# Sentimental Analysis of Online Product Reviews Using Decision Tree Algorithm

Mr. Ajith G L [1], Mr. Musheer Ahmed [2], Dr. Sanjay K S [3]

Assistant Professor, Dept. of MCA PESITM, Shimoga, Karnataka

Assistant Professor, Dept. of MCA PESITM, Shimoga, Karnataka

Associate Professor & Head, Dept. of MCA PESITM, Shimoga, Karnataka

## Abstract

Sentiment analysis or opinion classification aims in development of a system which can classify the reviews related to a product. The proposed model, supports to assess the product quality, marketers evaluations regarding the success of the newly launched product, also suggests that versions about the product is liked by the people. It is also helps us to the new features presents in the products and peoples most liked features related to the product. The term product mining also popularly known as opinion mining is using natural language processing – NLP, text mining, text analysis with computational linguistics for identifying and extraction of subjective information of the source materials, In this paper assessment the proposed work is measured with Multinomial Naive Bayes algorithm and Decision Tree algorithm.

## 1.INTRODUCTION

Sentiment analysis or opinion analysis is obtaining the sentiments by the aid of computers subjective in text data. It is one of the complex method to calculating the sentiment either as positive or negative from available multiple polarities. Hence it is very much important for real market understanding. Basically, Opinion is nothing but a person's attitude, feeling or sentiment towards a specific entity or it could be related to an specific topic. The main subjectivity of the analysis is to regulate the given sentence is subjective or objective. Consider a person is interested to purchase a product. In such situation the person would be interested to known the review of the other people. Review is nothing but the opinion of other people related to the product. But in real time situation, there will be huge amount of reviews for the given product. It would be difficult to come to a conclusion as good or bad by undergoing all the reviews. Hence sentiment mining algorithms will be deployed to find the rating about a product. Generally, opinion is developed to form goodness and badness of a product. This will help online customers to review the products. This will also help manufacturer to find the scope of its improvement.

Decision trees (DT) are considered as one of the majorly learning approaches. Noise in the data doesn't affect the algorithms efficiency. It uses the disjunctive expression. This is an advantage which leads to extend its application in the document classification. DT used as a text classifier it contains – nodes, leaves and branches. In the DT, terms are referred as internal nodes. Considering the test documents $dj$, weights are calculated and departed as branches. Categories are mentioned as the leaves in the tree. This classifier technique takes a document for testing and iteratively this process is done to attain the weights that could be labeled the nodes. The query vector representing document $d$this step is repeated till a leaf node is grasped.

## 2.LITERATURE SURVEY

Here feature selection method is adapted to find the sentiment classification rate, in terms of accuracy. Well known metrics like, accuracy, precision, recall will metrics will be considered. Five feature selection algorithms are considered for evaluation, they are Information gain, document frequency, gain, ratio, relief – f and chi squared along with popular three sentiment features Opinion Lexicon, GI, and HM are considered to design a sentiment analysis algorithm on review of movie corpus consisting of 2000 samples. (Anuj Sharma & Shubhamoy Dey, 2012). (Asuncion and Newman, 2007) machine learning data repository web site. The complete details of all the dataset which are considered can be found at UCI Machine Learning Repository. Another decision tree based method using IQ tree as data structure for classifying the reviews is proposed in (Bhanu Prakash Battula et.al, 2015). To obtain the attributes and its structure for set of objects is found at (Bhaskar Patel et.al, 2012). Sentiment analysis with opinion mining by stating “Sentiment analysis is the field of study that analyses people’s opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes” has been proposed (Bing Liu, 2012). concept of text classification is used to understand and classify the email content as spam or non-spam has been proposed (Blanzieri, and Anton, 2009). Centroid classification based approach for review classification is found here. In this approach, data set is divided into some clusters in the beginning and later 34 centroid is applied (Carlos Ordonez, 2003). A decision tree modified algorithm for user review classification using genetic algorithm for optimizing the DT results is presented here. (Liu and Fan , 2014). Majority of the approaches found here are at sentence level or term level. But here there is one document level sentiment classification approach with part of speech tagging using information gain can be found here. The approach is through shifting the gain to another dimension by synthesizing the products reviews. In the approach, essential features will be considered and aggregated for construction of final feature matrix. This approach is implemented and 26 tested for its effectiveness (Dave et al., 2003). An internet based application’s evaluation is done using large amount of review data for different applications is found here. These reviews were taken from wiki, blogs or forum websites and social media sites. It is implemented to check the efficiency of the system 31 on large data corpus. (Devika. et.al 2016). A probabilistic classifier also known as naïve bayes learning algorithm which works on the probability of the given input probability distribution from the set of classes in the database. It will consider set of classes distribution rather than 28 probability of single input. The approach categorize the review data as fake or not. It will provide the output for best distribution among the other classes. "What is the probability that a crime document D belongs to a given class C?" has been proposed (Eibe Frank et al., 2006). Logistic regression based model for the credit scorecard monitoring model is found here. Logistic regression is one of the common algorithm used in banking industry. It is implemented using real time data and evaluated for its performance. It is found at (Gang Dong et.al.,2010). Every machine 35 learning application is having some common things, i.e., rules definition (Guang-Hua Chen et.al, 2009). System learning the patterns on its own way by discovering new things by adapting specifically structural features is presented in (Ingo Mierswa et.al, 2006).

