2	Sentiment Analysis Using Product Review Data	Xing Fang, Justin Zhan [4]	2015	In this paper, experiments for both sentence-level categorization and review-level categorization for sentiment analysis.	Review-level categorization becomes difficult if we want to classify reviews to their specific star-scaled ratings.	Scikit-learn(software), Models like Naïve Bayesian, Random Forest, and Support Vector Machine.
3	Soft Computing Approaches to Classification Of Email for Sentimental Analysis	Mrs. Pranjal S. Bogawar and K.K. Bhoyar	2016	It's helpful to clustering and classifying emails into different categories.	Clustering algorithm performing well due to future in identifying negative emails.	Expectation Maximization (EM), K-Means, Agglomerative Algorithm.
4	Random Forest and Support Vector Machine based Hybrid Approach to Sentiment Analysis.	Yassine Al Amrani, Mohammed Lazaar, Kamal Eddine El Kadiri [5]	2018	Tracking the mood of the public about a particular product.	Using a hybrid classification methodology is better than use of the individual classification method.	Supervised learning: Random forest, SVM, RFSVM.
5	A Comparative Study of Support Vector Machine and Naïve Bayes Classifier for Sentimental Analysis on Amazon Product Reviews	Sanjay Dey, Sarhan Wasif, Dhinman Skider Tommoy, Subrina Sultana [6]	2020	Mutiple review factors, including product quality, content, time of the reviews will helpful for the customer.	Does not work with multiple-word phrases.	NLP, SVM and Naïve Bayes classifier.

To sum up, no research work has provided a comparison between the support vector machine, the Naïve Bayes, and the Logistic Regression classifier. Later, in this work a comparison between three machine learning approaches has been presented for analyzing the sentiment of the customers' reviews on product.

III. METHODOLOGY

Amazon, Flipkart are the largest e-commerce sites as it is possible to see countless reviews. The dataset was unlabelled and it must be labelled in order to use it in supervised learning model. Eventually, this research has only worked with the feedbacks of e-commerce product like Alexa and Mobile Phone. Almost 1946 reviews of the 1918 reviews of the Mobile Phones and 3106 reviews of the Alexa have been processes for the analyzing the polarization. Meanwhile, Fig.1 has been illustrated the overview of the system methodology.



A: Data Acquisition

Data Acquisition has been proceeded in the very first step for labelling the data. As the dataset consist of many reviews, manual labelling is quite impossible for a human being. Therefore, after the completion of pre-processing dataset, the active learner has been implemented to label the datasets. As product feedbacks come in 5-star ratings,3-star rating is usually regarded as neutral reviews meaning neither negative nor positive. So, the system has been discarded any review which contains a 3-star rating from dataset and has been taken the other reviews for proceeding to the next step leading to pre-processing step.

Star Level	General Meaning
★	I hate it.
★★	I don't like it.
★★★	It's okay.
★★★★	I like it.
★★★★★	I love it.

Fig 2- Rating System Parameter

B: Data Pre-processing

Pre-processing data comprises of three steps called tokenization, removing stop words and POS tagging

1) *Tokenization*: It is the method of separating a string sequence into individuals such as keywords, symbols, phrases, and other elements called tokens. Eventually, tokens can be phrases, words, or even entire sentences. However, some characters such as punctuation marks are removed in the tokenization phase. The tokens are the input for various procedures such as parsing and text mining.

