IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Volume-5, Issue-9 E-ISSN: 2347-2693

Sentiment Analysis of Movie Reviews: A Study of Machine Learning Algorithms with Various Feature Selection Methods

Rajwinder Kaur[1*], Prince Verma[2]

[1*]Dept. of Computer Science Engineering, CTIEMT, Jalandhar, India
[2]Dept. of Computer Science Engineering, CTIEMT, Jalandhar, India

[*]*Corresponding Author: dhillon.raj320@gmail.com*

Available online at: www.ijcseonline.org

Received: 19/Aug/2017, Revised: 28/Aug/2017, Accepted: 22/Sep/2017, Published: 30/Sep/2017

***Abstract*—** Nowadays, with rapid use of internet, a very large number of reviews are posted by visitors on different website related to the various movies that describe the polarity between movies. Customers share their feelings with others in the form of comments or reviews that describe their opinion as either in negative or in positive or in neutral. Such websites are essential to people for decision making. In this paper, the sentiment analysis is done in order to analyze the movie reviews, so we use the machine learning classifier Random Forest with Gini Index based Feature Selection and also compared it with another algorithm such as SVM. The results show that Gini Index method with Random Forest classifier has better performance in terms of Accuracy, Root Mean Square Error, Precision, Recall and F-Measure.

***Keywords*—**Sentiment Analysis, Related Work, Feature Selection, Classification Algorithms, Evaluation Matrices.

I. INTRODUCTION

Today with the tremendous help of technology, the internet has becomes a highly valuable place in which ideas has been exchanged easily, online learning, reviews for a service or product or movies. It makes hard to understand and record the emotions of the user because reviews on the internet are available for millions for a services or products [1].

Sentiments are emotions of the users regarding entities such as products, events, issues and services that may be good, excellent, bad or neutral. Analysis of people's emotions, reactions based on feedback from internet is known as sentiment analysis. Sentiment analysis is also called as opinion mining, sentiment mining, opinion extraction, subjectivity analysis, emotion analysis, affect analysis, review mining [2]. Flow Chart of Sentiment Analysis is shown in Figure1.

For research, Sentiment analysis is an emerging area to collect the subjective information from source material by applying Linguistics and text analytics, Natural Language processing, Computational and categorized the polarity of the sentiment or opinion. Sentiment analysis is language processing task which uses computational approach to identify the opinion of user and classify as positive, negative or neutral. The main aim of analysis of sentiment is to identify the attitude of a writer or a speaker with respect to some topic. This attitude of writer or speaker may be their evaluation, affective state which is the emotional state of the author when writing, or the intended emotional communication (the emotional effect the author wishes to have on the reader). We can distinguish poor content from high quality content with the help of opinion mining [3].

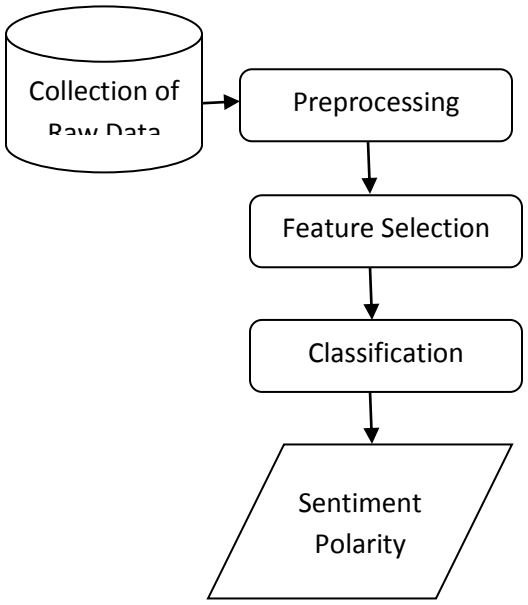


Figure 1. Sentiment Analysis Process Flow

Both organizations and customers can take advantage of opinion mining and sentiment analysis. When a customer

wants to buy a product or decides whether product has good quality or not, he/she has access a large number of user reviews, but reading and analyzing all of the reviews could be a lengthy and frustrating process. And when an organization wants to extract the public opinion about its products, identify new opportunities, market its products, predict sales trends, or manage its reputation, it needs to deal with a huge number of available customers comments. So sentiment analysis may help both customers and organization to achieve their goals.