## 3.PROPOSED METHODOLOGY

This section of the chapter presents the details of proposed system. Here the opinions are categorized using the following features. In the above algorithmic model, the opinions are classified as follows.

- Word here represent the frequency: These n-gram terms represent the frequency of appearance in the documents.

- Parts of speech (POS): These are the adjectives which conveys the necessary feelings in the reviews. Example good, very good, best.

- Outlook words or phrases: Terms and words considered to convey outlooks.

- Negation: Since the terms present the negative words changes the sense of the entire sentence. For example ‘not good’ is identical to ‘bad’.

### 3.1 Feature Extraction

The performance of the proposed algorithm is evaluated with the popular metrics like recall, precision and f-measure. These are all the widely accepted metric in the domain the machine learning. Among the considered metrics, precision is the fraction of observed results which are correct. But recall represents the amount of correct observed instances.

The proposed model and the stages involved in it is shown in Fig 1. The reviews are downloaded from web portals like amazon web portal using the internet using Internet Product Database–IPDb. Developing a sentiment classification application for product review classification is considered as a challenging task as compared to another types of mining applications.

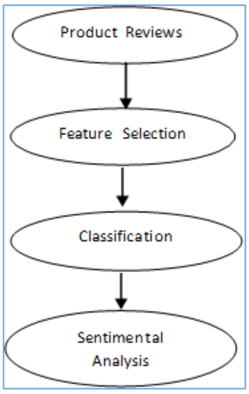


Fig 1. Block diagram of the sentimental Analysis

Developing product review application is one of the challenging application compared to other review applications, since human intervention is involved in it. In literature there are many works can be found on movie domain, but these two applications are different because of two major reasons.

First, the scope of reviews in product domain if fixed, ie. people will write it with respect to specific features. Means people may like few features and dislike some of the other features. Hence, product reviews will have both positive and negative polarity sentences. Means the algorithm need to handle both positive and negative polarities at the same time. One more interesting thing involved here is, majority of the people always discuss about other products in the review. Hence, the objective of detecting the target is important aspect of the sentiment analysis.

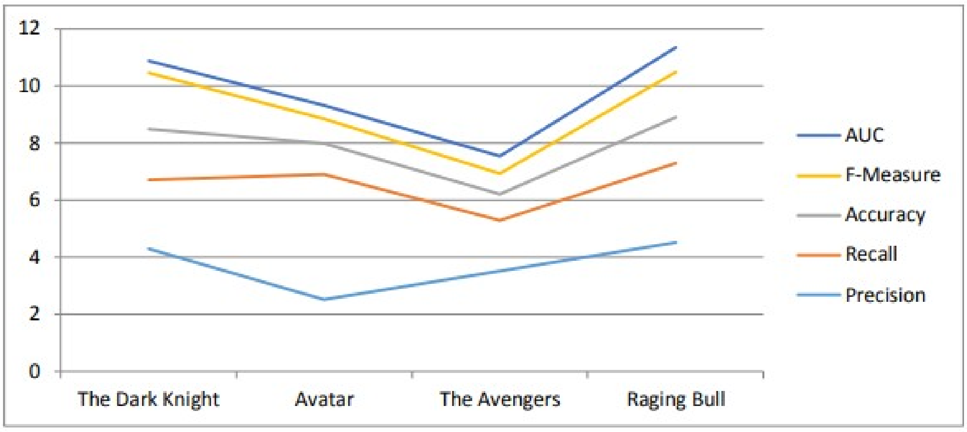


Figure 2. Sentimental Analysis for Movie Review data set using Accuracy

A comparative study of sentiment analysis using movie review datasets as show in Fig 2. In this picture, product review is from amazon. Manu methods are considered to check the performance. From the results obtained from these approaches it will be difficult to come to point which algorithm is suitable for these kind of applications since the inputs will have different set of documents. A different set with various other features are considered. Also many other feature selections are applied with variation in text granularity.

