Smart Recommendation System based on Product Reviews using Random Forest

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

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

***Abstract*— 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 is a growing field in natural language processing. Sentiment analysis is major focused on company's improvement. But sentiment analysis can be useful in recommendation system also. Based on various performance measures, this paper compares the results of machine learning algorithms like Multinomial Naive Bayes algorithm, Logistic Regression, SVM Classifier, Decision Tree and Random Forest. These algorithms are used for sentiment analysis of reviews and in turn for product recommendation. In proposed system, Random Forest shows outstanding performance. To create suitable recommendations using the analysis of emotions, there is a need to use polarity obtained through the reviews.**

Keywords*— *Random Forest, Sentiment Analysis, Product Recommendation, NLP, Reviews.

I. INTRODUCTION

Sentiment analysis is computational methodology. It extracts emotional insights from person's voice such as text, speech or dataset. Sentiment analysis refers to use Natural language Processing. Main aim of sentiment analysis is to find actual insight from piece of text. It can provide insights that can:

1) Deciding marketing strategy
2) Enhances campaign success
3) Reform product messaging
4) Improve the quality of customer service
5) Examine business KPIs
6) Generate leads [1][2]

Sentiment analysis classifies the subjective effect of the user's feelings, the author's perspective, the opinions given by the customer into polarity. In general, there are two categories of approaches to address this problem.

1) Machine Learning based approaches- This approach is one of the most well-known technique. This technique is more adaptable and accurate, hence it gaining interest of researchers. There are two machine learning approaches, supervised and unsupervised which performs sentiment classification. Algorithms used for classification are generally come into supervised approach where it uses supervised machine learning algorithms such as Decision tree, Naive Bayes, etc. whereas Neural network is unsupervised learning algorithm which is used for clustering.
2) Lexicon-based approaches- Lexicon-based approaches is easiest way for performing sentiment analysis. This technique makes use of a dictionary which contains pre-tagged lexicons. WordNet and SentiwordNet are publicly available dictionaries to perform sentiment analysis. As the size of the dictionary grows, the performance decreases, that is, it requires more time and becomes more complicated, the accuracy also decreases. [3]

Recommender system generates list of recommended products or services. Generating meaningful recommendation to user is the goal of recommendation system. There are three approaches of Recommendation system.

1) Content-based approach- - The content-based approach considers user's profile and finds user preferences. This approach generate recommendation by considering information of items and user preferences.
2) Collaborative-based approach- As compare to content-based approach, collaborative-based approach gives more accurate results as it analyses the user behavior and liking and accordingly finds the same relation and liking between people.
3) Hybrid approach- The hybrid approach combines two or more approaches for better recommendation. [4]

II. WORK DONE IN SENTIMENT ANALYSIS AND RECOMMENDATION SYSTEM

Table 1: Summary of papers reviewed for Sentiment Analysis

Sr. no.	Authors Name	Key Points	Algorithm used	Inference
1.	Abirami, A. M., and V. Gayathri [2]	Different approaches of sentiment analysis and problem faced by them.	This is a survey paper on sentiment analysis.	Polarity shift, and data sparsity are the issues of sentiment analysis which can be handled by different techniques.
2.	Thakkar, Harsh, and Dhiren Patel [3]	This is a survey paper on approaches of sentiment analysis. on Twitter dataset.	Lexical-based, machine learning and hybrid approaches	Most popular Machine learning approach gives best results. But without proper training of a classifier in machine learning approach results may deteriorate drastically.