In general, there are three levels where opinion mining is performed. They are:

A. Document-level sentiment analysis

Document-level analysis is the simplest form of sentiment analysis. In Document-level sentiment analysis, the entire document is classified as either in positive or in negative sentiment for a services or products. Mainly, there are two approaches that are used for document level sentiment analysis: supervised learning and unsupervised learning. The supervised approach that assumes there are a finite set of classes or groups in which the document should be classified and for each class, training data is available. The simplest case is when there are two classes: positive and negative. Unsupervised approaches for document-level sentiment analysis are within the document determining the semantic orientation (SO) of specific phrases. For the selection of the phrases, there are two main approaches: to select these phrases, a set of predefined part of speech patterns can be used or a lexicon of sentiment words and phrases can be used [3].

B. Sentence-level sentiment analysis

Sentence-level sentiment classification is used to determine whether each sentence indicates a positive, negative or neutral opinion for a particular product or service. Sentence level analysis is performed by two tasks: subjective and objective. Subjectivity is the classification which makes differentiation between objective sentences and subjective sentences. The objective sentences express true information [2]. The subjective sentences are express subjective views and opinions. Objective: few days ago, I purchase ABC mobile. Subjective: It is such a perfect phone.

C. Aspect-based sentiment analysis

The previous two approaches work well either for the whole document or for each individual sentence that refers to a single entity. However, in many cases people discuss about entities that have no. of many aspects or attributes and they have given a different opinion about each of the aspects. This happens in reviews about products or in discussion forums dedicated to specific product categories (such as cameras, cars, smart phones, and even pharmaceutical drugs). Aspect-based sentiment analysis is also called feature-based or entity-based sentiment analysis that is the research problem which focuses on the recognition of every sentiment expressions within a given document and the aspects to which they refer. The aspect level analysis is more difficult and challenging than the document level and sentence level classifications [2].

Aside from these three levels of classification, comparative opinions and regular opinions are two more categories of opinions. A sentiment that expressed only on a particular entity or an aspect of the entity is called regular opinion, e.g., "Manish perform very well his performance" expresses positive sentiment on the aspect of performance of Manish. A sentiment that expressed by differentiating multiple aspects based on some of their shared attributes is called comparative opinion. For example, "a Fruit cake tastes better than Chocolate cake", compares cakes based on their flavors (an aspect) and expresses feeling and preference for Fruit cake.

In this paper, Section II throws the light on brief summary of related work. Moving further towards Section III explains the problem formulation. Section IV depicts Proposed Technique that we used and shows implementation step by step. Section V shows the result and performance and at last Section VI provides a brief conclusion.

II. RELATED WORK

Cagatay CATAL et al. [4] introduced objective of paper is to investigate the potential benefit of multiple classifier systems concept on Turkish sentiment classification problem and propose a novel classification technique. Vote algorithm had been used in conjunction with three classifiers, namely Naive Bayes, Support Vector Machine (SVM), and Bagging. Parameters of the SVM have been optimized when it was used as an individual classifier. Experimental results showed that multiple classifier systems increase the performance of individual classifiers on Turkish sentiment classification datasets and Meta classifiers contribute to the power of these multiple classifier systems. The proposed approach achieved better performance than Naive Bayes, which was reported the best individual classifier for these datasets, and Support Vector Machines. Multiple classifier systems (MCS) are a good approach for sentiment classification, and parameter optimization of individual classifiers must be taken into account while developing MCS-based prediction systems.

In paper [5], Rajesh Piryani et al. presented an experimental work on aspect-level sentiment analysis of movie reviews. Movie reviews basically contain user opinion for various aspects such as direction, acting, choreography, cinematography, etc. They had formulate a linguistic rule-based approach which recognize the aspects from movie reviews, locates opinion about that aspect and enumerate the sentiment polarity of that opinion using linguistic approaches. The system generates an aspect-level opinion summary. The experimental design is evaluated on datasets

of two movies. The results achieved good accuracy and shows promise for deployment in an integrated opinion profiling system.

Asha S Manek et al. [6] implemented sentiment analysis for movie reviews using various feature selection methods with naive bayes and Support Vector Machine (SVM). The proposed work uses number of steps such as collection of movie reviews datasets, pre-processing, feature selection, classification techniques. Result shows that gini index method gives better performance with SVM for classification for large amount of dataset and Correlation based feature selection with SVM for small amount of dataset.

Deepa Anand et al. [7] contributed this paper is two-fold: Firstly, a two class classification scheme for plots and reviews without the need for labeled data is proposed. The overhead of constructing manually labeled data to build the classifier is avoided and the resulting classifier is shown to be effective using a small manually built test set. Secondly they proposed a scheme to detect aspects and the corresponding opinions using a set of hand crafted rules and aspect clue words. There are three schemes that helps for the selection of aspect clue words are explored - manual labeling (M), clustering(C) and review guided clustering (RC). The aspect and sentiment detection using all the three schemes is empirically evaluated against a manually constructed test set. The experiments establish the effectiveness of manual labeling over cluster based approaches but among the cluster based approaches, the ones utilizing the review guided clue words performed better.