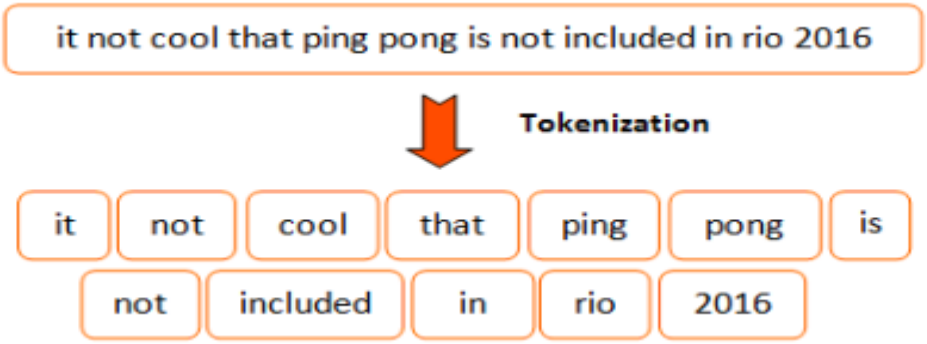


Fig 3 – Text Tokenization

2) *Removing stop words*: Stop words are items in a phrase that are not essential in text mining for any division. In order to improve the precision of the assessment, these words are usually ignored. There are distinct stop words in various formats depending on the realm, language, and so on. However, there are several stop words in English format.
3) *POS tagging*: The process of assigning one of the parts of speech to the given word is called part of speech tagging. It is generally referred to as POS tagging. Parts of speech generally contains nouns, verbs, adverbs, adjectives, pronoun, conjunction, and their sub-categories. Parts of speech tagger or POS tagger is a program that does this job.

C. Feature Extraction

Feature of the dataset has been extracted by the three method steps named term Bag of Words, Frequency-inverse document frequency (TF-IDF), Relevant noun removal.

1) *Bag Of Words:* Bag of words is a process of extracting features by representing simplified text or data, used in natural language processing and information retrieval. In this model, a text or a document is represented as the bag (multiple set) of its words. So, simply bag words in sentiment analysis is creating a list of useful words.
2) *TF-IDF: TF-IDF* is a technique for retrieving information that weighs the frequency of a team (TF) and the inverse frequency of document (IDF), Every word or phrase has its score TF and IDF. Meanwhile, A term's TF and IDF product results refer to that term's TF-IDF weight.

3) *Relevant noun removal*: The list of most frequent nouns does not have always cover all important features being commented on. There have been still some features mentioned in a smaller number of opinions that have been also important to be extracted by the system. Henceforth, refinement of the relevant noun, which has been selected as infrequent nouns denoted by candidate features based on opinion words.

Algorithm for the System

1) Taking Input
2) Pre-processing the data
    a) Removing blank rows
    b) Changing all text to lower case
    c) Tokenization will break data into words
    d) Removing stop word
3) Splitting the dataset into train-set and test-set
4) Vectorization the words by using TF-IDF vectorizer
5) Running different classifiers to classify the dataset
6) Labelling the encoded target value

IV: SPLITTING DATA

In this phase, we are splitting the dataset for two different purposes: training and testing. The training subset is for building the model and testing subset is to evaluate the performance of the model in this model. In this paper, we are splitting the dataset into 90% for training and 10% for testing.

V: APPLIED ALGORITHM/CLASSIFICATION

In this work, we will apply tree supervised learning algorithms

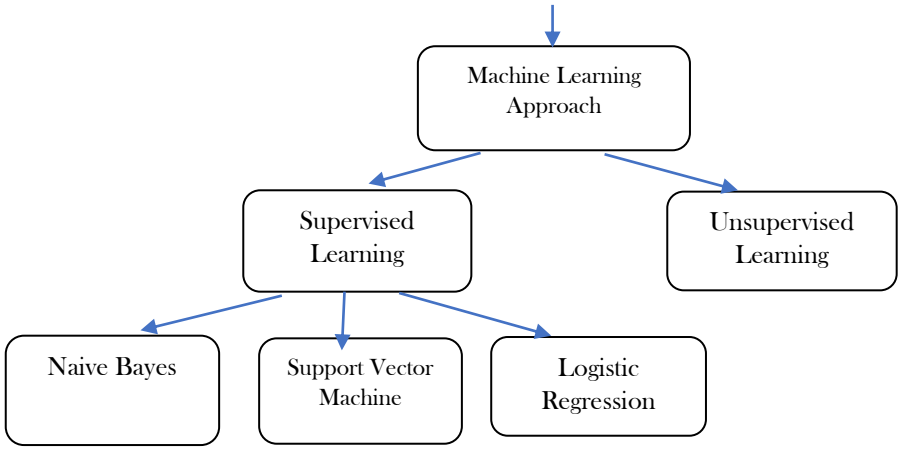


Fig 4- Classification of sentiment analysis

Recently, sentiment analysis has attracted an increasing interest. It is a hard challenge for language technologies, and achieving good results is much more difficult than some people think. The task of automatically classifying a text written in a natural language into positive or negative feeling, opinion, or subjectivity is sometimes so complicated that even different human annotators disagree on the classification to be assigned to given text**[7].** Personal interpretation by an individual is different from others, and this is also affected by cultural factors and each person's experience. And the shorter the text, and the worse written, the more difficult the task becomes, as in the case of messages on social networks **[8].**

