



## International Journal of Control Theory and Applications

ISSN : 0974-5572

© International Science Press

Volume 10 • Number 30 • 2017

### Sentiment Analysis on Online Reviews using Supervised Learning: A Survey

**B. Vamsi<sup>a</sup> N. Suneetha<sup>a</sup> Ch. Sudhakar<sup>a</sup> and K. Amaravati<sup>a</sup>**

*<sup>a</sup>Assistant Professor, Computer Science and Engineering, Vignan's Institute Of Information Technology, Visakhapatnam*

**Abstract:** Sentiment Analysis (also known as Opinion Mining) is an area of text classification which continuously contributes to the research arena. The primary objective of opinion mining is sentiment classification *i.e.* to classify the opinion into positive or negative classes. The real world scenario involves an opinionated approach on various issues like business, health, education, shopping, entertainment, etc. This paper presents an empirical study of classifying reviews on products, movies and Stock Market. The targeted levels of classification for these reviews are sentence level and feature level. Sentence level classifies the sentiment expressed in each sentence. Feature level classifies the sentiment expressed on each feature. Aspect level aims to classify the sentence with respect to the specific aspects of entries. Supervised learning techniques such as Naive Bayes Classifier are used. This study explores future prospects of sentiment analysis in myriad ways.

**Keywords:** *Sentiment Analysis, Supervised Learning, Data Mining.*

#### 1. INTRODUCTION

Organizations from diverse fields such as Finance, Entertainment, Education, Medicine, Commodities, etc. wish to obtain the opinion regarding their excellence in terms of performance, productivity and scalability. This is a manual and time-consuming task in this fast-paced world [2].

Opinions are mined to extract meaningful information from the raw text. Opinion Mining is the process of automating the input text, image, audio and video in an understandable format. It can also be defined as a subdiscipline of computational linguistics that is concerned with the opinion that a document expresses. Sentiment analysis also known as opinion mining analyses the document as a whole or a sentence or an aspect of a product. Sentiment classification [7] refers to determining the subjectivity, polarity (positive/negative) and polarity sentences (weakly positive, mildly positive or strongly positive) of the input text and the hierarchy is as follows.

Fig 1 shows the working procedure of Sentiment analysis

This paper is organized as follows: Section 2 describes the Methodology used to extract the opinions. Section 3 tabulates the results of each technique. Section 4 provides the conclusion.

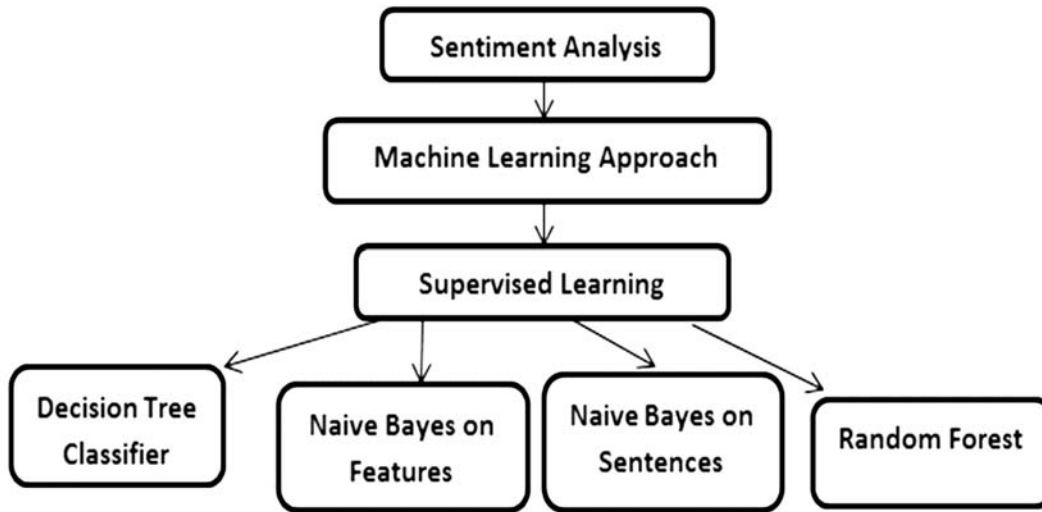


Figure 1: Hierarchy of Sentiment Analysis

## 2. METHODOLOGY

### 2.1. Supervised Learning

Data Mining refers to extracting the meaningful information from voluminous data. To mine the data, one of the mechanism used is Supervised Learning. Mapping the classified data to the predefined class label is known as Classification or Supervised Learning.