In paper [8] by Bogdon Batrinca et al. proposed an overview of software tool for social media, blogs, chats, newsfeeds etc. and how to use them for scraping, cleansing and analyzing. For scraping the social media it suggests the challenges such as Data cleansing, Data protection, Data analysis and Visualization and analytics Dashboard. This paper presents a survey on methodology of social media, data, providers and analytics techniques such as stream processing, sentimental analysis. An overview of different tools needed for social analysis purpose is also presented. There has been easy availability of APIs provided by Twitter, Facebook and News services which led to explosion of data services for the purpose of scraping and sentiment analysis.

Mrs. R.Nithyaet al. [9] represented Sentiment analysis that mainly on subjective and polarity detection. A proposed work include: (i) Feature Extract- Commonly, Sentiment analysis uses machine learning algorithm and a method to extract features from texts and then train the classifier. (ii) Preprocessing- stemming refers reducing words to their roots. Porter's stemming algorithm used for removing stop words. Mostly, adjective words have sentiment. (iii) Product aspects- Textstat is a freely available that can be used for extracting pattern. (iv) Find polarity of opinionated sentence- here SentiStrength lexicon-based classifier used to detect sentiment strength. Here, 575 reviews have been taken from shopping sites. Tanagra1.4 tool used for data mining. Naïve bayes classification done through this tool based on each individual features such as display, accessories, battery life, weight and cost. Results shows that 'battery life' have most positive value so it improves branding and 'cost' have very low positive value that indicate seller to concentrate more on reputation and product quality.

III. PROBLEM DEFINITION

Different type's knowledge is generated from specific Social media companies that need to be equipped and to observe person's perspective in the direction of products, objects, movie assessment etc. This database is accrued from distinctive social media websites for illustration Twitter, Facebook, online review, shopping web sites and so on. Text analytics and Sentiment analysis can support to advance priceless trade insights from text headquartered contents that could be within the type of word files, tweets, comments and news that concerning Social media. As it is known that the problem with information gain is the attributes with a large number of values. It is biased towards choosing attributes with a large number of values. This may result in over fitting (selection of an attribute that is non optimal for prediction).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.

Compare with SVM, RF is able to estimate feature importance during training for little additional time. It is faster to train and has fewer parameters. It is resistant to outliers and is able to automatically handle missing values and more importantly, it works better with large databases and a large number of features. Furthermore, RF is applicable to high-dimensional with a low number of observations. There are no real hyper parameters to tune in RF (maybe except for the number of trees, typically, the more trees we have the better. On the contrary, there are a lot of knobs to be turned in SVM; choosing the appropriate kernel function can be tricky. RF is better than SVM in terms of prediction accuracy.

IV. PROPOSED TECHNIQUE

Sentiment Analysis is a challenging task in Machine Learning. The analysis of the sentiment of text is not a trivial task because when people share their views or opinions regarding any particular topics by writing text, they can use sarcasm and some views can be ambivalent and also can use some common words to express their feelings. Sentiment label can be categorized as positive, negative or neutral by indicating the numbers to each and every review.

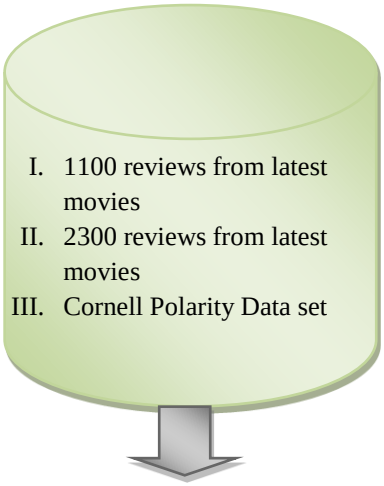
The Sentiment Analysis is difficult because words quite often take extraordinary meanings and are related to exclusive emotions depending on the area in which they are getting used. To minimize this, knowledge mining techniques has been used for extraction of elements from these datasets. To extract the sentiments, we propose classification tasks in following steps in Figure 2.

A. Data Source

In this phase, we contribute the perfunctory details about the datasets that are used in our implementation. We extricate the latest reviews of movies from IMDb websites. We have performed demonstration on above declare corpus and our own datasets. Table1 shows the dataset details that are used in proposed work.