978-1-5386-9166-3/19/$31.00 ©2019 IEEE

Sr. no.	Authors Name	Key Points	Algorithm used	Inference
3.	Rao, Shivani, and Misha Kakkar [5]	In this paper it sorts product/ service based on polarity of reviews written by user on different social networking sites which is achieved by using sentiment analysis.	Lexicon based approach	Sentiment analysis, its approaches and explanation of importance of bag of world model.
4.	Wan, Yun, and Qigang Gao. [6]	An ensemble sentiment classification strategy generally gets applied on multiple classification methods where it uses Majority Vote principle of it.	Bayesian Network, Naive Bayes, SVM, C4.5 and Random Forest	In business analysis Sentiment analysis plays vital role. Accuracy Evaluation of classification algorithms are based on f1-measures, precision, Recall etc.
5.	Hegde, Yashaswini, and S. K. Padma. [7]	The sentiment analysis for Kannada Language to identify polarity and measure performace of ML Classifiers	Naive Bayes and Random Forest	Preprocessing and sentiment extraction are important steps in sentiment analysis to achieve higher accuracy.
6.	Parmar, Hitesh, Sanjay Bhanderi, and Glory Shah. [8]	This paper focuses on tuning set of hyperparametes of Random Forest manually.	Random Forest	Randomness with respect to data and features, can be provided by Random Forest algorithm efficiently
7.	Bhavitha, B. K., Anisha P. Rodrigues, and Niranjan N. Chiplunkar. [9]	This a survey paper which compares accuracy, advantages and limitations of various machine learning techniques.	Lexicon based approach, Naive Bayes, random forest and SVM	Random Forest, classifier requires more processing power and more training time but gives result with greater accuracy.
8.	Pham, Binh Thai, Khabat Khosravi, and Indra Prakash. [10]	In this paper for prediction of local landslides they have used machine learning approach. In this they have also compared the results of decision trees which have been used in prediction.	Random Forest, Logistic Model Trees (LMT), Best First Decision Trees (BFDT) and CART	Random Forest overcomes the limitations of DT. i.e. it is able to deal with unbalanced data and over-fitting
9.	Dixit, Apurva, et al. [11]	Recognizing emotions from text available on social networking sites with the help of Sentence level classification from the tweets.	SVM and Decision Tree	Data labeling and pre-processing is an important phase for getting best results.
10.	Vaghela, Vimalkumar B., and Bhumika M. Jadav. [12]	Comparative study of sentiment classification of various approaches and algorithms. This is main contribution of paper.	SVM, Naive Bayes, Maximum Entropy and Lexicon based approach	Feature Selection plays vital role in sentiment analysis. Careful feature selection can give better accuracy.
11.	Kuzey, Cemil, Ali Uyar, and Dursun Delen. [13]	Several prediction models are developed and tested	C5, CART, CHIAD, SVM	Intuitiveness, expressiveness, transparency, efficiency, robustness, accuracy, and deploy ability are the main reasons for DT popularity.
12.	Cerňak, Miloš. [14]	Applying and comparing performance of different Decision tree Algorithms on data of computer for speech recognition.	CART with various splitting criteria and C4.5	To improve the classifier or predictor of the error made needs to decrease misclassification rate.s

Table 2: Summary of papers reviewed for Recommendation System

Sr. No.	Author Name	Key points	Inference
1.	Rosa, Renata L., Demsteneso Z. Rodriguez, and Graca Bressan [15]	The sentiments extraction from feedback posted on different social networking sites and the music recommendation system is performed using hybrid approach.	In recommendation system, user related information and analysis of writers' emotions are important factors but Sentiment Analysis do not differentiate emotions of users according to their profiles.
2.	Zheng, Xiaoyao, [16]	A recommender system of tourism destination uses opinion-mining technology to find user's sentiments.	There are different selection techniques used while generating recommendation, such latent factor-based recommender systems, review text analysis techniques and the temporal factor, which helps to improves the performance of system.
3.	Amel Ziani, Nabiha Azizi, Didier Schwab, Monther	A recommendation system which is based on multilingual sentiment analysis of online available	In case of lack of labelled dataset, Semi-supervised support vector machines is the best solution. This

Sr. No.	Author Name	Key points	Inference
	Aldwairi, Nassira Chekkai, Djamel Zenakhra, and Soraya Cheriguene [17]	products reviews, helps user to decide different products or services.	algorithm have been widely used in many classification problems. Superior performance with unlabeled datasets is the major advantage of the S3VM.
4.	Greg Linden,Brent Smith,and Jeremy York [18]	Comparison of all Recommendation system approaches.	A good recommendation algorithm is able to meet all challenge. With the help of item-to-item collaborative filtering.
5.	Aggarwal [19]	Information of popular examples of historical and current recommender systems, and basic approach of recommendation system.	memory-based algorithms and model-based algorithms are two types of collaborative filtering algorithms.
6.	Stefan Hauger Karen H. L. Tso and Lars Schmidt-Thieme [20]	The new-item problem, and the user-bias problem are the issues related to the recommendation system. Collaborative filtering algorithms are able to solve these problems.	With the help of selecting adequate attributes, issues like new-item and user-bias get solved and results can be improved.

III. PROPOSED SYSTEM

Sentiment Analysis and recommendation generation according to polarity obtained by sentiment analysis are the main tasks of proposed system. System will do prediction with the help of machine learning algorithms. The system consists of the Collection of data, Pre-processing of data, Bag-of-words, Apply machine learning algorithms, Prediction and accuracy evaluation of algorithms and Recommendation. This block diagram depicts all the components of the system. It has the following major activities:

Step 1: Collection of Data

The system used amazon dataset from www.kaggel.com where the total number of reviews extracted were more than 400,000 and total number of unlocked mobile phones (Products) sold on Amazon.com are 4,400. This dataset contains following Attributes:

1) Product Name
2) Brand Name
3) Price
4) Rating
5) Reviews
6) Review Votes [4] [21]

Step 2: Data pre-processing

Data pre-processing enhances the quality of data. The quality of data affects the final result. Data pre-processing includes data cleaning, data normalization, data transformation, etc.

