DOI: http://dx.doi.org/10.26483/ijarcs.v10i3.6417

ISSN No. 0976-5697

Volume 10, No. 3, May-June 2019

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

SENTIMENT ANALYSIS OF MOVIES REVIEWS USING IMPROVISED RANDOM FOREST WITH FEATURE SELECTION

Er.Manpreet Kaur
M.Tech Scholar, Department of Computer Science & Engineering,
Amritsar College of Engineering & Technology,
Amritsar, Punjab, India

Er.Ajay Sharma
Associate Professor, Department of Computer Science & Engineering,
Amritsar College of Engineering & Technology,
Amritsar, Punjab, India

Abstract: Human psychology has perpetually influenced by means of others suggestion and experiences. Our experiences about whatever are very a lot influenced by means of different experiences, and at any time when we have to make a choice or answer, we regularly search out different reports, so folks are excited to understand other's experiences for his or her profit because of this computerized sentiment analysis systems are required. Sentimental evaluation, also known as opinion mining, is a ordinary language processing system used to extract the feeling or perspective of common lots regarding a given subject or product.This paper proposes a prediction model for the sentiment analysis of movies review using improvised random forest classification algorithm. The proposed work consists of hybrid approach of feature selection with classification for prediction analysis in text mining. Gini Index feature selection technique is used for optimizing the features. The experimental results of the proposed model are evaluated with respect to existing techniques and showed that the proposed technique achieved 90.69% accuracy and precision, recall class parameters of propose technique are also better than the existing techniques.

Keywords: sentiment analysis, movies review, opinion mining, random forest, gini index.

I. INTRODUCTION

Social Network has developed and increased in the previous couple of years. It is a kind of mode of communication that can be utilized by any individual lives in any area of the world. With such all-inclusiveness and high-speed information, sentimental analysis on such systems has been most focused on research subjects in NLP in the previous decade. The fundamental point of sentimental analysis is to identify the extremity of the content. The web has drastically changed the manner in which individuals express their perspectives and sentiments [1]. Big information is trending analysis space in engineering science and sentiment analysis is one amongst the foremost necessary a part of this analysis space. Big information is taken into account as terribly great amount of information which might be found simply on internet, Social media, remote sensing information and medical records etc. in form of structured, semi-structured or unstructured data and we can use these data for sentiment analysis. Sentimental Analysis is all on the brink of get the $64000 voice of individuals towards specific product, services, organization, movies, news, events, issues and their attributes. Sentiment Analysis includes branches of engineering science like linguistic communication process, Machine Learning, Text Mining and Information Theory and Coding. By using approaches, methods, techniques and models of defined branches, we can categorize our data which is unstructured data may be in form of news articles, blogs, tweets, movie reviews, product reviews etc. into positive, negative or neutral sentiment according to the sentiment is expressed in them

1.1 Sentimental Analysis

Sentimental Analysis (SA) is a data mining technique. SA is a procedure of examining of emotional tone beyond a processing of words, that is utilized forgrabbing an appreciation of a perspectives, suppositions and sentiments conveyed inside an online determine to get an audit of the broader general conclusion beyond particular points Sentiment analysis is tremendously helpful in observing of social media. The utilizations of Sentimental Analysis are wide and strong. The ability to isolate social information from bits of knowledge is a training that is vast for the most part grasped by relationship over the world.

1.2 Working of Sentimental Analysis:

Sentimental Analysis (SA) is frequently determined by scoring the words, algorithms utilized alongside voice inflections which can show a man's basic feelings about the theme of a talk. Sentimental Analysis takes into consideration a more target translation of components that are generally hard to find or regularly estimated emotionally, such as,

- The measure of pressure or disappointment in a client's voice
- How quickly the individual is talking
- Changes in the level of pressure demonstrated by the individual's speech.

In client service call centre applications, SA is a profitable device for observing feelings and opinion in different client segments, for example, clients interfacing with a specific group of agents, during movements, clients calling in regards to a particular issue, item or administration lines, and other distinct gatherings.