Datasets	No. of Reviews
IMDb movie reviews (http://www.imdb.com/)	1100
IMDb movie reviews (http://www.imdb.com/)	2300
Cornell Polarity Data set v1.0 (http://www.cs.cornell.edu/people/pabo/movie-review-data/)	3400
Cornell Polarity Data set v1.0+ IMDb movie reviews (http://www.cs.cornell.edu/people/pabo/movie-review-data/ http://www.imdb.com/)	4500

Table1. Details of datasets that are used for implementation



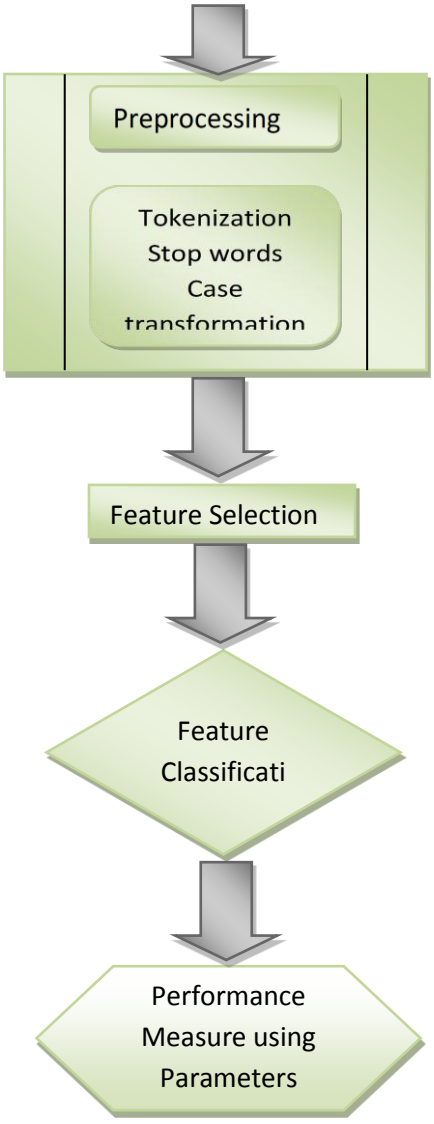


Figure 2. Flow diagram of proposed work

B. Data Preprocessing

For the pre-processing of the data will be applied some data filtering techniques to make that raw data into structured format. Pre-processing involves several steps such as tokenization, Removal of stop words and Case Normalization.

1) Tokenization: Tokenization is a process in which text of a document is cleaved into series of token. The data that is extracted from online reviews hold noise such as symbols, scripts etc which is not necessary and not used for machine learning. In order to retain only text, these

noises are to be removed so as to improve the performance of the classifier.

2) *Removal of stop words:* It is a process to minimize the size of document by eliminating the most usual words according to stop word list. Stop words are those words that are not compulsory for the sentences or opinions. Stop word list consist of preposition and determiners. For example: "Rahul is a good boy" will be processed to "Rahul is good boy".

3) *Case Normalization:* Most of the reviews are in the compound form that is uppercase and lowercase and it needs to convert whole document into uppercase or lowercase. Case Normalization is the process in which all the characters in a document convert either in the uppercase or lowercase.

C. Feature Selection

Feature selection is a process that performs the selection of features in the data, out of which majority data is related to the current predictive modeling problem. It is a process of choosing a reduced relevant features that improves classification by searching for the best feature subset, from the fixed set of original features according to a given classification accuracy [11]. It removes irrelevant or redundant features. Feature selection is also known as attribute selection, variable selection or variable subset selection, is the process of selecting a subset of relevant features (variables, predictors) for use in model construction. In this work, we are used Gini Index method for feature selection. The Gini Index of node impurity is the measure most common chosen for classification-type problem. If a dataset T contains examples from n classes,

Gini Index, Gini(T) is defined as:

$$\text{Gini(T)} = \sum_{j=1}^{n} (pj)\,2 \qquad (1)$$

D. Classification

Classification is a technique in data mining that assigns items in a set to target classes or categories. The purpose of classification is to get the forecast of the target class for every case in the data. The algorithm will try to find out relationships between the attributes/variables that will ensure it is possible to forecast the outcome. In the part of classification, SVM and RF algorithm is used for classify movie reviews that are collected from internet. In this work algorithms are used for identification and compared for the detailed evaluate the results. Classification is done by SVM and RF. Support Vector Machines are supervised learning models that are used hyperplanes or set of hyperplanes for separation of classes by evaluating maximum margin from both classes. A hyperplane is represented by the following equation [6]:

$$WX + b = 0 \qquad (2)$$

On the other hand, Random Forest algorithm works as a large collection of de-correlated decision trees. Random Forest is applicable to high-dimensional data with a low number of observations.

E. Performance Measure

After completing the classification of features, performance is measured by using various parameters such as Accuracy, Root Mean Square Error, Precision, Recall and F-Measure.

