Sentiment Analysis of Social Media Network Using Random Forest Algorithm

P. Karthika
Post Graduate Student
Department of Computer Science and Engineering
Kalasalingam Academy of Research and Education
Krishnankoil-626126, Tamilnadu
karthikaagila@gmail.com

Dr. R. Murugeswari
Associate Professor
Department of Computer Science and Engineering
Kalasalingam Academy of Research and Education
Krishnankoil-626126, Tamilnadu
r.murugeswari@klu.ac.in

Mrs. R. Manoranjithem
Assistant Professor
Department of Computer Science and Engineering
Kalasalingam Academy of Research and Education
Krishnankoil-626126, Tamilnadu
V.Manoranjithem@klu.ac.in

***Abstract*—Sentiment Analysis is the identification of sentiments or opinions from the given text. Social media generates large amount of sentiment loaded information in the form of reviews. Sentiment analysis is used to identify the customer's opinions from user reviews. Online purchasing have became more fashionable due to its varieties, low cost and immediate supply. In today's competitive ecommerce market ratings and reviews of various brand is used to understand how consumers really feel about the product. The feedback environment is developed to help the customers to buy the correct product and to guide the companies to enhance the features of product depending on the consumer's demand. The customer feels difficult to accurately find the review for a particular feature of a product that they intend to buy. Also, there is a mixture of positive and negative reviews. To avoid this confusion and to make this review system more transparent and user friendly, feature based opinion extraction is carried out. In this paper the rating from the online shopping website known as flipkart.com is analyzed, based on the aspects of the product the rating is classified as positive, neutral and negative. The proposed work is analyzed by using Machine Learning algorithm called Random Forest and simulated by using SPYDER. In our system the accuracy, precision, F-measure and recall is calculated for both Random Forest and Support Vector Machine (SVM) algorithm and then accuracy comparison is made these two algorithms. In which the Random Forest gives the best accuracy of 97% than the Support Vector Machine.**

Keywords*—Machine Learning (ML), Support Vector Machine (SVM), Random Forest, Sentiment Analysis.

I. INTRODUCTION

Rapid growth in the internet brought us to the era of ecommerce. Ecommerce is nothing but selling products on online. Many customers share their own opinions about products on online [1]. The user opinion provides a new way in decision making process to make impact on business model. Online shopping is an easiest way to purchase the products from owner over the internet by using various kinds of websites and apps. The users also consider the review of other online shoppers while buying the product on online websites which helps the new online shoppers a lot. Thus, we need to study and analyze customer reviews. [2-3]

In general ratings are provided in star format. Each and every product has number of ratings. It is very difficult to read each rating in detail. Many researches shown pictorial representation is more effective and can be memorized, understood easily rather than textual representations. So if we are it will enhance reliability in decision making [4]. Sentiment analysis for product review is analyzed for various online websites such as flipkart.com, amazon.com as well as for rediff.com. [5- 6]

Sentiment analysis helps both the provider as well as the buyer, it helps the provider to introduce new products and the buyer to find original product by using the user review on online websites. The user review is categorized on three different basis called positive, negative or neutral [7-14]. In this paper, the user rating for products is extracted from the online shopping website called flipkart.com and then the extracted dataset were stored in CSV file. The data extracted is processed by using Random Forest and the accuracy is measured.

II. RELATED WORK

Ajitha et al. [1] described a technique known as semantic orientation, which automatically identifies frequently used phrases of a product from online website. Firstly the product factor had been diagnosed, and then sentiment analysis is carried out. Additionally, other strategies like stop phrases removal, context based totally mining and stemming becomes employed. It estimates the overall performance of the product after checking the product overall performance whether good or bad based on evaluations.

Sukhchandan Randhawa et al. [`2] focused on the computational study on Sentiment Analysis. It retrieves information from given material. In this system about 4, 000, 00 opinions for mobile phones was classified into superb and poor sentiments. Classification process was done by, Support Vector Machine (SVM), Decision Tree and Naïve Bayes had been hired for evaluations. The evaluation of fashions was accomplished using Fold Cross Validation.

