# Product Recommendation using Sentiment Analysis of Reviews: A Random Forest Approach

**Gayatri Khanvilkar, Deepali Vora**

***Abstract- Nowadays people are attracted to social networking sites and e-commerce websites. Due to growth in social media all the fortune companies are working on Sentiment Analysis. In Natural Language Processing, Sentiment analysis has become a major area of research. This paper explores the performance of Machine Learning Algorithms such as Multinomial Nave Bayes algorithm, Logistic Regression, SVM Classifier, Decision Tree and Random Forest are used for sentiment analysis. Comparative tabulation of above mentioned classifiers is created to analyze the performance of sentiment analysis. Random Forest can produce a great result most of the time. It is most flexible and easy to use supervised machine learning algorithm. In proposed system, Random Forest shows outstanding performance. The polarity achieved by different algorithms is used to generate product recommendations to users.***

***Keywords - Sentiment analysis, Recommendation System, Machine learning, Random Forest, Content-based Recommendation***

## I. INTRODUCTION

### *A. Sentiment Analysis*

Social network, e-commerce sites, blogs are new emerging platforms for people to express their opinion. These sites contain huge amount of text which can be used for different purpose like Sentiment Analysis. Sentiment analysis uses Natural Language Processing technique. It is also known as Opinion Mining because it takes the writer's opinion. The feedback of consumer such as reviews, comments or response to survey is the main factor of Sentiment analysis. Sentiment analysis determines the attitude of customer, writer or speaker. The main task of sentiment analysis is to classify the given content based on polarity. Main aim of sentiment analysis is to find actual insight from piece of text. Social media is the important source of customer's voice. Using this information, can analyze consumer behavior, what customers want, what are customer's like and dislike about products, what their buying signals are, what their decision process looks like etc. Based on these factors can provide insights that can: [1][2][3]

- Fix marketing strategy
- Improve the success of the campaign
- Modify production messaging
- Increase customer service
- Test business KPIs
- Create leads

### *B. Random Forest*

Random Forest is meta-algorithm which uses several Decision trees. It combines prediction value of each decision trees and uses majority vote method which returns class with majority votes. Sometimes Decision tree grows deeply and faces problem of overfitting and learn irregular patterns. But Random forest solves this problem of overfitting. [4][5]

Random Forest classifier is best known for its randomness. It provides randomness in two ways:

1) data related randomness and
2) features related randomness.

Random Forest classifier uses the concept of Bagging and Bootstrapping.

Features of Random Forest

There are two features of random forest which are as follows:

- Robust:

Random forest combines the results of different decision trees which trains on different data. This is known as bootstrapping. In this each tree considers random subset of training data. So, it is much robust for noise.

- Accuracy:

Random forest uses the concept of bagging. In this average of all classifiers is calculated for final output. Giving huge data to single classifier will not return appropriate result but if those data can divide into number of classifiers then averaging of results of classifiers will give consistent solute on. [6] [15]

## II. LITERATURE REVIEW

Sentiment Analysis is analysis of customer's opinions, expressions, likes and dislikes towards different products, services, organizations, individuals etc. [1] Nowadays sentiment classification gaining more attention of researchers and decision makers as it helps to better understand customers' feedback. In Business analysis field Sentiment analysis can play vital role. [4] Sentiment Classification has two approaches: Machine Learning and Lexicon-based. [1] There are two types of methods for solving problem of categorizing sentiment analysis: 1) topic-based and 2) knowledge-based. For solving Sentiment Analysis as topic-based categorization problem need to use machine learning based approach. For solving

Revised Manuscript Received on December 22, 2018.