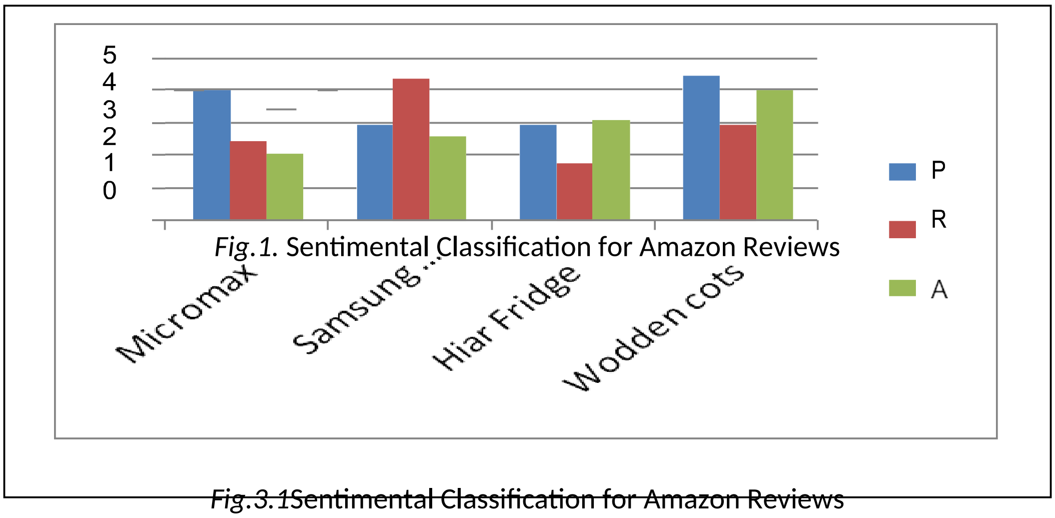


*Fig.1.* Sentimental Classification for Amazon Reviews

*Fig.3.1*Sentimental Classification for Amazon Reviews

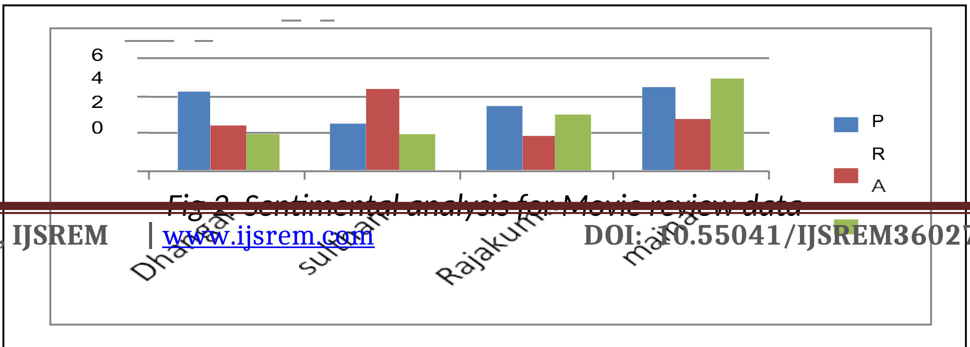


*Fig.2. Sentimental analysis for Movie review data*

*Fig. 3.2.* Sentimental Analysis for Movie Review data set

Classification of positive, negative and mixed reviews are performed using statistical features using equation (1)-(6). True achievement of the proposed model in terms of accuracy is shown in the Fig 3. Also the performance using widely accepted metrics recall, precession and f-measure are reveal in the Fig 1 and Fig 2.