Ahmad Hamy Hossny et al. [3] explained about the concept of sentiment analysis in social media networks. It provides various indictors in various domains known as industrial, medical and social. This survey presented sentiment analysis levels, enhancement methods, techniques, applications, lexicons, tools and existing research gaps.

Bhuvneshwar Kumar et al. [4] focused on the Natural Language Processing (NLP) troubles to differentiate the wonderful as well as the poor evaluations of the patron's for the objects on online market. Information utilized on this had been collected from Amazon.Com, Rediff.Com, Flipkart.Com. Here the item was isolated into high quality,

978-1-5386-9543-2/19/$31.00 ©2019 IEEE

bad and impartial classifications which show the sentiment of an item. In this system the product was divided into nice, negative and impartial categories which indicate the sentiment of a product. The analysis technique for the classification of product, which was done by way of characteristic extraction followed by the classification step. Many of the applications of Opinion Mining had been based totally on bag-of-phrases, which do not capture context that was crucial for Sentiment Analysis.

Janhavi et al. [5] projected on the basics of opinion mining. It consists of different techniques which include Extraction, Clustering and Classification. The flipkart product reviews were extracted by using product API. The author fetched the logo name, opinions, score and other associated matters for product, the clustering is done by the usage of ROCK and the usage of CART algorithm to categories critiques as nice and bad phrases from remarks and ultimately they arrive to recognize the product having more percentage of fine critiques. They have been categorized as fine and poor words from evaluations and from which the percentage of superb and negative phrases was calculated. The end result provides the assessment percentage allows the consumer to conclude primarily based on the high quality evaluates percent of the product.

Abhinash Singla et al. [6] proposed a hybrid set of rules by combining Decision Trees and Naive Bayes. The customer review comments about the product from flipkart.com was taken as records set after which the dataset had been labeled as subjectivity/objectivity and bad/high quality primarily based on the mind-set of buyer. This system included spelling correction in evaluate textual content, and then it classified the comments. The spelling correction was done to make the most realistic comment for understanding the polarity of words using Word Net dictionary. Then stemming was carried out to get rid of the stop words.

Neethu et al. [7] proposed Naïve Bayes and SVM algorithm which analyzes the twitter posts for digital products such as mobiles, laptops etc. Performing sentiment evaluation in particular area, it was viable to discover the impact of area information in opinion type. The characteristic vector was used for separating the tweets as high quality and bad. There are positive issues even when identifying emotional key-word from tweets. It became additionally tough to deal with misspellings and slang phrases.

Gurdeep Singh et al. [8] proposed Naïve Bayes algorithm to classify tweets into high quality, terrible neutral or negation. They focus on sentiment analysis by getting user opinion robotically from the twitter. Twitter circulation API was used to evaluate the Twitter records. Using this technique they constructed a sentiment classifier, it determines fantastic, bad and objective sentiments for a record. The author also presented experimental assessment of live review twitter dataset and class results.

Aswathy Wilson et al. [9] presented a study on challenges in sentiment analysis similar to their approaches and techniques. Here the author focused on three main classifiers called SVM, Naïve Bayes and Maximum Entropy. These sentiments become very useful for governments, business and as well as for individuals. There are of various challenges in evaluation process. These challenges become tough in evaluating accurate meaning of sentiments and identifying the suitable sentiment polarity. Text analysis techniques were used to evaluate and retrieve information from material.

Rastislav et al. [10] focused on the sentiment evaluation of posts on Facebook in Slovak language. This approach depend totally on device learning and multilevel textual pre-processing which address specifics of user generated social content. Here the method for sentiment evaluation primarily depend on complicated preprocessing to deal with language as well as content material style demanding situations of social networks. The most important troubles are excessive flections, complex morphology and syntax.

III. Sentiment Analysis

Sentiment analysis is contextual mining of textual content which identifies and extracts subjective data in supply material, and supporting an enterprise to recognize the social sentiment in their brand, services or products at the same time as monitoring on line conversations. Sentiment analysis is also called as subjective analysis, it classifies the text according to the polarity and orientation of the opinion expressed into positive, neutral and negative.