SA might be completely computerized, in light of on human examination, or some mix of the two. At times, SA is essentially computerized with a level of human oversight that powers machine learning and refines processes and algorithms, especially in the beginning times of implementations.SA is utilized over a variety of utilizations and for heap purposes. For example, SA might be performed on Twitter to decide in general conclusion on a specific inclining point. Organizations and brands frequently use assessment examination to screen mark notoriety crosswise over online networking stages or over the web overall. A standout amongst the most generally utilized applications for assumption investigation is for checking call focus and client support execution. As organizations try to keep a finger on the beat of their groups of audience, SA is progressively used for by and large brand observing purposes. SA frameworks to distinguish sarcastic remarks.

SA is an instrument that "implies for choosing an aura of speaker and after that again a creator on some point. Online life has charged up the client opinions in the network space as appraisals, surveys, remarks, and so on. The requirement for exact and dependable data about customer inclinations has prompted expanded enthusiasm towards investigation of web-based life content. For some organizations, on the web conclusion has transformed into a sort of virtual cash that can represent the deciding moment an item in the commercial centre. Conclusion investigation implies observing internet-based life posts and talks, at that point making sense of how members are responding to it.2

1.3 Classification of Sentiment Analysis

- Document Level-Document Level Sentiment analysis is performed for the whole document and then decide whether the document express positive or negative sentiment.

- Entity or Aspect Level- Entity or Aspect Level sentiment analysis performs finer-grained analysis. The goal of entity or side level sentiment analysis is to seek out sentiment on entities and/or side of these entities. For example, consider a statement “My HTC Wildfire S phone has good picture quality but it has low phone memory storage.” so sentiment on HTC‟s camera and display quality is positive however the sentiment on its phone memory storage is negative.

- Sentence Level -Sentence level sentiment analysis is related to find sentiment from sentences whether each sentence expressed a positive, negative or neutral sentiment. Sentence level sentiment analysis is closely related to subjectivity classification. Many of the statements regarding entities are factual in nature and however they still carry sentiment. Current sentiment analysis approaches express the sentiment of subjective statements and neglect such objective statements that carry sentiment. For Example, “I bought a Motorola phone two weeks ago. Everything was good initially. The voice was clear and therefore the battery life was long, although it is a bit bulky. Then, it stopped working yesterday. The first sentence expresses no opinion because it merely states a truth. All alternative sentences specific either express or implicit sentiments. The last sentence “Then, it stopped working yesterday” is objective sentences but current techniques can’t express sentiment for the above specified sentence even though it carries negative sentiment or undesirable sentiment.

II. BACKGROUND

Yadav P et al. (2017)Informal communication has bit by bit turned into a daily schedule for individuals to post their suppositions, perspectives and remarks on any item or individual. Individuals share their sentiments online in an exceptionally casual dialect. Hence, it is exceptionally troublesome undertaking to break down correct assumptions appended with that characteristic dialect. Notion Analysis is an investigation of individuals' mentality, feelings, and feelings to arrange whether it is certain, negative or unbiased. Utilization of emojis via web-based networking media has expanded quickly as of late. Consequently, we have concentrated more on how emojis assume an imperative job in estimation examination. Different variables that influence opinion examination are talked about quickly in this paper. Additionally different issues like sarcasm recognition, multilingualism, taking care of acronyms and slang dialect, lexical variety and dynamic word reference dealing with are examined [2].

Prasad AG et al. (2017) The development of web based life has been exponential in the ongoing years. Monstrous measure of information is being put out onto the general population area through online life. This colossal freely accessible information can be used for research and a grouping of utilizations. The objective of this paper is to counter issues with the web based life dataset, to be particular : short substance nature - the compelled measure of substance data constant spilling nature, utilization of short structures and present day slangs and extending use of mockery in messages and posts. Tweets on sarcasm can mislead data mining activities and result in wrong gathering.This paper make comparison of different grouping calculations, for example, Gradient Boosting, Random Forest, Logistic Regression, Decision Tree, Adaptive Boost. The slang and emoticon word reference being the clever thought presented in this paper. [3].