Step 2.1 Data Labelling- In this step read the data from csv file and add new column for labels. Current dataset is unlabeled. There is a need to label data for applying supervised learning algorithms on basis of ratings given by customer.

Step 2.2: Data Cleaning- The dataset has some missing values. All machine learning algorithms cannot handle missing values. Hence, remove all the rows containing blank cells.

Step 2.3 Data Transformation- Transformation of collected dataset is required to convert raw data to required form, so that feature extraction will be easy. It can remove html tags, remove non-character such as digits and symbols, convert to lower case, remove stop words and convert to root words by stemming, etc.

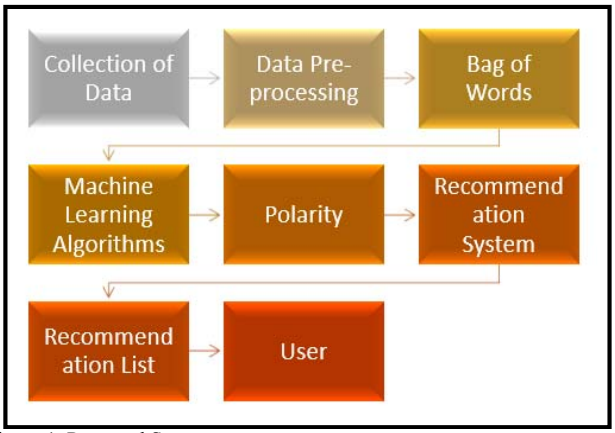

Figure 1. Proposed System

Step3: Bag of Words

To convert the reviews which are in natural language format into numerical representations for machine learning algorithm needs to use BoW model. BoW can be created using CountVectorizer and TfidfVectorizer.

1) CountVectorizer- It implements both tokenization and occurrence counting in a single class. The output is a sparse matrix representation of a document.
2) TfidfVectorizer- - It implements both tokenization and tf-idf weighted counting in a single class. This eliminates frequently occurring words which does not content much meaning.

Step 4: Enforce Machine Learning algorithms

On Pre-processed data perform classification using Machine learning algorithms such as Multinomial Naive Bayes algorithm, Logistic Regression, SVM Classifier, Decision Tree and Random Forest.

Step 5: Model Evaluation of algorithms

After applying machine learning algorithms on dataset, needs to compare each one's performance based on performance measures such as Accuracy score, Precision score, Recall score and F1-score score. This is known as model evaluation. There are multiple functions for model evaluation in scikit learn such as Accuracy, Precision, Recall and F1-score etc.

Step 6: Recommendation generation

The system considers polarity of Sentiment analysis and features of product for recommendation of product and gives the list of product recommendations.

IV. Results and Discussion

Dataset of product reviews is divided into training and test dataset but in different ratios such as 60:40, 70:30, 80:20 and 90:10. Result of all dataset ratio and algorithms are shown in following Table 3. All values are in percentage.

Dataset Ratio	**Param eters**	**Algorithms**				
		Multin-omial Naive Bayes	Logistic Regress ion	SVM	Deci-sion 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

A. Rusults of Algorithms

Results of implemented algorithms shown in figure. The percentage value of Accuracy, Recall, F1-score and Precision evaluated.

1) *Multinomial Naive Bayes:*

Following Figures shows results of Multinomial NB algorithm with respect to dataset splitting ratio and performance measures. It shows that Multinomial NB gives high accuracy for 70:30 ratio followed by 90:10.