It helps facts analysts inside massive organizations gauge public opinion, conduct nuanced marketplace studies, display brand and product popularity, and recognize patron experiences. The process of sentiment analysis consists of Sentiment Identification, Feature Selection, Sentiment Polarity and Sentiment Classification.

IV. Random Forest Algorithm

Random Forest is a supervised learning algorithm, it is used in regression and classification problems. Random Forest is simply called as collection of trees and each tree is different from one another. It constructs multiple decision trees and finally merges them together to gain absolute and stable value, which is mainly used at the time of training and outputting the class. The algorithm is given in the Table 1.

Table 1: Algorithm for Random Forest

Random Forest Algorithm
Step 1: Load flipkart dataset and apply random forest algorithm.
Step 2: The required records were selected and the decision tree is created depending on the record.
Step 3: The decision making process is done based on the class value.
Step 4: If the class value is less than the threshold value then it is considered as false or else it is considered as true.
Step 5: The performance of random forest algorithm is compared with SVM algorithm based on the Performance Metrics such as accuracy, precision, F-measure and recall.

Working of principle of Random Forest

1. Random Forest algorithm selects N random records from the given dataset. Depending on the N records the tree is constructed.
2. The decision tree is constructed based on the N records
3. The number of trees was selected according to the available dataset
4. In the case of regression problem, each tree in the forest predicts the value for Y for a new record.
5. The average of all values were predicted by all the trees in the forest in order to calculate the final value
6. In the case of classification problem, every tree inside the forest predicts the category to which the new record belongs.
7. Finally, the new record was assigned to the category.

IV. Proposed System Architecture

Review analysis helps the online shoppers to gain information about the product they want to buy. Usually if the user want to buy products from online they will first check reviews for the particular product that we want to buy depending on the star provided we will decide whether to buy or not.

This paper shows the review analysis of flipkart products to help the online buyers to make the fine decision and also the accuracy level of the reviews were also determined using Random Forest algorithm. The detailed work of the system is shown in Fig1.It consists of various process known as Data Collection, Data Preprocessing, Feature Selection, Detection Process, and Sentiment Classification.

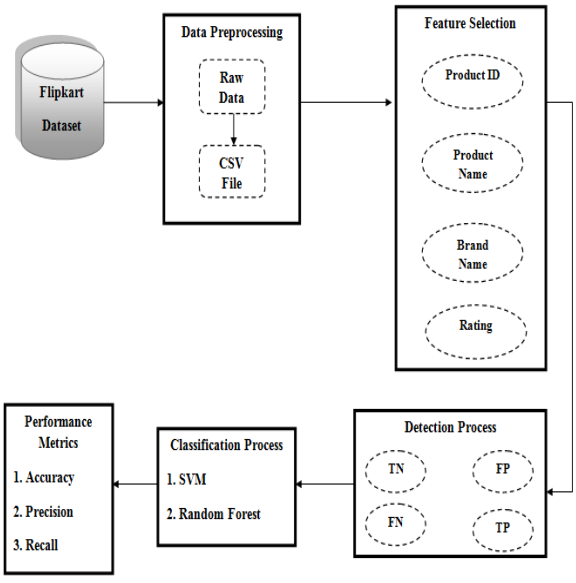

Fig. 1 System Architecture

A) Data Collection

The proposed system collects the flipkart data set from the website called kaggle.com. It consists of 20,000 records out of which 50% were taken as training data and 50% were taken as testing data, i.e. 10000 records were used for training process and 10000 dataset used for testing process and also the irrelevant dataset were removed from the samples as there is no usage of those data.

B) Data Preprocessing

In data preprocessing the raw data is converted into the required format. All the impurities were neglected from the dataset and then the unwanted attributes were also removed.

C) Feature Selection

After completing the preprocessing the feature selection is done here the feature needed for the analysis is selected is shown in the Table 2. It shows the description for the selected features in the dataset.

TABLE 2: SELECTED FEATURES FROM THE DATASET

Features	Description
Product ID	Describes the ID provided for the particular product
Product name	Describes the name provided for the particular product
Brand name	Describes the name of the manufacturing company
Rating	Customer star rating between 1 to 5

D) Detection Process

After completing the feature selection process the next process is to detect the sentiment by using the following attributes.