V. Results And Performance Analysis

Our proposed model is implemented on movie reviews datasets. There are four datasets. The first and second dataset consider 1100 and 2300 reviews of latest movies that are collected from IMDb web site. The third dataset is used from Cornell website that consist 3400 reviews. In last, the fourth dataset is constructed by combining the reviews from first three datasets with 4500 reviews. These all datasets are having both positive and negative reviews that are delivered by viewers on websites. After applying string to vector filtration for preprocessing of data, feature selection methods are used to select the most relevant features. Random Forest and Support Vector Machine (Linear) algorithms are applying for classification of movie reviews. We are using six parameters to comparing the results that is Accuracy, Root Mean Square Error, Precision, Recall, and F-Measure.

A. Evaluation Metrics

1) *Accuracy :* Accuracy is the performance evaluation parameters in which the true outcomes such as true positive and true negative among the total cases are examined such as true positive, true negative, false positive and false negative [11].

$$\textbf{Accuracy} = \frac{TN+TP}{TN+TP+FP+FN} \qquad (3)$$

2) *Root Mean Square Error:* The square root of the arithmetic mean of the squares of a set of the values. The Root Mean Square Error (RMSE) (also called root mean square deviation, RMSD) is the frequently used measure of the difference between values predicted by the model (y) and the values actually observed from the environment (yi) [6]. It can be calculated as:

$$\boldsymbol{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(y-yi)^2}{n}} \qquad (4)$$

3) Precision: Precision is defined as the fraction of documents that are retrieved and that are relevant from that retrieved documents to the Query [11]. Precision is defined as:

$$\textbf{Precision} = \frac{\textbf{relevant documentss} \cap \textbf{retrived documents}}{\textbf{retrived documents}} \quad (5)$$

4) *Recall:* Recall is defined as the division of the documents that are matches to the query that are retrieved successfully [11].

$$\textbf{Recall} = \frac{\textbf{relevant } \boldsymbol{documentss} \cap \textbf{retrived documents}}{\textbf{relevant documents}} \quad (6)$$

5) *F-Measure:* The F1 score (also F-score or F-measure), in statistical analysis of binary classification is a measure of a test's accuracy [11]. It computes the score considering both precision p and recall r as follows:

$$\boldsymbol{F - Measure = 2\frac{Precision.Recall}{Precision+Recall}} \quad (7)$$

B. Result of Algorithms For Movies Reviews Datasets

1) *Accuracy:* For Accuracy, table2 shows that overall result and performance are best proved for different number of movie reviews with Random Forest algorithm with Gini Index based feature selection than other techniques with 63.6364 (in case of 1100 reviews), 71.6087(in case of 2300 movie reviews), 78.0588(in case of 3400 reviews) and 80.9111(in case of 4500 reviews. This shows that the proposed model is performing better in terms of Accuracy, as shown in Figure3.

Table2. Comparing various algorithms by Accuracy for different Movie Reviews Dataset

ALGORITHMS	1100 Reviews	2300 Reviews	3400 Reviews	4500 Reviews
GI+SVM	63.0909	69.4348	71.2941	75.2444
IG+RF	62.2727	71.1304	76.0588	79.8667
GI+RF	63.6364	71.6087	78.0588	80.9111

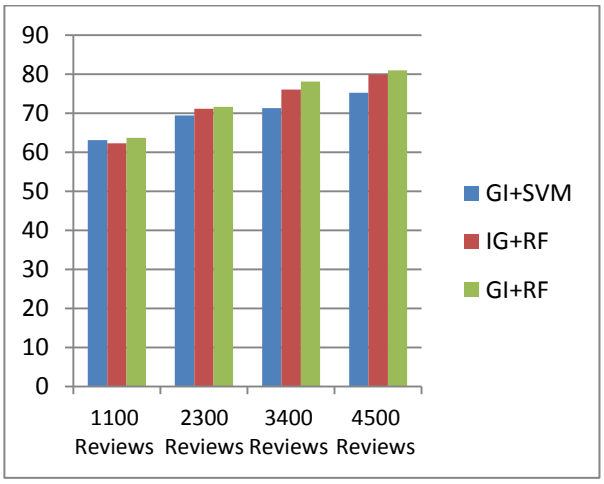


Figure 3. Graphical Representation of Accuracy for different Movie Reviews dataset

2) *Root Mean Square Error (RMSE):* From the Table3, it is depicted that the Random Forest with Gini Index model outperforms well in case of root mean square error. The proposed model has a minimal error rate than the other models. The graphical representation for root mean square error is shown in Figure4.