Sharma p. et al. (2016) the author in this paper have proposed a system which classifies the polarity of the movie reviews on the basis of features by handling negation, intensifier, conjunction and synonyms with appropriate pre-processing steps. SentiWordNet tool is used for calculating the scores of reviews.The GUI include Negation,Intensifier, coordinating conjunction and synonyms handling. The author consideredtwo types of cases, as if to give single review as input then,only negation, intensifier and conjunction handling is done,but if to give multiple reviews as input then along with them,synonyms handling also done [4].

Ravi K et al. (2015) With the methodology of Web 2.0, people ended up being more on edge to express and offer their sentiments on web concerning ordinary activities and overall issues as well. Advancement of web-based systems administration has furthermore contributed colossally to these activities, in this way giving us a direct stage to share sees over the world. These electronic Word of Mouth (eWOM) clarifications conveyed on the web are much normal in business and organization industry to enable customer to share his/her point of view.In the last one and half decades, look at systems, the insightful world, open and organization organizations are working completely on SA, generally called, OM, to isolate and research open tendency and points of view. In such manner, this paper exhibits a thorough study on SA, which depicts sees introduced by more than one hundred articles distributed in the most recent decade with respect to important assignments, methodologies, and uses of supposition examination. A few sub-assignments should be performed for examination which thusly can be refined utilizing different methodologies and systems. This overview covering distributed writing during 2002– 2015, is sorted out based on sub-errands to be performed, machine learning and regular dialect handling procedures utilized and uses of assessment investigation [5].

Augustyaniak et al. (2014), examines two ways to deal with sentiment analysis: lexicon based versus supervised learning in the space of reviews of movies. In assessment, the methodologies were looked at by utilizing a test collection of standard movie review. The outcomes demonstrate that approach based on lexicon is effortlessly outperformed by approach of classification. [6]

Nehra et al. (2014), gives a general study about opinion mining or sentiment analysis identified with reviews of movies. With the assistance of innovation, the web turns into a profitable spot for trading thoughts, web learning, surveys for an item or movies or administration. It makes hard to record and comprehend the emotion of client in light of the fact that surveys over the web are accessible for millions for an item or administrations. Sentiment analysis is a developing area for exploration to gather the subjective data in source material by applying Natural Language handling, Computational Linguistics and content examination and arranged the extremity of the assessment or assumption. In basic words it is said that for process of making of decision, sentimental analysis is essential. [7]

Jalaj et al. (2013), this paper presents a new approach for classifying and handling subjective as well as objective sentence of sentiment analysis. In proposed approach, it includes four steps: (a) first of all classify the sentence into two categories i.e. opinionated and non-opinionated, without regarding whether it is subjective or objective. (b) As having Opinionated sentence classify them as subjective or objective. (c) Classify subjective sentence into positive, negative and neutral. (d) classify objective sentences into positive, negative and neutral, providing semantic orientation for complex one. In this paper, Support Vector Machine (SVM), Naïve bayes, Bag of Words (BOW), POS (Part-of Speech), SentiwordNet, N-gram , Text mining and grammar rules technology used for classification and sentiment analysis. Here, domain used is Indian Political news article and preparing dictionary for this domain [8].

Javier et al. (2013), paper presents a COSMOS platform for sentiment and tension analysis on twitter dataset. Tool used for sentiment analysis is SeniStrength. To run application based on cloud environment, it uses virtualized Hadoop Clusters in Open Nebula. This system configuration used for performance aspects which shows how virtual server needs to be distributed as to reduce variability in the analysis performance. It also presents the architecture for data processing of COSMOS using Open Nebula and Hadoop. Processing performance comparison is done over Cardiff Cloud Tweet and UCLM Cloud Tweet which shows Cardiff Cloud have better performance due to its compute node has been more powerful than UCLM Cloud compute node. This paper involves future work to evaluate on bigger cloud environment and increase number of virtual cluster and Twitter message and improve performance with multiple concurrent users using the same cloud service. As using COSMOS we can add more nodes and workers to the problem and bring processing time down further [9].