**Gayatri Khanvilkar**, Department of Information Technology, Vidyalankar Institute of Technology, Mumbai, India (Email: khanvilkarg7@gmail.com)

**Deepali Vora**, Department of Information Technology, Vidyalankar Institute of Technology, Mumbai, India (Email: deepali.vora@vit.edu.in)

Sentiment Analysis as knowledge-based categorization problem need to use lexicon-based approach. It contains pre-defined lexicons i.e. sentiment polarity for each word, to label the sentiments of words. The machine learning does not need predefined semantic rules but requires a labelled dataset. [12]

In Machine Learning technique to classify emotions, first phase is cleaning data. Second phase is training a model with the help of cleaned trained data to classify test data. In this, the text is given the weight using Term Frequency-Inversed Document Frequency (TF-IDF) algorithm. Some of the pre-processing steps that have been carried out are-Tokenization, Filtering, Lemmatization, Stemming, etc. [9] In Natural Language Processing needs to study vocabulary which can be done efficiently by Bag-of-words model. The bag-of-words is mostly used in document classification. [1]

Generally, SVM gives good accuracy and can be improved by modifying hyper-parameters. [10] There are three most popular CART styles: using Gini index, information gain and twoing. The identity of best classifier (predictor) is lower misclassification rate. [13] The main reasons for DT popularity includes intuitiveness, expressiveness, transparency, efficiency, robustness, accuracy, and deploy ability. [11] Random Forest Classifier requires high processing power and training time to achieve greater accuracy and performance. [7] Random Forest classifier provides best randomness with respect to data and features. Random Forest is robust classifier which achieves great accuracy. [6] RF method has many advantages. It does not need assumptions on the distribution of explicative factors; it can properly analyse interaction between factors; In Random Forest the random predictor selection holds low bias; and Random Forest can handle problem of over-fitting and unbalanced data. [8]

Careful feature selection can give better accuracy. [10] Generally, steps like pre-processing, sentiment extraction, feature selection get neglected, but these are important steps to achieve great Accuracy in sentiment. Efficiency of classifier is important factor as well. [5] F1-measure, Recall, Precision are the main Accuracy Evaluation parameters. [4]

## III. PROPOSED METHODOLOGY

Proposed methodology uses machine learning approach for sentiment analysis. Proposed system gives polarity of reviews and recommendation list. The flow of proposed methodology is shown in fig1.

Step 1: Collection of Data

Dataset downloaded from www.kaggel.com where PromptCloud extracted 400 thousand reviews of unlocked mobile phones sold on Amazon.com. This dataset contains following information:

- Name of Product
- Name of Brand
- Price of product
- Rating given by customers
- Reviews written by customers
- Number of people who found the review helpful

The total number of reviews extracted were more than 400,000 covering close to 4,400 unlocked mobile phones. [14] [15]

Step 2: Data Labelling

In this step we clean the data and label the data. Read the data from csv file and add new column for labels.

Step 3: Data Cleaning

Remove all the rows containing blank cells. After Labeling and Cleaning dataset the resultant data is labelled dataset and is stored in new csv file.

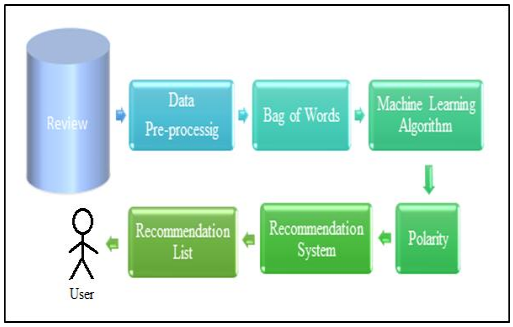



**Fig 1. Proposed System**

Step 4: Data pre-processing

The text pre-processing is important to convert raw data to cleaned data. Hence, following steps are implemented:

- remove html tags
- remove non-character
- convert to lower case
- remove stop words
- convert to root words by stemming

Step 5: Bag of Words

Now we have cleaned reviews, the next step is to convert the reviews into numerical representations for machine learning algorithm. Create BoW using CountVectorizer / Tfidfvectorizer.

CountVectorizer which implements both tokenization and occurrence counting in a single class. The output is a sparse matrix representation of a document.

TfidfVectorizer which implements both tokenization and tf-idf weighted counting in a single class.

Step 6: Train Machine Learning algorithm

Train Machine Learning algorithms such as Multinomial NB, Logistic Regression, Decision Tree, SVM and Random Forest using training dataset and test it on validation dataset. It will give polarity of reviews.

Step 7: Model Evaluation

There are multiple functions for model evaluation in scikit learn such as Accuracy score, Recall, F_score, Confusion Matrix, etc.