Table3. Comparing various algorithms by RMSE for different Movie Reviews Dataset

ALGORITHMS	1100 Reviews	2300 Reviews	3400 Reviews	4500 Reviews
GI+SVM	0.6075	0.5529	0.5358	0.4975
IG+RF	0.4735	0.4325	0.396	0.3609
GI+RF	0.4674	0.4274	0.3905	0.3549

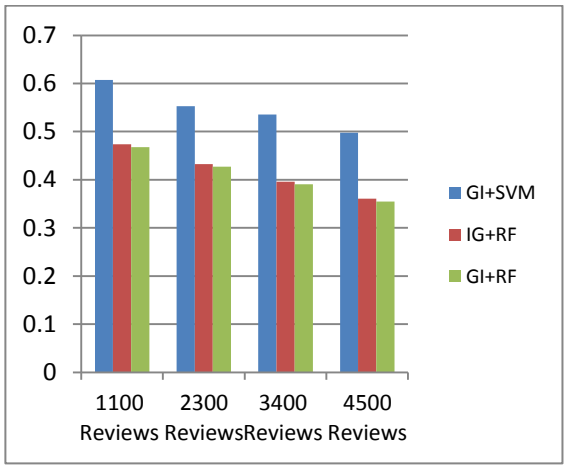


Figure 4. Graphical Representation of RMSE for different Movie Reviews datasets

The proposed model is outperforming with minimal error of 0.4674(in case of 1100 reviews), 0.4274(in case of 2300 movie reviews), 0.3905(in case of 3400 reviews) and 0.3549(in case of 4500 reviews).

3) *Precision:* For Precision, the results are presented in Table4; it demonstrates that the proposed model is performing better than the base models with better precision. The proposed model has greater precision with value 0.633 for 1100 movie reviews, 0.721 for 2300 movie reviews, 0.783 for 3400 movie reviews and 0.807 for 4500 movie reviews dataset. Thus, it has been seen that the proposed model is performing better than base model in terms of precision measure.

Table4. Comparing various algorithms by Precision for different Movie Reviews Dataset

ALGORITHMS	1100 Reviews	2300 Reviews	3400 Reviews	4500 Reviews
GI+SVM	0.695	0.694	0.713	0.764
IG+RF	0.619	0.719	0.764	0.797
GI+RF	0.633	0.721	0.783	0.807

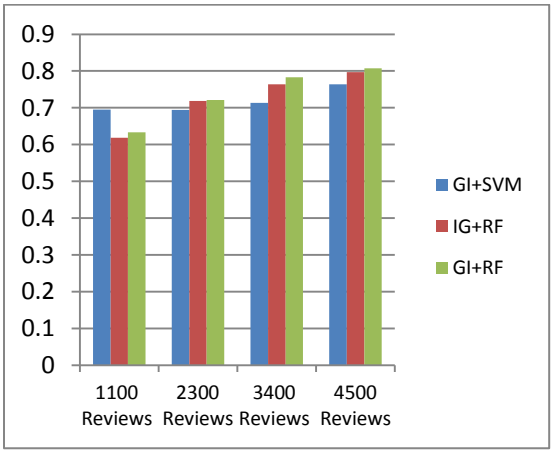


Figure 5. Graphical Representation of Precision for different Movie Reviews datasets

4) *Recall:* For Recall, Table 5 shows that the proposed Model performs better than base models with 0.636(in case of 1100 reviews), 0.716(in case of 2300 reviews), 0.781(in case of 3400 reviews) and 0.809(in case of 4500 reviews). This shows that the proposed model is performing better in terms of higher recall, as shown in Figure6.

Table5. Comparing various algorithms by Accuracy for different Movie Reviews Dataset

ALGORITHMS	1100 Reviews	2300 Reviews	3400 Reviews	4500 Reviews
GI+SVM	0.631	0.694	0.713	0.752
IG+RF	0.623	0.711	0.761	0.799
GI+RF	0.636	0.716	0.781	0.809

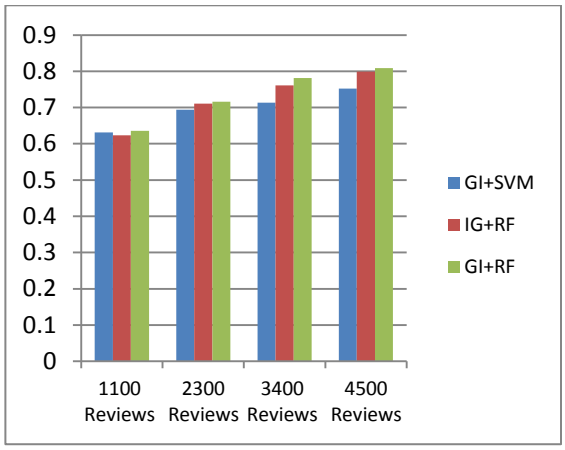


Figure 6. Graphical Representation of Recall for different Movie Reviews datasets

5) *F-Measure:* From Table6, it is depicted that the proposed technique has better performance among other techniques. The Random Forest using Gini Index has greater f-measure rate with value of 0.634 for 1100 movie reviews, 0.715 for 2300 movie reviews, 0.78 for 3400 movie reviews and 0.808 for 4500 movie reviews dataset. Thus, it can be seen that the proposed technique is performing better than base model in terms of f-measure parameter. Figure7 shows the graphical representation of F-Measure.