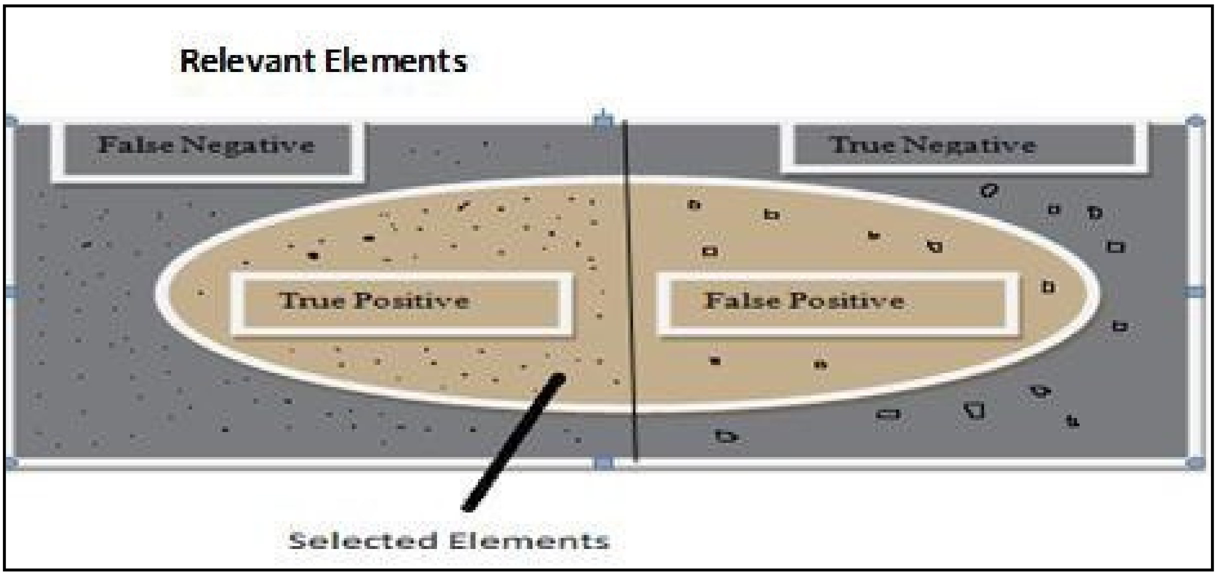


Fig. 3 : Pictorial representation of accuracy of result

## 4.EXPERIMENTAL RESULTS

In this paper, POECS(Platform for Opinion Extraction, Classification and Summarization) and CSV (Comma Separated Value) files to generate the relationship with terms to their orientations. During the process of sentiment classification at word level, the file fulfilled of the seed list.csv file is upload into the system using hash data structure. Means it will be in the form of stopword list with contains the following data. During observation of the review, if the sentiment (opinion) term is not present in the seed list, (list containing opinion term list) the searching process will be carried out on the remaining (ie., unmatched) partition to check for its presence.

In this paper the procedure is illustrated with the supporting reviews present in GSMArena.com for comparing the mobile phones. Unknown persons have clearly mentioned in his review about a Sony mobile phone that “Sony has a (possibly) latest superliner in passage and it's very equivelent to the Xperia X Performance”. From this review it we can come to a conclusion that, this review is about the particular product SONY and it is related to its performance. The details number of reviews about different product can be found in Table 1

**Table 1 List of reviews for different products**

| **Product** | **Best Ratings** |
|---|---|

| | |
|---|---|
| Samsung Galaxy S6 Edge | 1,416 |
| XiaomiMi 5 | 1,193 |
| Sony Xperia Z5 Premium | 1,172 |
| Samsung Galaxy S7 | 1,102 |
| Samsung Galaxy S7 edge | 1,109 |
| Samsung Galaxy Note5 | 969 |
| HTC 10 | 947 |
| XiaomiRedmi Note 3 | 920 |
| Huawei Nexus 6P | 909 |
| LG G5 | 862 |

In this proposed model, WEKA tool is considered for obtaining the data and setting it up ready for sentiment classification. Many reviews samples are collected related to different mobile phones. Reviews are collected based on the review samples like good, bad, worst. The details of this is shown in the following table. It is on the different cell phones and accuracy and f measures are shown below.

**Table 2 Product Reviews on different parameters**

| Features<br>Product name | True Positive | True Negative | False Positive | False Negative |
|---|---|---|---|---|
| Sony Xperia XA | 2245<br>(Excellent) | 665<br>(ok but not Good) | 1143<br>(ok but not Bad) | 771<br>(Not user friendly) |

| | | | | |
|---|---|---|---|---|
| Google Nexus 5X | 1498<br>(Excellent) | 788<br>(ok but not Good) | 231<br>(ok but not Bad) | 291<br>(Not good) |
| Samsung Galaxy A5 (2016) | 10000<br>(Excellent) | 2891<br>(ok but not Good) | 439<br>(ok but not Bad) | 463<br>(Same features, getting bore) |