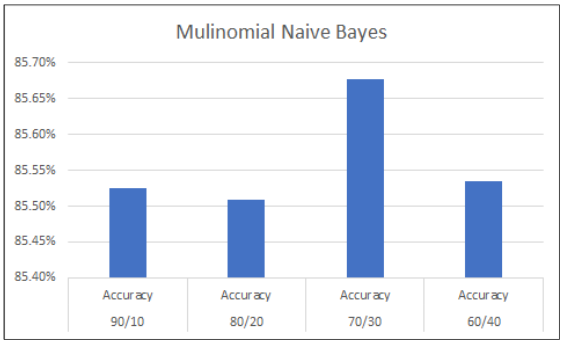

Figure 2. Accuracy of Multinomial NB

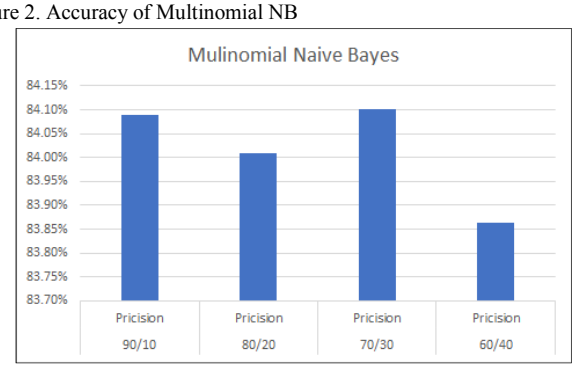

Figure 3. Precision of Multinomial NB

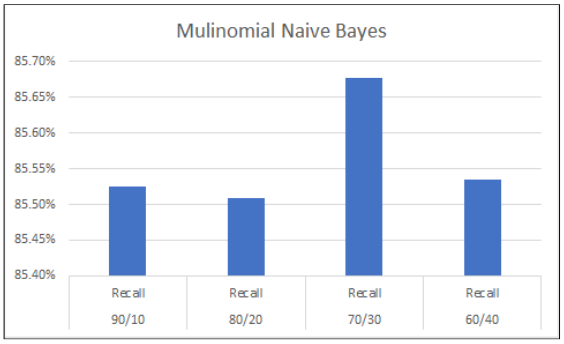

Figure 4. Recall of Multinomial NB

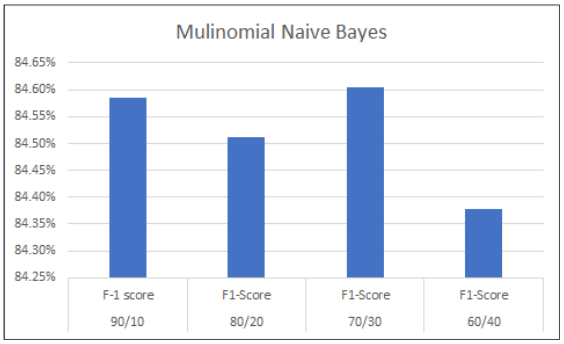

Figure 5. Recall of Multinomial NB

2) *Logistic Regression:*

Following Figures shows results of Logistic Regression algorithm with respect to dataset splitting ratio and performance measures. Like Multinomial NB it shows that, Logistic Regression gives high accuracy for 70:30 ratio followed by 90:10.