Step 8: Recommendation system

Consider Sentiment analysis for recommendation of product

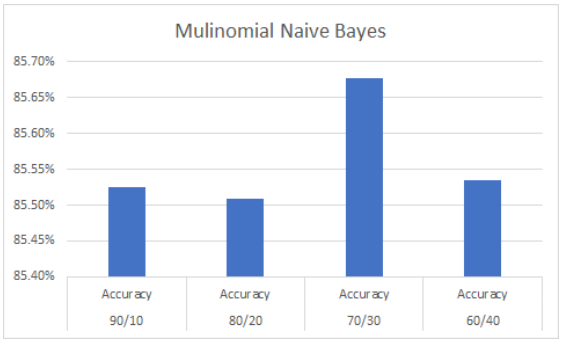
## IV. RESULTS AND DISCUSSION

Dataset divided into training and test dataset. It is divided into four types of ratio. Result of all dataset ratio and algorithms are shown in following Table 1.

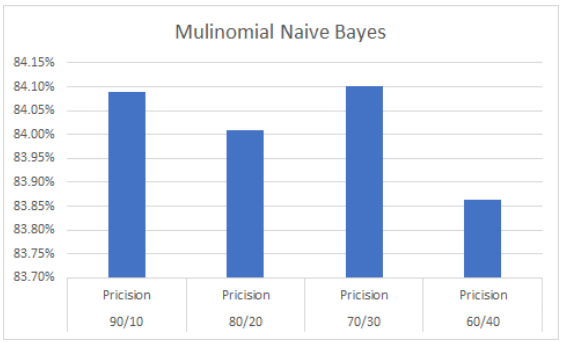
**Table 1. Overall Result of Algorithms**

| Dataset Ratio | Parameters | Algorithms | | | | |
|---|---|---|---|---|---|---|
| | | Multinomial Naive Bayes | Logistic Regression | LinearSVC | Decision Tree | Random Forest |
| 90/10 | Accuracy | 85.52% | 88.13% | 89.17% | 92.65% | 95.03% |
| | Precision | 84.09% | 86.73% | 88.34% | 92.50% | 95.12% |
| | Recall | 85.52% | 88.13% | 89.17% | 92.65% | 95.03% |
| | F-1 score | 84.58% | 86.19% | 87.73% | 92.55% | 94.78% |
| 80/20 | Accuracy | 85.51% | 88.10% | 89.06% | 92.27% | 94.54% |
| | Precision | 84.01% | 86.66% | 88.13% | 92.27% | 94.64% |
| | Recall | 85.51% | 88.10% | 89.06% | 92.27% | 94.54% |
| | F1-Score | 84.51% | 86.13% | 87.58% | 92.17% | 94.27% |
| 70/30 | Accuracy | 85.68% | 88.17% | 89.05% | 91.68% | 94.15% |
| | Precision | 84.10% | 86.65% | 88.03% | 91.50% | 94.26% |
| | Recall | 85.68% | 88.17% | 89.05% | 91.68% | 94.15% |
| | F1-Score | 84.60% | 86.15% | 87.56% | 91.57% | 93.82% |
| 60/40 | Accuracy | 85.53% | 88.03% | 88.83% | 90.66% | 93.64% |
| | Precision | 83.86% | 86.46% | 87.74% | 90.48% | 93.76% |
| | Recall | 85.53% | 88.03% | 88.83% | 90.66% | 93.64% |
| | F1-Score | 84.38% | 85.97% | 87.26% | 90.55% | 93.25% |

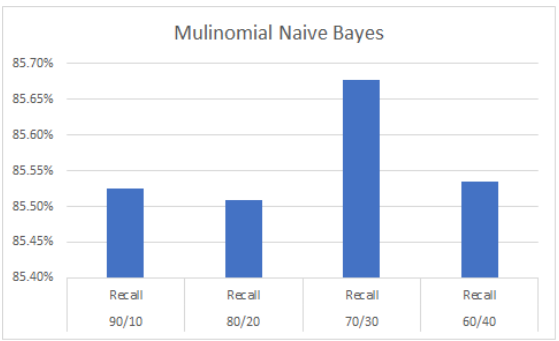
Analysing results of every algorithm with respect to Accuracy, Precision, Recall and F-1 score for each ratio.