**Table 3. Statistical Measures**

| Products | Precision (p) | Recall (r) | Accuracy (A) | F-measure (f) | AUC |
|---|---|---|---|---|---|
| Sony Xperia XA | 1.24 | 0.74 | 4.82 | 0.92 | 0.55 |
| Google Nexus 5X | 1.37 | 0.83 | 2.80 | 1.03 | 0.8 |
| Samsung Galaxy A5 (2016) | 3.00 | 0.95 | 4.86 | 1.44 | 0.95 |
| Xiaomi Redmi Note 3 | 0.50 | 0.76 | 3.44 | 1.39 | 0.81 |

For positive or negative or mixed opinions, categorized are finished based on statistical measures as given below.

$$F - measuref = \frac{(\beta 2 + 1) * precision * recall}{\beta 2 * Precision + recall} \qquad (1)$$

The F-measure will be stabilized similarly only when β=1 habitually, it habitually influence good for precision when β > 1and recall or else.

Where

$$Precisionp = \frac{Tp}{Tp + Fp} \qquad (2)$$

$$Recall/Sensitivity \;\; r = \frac{Tp}{Tp+Fn} \qquad (3)$$

$$\text{Accuracy A}= \frac{Tp+Tn}{Tp+Fp+Fn+Tn} \qquad (4)$$

Where:

*Tp=True Positives,* *Tn=True Negative,*

*Fp=False positive,* *Fn=False Negative,*

## 5.QUANTITATIVE ANALYSIS

In comparison to the proposed machine learning model for sentiment analysis of online products, this section discusses the similar models developed for comparing and demonstrating the effectiveness. In this current chapter, machine learning approach is applied for addressing sentiment classification problem. Our proposed model provides the rating to the input and address the issue. The proposed model is tested and results presents the effectiveness of the model presented here. This model will support to assess the quality of a product, evaluation of the market, new product success and also examine which version of the product is accepted by the public and popularized. It also select important features associated with it.

### 5.1 Multinomial Naive Bayes algorithm for Sentiment Analysis

It is one of the most popular learning technique, specially used for addressing text mining application. This technique has shown good performance for other text mining applications like text classification, email spam filtering, etc. Basically it is probability based model, where test sample will be classified based on the class probability and prior probability of the training sample. The reason for selecting this technique among available wide variety of algorithms is, in Multinomial Naive Bayes technique the problem will converge to a point and provide good considerable accuracy.

### 5.2 Algorithm: Multinomial Naive Bayes algorithm

```
Step 1: Start
Step 2: Declare variables Word, Training Sentence, Test Sentence.
Step 3: Initialize Text Vectorizer (n=2 and n=3)
Step 4: Create multinomial naive bayes instance.
Step 5: Train the algorithm with train sentence.
Step 6: Test the algorithm with test sentence.
Step 7: Evaluate the algorithm with f-measure, accuracy.
```

| Table 4 Statistical Measures (Multinomial Naive Bayes algorithm ) | | | | | |
|---|---|---|---|---|---|
| Products | Precision (p) | Recall (r) | Accuracy (A) | F-measure (f) | AUC |
| Sony<br>Xperia XA | 0.89 | 0.62 | 2.8 | 0.4 | 0.34 |
| Google<br>Nexus 5X | 0.95 | 0.75 | 1.35 | 0.8 | 0.4 |
| Samsung Galaxy A5<br>-2016 | 1.98 | 0.68 | 2.68 | 0.36 | 0.65 |
| Xiaomi<br>Redmi Note 3 | 0.2 | 0.4 | 1.65 | 0.34 | 0.36 |

