New Ensemble Approach to Analyze User Sentiments from Social Media Twitter Data

E. Prabhakar* & K. Sugashini**

*Assistant Professor, Department of Computer Science & Engineering, Nandha College of Technology, Erode, Tamilnadu, INDIA.
E-Mail: prabhakarit10{at}gmail{dot}com
**Freelancer, Tiruppur, Tamilnadu, INDIA. E-Mail: sugashinikalimuthu26{at}gmail{dot}com

***Abstract*—**The Sentiment analysis helps to identify and classify the opinions or the sentiments which are expressed by means of source text. Social media have gained more attention nowadays. Public and private opinions about varied subjects are expressed and spread continually via numerous social media. Twitter is one of the social media that is gaining popularity. Thus the twitter is widely used for analyzing the sentiment of the huge groups. It is difficult to perform sentiment analysis in twitter in comparison with general sentiment analysis since the twitter contains many special characters, symbols, slang words and also has the limitation of 140 words. Knowledge base approach and Machine Learning approach are the two paths for analyzing the text. This paper proposes a twitter sentiment analysis that can spot the interest and opinion of people in regard to the products such as e-Commerce sites, network operators and social media. Feature for classification of the tweets as negative, neutral and positive that depends on views of people and the identification of accuracy by various algorithmic approaches. It also involves the emotional classification of tweets into joy, fear, sadness, disgust, anger, surprise and unknown. The word cloud has been identified for the usage of most repeated word. This paper proposes new ensemble approach for sentiment analysis. Then the algorithms such as Naïve Bayes, Random Forest and kNN have been employed in Rapidminer to find the accuracy of the sentiment. Results indicate that the proposed approach gives good accuracy than traditional algorithms.

***Keywords*—**Big Data; Ensemble Learning; Sentiment Analysis; Social Media Data.

***Abbreviations*—**False Positive (FP); False Negative (FN); True Positive (TP); True Negative (TN).

I. INTRODUCTION

THIS work with twitter obtains hopeful approximation of the public sentiment which is opposed to the conventional internet articles and the web blogs. The reason behind is number of relevant data records are as much large for the twitter and compares to the various traditional based blogging sites. In addition to that, response on the twitter is as more prompt in nature and could be general since amount of the users who are tweet is substantially higher than the one writing web blogs on the daily basis. Sentiment analysis of the public is as highly critical in the macro-scale socioeconomic methods which are same as predicting market rate for particular commodity. This could be carried out by analysing the overall sentiment of public which are towards the product based on time and tools for identifying the public sentiment. This could be find how the product has respond in market by means of classifying the tweets as positive, negative and neutral. By this information, the efficiency of product could get enhanced. Predicting the output of the political elections and polls which are popular is also an rising application for sentiment analysis.

The Sentiment Analysis in twitter is performed on the various products like e-Commerce sites, operators and social media. This involves the various modules likely data collection, data preparation such as stop words removal, emotional classification and ensemble approach. The data collection from twitter is the first phase in which the R programming and Rapidminer is used to fetch data from twitter.

Then the collected data are classified into positive, negative and neutral sentiments and stored as CSV file. Then the CSV file is provided as input in tableau for the visualization of classified sentiments. The polarity score has been calculated for each sentiment in the R Programming and the comparative visualization is provided between e-commerce sites, operators and social media. Sentiment classification is a way to analyse the subjective information in the text and then mine the opinion. Sentiment analysis is the procedure by which information is extracted from the

opinions, appraisals and emotions of people in regards to entities, events and their attributes.

In R programming the collected data is processed such as stop word removal and punctuation removal which would be the data preparation and five thousand records are involved in this process. Then the emotions of the people are classified such as joy, sad and neutral based on the data.

The frequently used words in the tweets have been obtained by processing the data which would provide the insight of the data available. Then the ensemble approach is applied in the Rapidminer to the processed data in which various algorithms are used and comparison is performed between the algorithms such as which would provide the better result.

Tracking of user interest depending on tweets are intent to providing the better understanding and the people view on particular commodity by means of displaying the user sentiment which are classified at high accuracy rate. This is helpful for purchase a better product for people and additionally product developers would also enhance their feature of product.