True Positive: This defines the correct Real Positive Reviews in the testing data

False Positive: It defines the incorrect Fake Positive Reviews in the testing data.

True Negative: This defines the correct Real Negative Reviews in the testing data.

False Negative: It defines the incorrect Fake Negative Reviews in the testing data.

F) Sentiment Classification

The final process is Sentiment classification, which is done using Random Forest algorithm.

The foremost source of facts used is the flipkart product reviews from the website called kaggle.com. Here first the raw data is cleaned and then stored as CSV. The CSV is loaded into SPYDER. The product reviews presented in the dataset is divided into five different classes. Here the class 0, 1 denotes the negative reviews and class 2, 3 denotes the neutral reviews and finally class 4, 5 denotes the positive reviews. This class separation makes the process very simple and it become very easy during the classification process. The flipkart dataset consists of reviews for various products from which some of the products is randomly selected as sample. For example Clothing, Toys and School Supplies, Tools and Hardware, Baby Care, Mobile and Accessories are selected. The dataset consists of the following attributes Product Name, Product ID, Retail Price, Discounted Price, Product Rating, Overall Rating, and Brand.

After performing all the above process the classification process is made by using the Machine Learning algorithm called Random Forest which detects the accuracy for the loaded data. Then the Receiver Characteristic Operator (ROC) curve and the confusion matrix is also evaluated, it is evaluated by using the classes of product reviews. Finally the Random Forest is compared with Support Vector Machine (SVM) to ensure Random Forest provides the best result.

VIII. Result And Discussion

The flipkart product reviews are analyzed by using the Random Forest algorithm. The proposed work is implemented by using SPYDER. In this module the following performance evaluations have been introduced. It includes Real Positive Reviews, Fake Positive Reviews, Real Negative Reviews and Fake Negative Reviews.

Performance Metrics

A) Accuracy

In general, performance of the classification algorithm is measured as accuracy. Accuracy means the ratio of number of accurately estimated examples to the total number of predicted examples. The user can occur high accuracy when they label all the examples as dominant class.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

B) Precision and Recall

Precision and Recall compares the overall performance of textual content mining. Precision is used to evaluate correctness and recall is used to evaluate the completeness. Precision means the proportion of number of examples correctly labeled as positive to the number of examples classified as positive labels. Recall means the proportion of number of examples correctly labeled as positive to the total number of examples labeled as positive.

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

C) F-measure

F-measure means the arithmetic mean of precision and recall. F-measure is used as assessment metric to analyze the views of sentiment classification.

$$F - measure = 2 * \frac{Precision * Recall}{Precision + Recall}$$

The result of accuracy comparison of Support Vector Machine (SVM) and Random Forest is shown in Fig 8

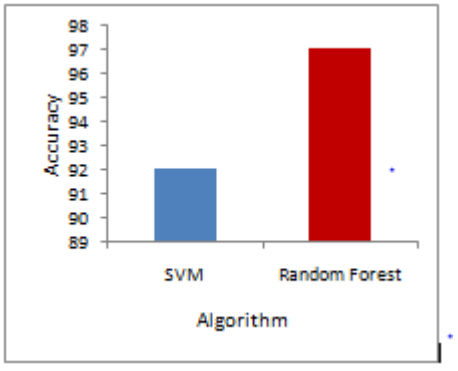

Fig. 2 Accuracy Comparison

In Fig 2 the graphical representation shows the accuracy comparison is made between Random Forest and Support Vector Machine (SVM).

Random Forest gives the best accuracy among other algorithms with 97% and the Support Vector Machine (SVM) gives the accuracy of 92%. I n Random Forest algorithm the implicit feature selection is made during the analysis of dataset. The Random Forest algorithm also includes the unbalanced data into the process and it limits over fitting without increasing the error rate. This leads Random Forest to achieve the greater accuracy. Fig 3 represents the comparison of precision, F-measure and recall for Random Forest and Support Vector Machine (SVM).