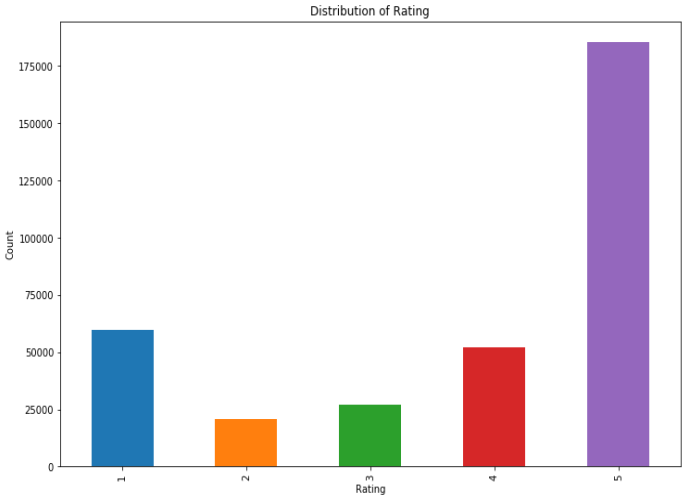
After pre-processing and feature extraction, we start to classify the reviews. In this paper, after we are splitting the dataset in both training and testing set, we can input them in the following classifier's:

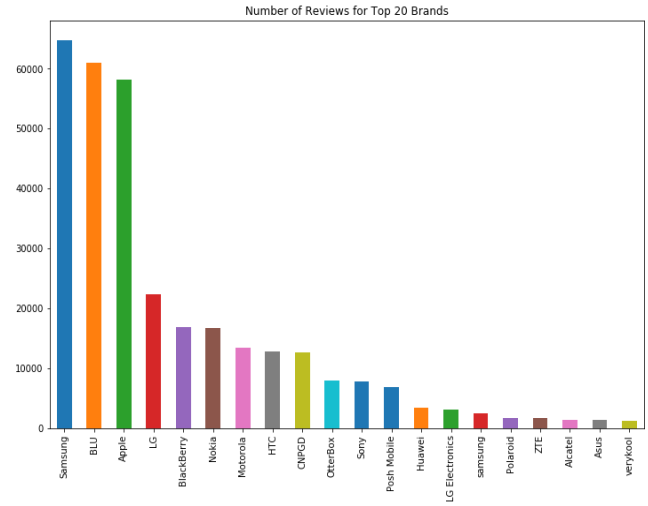
A) Naïve Bayes: The main idea of the idea of this classifier is hypothesis that predictor variables are autonomous which substantially minimizes the computation of probabilities. It gives excellent results, and it was used in most research **[9].**

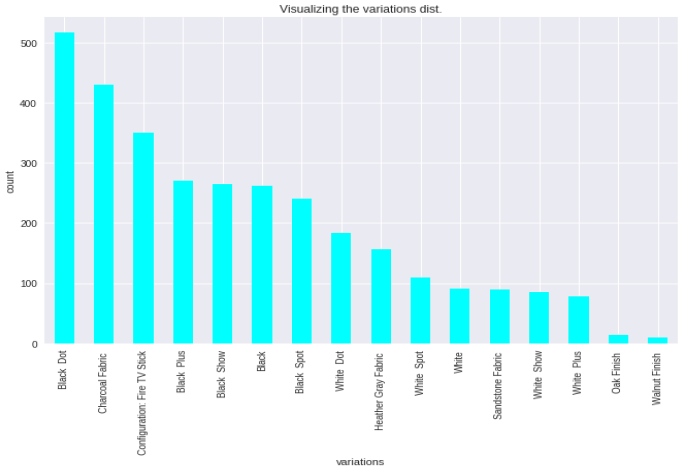
B) Support Vector Machine: It is potential classification technique, and is a supervised learning model with associated learning algorithms that analyze data for classification.