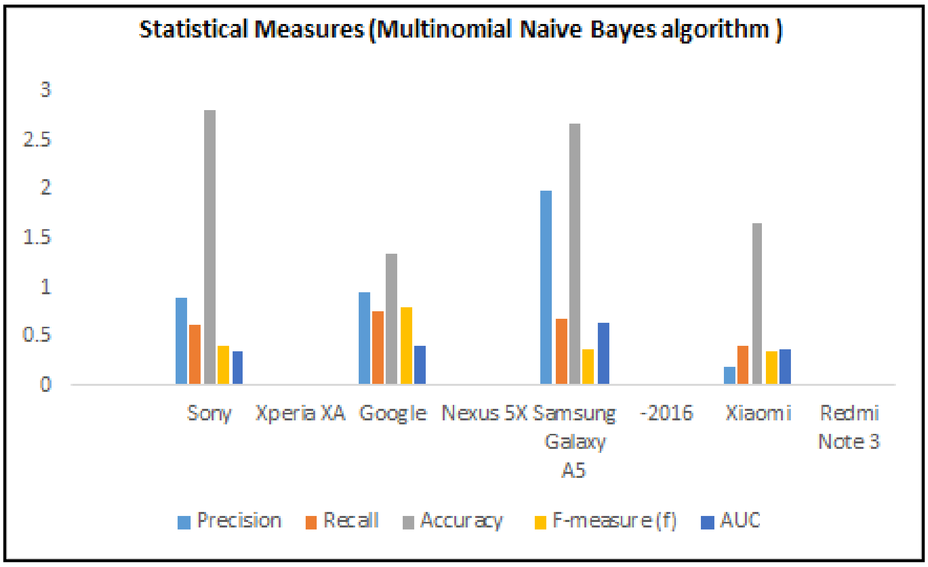


Fig 4. Statistical Measure using Multinomial Naive Bayes algorithm

### 5.3 Decision Tree algorithm for Sentiment Analysis:

It is one of the most popular supervised learning technique, specially used for addressing regression and classification problems. Since, it is sentiment classification problem a decision tree of continuous variable is built. Working principle of decision tree is, the problem starts at the root level. Based on the entropy value the root will split into two. Again the same process will carried out. Some of the techniques popularly used in decision tree are as follows: CART = classification and regression tree.CHAID = Chi-Square automatic interaction detection (this techniques multi- level splits for classification problem).

Apart of these techniques, there are many approaches are available. Based on our study we are restricting to these techniques.

### 5.4 Algorithm: Decision Tree based algorithm

Step 1: Start

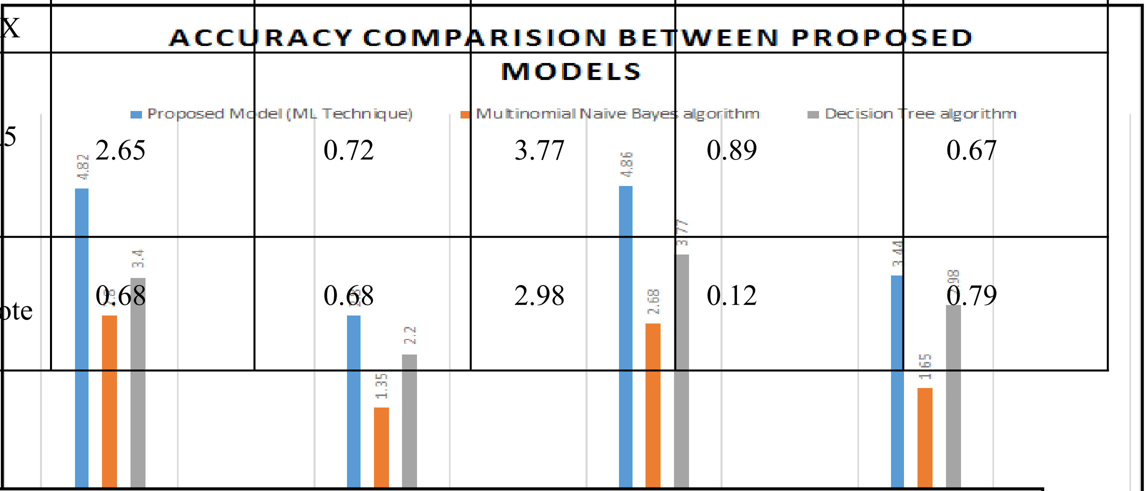
Step 2: Declare variables Word, Training Sentence, Test Sentence. Step 3: Initialize Text Vectorizer (*n=2* and *n=3*)
Step 4: Create Decision Tree instance.