The entire overview of social media sentiment analysis has provided in this particular section. The proposed work would provide in the section II. The methodologies are represented in section III. Finally the implementation has provided in the section IV. Thus conclusion and the future work of this review had provided in specific section V.

II. LITERATURE REVIEW

Different methods are proposed for sentiment analysis, finding sentiment in words, sentences and sentiment in topics. The Analysis of User Interest Based on Tweets has analysed the user views and classifying the tweets as positive, negative and neutral based on user interest and location interest. They have used NLP algorithm for classification and analysis [Nimita Mangal et al., 1]. Today the Analysing of sentiments on entertainment areas such as Movies are encouraging one. Twitter Sentiment Analysis for Movie Reviews by Machine Learning approaches would analyses movie reviews and then tweets are grouped as positive, neutral and negative with the help of Naïve Bayes classifier approach [Amolik et al., 2].

Analysing user emotions were carried out by using various techniques.Using Emotions to predict User Interest Areas in Online Social Networks that predicts the user emotions by using predefined hash tags. The Regression Algorithm has been used and positive, negative and neutral tweets were identified [Rishab Soni & James Mathai, 3].The sentiment analysis in social media is a trending one. The Sentiment Analysis on YouTube: A Brief Survey that predicts the sentiment of the user based on the sharing of comments and video by them. The events are categorised as low, medium, high, noise. The Precision and Recall is used for accuracy of the sentiment. The techniques such as User sentiment detection, Event classification and predicting YouTube comments are also employed for the command polarity [Dr. Muhammad Zubair Asghar et al., 4].

The user comments on products have been analysed. Sentiment Analysis for Social media and involves the unique process of languages like Tamil and Sinhala. The emotion symbols are also used for processing and classification of the sentiment. The user interest varied on age, profession are involved to predict the accuracy. The machine learning algorithm NLP is used for the entire processing. The reviews play a vital role in analysing a product [Sachira Chinthana Jayasanka & Saminda Premaratne, 5]. Consumer's sentiment Analysis of Popular phone brands and Operating system Preference depicts analyse of reviews which would be helpful for e-Commerce organizations. The Bag of words techniques is employed for better understanding. The supervised, unsupervised and semi-supervised machine learning techniques were used. The Naive Bayes is used for star ratings [Ashna Bali et al., 6].

Towards Sentiment analysis for Mobile devices involves the analysis on Android OS and the mobile properties such as battery, memory space, working speed are taken into consideration. The algorithm employed are NRC Hash tag and Opinion Lexicon for the opinion mining of the user. Thus it helps to improve the efficiency of the particular mobile products in terms of market rate [Johnnatan Messias et al., 7]. Sentiment Analysis on Social media- Facebook involves the analysis on newscast, Italian public broadcasting service and La7. The NLP algorithm is applied for analysis. This provides the result such that the classification of the positive, negative and neutral scores in the form of percentage. The analysis is also performed based on the feedback mechanism. The various mobile products such as IPhone, Samsung, HTC, Motorola and Google Nexus are comparatively analysed [Federico Neri Carlo Aliprandi et al., 8].

III. METHODOLOGY

Analysing the interest of twitter user in terms of tweets which are on various products. These products may be from e-Commerce sites involving Flipkart, Amazon, and Snapdeal, and for Mobile Operating System namely Bada, Android, IOS and then Social Medias involving YouTube, Whatsapp and Tumblr. New ensemble approach is proposed in this paper to analyze user sentiments.

Technique/Algorithm Used:

k-NN, Random forest, Naive Bayes, New Ensemble

3.1. Data Collection

The dataset used in this review has been collected from twitter by means of R tool and the RapidMiner specifically for e-Commerce sites, the Mobile Operating systems and the Social Medias. Then details are gathered for top 3 e-Commerce sites namely Flipkart, Amazon and Snapdeal, and then for Mobile Operating System including Bada, Android, IOS and for Social Medias involves YouTube, Whatsapp and Tumblr. There are 15 columns like tweet_id, text, tweet_location, etc. And 5000 records of tweets by different people for every product. The tweets are stored as a CSV files.