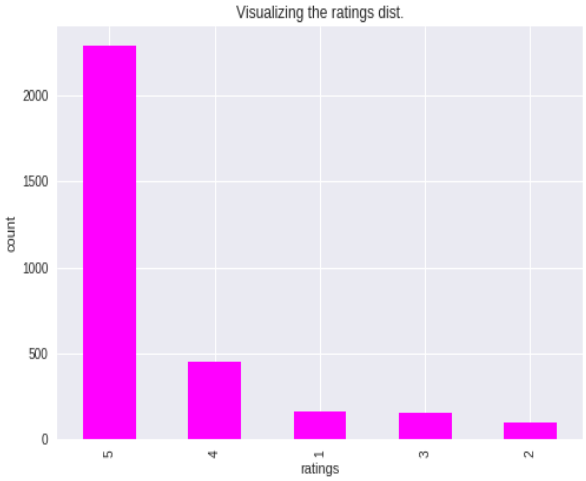
C) Logistic Regression: It is a supervised machine learning technique for classification problems. The goal of the model is to learn and approximate a mapping function f(Xi)=Y from input variable {x1, x2, xn} to output variable(Y). It is called supervised because the model predictions are iteratively evaluated and corrected against the output values, util an acceptable performance is achieved.

VI: EXPERIMENT RESULT









This section intends to evaluate the performance of these three machine learning models through several experiments. Evaluating metrics play a significant role in measuring the efficiency of classification where measuring accuracy of classification where measuring accuracy is the most convenient for this purpose. Eventually, a classifier's precision on a given test dataset is the percentage of those datasets that the classifier properly categorizes. The system is evaluated by three commonly used statistical measure named recall, precision, and F-measure from with the help of confusion matrix.

The data from the confusion matrix are classified into four categories named as True Positive (TP), True Negative (TN), False Positive(FP) and False Negative(FN). True positive present an outcome where the system perfectly predicts the positive class. On the hand FP identifies an outcome where the positive class is incorrectly predicted by the scheme. Meanwhile, TN is the result where the system divines the adverse class exactly. On the contrary, FN is an outcome where the negative class is incorrectly predicted by the system.

TABLE I

CONFUSION MATRIX FOR SUPPORT VECTOR MACHINE OF ALEXA DATASET

Linear SVM	Positive	Negative
Positive	TP: 1721	FP: 325
Negative	FN: 315	TN: 1639

Precision is the predictive ratio among the total Positive instance denoted by the equation (1).

$$\text{Precision} = TP / TP+FP \quad (1)$$

TABLE II

CONFUSION MATRIX FOR NAÏVE BAYES CLASSIFIERS OF ALEXA DATASET

Naïve Bayes	Positive	Negative
Positive	TP:1711	FP:360
Negative	FN:325	TN:1604

TABLE III

CONFUSION MATRIX FOR LOGISTIC REGRESSION OF ALEXA DATASET

Logistic Regression	Positive	Negative
Positive	TP: 1701	FP: 350
Negative	FN:330	TN:1600

TABLE IV

CONFUSION MATRIX FOR SUPPORT VECTOR MACHINE OF MOBILE DATASET

Linear SVM	Positive	Negative
Positive	TP:1775	FP:33
Negative	FN:306	TN:7477

TABLE VI

CONSFUSION MATRIX FOR NAÏVE BAYES CLASSIFIERS OF MOBILE DATASET

Naïve Bayes	Positive	Negative
Positive	TP:724	FP:370
Negative	FN:720	TN:6594

TABLE VII

CONFUSION MATRIX FOR LOGISTIC REGRESSION OF MOBILE DATASET

Logistic Regression	Positive	Negative
Positive	TP:448	FP:104
Negative	FN:896	TN:7123

The equation (2) which introduces the scheme properly classifies the beneficial classes out of all classes denotes recall.