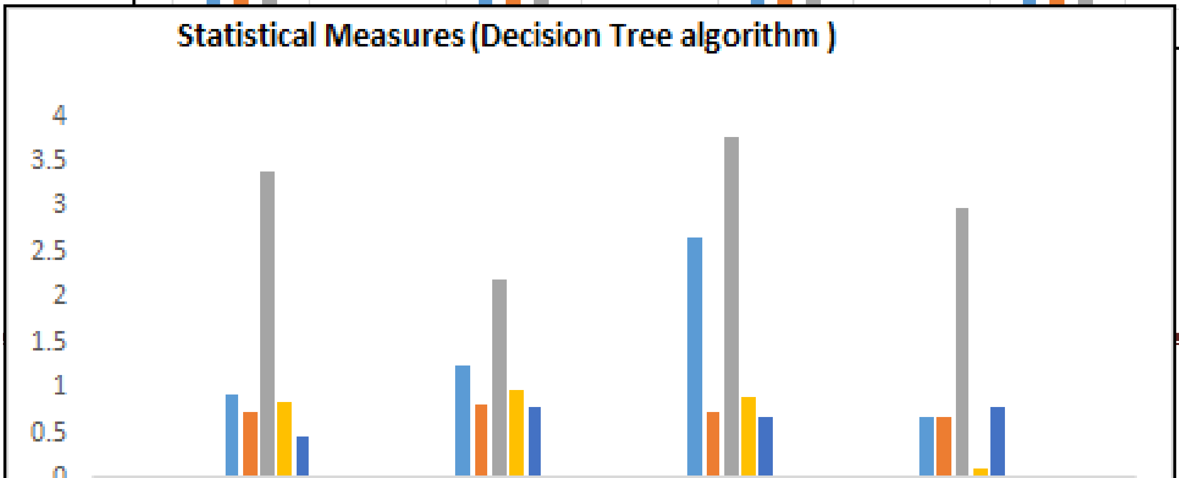
Step 5: Train the algorithm with train sentence. Obtain the Gini Index and generate a decision tree.
Step 6: Test the algorithm with test sentence.

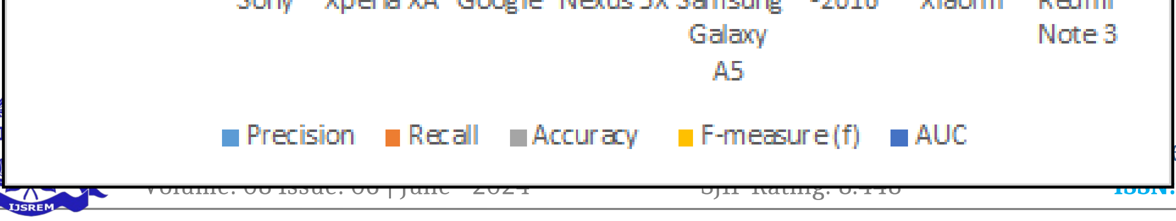
Step 7: Evaluate the algorithm with f-measure, accuracy.

**Table 5 Statistical Measures (Decision Tree algorithm )**

| Products | Precision<br>(p) | Recall<br>(r) | Accuracy<br>(A) | F-measure (f) | AUC |
|---|---|---|---|---|---|
| Sony<br>Xperia XA | 0.92 | 0.73 | 3.4 | 0.85 | 0.47 |
| Google<br>Nexus 5X | 1.25 | 0.8 | 2.2 | 0.97 | 0.78 |
| Samsung Galaxy A5<br>-2016 | 2.65 | 0.72 | 3.77 | 0.89 | 0.67 |
| Xiaomi<br>Redmi Note 3 | 0.68 | 0.68 | 2.98 | 0.12 | 0.79 |







**5.5 COMPARISON BETWEEN THE PROPOSED MODELS:**

In this chapter three models were discussed for Sentimental analysis of online product reviews using machine learning algorithms. Variety of algorithms are considered to examine the superiority between the algorithms. It is understood the results, decision tree performs well compared Multinomial Naive Bayes algorithm. Though Multinomial Naive Bayes algorithm is one of the popular technique, due to the input data and feature distribution decision tree has performed well compared to other two algorithms. Following graph on comparing the accuracy metric will provide the clear observations of the performance between the proposed models.

## 6.CONCLUSIONS.

It is open from the above tables, the result of the proposed approach is having good, appropriate results. Means it is having good precision, recall, accuracy and f-measure. Hence the proposed model is having higher performance, it can be considered for sentiment classification task of decision making of

Surveys are accessible in the web. The experiments is related to the mobile phone named Samsung Galaxy A5 manufactured year 2016. This android mobile product has obtained a good review in terms of the above mentioned metrics. The proposed model is compared with Samsung Galaxy and sonyxperia XA models. Among them, Samsung has got majority positive votings and Sony has got less votings. It means, Samsung has got more positive reviews from the public compared to Sonyxperia. All the proposed models can be enhanced and developed in several others languages also. This is one of the better enhancement direction one can think of extending it further. Another one scope is to build up the achievement of the system. Since this domain emerging fast, still lot of scope and issues are there to address.