Data Classification is a two stage process involving learning stage and classification stage.

The paper[4] discusses Naive Bayes Classifier which depends on the conditional probability of an event occurrence.

### 2.2. Naive Bayes Classifier

It is defined as the probability that event ‘X’ will occur given the evidence ‘Y’. This is represented as follows:

$$P(X | Y) = (P(X) P(Y | X)) / P(Y).$$

This paper emphasizes on sentence level and feature level (or) aspect level classification.

1. **Sentence level classification:** It is defined as the conditional probability that the given label occurs if the input sentences exist[6].

$$P(\text{label}|\text{sentence}) = \frac{P(\text{label}) * P(\text{sentences}|\text{label})}{P(\text{sentences})}$$

$$P(l_k | S_1, S_2, S_3, \dots, S_n) \propto P(l_k) \prod_{i=1}^n P(S_i | l_k)$$

where

$$S_i = S_1, S_2, S_3, \dots, S_n$$

$$P(l_k | S_i) = \frac{P(l_k) * P(S_i | l_k)}{P(f_i)}$$

**Example:**

<b>best browser</b>	in 16,587 reviews
* Just the best browser for android. If you are using soothing else you are doing it wrong*.	
<b>easy to use</b>	in 6,549 reviews
* ... I urge people to try both browsers though. Both are blisteringly fast and easy to use!*	
<b>useful</b>	in 6,154 reviews
* This app is very useful in our day to day life. you should download it without fear...*	
<b>full screen ads</b>	in 667 reviews
* Reading an article on a site no is getting to be a challenge with to pop-up ads...*	
<b>great search engine</b>	in 540 reviews
* Its great search engine with extra features. I love it and its my favorite browser.*	
<b>offline reading</b>	in 418 reviews
* ... It has full screen ads that are useful and can be saved for offline reading...*	

**Figure 2: Reviews on Google Play Store**

Figure 2 presents the reviews made by various users on a particular app. Sentence level sentiment analysis has been applied on these reviews. To map the class label as positive to the given sentence such as “Its great search engine with extra features [10]. I love it & its my favourite browser” – Great Search Engine and mapping the given sentence as negative such as “Reading an article on a site now is getting to be a challenge with to pop-up ads” – Full Screen Ads[4].

- Aspect (or) Feature level classification:** The occurrence of the prescribed label is dependent on the existence of set of features.

$$P(\text{label} | \text{features}) = \frac{P(\text{label}) * P(\text{features} | \text{label})}{P(\text{features})}$$

$$P(l_k | f_1, f_2, f_3, \dots, f_n) \propto P(l_k) \prod_{i=1}^n P(f_i | l_k)$$

where

$$f_i = f_1, f_2, f_3, \dots, f_n$$

$$P(l_k | f_i) = \frac{P(l_k) * P(f_i | l_k)}{P(f_i)}$$

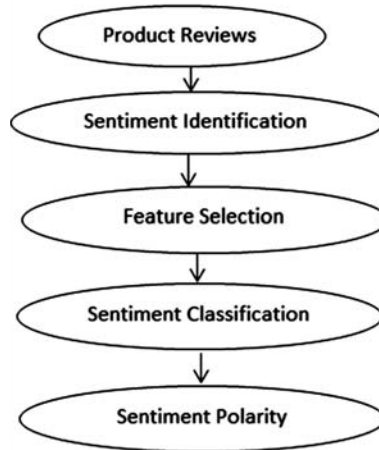


Figure 3: Opinion Mining process on product reviews

Fig 3 describes the framework of Feature Level Sentiment Classification. From the obtained product reviews, sentiments are identified as good, bad or neutral. Feature selection is applied on the parallel products and then sentiment classification is performed. The polarity of the sentiment is labelled as positive and negative [14].