ALGORITHMS	1100 Reviews	2300 Reviews	3400 Reviews	4500 Reviews
GI+SVM	0.517	0.694	0.713	0.729
IG+RF	0.621	0.709	0.76	0.797
GI+RF	0.634	0.715	0.78	0.808

Table6 Comparing various algorithms by F-Measure for different Movie Reviews Datasets

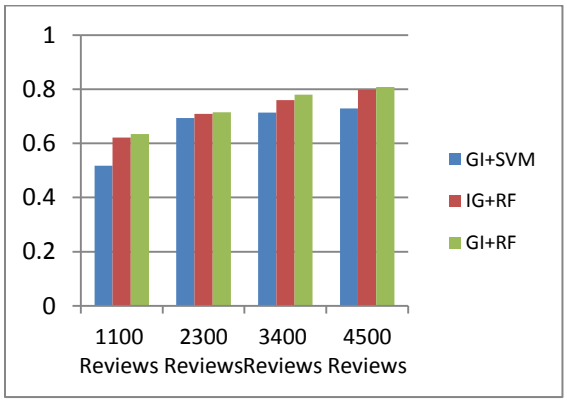


Figure 7. Graphical Representation of F-Measure for different Movie Reviews datasets

VI. CONCLUSION

In this paper, Random Forest algorithm and Gini Index feature selection method are fused to predict sentiment analysis on Movie reviews. Gini Index is used to select the features that are relevant to the task and Random Forest is used to classify the selected items as positive and negative. The performance of proposed technique is compared with the existing algorithms by using Accuracy, RMSE, Precision, Recall and F-Measure. The work can further be extended by including other algorithms and feature selection methods and can be analysis on other domains of opinion mining namely product reviews, political discussion, newspaper articles, social media sites etc.

REFERENCES

[1] V. Krishnaiah, Dr.G.Narsimha and Dr.N.Subhash Chandra, *"Survey of Classification Techniques in Data Mining"*, International Journal of Computer Sciences and Engineering Vol.2, Issue.9, pp 65-74, 2014.

[2] N. Nehra, *"A Survey On Sentiment Analysis Of Movie Reviews"*, International Journal Of Innovative Research In Technology (IJIRT), Vol.1, Issue.7, pp 36-40, 2014 .

[3] R. Feldman, *"Techniques and Applications for Sentiment Analysis"*,Communications of the ACM, Vol.56, Issue.4, pp 82-89, 2013.

[4] C. Catal, M. Nangir, *"A Sentiment Classification Model Based On Multiple Classifiers"*, Applied Soft Computing Elsevier, Vol.50, pp 135–141, 2017.

[5] R. Piryani, V. Gupta, V. K. Singh and U. Ghose, *"A Linguistic Rule-Based Approach for Aspect-Level Sentiment Analysis of Movie Reviews"*, Advances in Computer and Computational Sciences, Springer Nature Singapore Pte Ltd, Vol 1, pp 201-209, 2017.

[6] A.S. Manek, P.D. Shenoy,M.C. Mohan and Venugopal K R, *"Aspect term extraction for sentiment analysis in large movie reviews using Gini Index feature selection method and SVM classifier"*, World Wide Web Internet and Web Information Systems Springer, Volume 20, Issue 2, pp 135–154, 2016 .

[7] D. Anand, D. Naorem, *"Semi-supervised Aspect Based Sentiment Analysis for Movies using Review Filtering"*, 7th International conference on Intelligent Human Computer Interaction, IHCI, Science Direct Elsevier, Vol 84, pp 86-93, 2016.

[8] B. Batrinca, P. C. Treleaven, *"Social media analytics: a survey of techniques, tools and platform"*,AI & Society Springer, Vol.30, Issue.1, pp 89-116, 2015.

[9] Mrs. R.Nithya, Dr. D.Maheshwari, *"Sentiment Analysis on Unstructured Review"*, 14 Proceedings of the International Conference on Intelligent Computing Applications IEEE, , pp 367-371, 2014.

[10] I. Maks, P. Vossen , *"A lexicon model for deep sentiment analysis and opinion mining applications"*, Decision Support Systems Elsevier, Vol.53, Issue.4, pp 680-688, 2012.