## REFERENCES

1.Anuj Sharma & Shubhamoy Dey. “Performance Investigation of 147 Feature Selection Methods and Sentiment Lexicons for Sentiment Analysis”, Special Issue of International Journal of Computer Applications (0975–8887) on Advanced Computing and Communication Technologies for PC Applications – ACCTHPCANumber 3, June 2012.

2.Asuncion and D. Newman. “UCI Repository of Machine Learning Database (School of Information and Computer Science)”, Irvine, CA: Univ. of California , Online: http://www.ics.uci.edu/∼mlearn/MLRepository.htmJ. R. Quinlan , Available:http://www.ics.uci.edu/~mlearn/MLRepository.html, 2007.

3.Bhanu Prakash Battula, KVSS Rama Krishna and Tai-hoon Kim .“ An Efficient Approach for Knowledge Discovery in Decision Trees using Inter Quartile Range Transform”. International Journal of Control and Automation(IJCA) , Vol. 8, No. 7, ISSN: 2005-4297, Pages 325-334, 148 2015.

4. Bhaskar N Patel, Satish G. Prajapati and Dr. Kamaljit I. Lakhtaria. “Efficient Classification of Data Using Decision Tree”, Bonfring International Journal of Data Mining, Vol. 2, No. 1, ISSN 2277 – 5048, March 2012..

5. Bing Liu , Xindong Wu, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan, Angus Ng, and Philip S. Yu, et al., "Top 10 algorithms in data mining", Knowledge and Information Systems, Vol. 14, No. 1, 1-37, DOI: 10.1007/s10115-007- 0114-2, 2007.

6.Blanzieri, Enrico, and Anton Bryl. "A survey of learning-based techniques of email spam filtering." Artificial Intelligence Review 29, no. 1 , DOI: 10.1007/s10462-009-9109-6, Pages 63-92, 2009.

7. Carlos Ordonez. "Clustering binary data streams with K-means", DMKD, Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery, DOI: :10.1145/882082.882087, Pages 12-19, 2003.

8. D.-s. Liu and S.-j. Fan. "A Modified Decision Tree Algorithm Based on Genetic Algorithm for Mobile User Classification Problem", The Scientific World Journal, DOI: 10.1155/2014/468324, 2014.

9. Dave K, Lawrence S, Pennock D. M. "Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews", in Proceedings of the 12th International Conference on World Wide Web, ACM Journal, Pages 519–528, 2003.

10. Devika M D, Sunitha C, Amal Ganesh. "Sentiment Analysis:A Comparative Study On Different Approaches", Fourth International Conference on Recent Trends in Computer Science & Engineering. Chennai, Tamil Nadu, India, Procedia Computer Science 87, www.sciencedirect.com , , Pages 44 – 49, 2016.

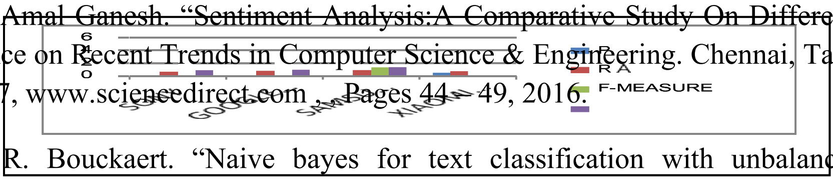


Fig. 3.7. Graphical analysis of products.

11.Eibe Frank and Remco R. Bouckaert. "Naive bayes for text classification with unbalanced classes", In Proceedings of the 10th European Conference on Principle and Practice of Knowledge Discovery in Databases, PKDD'06, Berlin, Heidelberg, Springer Journal, pages 503– 510, 2006.

12 .Gang Dong,Kin Keung Lai,Jerome Yen. "Credit scorecard based on logistic regression with random coefficients" , International Conference on Computational Science, ICCS 2010, ELSEVIER Journal, DOI: 10.1016/j.procs.2010.04.278 Pages 2463–2468, 2010.

13. Guang-Hua Chen; Zheng-Qun Wang; Zhen-Zhou Yu. "Constructing Decision Tree by Integrating Multiple Information Metrics," Chinese Conference on Pattern Recognition ,CCPR 2009, DOI: 10.1109/CCPR.2009.5344133, Nov 2009.

14. Ingo Mierswa, Michael Wurst, Ralf Klinkenberg, Martin Scholz, Timm Euler. "YALE: rapid prototyping for complex data mining tasks", Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, DOI 10.1145/1150402.1150531, 2006.