Example:



Comparison Criteria		
<input type="checkbox"/> Show differences only	 Moto X 201 7	 Moto G4
<b>SUMMARY</b>		
Display	5.2" (13.21 cm)	5.5" (13.97 cm)
Storage	32 GB	16 GB
Camera	13 MP	13 MP
Battery	3500 mAh	3000 mAh
Ram	4 GB	2 GB
<b>SPECIAL FEATURES</b>		
Other Sensors	Light sensor, Proximity sensor, Accelerometer	Light sensor, Proximity sensor, Accelerometer, Compass, Gyroscope
<b>GENERAL</b>		
Operating System	Android v7.0 (Nougat)	Android v6.0.1 (Marshmallow)

Figure 4: Product Reviews on Flipkart

Figure 4 presents the customer reviews made on various products. Feature level sentiment analysis has been applied on these reviews. Features such as Display, Storage, Camera, Battery etc. are considered for the mobile phone brand named Moto with its variants. Moto X and Moto G4 vary in their appearance. The aspects of Moto X like storage capacity and RAM size are having high functionality than Moto G4. The class label positive is mapped to it. Negative class label is mapped to Moto X for the display aspect as it is exhibiting less performance in comparison with Moto G4.

3. **Decision Tree Classifier:** A decision tree represents a tree like structure where each non-terminal node denotes a test on an attribute, each branch represents the outcome of the test and each terminal node holds a class label. It is the predictive model which maps observations about an entity to conclusions regarding the entities target value[8]. This paper describes an empirical study of decision tree on “Stock Market” decisions such as steep – rise and sudden downfall.