Mohsen et al. (2013), this paper represents improved method for aspect level sentiment analysis. It also proposes to use bag of noun instead of bag of words as to improve the clustering result for the aspect identification. By illustrating an example it is proven that bag of noun is able to find similar sentences using clustering algorithm as compared to bag of words. After identifying the aspects, the sentiment of each sentence need to be identified which contain one of those aspects. Here, SVM (Support Vector Machine) classifier has been used. Two type of representation i.e.(a) Bag of words (b) Score . In this paper, two experiments were done i.e. firstly, comparing bag of noun and bag of words clustering and secondly, classifying each sentence to positive, neutral or negative sentiment. Here, TripAdvisor.com used to create corpus. For 3-class classification, one-against-all scheme have been used. It use 5-fold cross validation. Results prove that clustering with bag of noun yields better and meaningful aspect than bag of words approach. By using 3-class sentiment analysis, we can improve the performance by 20% in terms of average f1-score [10].

III. PROPOSED WORK

The proposed work consists of hybrid model comprising Gini index-based feature selection with balanced random forest as a classification technique for predicting the sentiments of movie reviews.The proposed Gini index feature selection addresses the issues of uneven distribution of prior class probability and global goodness of a feature in two stages. First, it transforms the samples space into a feature specific normalized samples space without compromising the intra-class feature distribution. In the second stage of the framework, it identifies the features that discriminates the classes most by applying gini coefficient of inequality. Also, Balanced **Random Forest Algorithm** is used for classification which handle missing values using median for numerical values or mode for categorical values. The proposed work leads to the selection of relevant attributes for prediction and less error rate and accuracy in the result

Firstly, the collection of raw data from Imdb movies review site and then filtering techniques are applied to make that raw data into structured format. Filtration of textual movie reviews consist of following phases:

- Word parsing and tokenization: In this phase, each user review splits into words of any natural processing language. As movie review contains block of character which are referred to as token.
- Removal of stop words: Stop words are the words that contain little information so needed to be

removed. As by removing them, performance increases. Here, we made a list of around 320 words and created a text file for it. So, at the time of pre-processing we have concluded this stop word so all the words are removed from our dataset i.e. filtered.

- Stemming: It is defined as a process to reduce the derived words to their original word stem. For example, “talked”, “talking”, “talks” as based on the root word “talk”. We have used Snowball stemmer to reduce the derived word to their origin.

After pre-processing and filtration feature selection using Gini Index is done to rank the words according to gini index functionality.