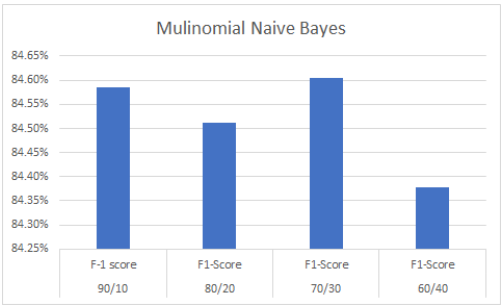

**Graph 1(a) Accuracy of Multinomial NB**




**Graph 1(b) Precision of Multinomial NB**

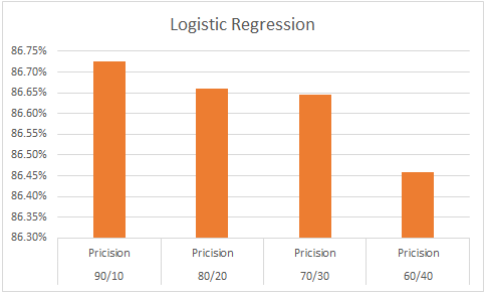



**Graph 1(c) Recall of Multinomial NB**



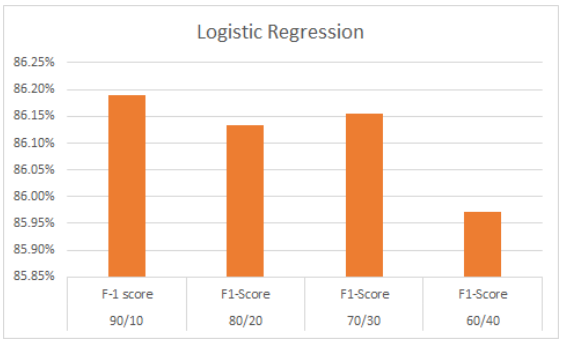

**Graph 1(d) Recall of Multinomial NB**


**Graph 2(a) Accuracy of Logistic Regression**

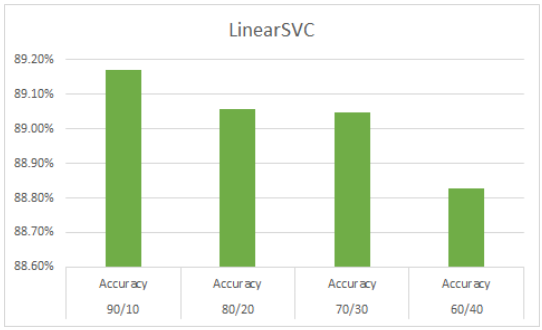



**Graph 2(b) Precision of Logistic Regression**

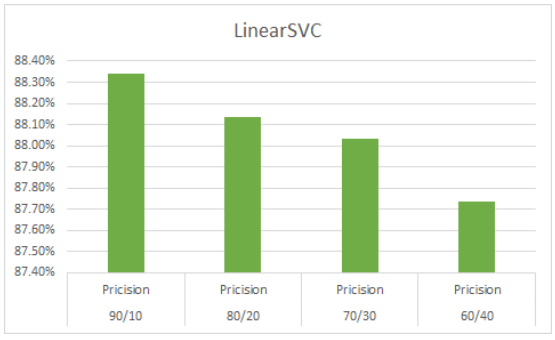

**Graph 2(c) Recall of Logistic Regression**



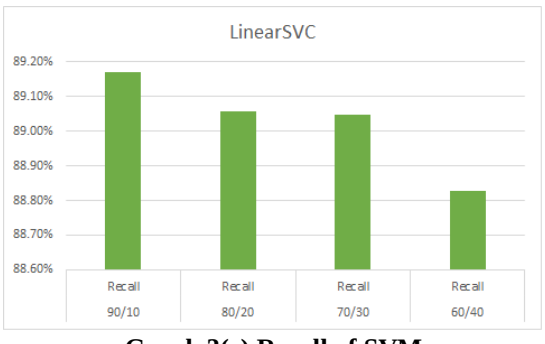

**Graph 2(d) F1-Score of Logistic Regression**



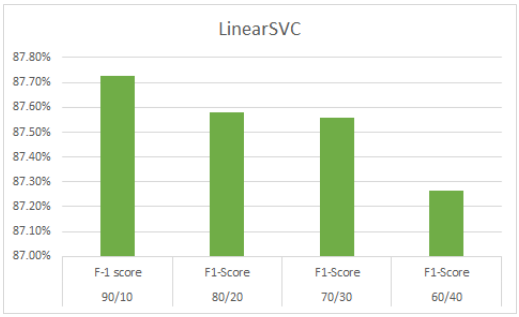