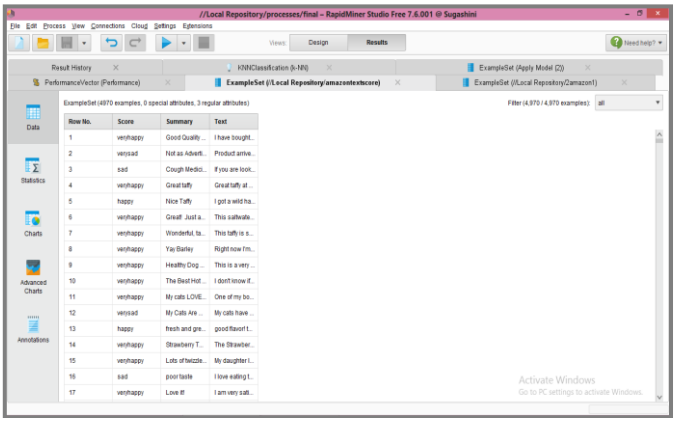

Figure 1: Data Collection using Rapidminer

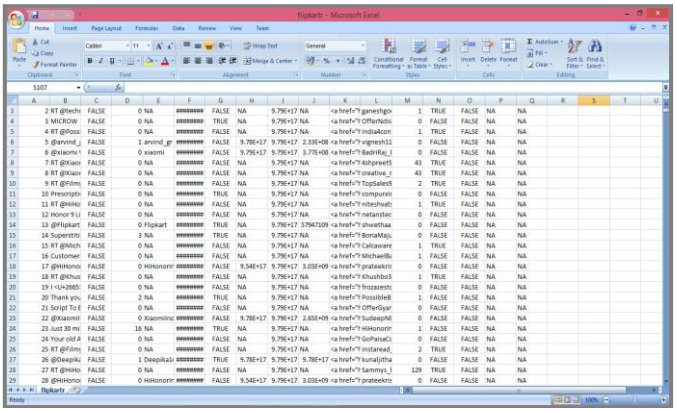

Figure 2: Data Collection using R

3.2. *Data Preparation*

In this process of data preparation re-tweets, stop words, links, URLs, mentions, punctuation, unnecessary spaces, null values, conjunctions, pronouns, articles and accentuation were removed so that data set could be standardized. This makes the process easier and the time taken for process will also be reduced.

3.3. *Emotional Classification*

In this process the Bayes algorithm has been used for emotional classification such as joy, sadness, disgust, fear and unknown. These emotions help the user to estimate the data in the better way. The classification of emotions has been provided in the bar graph for visualization.

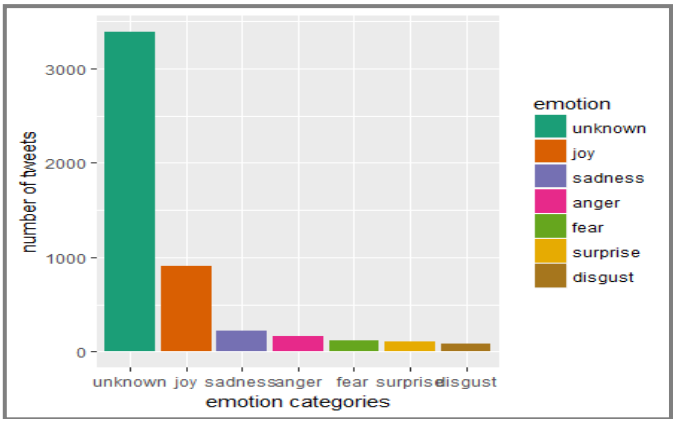

Figure 3: Emotional Classification Graph

The next process in this classification involves the polarity calculation in which the tweets as classified into positive, negative and neutral.