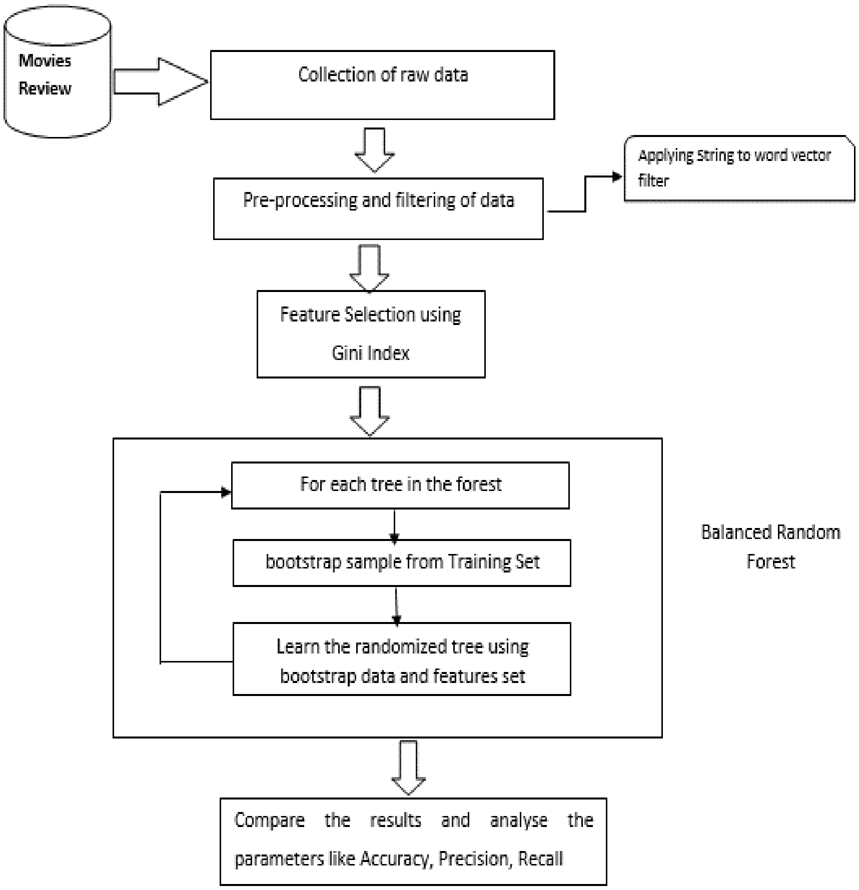

Movies Review
Collection of raw data
Pre-processing and filtering of data
Applying String to word vector filter
Feature Selection using Gini Index
For each tree in the forest
bootstrap sample from Training Set
Learn the randomized tree using bootstrap data and features set
Balanced Random Forest
Compare the results and analyse the parameters like Accuracy, Precision, Recall

Figure 1: Flowchart of proposed methodology

Selected features data is then classified by using Balanced Random Forest

- For each iteration in random forest, draw a bootstrap sample from the minority class. Randomly draw the same number of cases, with replacement, from the majority class.
- I

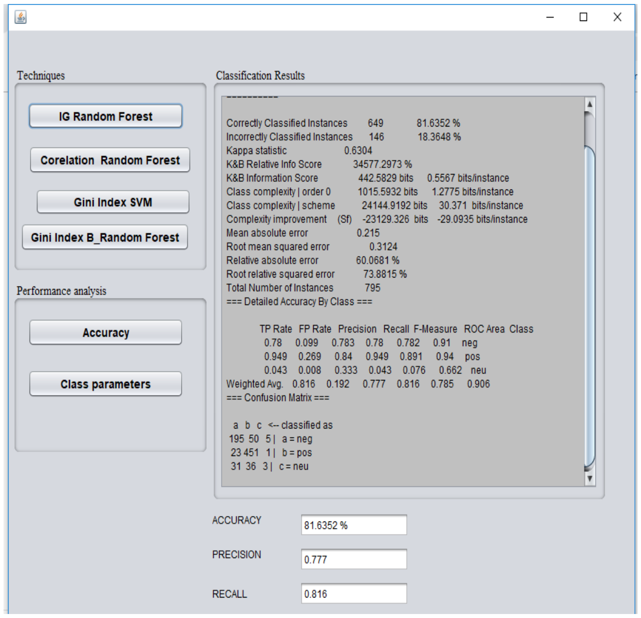
nduce a classification tree from the data to maximum size, without pruning. The tree is induced with the Random Tree algorithm, with the following modification: At each node, instead of searching through all variables for the optimal split, only search through a set of m- try randomly selected variables.

- Repeat the two steps above for the number of times desired. Aggregate the predictions of the ensemble and make the final prediction.

Evaluate and analyze the performance using k cross validation model on the basis of Recall and Precision of existing algorithm and new proposed algorithm.