**Graph 3(a) Accuracy of SVM**



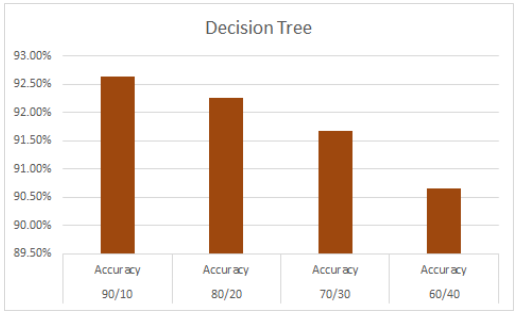

**Graph 3(b) Precision of SVM**



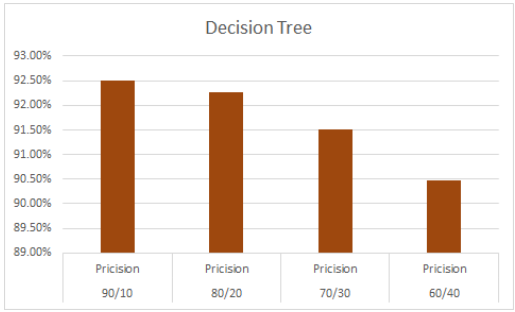

**Graph 3(c) Recall of SVM**



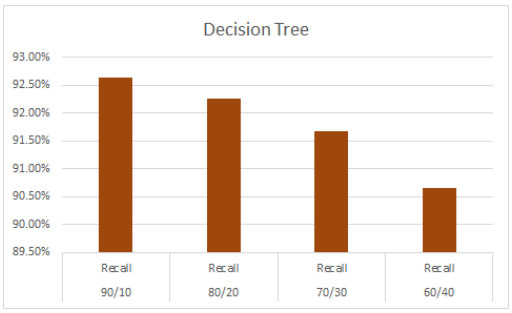

**Graph 3(d) F1-Score of SVM**



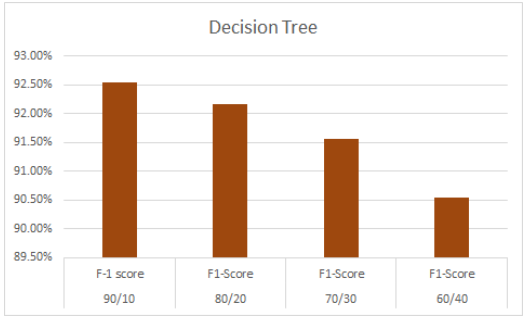

**Graph 4(a) Accuracy of Decision Tree**




**Graph 4(b) Precision of Decision Tree**

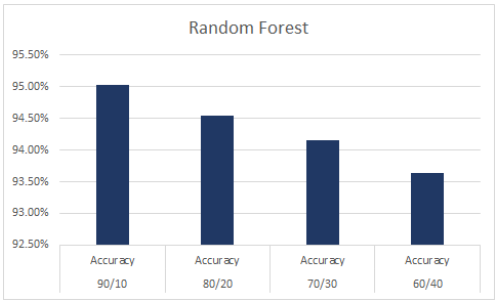



**Graph 4(c) Recall of Decision Tree**

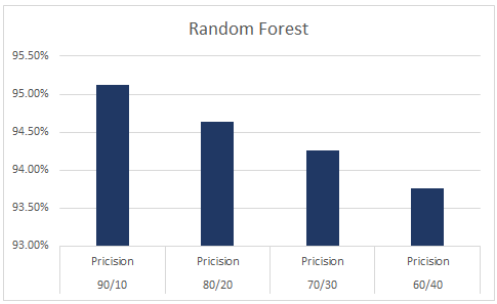



**Graph 4(d) F1-Score of Decision Tree**

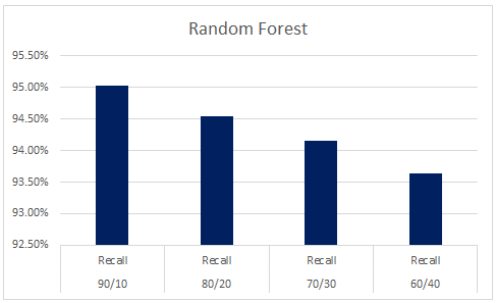

*Published By:*