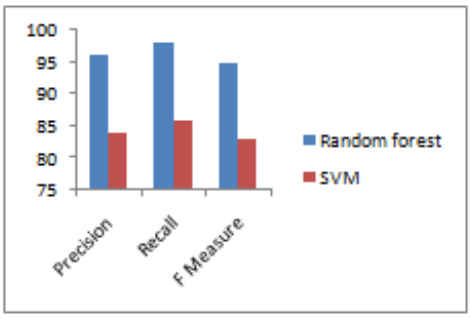

Fig. 3 Comparison of Precision, Recall and F-measure between Random Forest and SVM

The accuracy for Clothing, Toys and School Supplies, Tools and Hardware, Baby Care, Mobile and Accessories were determined using Support Vector Machine (SVM) and Random Forest. The dataset consists of various products but here we selected only 7 products randomly as samples. In which the random forest shows the greater accuracy than the SVM. In Fig 4 the accuracy of product rating using SVM and Random Forest is shown.

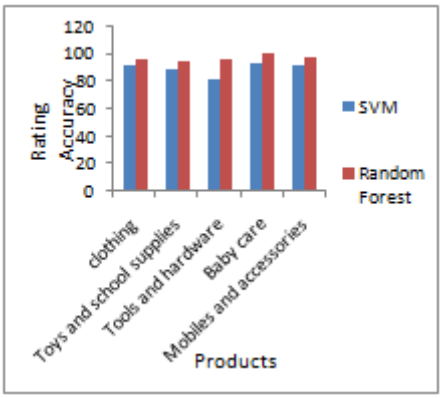

Fig. 4 Accuracy of Product Rating using Random Forest and SVM

D) Receiver Operating Characteristic (ROC) curve to evaluate classifier output

ROC curves are normally utilized in binary class to study the output of a classifier. In order to increase ROC curve and ROC area to multi-magnificence or multi-label category, it's miles important to binarize the output. One ROC curve can be drawn consistent with label, however one also can draw a ROC curve by using indicator matrix as a binary prediction also called as micro-averaging. The probabilities are computed as the predicted classifiers in the ensemble.

ROC curve represents true positive rate at the Y axis and false positive rate at the X axis. This defines the top left corner of the plot is the “ideal” point that represents false positive rate as zero and a true positive rate as one. This module indicates the ROC reaction for flipkart product dataset. The ROC curve is created for class 1, class 2, class 3, class 4 and class 5.

The “steepness” of ROC curve is used to increase the true positive rate while reducing the false positive rate. The Fig 5 shows the ROC curve for multiclass

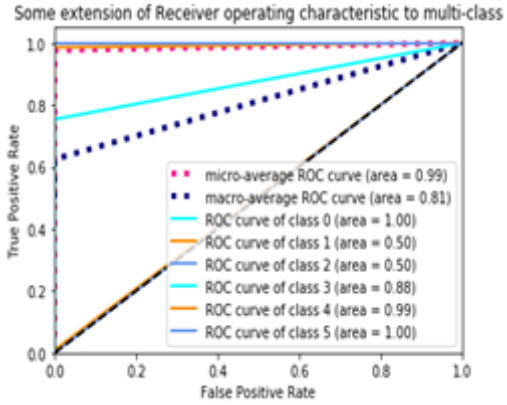

Fig. 5 ROC Curve for Multiclass

This graph represents the ROC response of different datasets, from K-fold cross-validation. This curve determines the mean area and variance when the training set is split into different subsets. This represents how the classifier output is affected when there is a change in the training data.

F) Confusion Matrix

A confusion matrix consists of detailed information about the predicted and actual values. Performance of the classification values is evaluated using the data in the matrix. In confusion matrix the X axis consists of predicted labels and the Y axis consists of true labels.

Here the matrix is performed for each row and column using the predicted value and true value. If there is no value in the diagonal then it is states that true label and the predicted label consists of same value. If there is some value in the diagonal, it determines there are some differences between true labels and predicted labels.

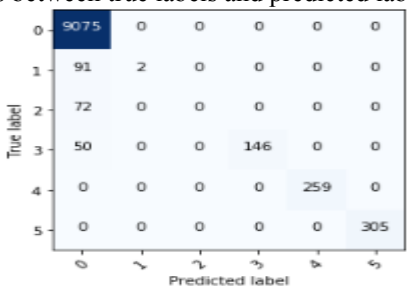

Fig. 5 Confusion Matrix for Random Forest

In Fig 5 the confusion matrix for Random Forest is represented.