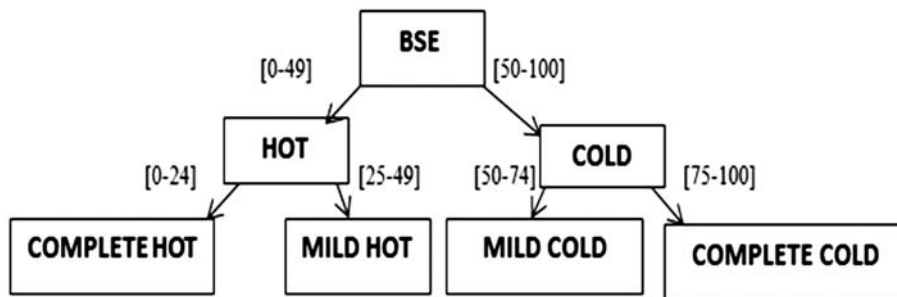


Figure 5: Decision Tree Structure for BSE

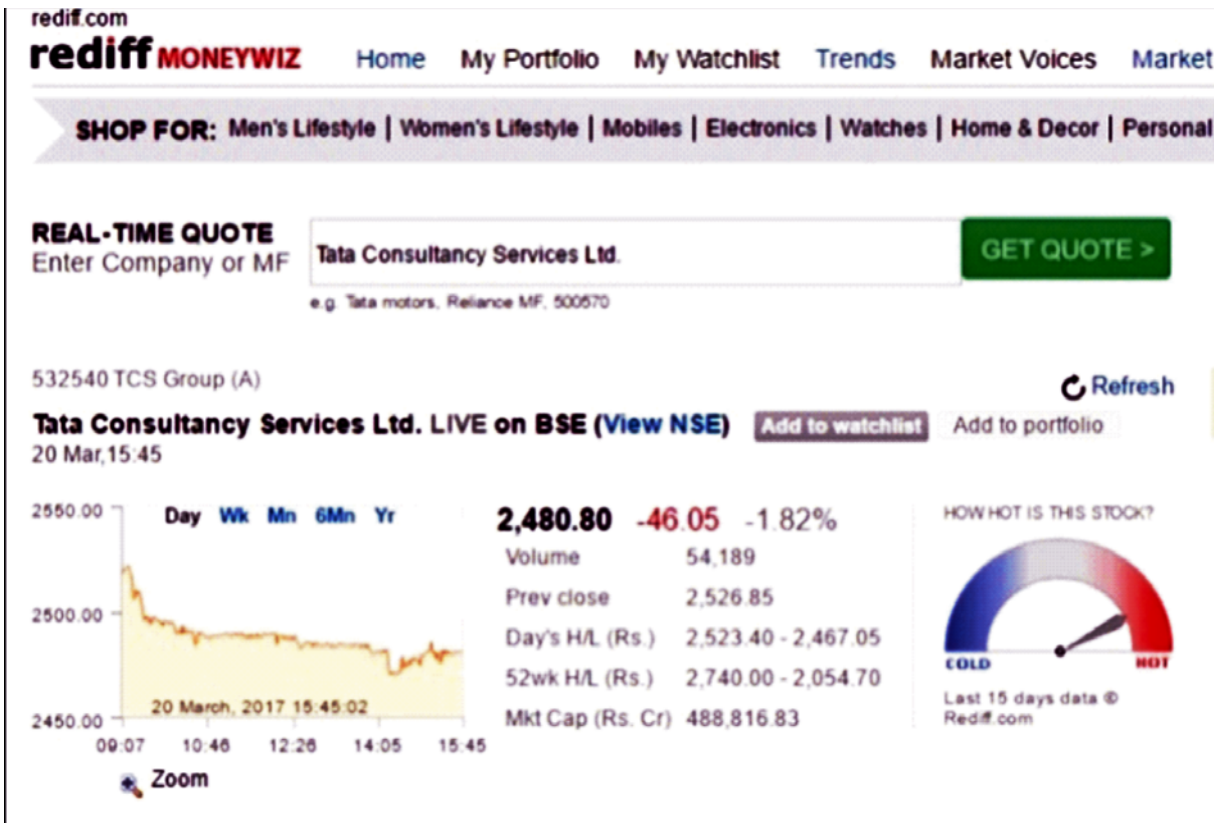


Figure 6: MILD HOT Representation of BSE for ‘TCS’ company

BSE is an Indian Stock Exchange that shows the trading functionality of various companies that participate in the share market. Figure 5 describes the representation of decision tree in which the parent node is BSE and the class labels are 'HOT' and 'COLD'. For a company to readily buy a share is evidently visible through the class label 'COLD' and 'HOT' determines the down fall of the particular company's share[12].

The variants of 'COLD' are 'COMPLETE COLD' which denotes perfect share value and 'MILD COLD' indicates satisfactory share value.

The variants of 'HOT' are 'COMPLETE HOT' which denote the imperfect share value and 'MILD HOT' indicates possible growth in the share value.

Example:

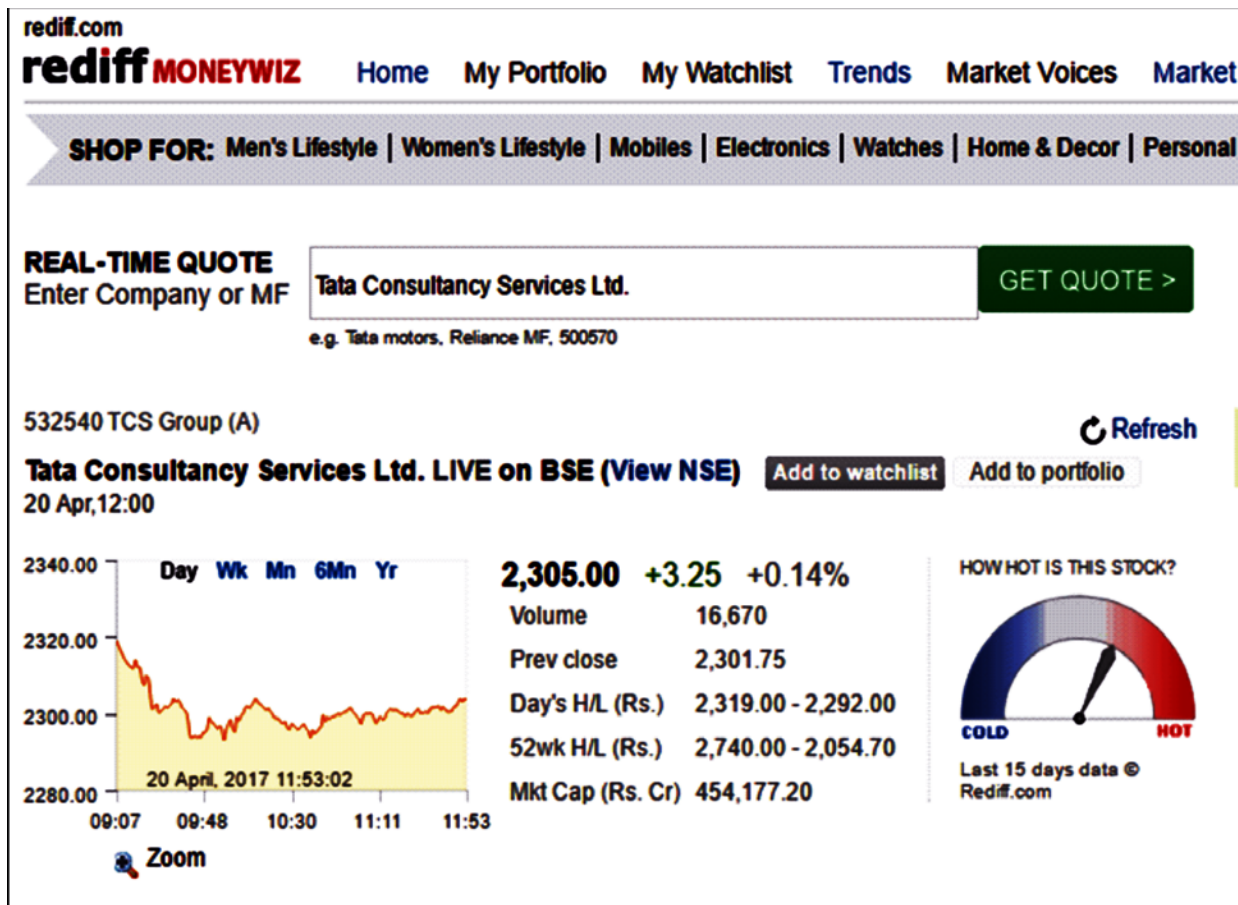


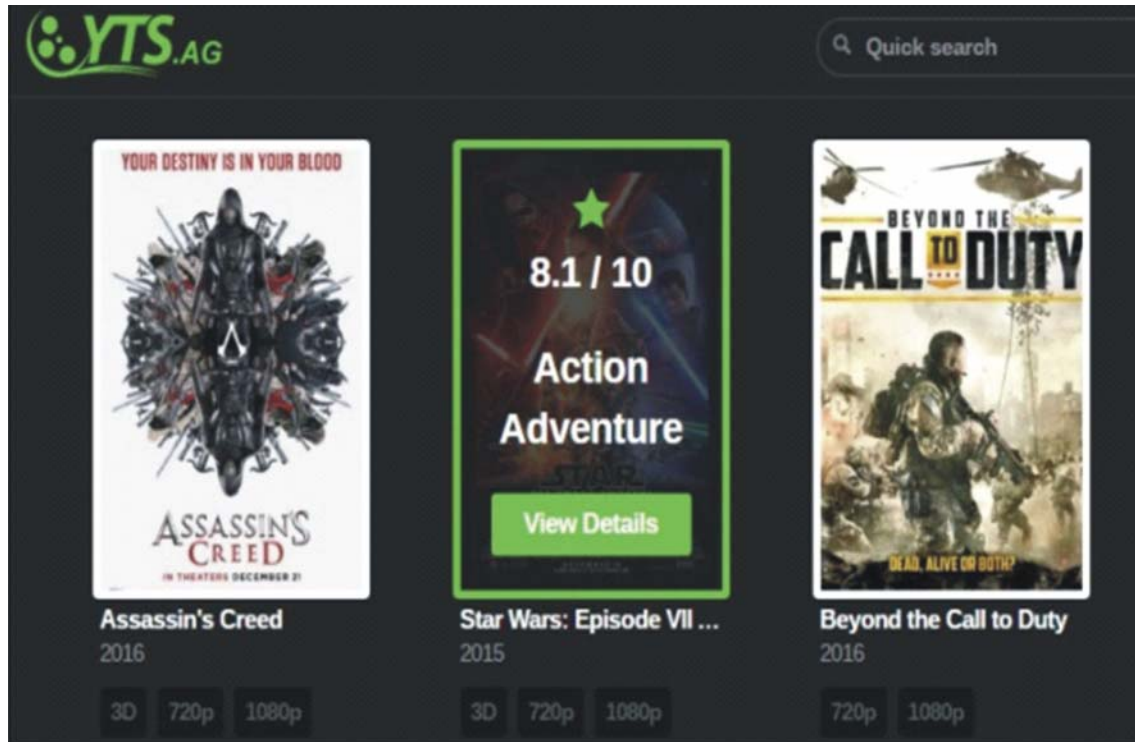
Figure 7: HOT Representation of BSE for 'TCS' company

The above Figures 6 and 7 illustrate the fluctuations of TCS Company's share value in terms of MILD HOT for the range 50-74 and HOT for the range 75-100.