*Blue Eyes Intelligence Engineering*
*& Sciences Publication*



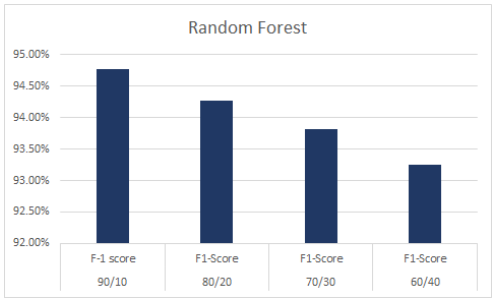

**Graph 5(a) Accuracy of Random Forest**




**Graph 5(b) Precision of Random Forest**

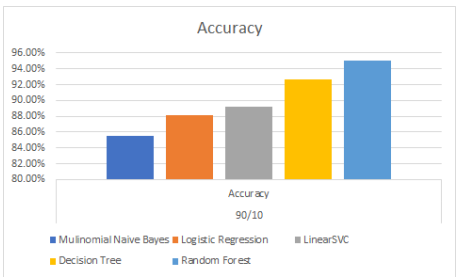



**Graph 5(c) Recall of Random Forest**

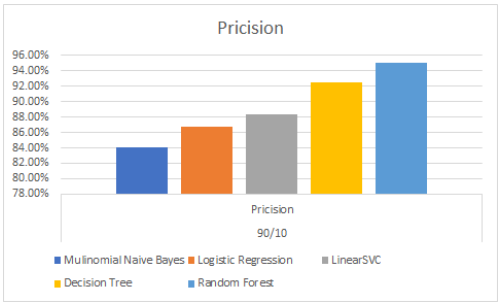



**Graph 5(d) F1-Score of Random Forest**

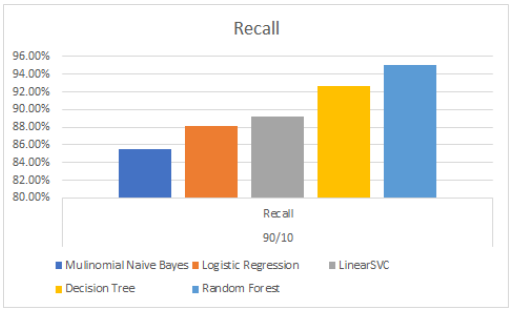
From the above graphs it is proved that data splitting in 90:10 ratio gives best results as compare to other dataset splitting ratios. Hence graph of all algorithms with respect to Accuracy, Precision, Recall and F-1 score for 90:10 ratio are as follows:



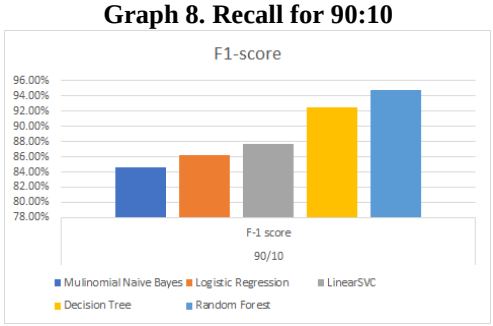

**Graph 6. Accuracy for 90:10**




**Graph 7. Precision for 90:10**

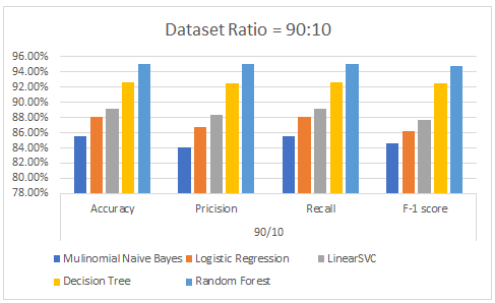



**Graph 8. Recall for 90:10**

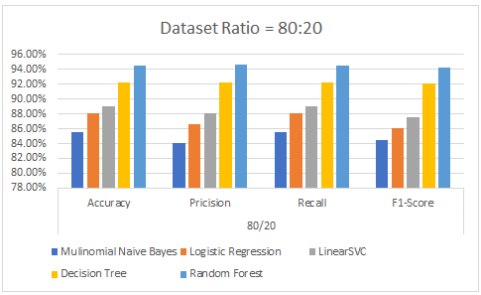



**Graph 9. F1-score for 90:10**

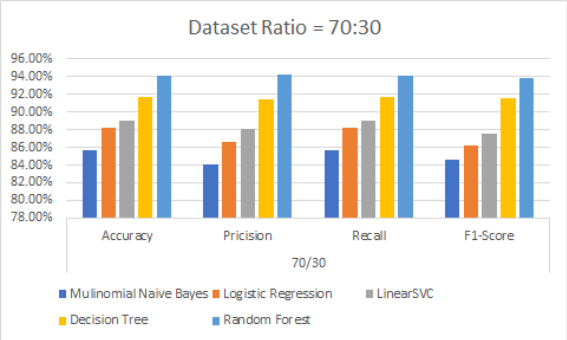
Graphical representation of all algorithms according to dataset splitting ratio is given below:



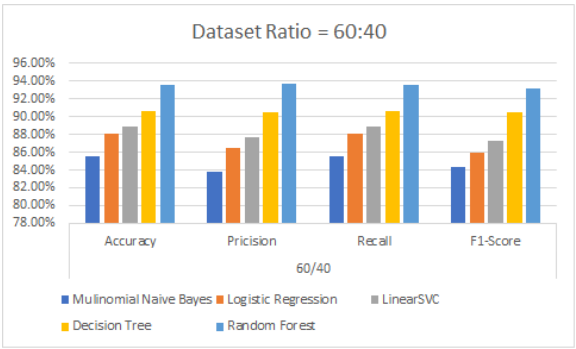

**Graph 10. Dataset Ratio 90:10**




**Graph 11. Dataset Ratio 80:20**




**Graph 12. Dataset Ratio 70:30**

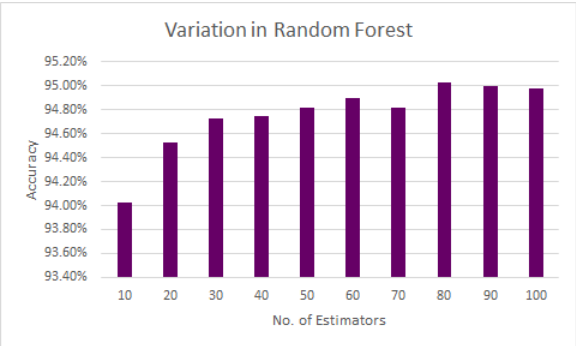



**Graph 13. Dataset Ratio 60:40**

From all above results we get that Random forest is showing high accuracy and result for all Dataset splitting ratios. Performance of random forest is depending on many parameters but it is mainly depending on number of estimators used. As the number of decision tree increases accuracy of random forest is also increases. Hence variation in random forest by changing no. of estimators is showed in Table 2. and its graphical representation is given below:

**Table 2. Variation in Random Forest**

| No. of Estimators | Accuracy |
|---|---|
| 10 | 94.03% |
| 20 | 94.53% |
| 30 | 94.72% |
| 40 | 94.75% |
| 50 | 94.82% |
| 60 | 94.90% |
| 70 | 94.82% |
| 80 | 95.03% |
| 90 | 95.00% |
| 100 | 94.98% |




**Graph 14. Variation in Random Forest**

From the graph 14 it shows that Random forest gives best result for data splitting ratio 90:10 when number of estimators i.e. number of trees in random forest are 80.