IX. Conclusion

Masses of customer share their feedback (or) reviews on social media, this helps the provider as well as the customer to track the sentiment about the product. Using customer feedback the provider can enrich their brand and also it helps the new user to gain more information about the product. In this paper the classification and analysis were done for flipkart product reviews. This work classifies the star rating provided by the customer for the product, here the classification process is done using Random Forest algorithm. The accuracy comparison is made for the product between the Random Forest and Support Vector Machine algorithm. In which the random forest shows the best accuracy and also the Receiver Operating Characteristic (ROC) curve is measured for multiclass.

References

[1] P.Ajitha and R.Jenitha Sri "Survey of Product Reviews Using Sentiment Analyis, Indian Journal of Science and Technology, vol 9(21), pg no 1-4, 2016.

[2] Sukhchandan Randhawa, Sushma Jain and Zeenia Singla, "Sentiment Analysis of Customer Product Reviews Using Machine Learning", International Conference on Intelligent Computing and Control (12C2),volume 3, pg no 1-5, 2017.

[3] Ahmad Hamy Hossny, Khaled Ahmed and Neamat EI Tazi," Sentiment Analysis over Social Networks: An Overview, IEEE International Conference on Systems, Man and Cybernetics, pg no 2170-2179, 2015.

[4] Bhuvneshwar Kumar, Krutika Wase, Pranali Rushabh Bandewar and Nadim Badole, "Sentiment Analysis of Product Review", International Journal of Innovations in Engineering and Science, vol 3, pg no 8-13, 2018.

[5] Janhavi N L, Jharna Majumdar and Santhosh Kumar, "Sentiment Analysis of Customer Reviews on Laptop Products for Flipkart", International Research Journal of Engineering and Technology (IRJET) ,volume 05,Issue 03, pg no 629-634,2018.

[6] Abhinash Singla and Gurneet Kaur," Sentimental Analysis of Flipkart reviews using Naïve Bayes and Decision Tree algorithm", International Journal of Advanced Research in Computer Engineering and Technology (IJARCET), Volume 5 Issue 1, pg no 148-153, 2016.

[7] Neethu M S and Rajasree R, "Sentiment Analysis in Twitter using Machine Learning Techniques", 4th ICCCNT, pg no 4-6, 2013.

[8] Gurdeep Singh, T.Praveen, Rohit Kumar Sharma, and Ram Shankar, "Sentiment analysis in social media using Machine Learning technique with R Language", Journal of Chemical and Phamaceutical Sciences (JCPS), ISSN-0974-2115, Issue:06, pg no 124-128, Mar-2017.

[9] Aswathy Wilson, Mejo Antony, Minara P Anto, Nivya Johny and Vinay James, "Product Rating Using Sentiment Analysis", International Conference on Electrical, Electronics and Optimization Technique (ICEEOT), IEEE-978-1-4673-9939-5, pg no 3458-3462, 2016.

[10] Rastislav, Krchnavy and Marian Simko, Sentiment Analysis of Social Network Posts in Slovak Language,978-1-5386-0756-5/17,IEEE International Conference on Semantic and Social media Adaption, pg no 20-25, 2017.

[11] Aashutosh Bhatt, Ankit Patel, "Amazon Review Classification and Sentiment Analysis", International Journal of Computer Science and Information Technologies (IJCSIT), Volume 6(6), pg no 5107-5110, 2015.

[12] Narsimha Reddy and U Ravi Babu, "Sentiment Analysis of Reviews for E-Shopping Websites", International Journal of Engineering and Computer Science, ISS: 23197242, volume 6, Issue 1, 2017.

[13] R. Chandeasekaran and G. Vinodhini, "Sentiment Analysis and Opinion Mining: A Survey," International Journal, vol 2, pg no 1-6, 2012.

[14] G.Angulakshmi, R.ManickaChezian, "An Analysis on Opinion Mining: Techniques and Tools", International Journal of Advanced Research in Computer and Communication Engineering, Vol 3, Issue 7, 2014.