$$\text{Recall} = TP / TP+FN \quad (2)$$

F1 score analyses accuracy and recall at the same time using harmonic mean instead of the arithmetic mean denoted by equation (3).

$$\text{F1 score} = 2 \times \text{Recall} \times \text{Precision} / \text{Recall} + \text{Precision} \quad (3)$$

Accuracy is the proportion of all specimens of the class and the cumulative number of samples denoted by equation (4).

$$\text{Accuracy} = TP + TN / TP + TN + FP + FN \quad (4)$$

TABLE I

EVALUATION PARAMETERS FOR CLASSIFIERS OF ALEXA DATASET

Classifiers	Precision	Recall	F1 Score	Accuracy (%)
Naïve Bayes	0.828	0.839	0.839	82.87
Logistic Regression	0.820	0.870	0.880	80
Linear SVM	0.839	0.890	0.839	84

TABLE 2

EVALUATION PARAMETERS FOR CLASSIFIERS OF MOBILE DATASET

Classifiers	Precision	Recall	F1 Score	Accuracy (%)
Naïve Bayes	0.84	0.86	0.80	85.52
Logistic Regression	0.87	0.88	0.86	88.12
Linear SVM	0.941	0.940	0.938	94.09

VII: CONCLUSION AND FUTURE WORK

This research was able to present a comparison study among three algorithms SVM, Naïve Bayes and Logistic Regression to analyse the polarization of the sentiment of product reviews. In this work, the models were trained by almost 2500 features with almost 6000 datasets after the pre-processing procedure. In the meantime, almost 4000 test set have been passed through the models for the statistical measurement. Experimental results have confirmed that the support vector machine can polarize the feedback of the product reviews with a high accuracy rate.

Some future works which can be included to improve the model and to make it more effective in practical cases. Our future works include applying PCA (Principal Component Analysis) **[10]** in active learning process to fully automate data labelling process with less assistance from the oracle. And lastly, we will try to continue this research until we generalize this model to all kinds of text-based reviews and comments.

VIII: REFERENCES

[1] T.U Haque, N.N. Saber and F.M. Shah, "Sentiment analysis on large scale amazon product reviews," in 2018 IEEE International Conference on Innovative Research and Development (ICIRD).

[2] J.H.W.S.o.E. Department of CSE. Multi-domain sentiment dataset, Available: http://www.cs.jhu.edu/mdredze/datasets/sentiment/

[3] Jing Tian, Wushour Slamu and Miaomiao Xu, "Research on Aspect-Level Sentiment Analysis Based on Text Comments", in 2022, Available: https//www.mdpi.com/journal/symmetry.

[4] Xing Fang and Justin Zhan, "Sentiment analysis of Product Review", in 2018 International Journal of Innovations in Engineering and Science.

[5] Yassine Al Amrani, Mohammed Lazaar and Kamal Eddine El Kadiri," Random Forest and Vector Machine based Hybrid Approach to Sentiment Analysis", in 2018, Available at www.sciencedirect.com.

[6] Samjay Dey, Sarhan Wasif and Dhiman Sikder Tonmoy," A Comparative Study Of Support Vector Machine and Naïve Bayes Classfier for Sentiment Analysis on Amazon Product Reviews", in 2020 IEEE International Conference on Contempory Computing and Application (IC3A).

[7] Xia, Rui, Chengquing Zong, and Shoushan Li. "Ensemble of feature sets and classification algorithms for sentiment classification," Information Science vol.181, issue 6, 15 Mar. 2011.

[8] Melville, Prem, Wojeiech Grye and Richard D. Lawrence. "Sentiment analysis of blogs by combining lexical knowledge with lexical knowledge with text classification," Proceedings of the 15th ACM SIGKDD International conference on knowledge discovery and data mining, 2009.

[9] Huma Parveen and Prof. Shikha Pandey," Sentimental analysis on twitter data-set using naïve bayes algorithm", Published in 2016 International conference on applied and theoretical computing and communication technology.

[10] G. Vinodhini, RM. Chandrasekaran," Sentiment classification using principal component analysis based neural network model", International conference on information communication and embedded system in 2014.