IV. RESULTS AND DISCUSSION

The parameters for the assessment of estimation investigation incorporate different terms. The terms are True positives, genuine negatives, false negatives and false positives. These are the terms that are used to differentiate the class marks selected with documents with the classes the things truly have a place with by a classifier. Genuine positive terms are really delegated positive terms. False positive are not named by the classifier as positive class but rather ought to have been. Genuine negative terms are effectively named as in negative class by the classifier. False negative terms are those terms that are not marked by the classifier as having a place with negative class yet ought to have be ordered. Disarray Matrix contains these terms that are utilized for assessment.



```
----------

Correctly Classified Instances       649           81.6352 %
Incorrectly Classified Instances     146           18.3648 %
Kappa statistic                        0.6304
K&B Relative Info Score            34577.2973 %
K&B Information Score                442.5829 bits     0.5567 bits/instance
Class complexity | order 0          1015.5932 bits     1.2775 bits/instance
Class complexity | scheme          24144.9192 bits    30.371  bits/instance
Complexity improvement     (Sf)   -23129.326  bits   -29.0935 bits/instance
Mean absolute error                    0.215
Root mean squared error                0.3124
Relative absolute error               60.0681 %
Root relative squared error           73.8815 %
Total Number of Instances            795
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.78      0.099     0.783      0.78      0.782      0.91     neg
                 0.949     0.269     0.84       0.949     0.891      0.94     pos
                 0.043     0.008     0.333      0.043     0.076      0.662    neu
Weighted Avg.    0.816     0.192     0.777      0.816     0.785      0.906
=== Confusion Matrix ===

   a   b   c   <-- classified as
 195  50   5 |   a = neg
  23 451   1 |   b = pos
  31  36   3 |   c = neu
```

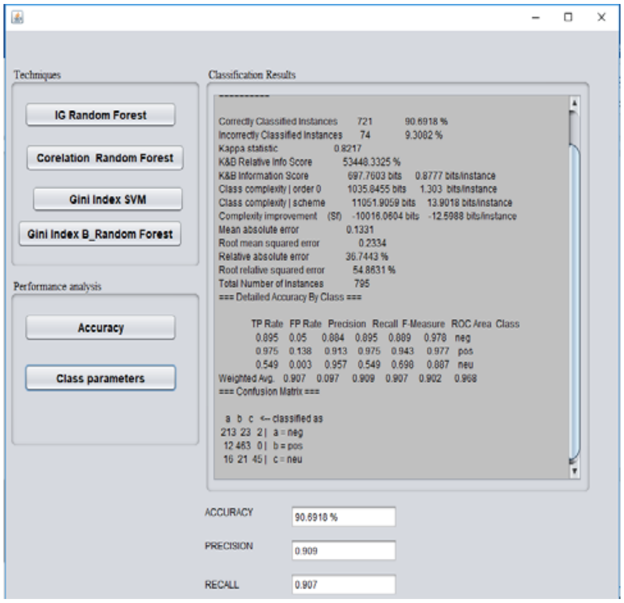
ACCURACY 81.6352 %

PRECISION 0.777

RECALL 0.816

Figure 2: Information gain with Random Forest

The figure above shows the existing Information gain with random forest algorithm results. The classification accuracy achieved in this is 81%