Figure 6. Accuracy of Logistic Regression

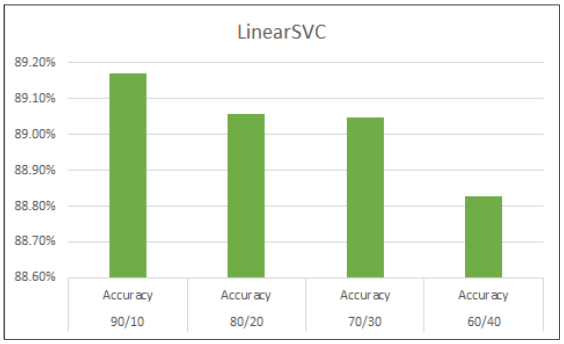

Figure 10. Accuracy of SVM

Figure 7. Precision of Logistic Regression

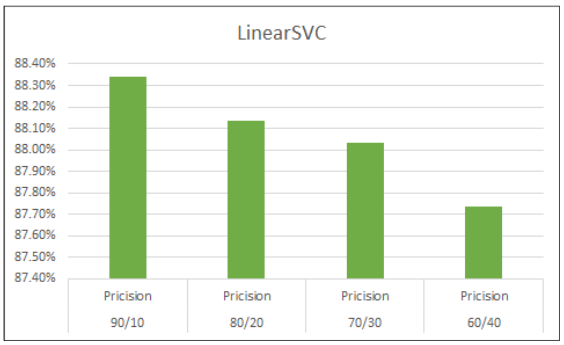

Figure 11. Precision of SVM

Figure 8. Recall of Logistic Regression

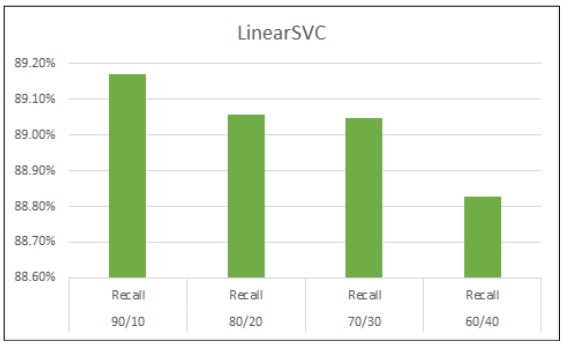

Figure 12. Recall of SVM

Figure9. F1-Score of Logistic Regression

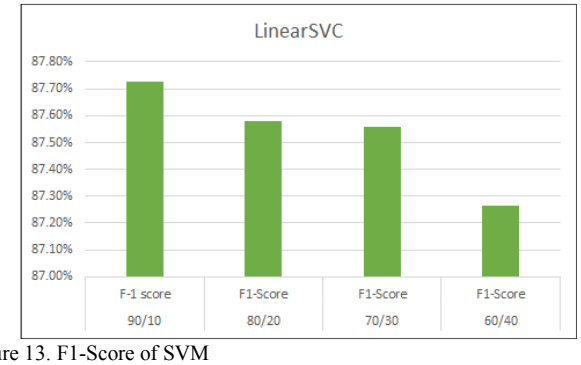

Figure 13. F1-Score of SVM

3) SVM:

Following Figures shows results of SVM algorithm with respect to dataset splitting ratio and performance measures. Unlike Multinomial NB and Logistic Regression, it shows that, SVM gives high accuracy for 90:10 ratio.

4) Decision Tree:

DT are commonly used machine learning algorithm. Following Figures shows results of Decision Tree algorithm with respect to dataset splitting ratio and performance

measures. It shows that Decision Tree gives high accuracy for 90:10 ratio followed by 80:20 ratio.

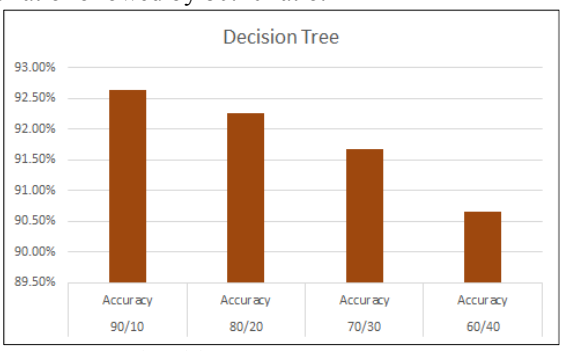

Figure 14. Accuracy of Decision Tree

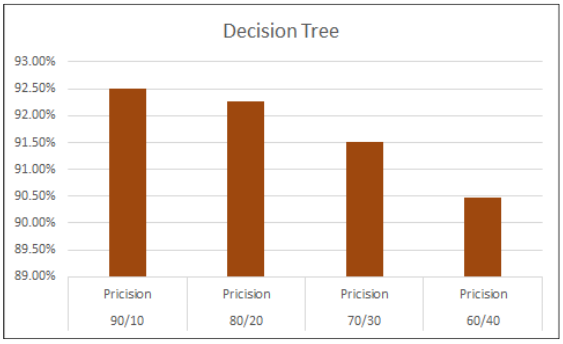

Figure 15. Precision of Decision Tree

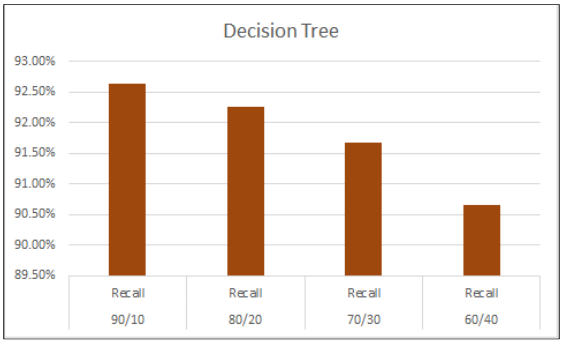

Figure 16. Recall of Decision Tree

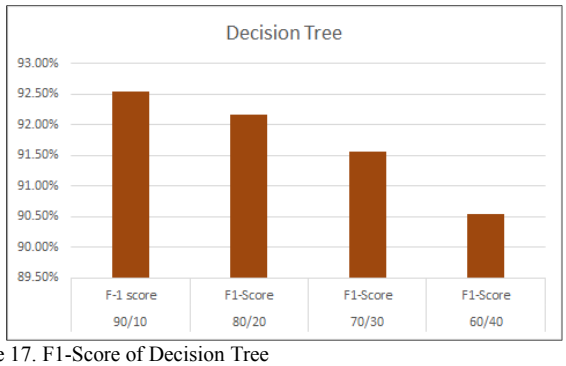

Figure 17. F1-Score of Decision Tree

5) Random Forest:

Following Figures shows results of Random forest algorithm with respect to dataset splitting ratio and performance measures. Like Decision Tree it shows that, Random forest gives high accuracy for 90:10 ratio followed by 80:20 ratio.