4. **Random Forest:** To make a classification decision, number of decision trees has to be learnt. At every step of the learning process an attribute is selected to split into two or more different parts and this process is repeated until a pure classification split is reached. A pure classification split is when the split parts represent only one class that they belong to. At every split we try to reach a local optimum solution.

Random Forest is the combination of decision trees which deal with multiple number of parameters such as:

- a) Number of trees to construct for the decision forest.
- b) Number of features to select at random.
- c) Depth of each tree.



**Figure 8: IMDB Representation for movie rating**

IMDB data search engine describes the rating of various movies given in the figure 8. Random Forest algorithm is applied on this search engine which pre-processes the data first and then cross validation function is applied to evaluate the performance of the prescribed model. It divides the data into two parts. First part is used to train the model and the classifier is built. Second part is used to test the accuracy of the predictions of the model. This process is done iteratively and different subsets of the data are used for training and testing at each iteration[20].

To have a quick access of the Hollywood movies, viewers browse the website named YTS.AG. By using IMDB this website provides the rating of every High Definition (HD) movie [16]. If the rating is greater than 6.5 it's regarded as good movie labelled as positive class and these movies can be downloaded with fast pace capability. Negative class label is assigned to the movies with the rating less than 6.5.

### **3. RESULTS AND DISCUSSION:**

In this study, we used three classifiers, namely Naive Bayes, Decision Tree and Random Forest in order to compare their performance.

The following tables summarize the results:

**Table 1**  
**Data Distribution**

<i>Class Label</i>	<i>Training</i>	<i>Testing</i>
Positive	1000	1000
Negative	1000	1000

Table 1 shows the dataset for each partition was randomly selected. The same training and testing set were used for each classifier. There are three metrics namely precision, recall and F-score were used to measure the performance of the three classifiers as shown below:

$$\text{Precision} = \frac{tp}{tp + fp}$$

$$\text{Recall} = \frac{tp}{tp + fn}$$

$$\text{F - score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

**Confusion Matrix (CM):** This matrix is defined as for the given ‘m’ classes ( $m \geq 2$ ), it is a table of size  $m \times m$ . An entry,  $CM_{ij}$  in the first ‘m’ rows and ‘m’ columns indicates the number of tuples of class ‘i’ that were labelled by the classifiers as class ‘j’.

**1. Naive Bayes Sentence Level Classifier:**

**Table 2**  
**Naive Bayes Sentence Classifier**

<i>Naive Bayes Classifier</i>	<i>Actual Positive</i>	<i>Actual Negative</i>
Predicted Positive	0.76	0.23
Predicted Negative	0.35	0.65

**2. Naive Bayes Feature Level Classifier:**

**Table 3**  
**Naive Bayes Feature Classifier**

<i>Naive Bayes Classifier</i>	<i>Actual Positive</i>	<i>Actual Negative</i>
Predicted Positive	0.73	0.24
Predicted Negative	0.27	0.76

**3. Decision Tree Classifier:**

**Table 4**  
**Decision Tree Classifier**

<i>Decision Tree Classifier</i>	<i>Actual Positive</i>	<i>Actual Negative</i>
Predicted Positive	0.65	0.35
Predicted Negative	0.52	0.48



#### 4. Random Forest Classifier:

**Table 5**  
**Random Forest Classifier**

<i>Random Forest Classifier</i>	<i>Actual Positive</i>	<i>Actual Negative</i>
Predicted Positive	0.6	0.4
Predicted Negative	0.7	0.93

**Table 6**  
**Comparison table**

<i>Accuracy Metric</i>	<i>Naive Bayes Classifier (Sentence)</i>	<i>Naive Bayes Classifier (Feature)</i>	<i>Decision Tree</i>	<i>Random Forest</i>
Precision	0.77	0.81	0.75	0.84
Recall	0.68	0.71	0.73	0.72
F-score	0.72	0.68	0.74	0.78

The above tables illustrate the accuracy of the classifiers through the Confusion Matrix. From Table 6, it is visible that Random Forest classifier is exhibiting high performance through precision matrix when compared to Recall. The F-score metric is high for Decision Tree classifier when compared to Naive Bayes classifier[9].