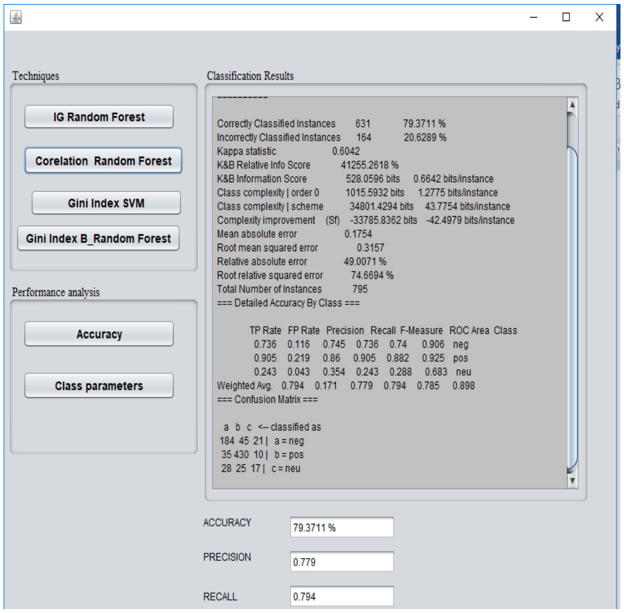
```
----------

Correctly Classified Instances       721           90.6918 %
Incorrectly Classified Instances      74            9.3082 %
Kappa statistic                        0.8217
K&B Relative Info Score            53448.3325 %
K&B Information Score                697.7603 bits     0.8777 bits/instance
Class complexity | order 0          1035.8455 bits     1.303  bits/instance
Class complexity | scheme          11051.9059 bits    13.9018 bits/instance
Complexity improvement     (Sf)   -10016.0604 bits   -12.5988 bits/instance
Mean absolute error                    0.1331
Root mean squared error                0.2334
Relative absolute error               36.7443 %
Root relative squared error           54.8631 %
Total Number of Instances            795
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.895     0.05      0.884      0.895     0.889      0.978    neg
                 0.975     0.138     0.913      0.975     0.943      0.977    pos
                 0.549     0.003     0.957      0.549     0.698      0.887    neu
Weighted Avg.    0.907     0.097     0.909      0.907     0.902      0.968
=== Confusion Matrix ===

   a   b   c   <-- classified as
 213  23   2 |   a = neg
  12 463   0 |   b = pos
  16  21  45 |   c = neu
```

ACCURACY 90.6918 %

PRECISION 0.909

RECALL 0.907

Figure 3: Correlation with Random Forest



```
----------

Correctly Classified Instances       631           79.3711 %
Incorrectly Classified Instances     164           20.6289 %
Kappa statistic                        0.6042
K&B Relative Info Score            41255.2618 %
K&B Information Score                528.0596 bits     0.6642 bits/instance
Class complexity | order 0          1015.5932 bits     1.2775 bits/instance
Class complexity | scheme          34801.4294 bits    43.7754 bits/instance
Complexity improvement     (Sf)   -33785.8362 bits   -42.4979 bits/instance
Mean absolute error                    0.1754
Root mean squared error                0.3157
Relative absolute error               49.0071 %
Root relative squared error           74.6694 %
Total Number of Instances            795
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.736     0.116     0.745      0.736     0.74       0.906    neg
                 0.905     0.219     0.86       0.905     0.882      0.925    pos
                 0.243     0.043     0.354      0.243     0.288      0.683    neu
Weighted Avg.    0.794     0.171     0.779      0.794     0.785      0.898
=== Confusion Matrix ===

   a   b   c   <-- classified as
 184  45  21 |   a = neg
  35 430  10 |   b = pos
  28  25  17 |   c = neu
```

ACCURACY 79.3711 %

PRECISION 0.779

RECALL 0.794

Figure 4: Gini Index with Balanced Random Forest

Evaluation parameters are:

Precision and Recall:

Precision and recall are the two measurements that are generally to evaluate execution in content mining, and in content examination field like data recovery. These parameters are utilized for estimating exactness and completeness respectively.

$$\text{Precision}=\frac{\text{True Positive} \qquad \text{True Positive}}{\text{True Positive}+\text{False Positive}\text{True Positive}+\text{False Positive}}$$

$$\text{Recall}=\frac{\text{True Positive} \qquad \text{True Positive}}{\text{True Positive}+\text{False Negative}\text{True Positive}+\text{False Negative}}$$

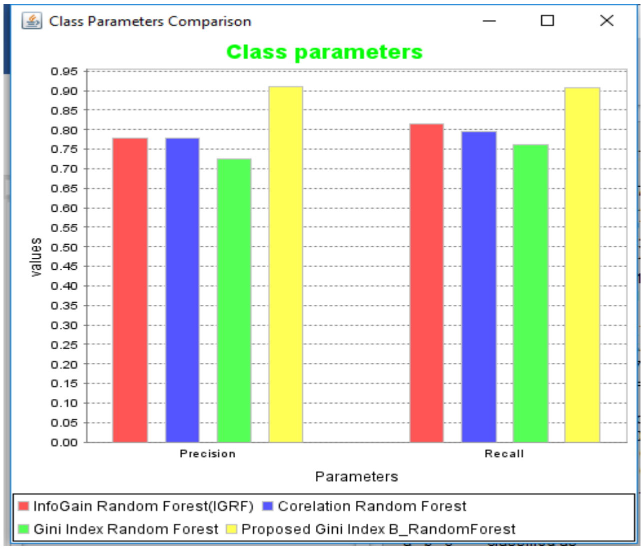


Figure 5: Comparison of class parameters of the proposed technique with the existing techniques