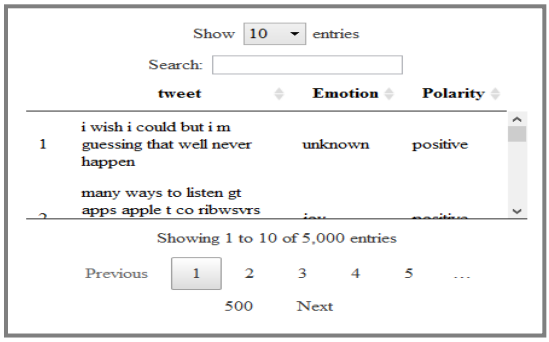

Figure 4: Table for Polarity Classification

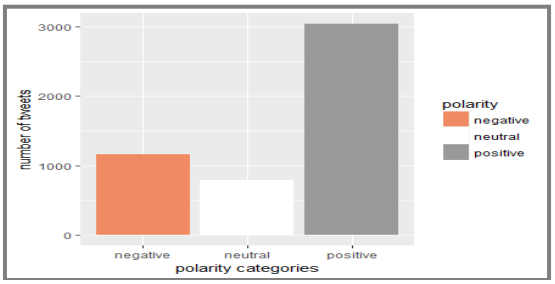

Figure 5: Graph for Polarity Calculation

The words obtained in the process are put together in the word cloud to identify the word which is used most of the times.

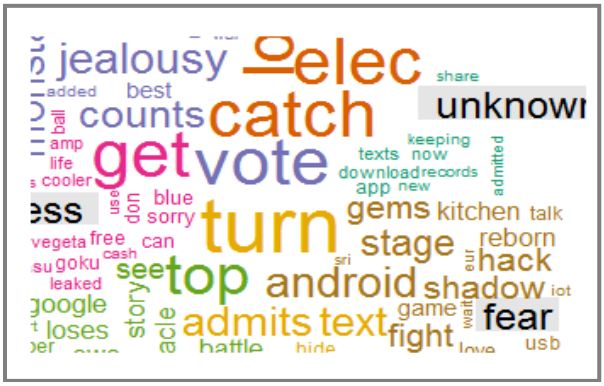

Figure 6: Word Cloud

IV. PERFORMANCE ANALYSIS

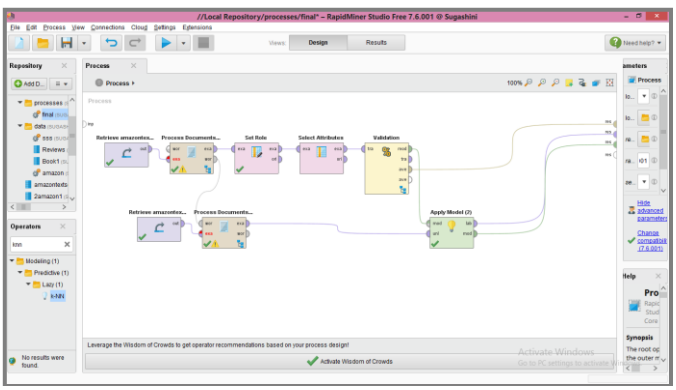

Figure 7: Algorithm Implementation

4.1. k-NN Classification

The k-Nearest Neighbor algorithm is based on comparing an unknown Example with the *k* training Examples which are the nearest neighbors of the unknown Example. The first step of the application of the k-Nearest Neighbor algorithm on a new Example is to find the k closest training Examples. "Closeness" is defined in terms of a distance in the n-dimensional space, defined by the n Attributes in the training Example Set.

4.2. Random Forest

Random forest is similar as bootstrapping algorithm with Decision tree namely CART model. In this review, there are 1000 observation in complete population having 10 variables. Random forest would tries to build the multiple CART model having various sample and initial variables. For examples, it would take randomly 100 observation and 5 initial variables as sample that builds a particular CART model. The process would take place by 10 times repeatedly and final prediction has been made on every observation. Final prediction is the function of each and every prediction. This final prediction could also simply be said as mean of every prediction.

4.3. Naive Bayes

Naive Bayes is one of the probabilistic techniques in constructing the classifiers. The characteristic are assumed for naive Bayes classifier has been considered where value of a specific feature is found as independent of value of any of other feature, namely given class variable. Inspite of oversimplified assumptions are mentioned previously and naive Bayes classifiers are found as having hopeful results in the complex real-world scenarios. An advantage of the naive Bayes is only a requirement of little amount of those training data which estimates the necessary parameters for the classification and finally classifier could trained by incrementally.

4.4. New Ensemble

Ensemble classifier uses a combination of models to increase the accuracy of the classification. It combines a series of learning models with the aim of creating new improved model. These ensemble classifiers are based on the weighting scheme. In this scheme weights are assigned to each example in the data set. The mean and median for the record values are calculated by using formulas and analyzed for better efficiency, similarly the percentage of the records are calculated from the total available records. After this process ensemble algorithms are applied.