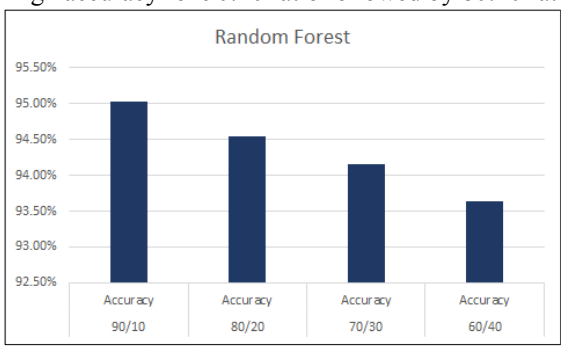

Figure 18. Accuracy of Random Forest

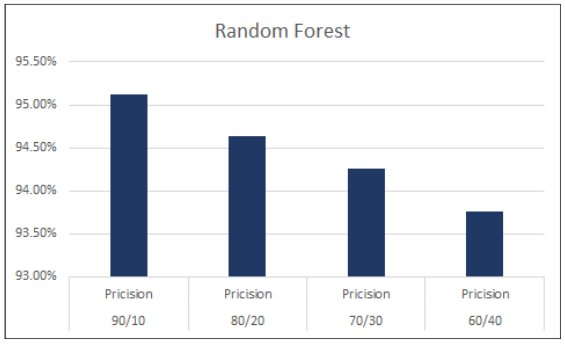

Figure 19. Precision of Random Forest

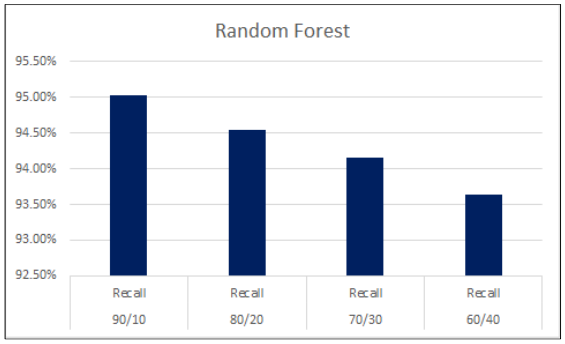

Figure 20. Recall of Random Forest

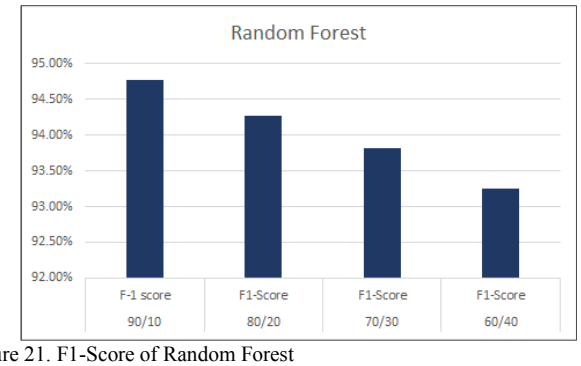

Figure 21. F1-Score of Random Forest

From the above Figures it is proved that data splitting in 90:10 ratio gives best results as compare to other dataset splitting ratios. Hence Figure of all algorithms with respect to

Accuracy, Precision, Recall and F-1 score for 90:10 ratio are as follows:

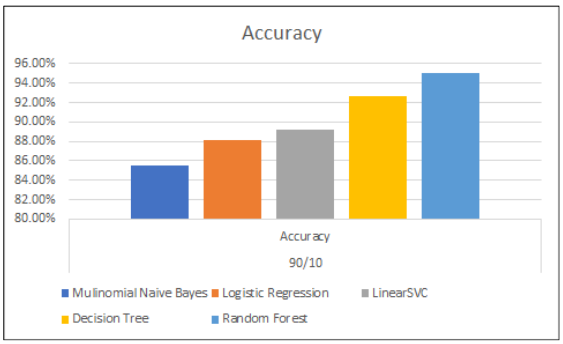

Figure 22. Accuracy for 90:10

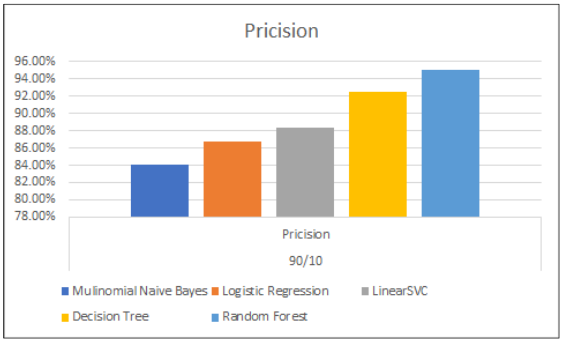

Figure 23. Precision for 90:10

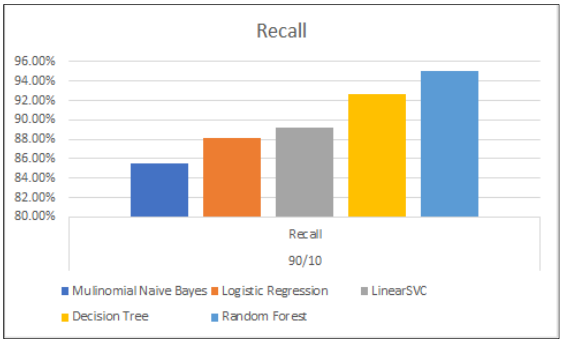

Figure 24. Recall for 90:10

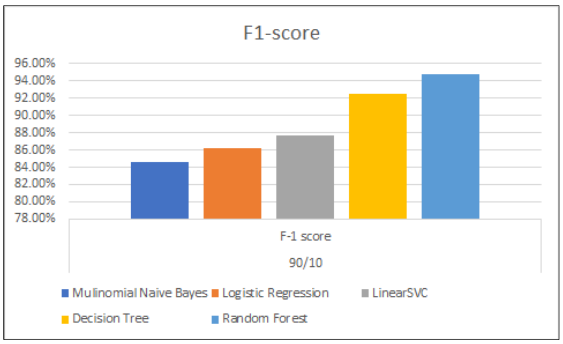

Figure 25. F1-score for 90:10

B. Overall preformance analysis of algorithms considered

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

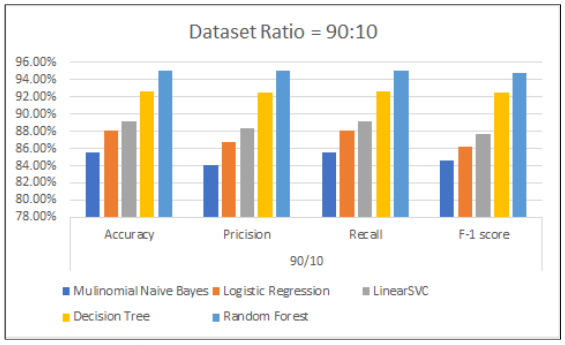

Figure 26. Dataset Ratio 90:10

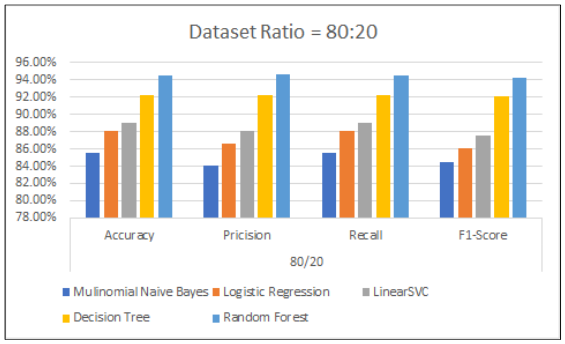

Figure 27. Dataset Ratio 80:20

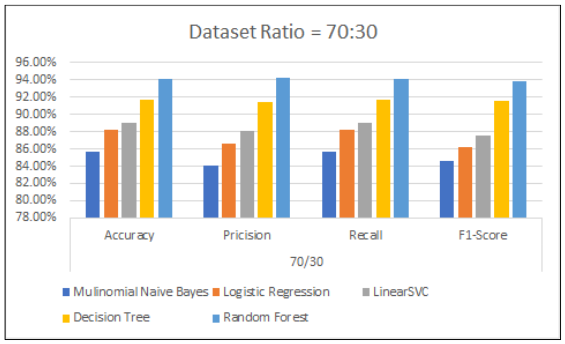

Figure 28. Dataset Ratio 70:30

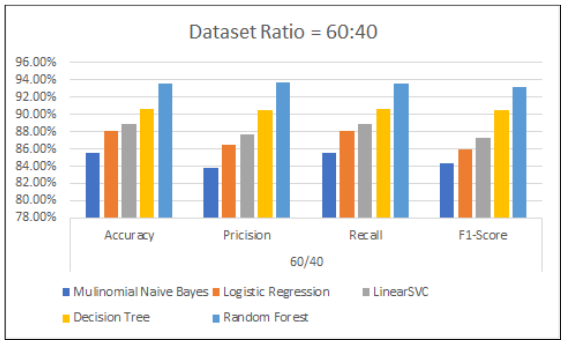

Figure 29. Dataset Ratio 60:40

C. *Variation in random Forest*

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 number of estimators is showen in Table 4. and its Graphical representation is given below:

Table 4. Variation in Random Forest

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

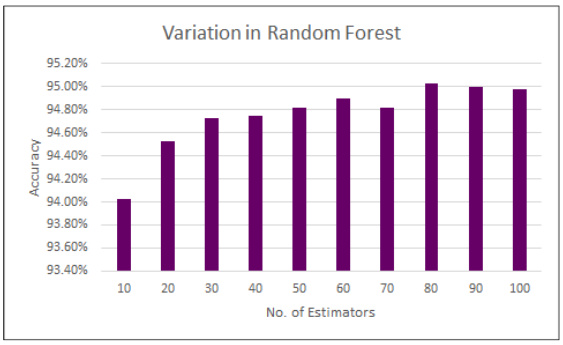

Figure 30. Variation in Random forest

From the Figure 30. 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

Sentiment analysis is important to know people's attitude behind text. Sentiment Analysis can be useful in product recommendations. Machine learning techniques such as Multinomial Naive Bayes classifier, Logistic regression, decision tree, SVM classifier and Random Forest are used to perform sentiment analysis. This paper mainly focused on performance of Random Forest algorithm. On the basis of experimental results, random forest performed well on the dataset. It provided very promising results with accuracy of 95.03%. Most of the previous research has focused on mainly SVM and Naive Bayes for the sentiment classification. By using classification techniques it's been concluded that Random forest can also provide good and competitive results and can provide better 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 while performing sentiment analysis and using sentiment polarity for recommendation.

VI. FUTURE SCOPE

In this area, mainly English language is considered but in future, additional languages can be explored for sentiment analysis. Here supervised machine learning approach has been used, but can try to get results from other machine learning algorithms or by using lexicon Based Approach or hybrid approach to improve results. Instead of using only binary classification for sentiment analysis can use Scale or can identify expressions for reviews. In case of recommendation system along with polarity of sentiments, can consider product related information like products features and user related information.

REFERENCES

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

[2] Abirami, A. M., and V. Gayathri. "A survey on sentiment analysis methods and approach." Advanced Computing (ICoAC), 2016 Eighth International Conference on. IEEE, 2017.

[3] Thakkar, Harsh, and Dhiren Patel. "Approaches for sentiment analysis on twitter: A state-of-art study." arXiv preprint arXiv:1512.01043 (2015).

[4] Gayatri Khanvilkar, Prof. Deepali Vora. "Sentiment Analysis for Product Recommendation Using Random Forest", International Journal of Engineering &Technology, 2018.

[5] 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.

[6] 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.

[7] 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.

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

[9] 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.

[10] 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.

[11] Dixit, Apurva, et al. "Emotion Detection Using Decision Tree." Development 4.2 (2017).

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

[13] 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).

[14] Cerňak, Miloš. "A comparison of decision tree classifiers for automatic diagnosis of speech recognition errors." Computing and Informatics 29.3 (2012): 489-501

[15] Rosa, Renata L., Demsteneso Z. Rodriguez, and Graca Bressan. "Music recommendation system based on user's sentiments extracted from social networks." IEEE Transactions on Consumer Electronics 61.3 (2015): 359-367.

[16] Zheng, Xiaoyao, et al. "A tourism destination recommender system using users' sentiment and temporal dynamics." Journal of Intelligent Information Systems (2018): 1-22.

[17] Amel Ziani, Nabiha Azizi, Didier Schwab, Monther Aldwairi, Nassira Chekkai, et al.. Recommender System Through Sentiment Analysis. 2nd International Conference on Automatic Control, Telecommunications and Signals, Dec 2017, Annaba, Algeria. <hal-01683511>

[18] Linden, Greg, Brent Smith, and Jeremy York. "Amazon. com recommendations: Item-to-item collaborative filtering." IEEE Internet computing 1 (2003): 76-80.

[19] Aggarwal, Charu C. "An introduction to recommender systems." Recommender Systems. Springer, Cham, 2016. 1-28.

[20] Hauger, Stefan, Karen HL Tso, and Lars Schmidt-Thieme. "Comparison of recommender system algorithms focusing on the new-item and user-bias problem." Data Analysis, Machine Learning and Applications. Springer, Berlin, Heidelberg, 2008. 525-532.

[21] Amazon Reviews: Unlocked Mobile Phones, https://www.kaggle.com/PromptCloudHQ/amazon-reviews-unlocked-mobile-phones, April 2018.

[22] Gayatri Khanvilkar, Deepali Vora(2018). "Activation Functions and Training Algorithms for Deep Neural Network", UGC approved journal, International Journal of Computer Engineering In Research trends, 5(4), 98-104