The figures shown above figure 3 and 4 are showing the results of existing correlation with random forest algorithm contributing 79 % accuracy and proposes gini index with random forest contributing highest accuracy of 90%. Figure 5 showing the comparison of class parameters precision and recall, figure clearly showing that the proposed algorithm performs better in the mentioned parameters.

V. CONCLUSION

Various forms of knowledge are produced from specific Social media companies that should be equipped and to observe person's perspective in the direction of products, objects, movie assessment etc. There are millions of users on social media giants like online sources, you tube, Twitter and Facebook. Apart from social media e-commerce sites also have millions of users. The sentiment analysis using different kind of reviews can give new insights into the business model that the different companies follow and make the company more profitable. The major issue with sentiment analysis is that the mood of the user cannot be known and it causes a big difference in the analysis of what the user wrote and what the user really meant. The problem with information is the attributes with a vast number of qualities. It is one-sided towards picking properties with an expansive number of qualities. In this paper, based on the drawbacks of random forest, hybrid model using gini index feature selection and balanced random forest is implemented and performance is analysed with the existing techniques. The results shown that the proposed technique achieved the highest accuracy rate of 90.69 %. In future, Ensemble of classifiers can be use to further improve the predictions and also sarcasm sentences can be included to further improve the performance.

VI. REFERENCES

[1] Abinash Tripathy, Ankit Agrawal, Santanu Kumar Rath, "Classification of Sentiment Reviews using N-gram Machine Learning Approach".

[2] Yadav P, Pandya D. SentiReview: sentiment analysis based on text and emoticons. InInnovative Mechanisms for Industry Applications (ICIMIA), 2017 International Conference on 2017 Feb 21 (pp. 467-472). IEEE

[3] Pandey AC, Rajpoot DS, Saraswat M. Twitter sentiment analysis using hybrid cuckoo search method. Information Processing & Management. 2017 Jul 1;53(4):764-79.

[4] Sharma P and Mishra N. , "Feature level sentiment analysis on movie reviews" , 2nd International Conference on Next Generation Computing Technologies (NGCT). IEEE 2016.

[5] Ravi K, Ravi V. A survey on opinion mining and sentiment analysis: tasks, approaches and applications. Knowledge-Based Systems. 2015 Nov 1;89:14-46.

[6] Lukasz Augustyaniak, Tomasz Kajdanowicz, Przemyslaw Kazienko, Marcin Kulisiewicz, Wlodzimierz Tuliglowicz, "An Approach to Sentiment Analysis of Movie Reviews: Lexicon Based vs. Classification", Springer, Vol. 8480, pp. 168-178, 2014.

[7] NehaNehra, "A Survey on Sentiment Analysis of Movie Reviews", International Journal of Innovative Research in Technology, Vol. 1, Issue 7, 2014.

[8] Jalaj S. Modha, Gayatri S. Pandi, Sandip J. Modha. (2013) "Automatic Sentiment Analysis for Unstructured Data", International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 3, Issue 12, December 2013.

[9] Javier Conejero, Peter Burnap, Omer Rana, Jeffery Morgan (2013) "Scaling Archied Social Media Data Analysis Using a Hadoop Cloud" , 6th international conference on cloud computing, IEEE.

[10] Mohsen Farhadloo, Erik Rolland. (2013) "Multi-class Sentiment analysis with clustering and score representation", 13th International Conference on Data mining Worshops, IEEE, pp. 904-912, December 2013.