[11] T. P. Sahu and S. Ahuja, *"Sentiment Analysis of Movie Reviews: A study on Feature Selection & Classification Algorithms"*, IEEE International Conference of Microelectronics, Computing and Communications (MicroCom) Durgapur, India, 2016.

[12] V. kumar, B. Vaghela and B. M. Jadav, *"Analysis of Various Sentiment Classification Techniques"*, International Journal of Computer Applications, Vol.140 , Issue.3, pp 22-27, 2016.

[13] F.H. Khan, U. Qamar, S. Bashir, *"SentiMI: Introducing point-wise information with SentiWordNet to improve sentiment polarity detection"*, Applied Soft Computing, Elsevier, Vol.39, pp 140-153, 2016.

[14] A. Tripathy, A. Agrawal, S.K. Rath, *"Classification of sentiment reviews using n-gram machine learning approach"*, Expert Systems with Applications, Elsevier, Vol.57, pp 117-126 , 2016.

[15] S.H. Bhojani and Dr. N. Bhatt, *"Data Mining Techniques and Trends – A Review"*, Global Journal For Research Analysis, Vol.5, Issue.5, pp 252-254 , 2016.

[16] Y.S. You, S. Lee and J. Kim, *"Design and Development of Visualization Tool for Movie Review and Sentiment Analysis"*, Sixth International Conference on Emerging Databases: Technologies, Applications, and Theory, Jeju, Repbulic of Korea, pp 117-123, 2016.

[17] P. Gupta, A. Sharma, J. Grover, *"Rating based Mechanism to Contrast Abnormal Posts on Movies Reviews using MapReduce Paradigm"*, 5th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO),IEEE, pp 262-266, 2016.

[18] M. Kannvdiya, K. Patidar and R.S. Kushwaha , *"A Survey On: Different Techniques And Features Of Data Classification"*, International Journal of Research In Computer Applications And Robotics, Vol.4, Issue.6, pp 1-6, 2016.

[19] S. Kaur and A.K. Grewal, *"A Review Paper on Data Mining Classification Techniques for Detection of Lung Cancer"*, International Research Journal of Engineering and Technology (IRJET), Vol.03, Issue11, pp 1334-1338, 2016.

[20] C.H. Chu, C.A. Wang, Y.C. Chang, Y.W. Wu, Y.L. Hsieh and W.L. Hsu, *"Sentiment Analysis on Chinese Movie Review with Distributed Keyword Vector Representation "*, Technologies and Applications of Artificial Intelligence (TAAI), IEEE Conference, pp 84-89, 2016.

[21] Z. Teng, D.T. Vo and Y. Zhang, *"Context-Sensitive Lexicon Features for Neural Sentiment Analysis"*, Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp 1629–1638, 2016.

[22] P. Chikersal, S. Poria, E. Cambria, *"SeNTU: Sentiment Analysis of Tweets by Combining a Rule-based Classifier with Supervised Learning"*, Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval), pp 647–651, 2015.

[23] A. Mukhopadhyay, U. Maulik, S. Bandyopadhyay and C. A. Coello, *"A Survey of Multiobjective Evolutionary Algorithms for Data Mining: Part I"*, IEEE Transactions on Evolutionary Computation, Vol.18, Issue.1, pp 4-19, 2014.

[24] V.K. Singh, R.Piryani, A. Uddin, P.Waila, *"Sentiment analysis of movie reviews: A new feature-based heuristic for aspect-level sentiment classification"*, International Multi-Conference on

Automation, Computing, Communication, Control and Compressed Sensing (iMac4s), IEEE, pp 712-717, 2013.

[25] J.S. Modha, G.S. Pandi, S.J. Modha, *"Automatic Sentiment Analysis for Unstructured Data",* International Journal of Advanced Research in Computer Science and Software Engineering, Vol.3, Issue.12, pp 91-97, 2013 .

AUTHORS PROFILE

Ms. Rajwinder Kaur, She received the B.Tech degree in Information Technology from CTIEMT, Jalandhar, Punjab, India in 2013 and currently, she is pursuing her master of Engineering in Computer Science from CTIEMT, Jalandhar, Punjab, India. Her research interest lies in Data Mining and its algorithms.

Mr. Prince Verma, He received the B.Tech degree in Computer Science from MIMIT, Malout, Punjab, India in 2008 and M.Tech in Computer Science in 2013 from DAVIET, Jalandhar, Punjab, India and Currently, He is the Assistant Professor in the department of Computer Science at CTIEMT, Jalandhar, Punjab, India. His research interest lies in Data Mining, Algorithm optimization techniques.