#### 4. CONCLUSION AND FUTURE WORK

Opinion Mining has become a fascinating research area due to the availability of voluminous user-generated content in review sides, forums and blogs. It has its applications in diverse fields ranging from education to advertising. In this paper, supervised learning techniques have been used to estimate the accuracy of various classifiers for several reviews. In future, to have a greater insight of sentiment analysis unsupervised learning can be implemented for online reviews as well as offline reviews to obtain higher accuracy levels.

#### REFERENCES

- [1] Shunsuke Doi, Shinya Hara, Yoshiro Imai "A Study of Sentimental Value Analysis for Tweeting Message" 2017.
- [2] Hongning Wang and ChengXiang Zhai "Generative Models for Sentiment Analysis and Opinion Mining" 2017.
- [3] V.Lakshmi, K.Harika , H.Bavishya, Ch.Sri Harsha "Sentiment Analysis Of Twitter Data" Volume: 04 Issue: 02" 2017.
- [4] Necmiye Genc-Nayebi, Alain Abran "A Systematic Literature Review: Opinion Mining Studies from Mobile App Store User Reviews" 2016.
- [5] Yang Liu, Jian-Wu Bi, Zhi-Ping Fan "Ranking products through online reviews: A method based on sentiment analysis technique" 2016.
- [6] Gurmeet Kaur, Abhinash Singla "Sentimental Analysis of Flipkart reviews using Naïve Bayes and Decision Tree algorithm" Volume 5 Issue 1, 2016.
- [7] Lei Zhang, Bing Liu "Sentiment Analysis and Opinion Mining" 2016.
- [8] A. Suresh and C.R. Bharathi "Sentiment Classification using Decision Tree Based Feature Selection" pp. 419-425, 2016.
- [9] Suchita V. Wawre, Sachin N. Deshmukh "Sentimental Analysis of Movie Review using Machine Learning Algorithm with Tuned Hyper parameter" Vol. 4, Issue 6, June 2016.
- [10] Tao Chena, Ruifeng Xua, Yulan Hec, Xuan Wanga "Improving Sentiment Analysis via Sentence Type Classification" 2016.
- [11] Shiliang Sun, Chen Luo, Junyu Chen "A Review of Natural Language Processing Techniques for Opinion Mining Systems" 2016.

- [12] Mariana Daniela, Rui Ferreira, Nevesa Nuno Hortaa “Company Event Popularity For Financial Markets Using Twitter And Sentiment Analysis” 2016.
- [13] Amrita Kaur, Neelam Duhan “A Survey on Sentiment Analysis and Opinion Mining” Volume 4, May 2015.
- [14] A. Sravani, D. N. D. Harini, D. Lalitha Bhaskari “A Comparative Study of the Classification Algorithms Based on Feature Selection” 2014.
- [15] Christos Troussas, Maria Virvou, Kurt Junshean Espinosa, Kevin Llaguno, Jaime Caro “Sentiment analysis of Facebook statuses using Naive Bayes classifier for language learning” 2013.
- [16] V.K. Singh, R. Piryani, A. Uddin “Sentiment Analysis of Movie Reviews” 2013.
- [17] Alexandra Balahur, Ralf Steinberger “Sentiment Analysis in the News” 2013.
- [18] Mostafa Karamibekr, Ali A. Ghorbani “Sentiment Analysis of Social Issues” 2012.
- [19] Apoorv Agarwal, Boyi Xie Ilia Vovsha, Owen Rambow, Rebecca Passonneau “Sentiment Analysis of Twitter Data” pages 30–38, 2011.
- [20] Tun Thura Thet, Jin-Cheon Na, Christopher S.G. Khoo “Sentiment Analysis of Movie Reviews on Discussion Boards using a Linguistic Approach” 2009.
- [21] Haji Binali, Vidyasagar Potdar, Chen Wu “A State Of The Art Opinion Mining and Its Application Domains” 2008.
- [22] Bo Pang and Lillian Lee “Opinion mining and sentiment analysis” Vol. 2, No 1-2, 2008.