## V. CONCLUSION

In this paper, we have used machine learning techniques such as multinomial naive Bayes classifier, Logistic regression, decision tree, SVM classifier and Random Forest to perform sentiment analysis. This paper mainly focused on Random Forest to carry out the sentiment analysis of mobile product reviews. On the basis of experimental results, it has been proved that random forest performed well on the dataset. It provided very favourable results with accuracy of 95.03%. Most of the previous sentiment classification work has focused on mainly SVM and Naive Bayes algorithms. By considering achieved results it's been concluded that Random forest can perform well. Random forest can improve results if data splitting ratio and number of estimators are perfectly tuned. The main contribution of this work is the performance investigation of different machine learning methods in terms of accuracy and using sentiment polarity for recommendation.

## REFERENCES


1 Rao, Shivani, and Misha Kakkar. "A rating approach based on sentiment analysis." *Cloud Computing, Data Science & Engineering-Confluence, 2017 7th International Conference on*. IEEE, 2017.

2 "The Importance of Sentiment Analysis in Social Media Analysis", _https://www.linkedin.com/pulse/importance-sentiment-analysis-social-media-christine-day,__October 2017.

3 "Why is Sentiment Analysis important from a business perspective?", http://blog.aylien.com/why-is-sentiment-analysis-important-from-a/, October 2017.

4 Wan, Yun, and Qigang Gao. "An ensemble sentiment classification system of twitter data for airline services analysis." *Data Mining Workshop (ICDMW), 2015 IEEE International Conference on*. IEEE, 2015.

5 Hegde, Yashaswini, and S. K. Padma. "Sentiment Analysis Using Random Forest Ensemble for Mobile Product Reviews in Kannada." *Advance Computing Conference (IACC), 2017 IEEE 7th International*. IEEE, 2017.

6 Parmar, Hitesh, Sanjay Bhanderi, and Glory Shah. "Sentiment Mining of Movie Reviews using Random Forest with Tuned Hyperparameters." (2014).

7 Bhavitha, B. K., Anisha P. Rodrigues, and Niranjan N. Chiplunkar. "Comparative study of machine learning techniques in sentimental analysis." *Inventive Communication and Computational Technologies (ICICCT), 2017 International Conference on*. IEEE, 2017.

8 Pham, Binh Thai, Khabat Khosravi, and Indra Prakash. "Application and comparison of decision tree-based machine learning methods in landside susceptibility assessment at Pauri Garhwal area, Uttarakhand, India." *Environmental Processes* 4.3 (2017): 711-730.

9 Dixit, Apurva, et al. "Emotion Detection Using Decision Tree." *Development* 4.2 (2017).

10 Vaghela, Vimalkumar B., and Bhumika M. Jadav. "Analysis of Various Sentiment Classification Techniques." *Analysis* 140.3 (2016)

11 Kuzey, Cemil, Ali Uyar, and Dursun Delen. "An Investigation of the Factors Influencing Cost System Functionality Using Decision Trees, Support Vector Machines and Logistic Regression." (2018).
12 Peng, Haiyun, Erik Cambria, and Amir Hussain. "A review of sentiment analysis research in Chinese language." *Cognitive Computation* 9.4 (2017): 423-435.
13 Ceňak, Miloš. "A comparison of decision tree classifiers for automatic diagnosis of speech recognition errors." *Computing and Informatics* 29.3 (2012): 489-501.
14 Amazon Reviews: Unlocked Mobile Phones, https://www.kaggle.com/PromptCloudHQ/amazon-reviews-unlocked-mobile-phones, April 2018. 15.
15 Gayatri Khanvilkar, Deepali Vora, "Sentiment Analysis for product recommendation using Random Forest", SCOPUS indexed, International Journal of Engineering and Technology (IJET-UAE), ISSN:2227-524X, Vol:7, No 3.3(2018), pp 87-89, 2018.