Mean=Sum of number of records in each class/Total number of classes
Median(if length is even)=a[a.length/2]+a[a.length/2-1])/2;
Median(if length is odd)=a[a.length/2]; (Thus a = array of records in the classes)
Percentage of records = (Records in every class/Total number of records) *100
Percentage of accuracy = (Accuracy in each class/Total number of accuracy) *100

4.5. Performance Metrics

4.5.1. True Positive Rate

Proportion of the positive instances which are classified correctly in the majority class and it is said to be as sensitivity.

TPrate = TP/(TP+FN)

4.5.2. True Negative Rate

Proportion of the negative instances are made to classify correctly in available minority class and could be known as specificity.

TNrate = TN/(TN+FP)

4.5.3. False Positive Rate

Proportion of the negative instances are to misclassified in minority class.

FP rate = FP/(FP+TN)

4.5.4. False Negative Rate

Proportion of availing positive instances could misclassified in the majority class.

FN rate =FN/(TP+FN)

4.5.5. Overall Accuracy

Overall accuracy can be calculated using every four of performance metrics are mentioned above.

Accuracy = (TP+TN)/(TP+TN+FP+FN)

4.6. Comparison of Algorithms

The comparison among three algorithms such as Naive Bayes, KNN and Random forest has been carried out in the aim of finding accuracy of sentiment based on specific tweets.

Table 1: Accuracy

Algorithm	Accuracy
k-NN	63.55%
Random forest	63.31%
Naive Bayes	53.14%
New Ensemble	78.67%

V. CONCLUSION

This paper has intended on analysing the sentiment of customer tweets made by them on various products in e-Commerce sites such as Flipkart, Amazon, and Snapdeal, and Mobile Operating System such as Bada, Android, IOS and Social Medias such as YouTube, Whatsapp and Tumbler. Thus by comparing the performance of different popular sentiment classification approaches and developing an ensemble approach, which further improves the sentiment classification performance. The previous work compares several different traditional classification methods and selects the most accurate individual classification method to implement sentiment classification. The graphical representation of sentiment by means of tools used for business intelligence which makes the evaluate process as

much easier and obtains better performance. Result obtained shows that the proposed new ensemble approach provides better accuracy than existing algorithms.

REFERENCES

[1] Nimita Mangal, Rajdeep Niyogi & Alfredo Milani (2016), "Analysis of User's Interest based on Tweets", *Springer International Publishing*, Switzerland.

[2] A. Amolik, Niketan Jivane, Mahavir Bhandari, Dr. M. Venkatesan (2016), "Twitter Sentiment Analysis of Movie Reviews using Machine Learning Techniques", *International Journal of Engineering and Technology (IJET)*, Vol. 7, Dec 2015-Jan 2016.

[3] Rishab Soni & K. James Mathai (2015), "Improved Twitter Sentiment Prediction through Cluster-then-Predict Model" International Journal of Computer Science and Network, Vol. 4, No. 4, Pp. 559–563.

[4] Dr. Muhammad Zubair Asghar, Shakeel Ahmad & Afsana Marwat (2015), "Sentiment Analysis on YouTube: A Brief Survey", *MAGNT Research Report*, Vol. 3.

[5] Sachira Chinthana Jayasanka & Saminda Premaratne (2013), "Sentiment Analysis for Social Media", *Springer International Publishing*, Switzerland.

[6] Ashna Bali, Payal Agarwal, Gayatri Poddar, Devyani Harsole & Nilufar M. Zaman (2016), "Consumer's Sentiment Analysis of Popular Phone Brands and Operating System Preference", *International Journal of Engineering And Technology (IJET)*, Vol. 155.

[7] Johnnatan Messias, Joao P. Diniz, Elias Soares, Miller Ferreira, Matheus Araujo, Lucas Bastos, Manoel Miranda & Fabricio Benevenuto (2016), "Towards Sentiment Analysis for Mobile Devices", *IEEE/ACM*.

[8] Federico Neri Carlo Aliprandi, Federico Capeci & Montserrat Cuadros Tomas, "Sentiment Analysis on Social Media", *IEEE/ACM*.