http://dx.doi.org/10.21172/ijiet.91.03

Implementation of Feature Selection and balanced random forest for Sentimental Analysis of Text Databases

Er. Babita
M.Tech Reasearch Scholar, CSE Department
Ramgarhia Institute of Engineering And Technology, Phagwara

Er. .Parminder Singh
Head of CSE Department,
Ramgarhia Institute of Engineering And Technology, Phagwara

Abstract*:* Our reviews and the assessments of others play a vital part in our basic leadership process also, even impact our conduct. As of late, an expanding number of individuals have taken to communicating their conclusions on a wide assortment of subjects. With innovation's expanding capacities, sentiment analysis turns into a more used apparatus for organizations. Social media monitoring tools utilize it to give their clients bits of knowledge about how the general population feels as to their business, items, or subjects of intrigue. It's broadly utilized by email administrations to keep spam away from your inbox and by survey sites to suggest new substance like movies or TV shows. When somebody writings you with a snide remark (without emoticons) would you be able to tell if it's snide? In the event that they're really cheerful, furious or nonpartisan? That is the thing that makes sentiment analysis such a broad and intriguing field. Sentiment analysis– additionally called conclusion mining– is the way toward characterizing and ordering suppositions in a given bit of content as positive, negative, or impartial. In this paper, we have proposed Correlation based feature selection algorithm with random forest classification. The proposed technique has been tested on various text datasets and the experimental results shown that proposed technique performs better than the existing classification technique.

***Index Terms*—Data Mining, Movie Reviews, Naïve Bayes Classifier, Correlation based feature selection algorithm**

I. INTRODUCTION

Social Media Monitoring or monitoring is foremost subject in today's present state of affairs. In today many businesses have been utilizing Social Media advertising and marketing to advertise their merchandise or manufacturers, so it becomes fundamental for them that they may be able to be equipped to calculate the success and usability of each product [1].

For constructing a Social Media Monitoring, various tool has been required which entails two accessories: one to evaluate what number of person of their brand are attracted as a result of their advertising and 2nd to discover what humans thinks in regards to the special manufacturer. Social media platforms are refers to the websites and applications that enable people for interacting, building, sharing and exchange information. Every day total world post 400 million tweets on Twitter, 350 million photos on Face book and 4 billion videos on YouTube. This has encouraged developer to develop of new techniques and methodological approaches for capturing, processing and analyzing large and complex data. Big data approaches for analyzing social media data can increase understanding of how people thoughts and act towards a particular topic. Companies can use this information and try to influence and knows users' behaviours in the future [2].

II. RELATED WORK

Asha S Manek, P Deepa Shenoy,M Chandra Mohan and Venugopal K R(2016), This paper implemented sentiment analysis for movie reviews using various feature selection methods with naive bayes and Support Vector Machine(SVM) . The Result shows that gini index method gives better performance with SVM [1]**.**

Tirath Prasad Sahu and Sanjeev Ahuja (2016), This paper analyzed the sentiment analysis on movie review by preprocessing of data ,then performed feature selection method applied and comparison of different classification techniques done. Highest accuracy was given by random forest with an accuracy of 88.95% [2].

Prerna,Soujanya Poria, ErikCambria (2015),have implemented a system for sentiment analysis by combining a Rule-based Classifier with Supervised Learning. The rule-based classifier is based on rules that are dependent on the occurrences of emoticons and opinion words in tweets. Whereas, the Support Vector Machine (SVM) is trained on semantic, dependency, and sentiment lexicon based features. The tweets are classified as positive, negative or unknown by the rule-based classifier, and as positive, negative or neutral by the SVM [3].

BogdonBatrinca, Philip C. Tr.eleaven, (2014),states 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 [4]**.**

III. PRELIMINARIES

Applying different mining techniques to derive usefulness about stored information. Specific mining approaches are classification, clustering, statistical evaluation, traditional language processing and so on. In text analytics, most likely classification process is used. Classification is a supervised studying procedure that helps in assigning a class label to an unclassified tuple in keeping with an already classified example set.

Naïve Bayes Classifier

The Naive Bayes classifier is a simple one for calculating the probabilities that is based on Bayes theorem that have strong and naïve independency.

Different types of Naive Bayes Variation

There are several number of variations for Naive Bayes i.e. (i) Multinomial, (ii) the BinarizedMultinomial and (iii) the Bernoulli.

i) Multinomial - It is used to refer when the number i.e. count of words means a lot in the technique for doing classification. Example for this type is when need for performing Topic Classification.
ii) The Binarized - It is referred when the occurrences of the words don't play an important role in classification. For example, in Sentiment Analysis, where it is no need as how much times we use the word "good" not considering the fact that he does.
iii) The Bernoulli - It is referred to use when the problem of not having the particular word matters a lot.

Naive Bayes classifier use to calculate each word probability

a. Estimating the probability P(c) for each class c i.e. positive, negative and neutral by dividing the total number of words in documents in c by the total number of words in the corpus.
b. Estimating the probability distribution P (w | c) for all words w and classes c where w is the number of tokens. This can be calculated by dividing the number of tokens of the words w in documents in c by the total number of words in c.
c. For scoring a document doc for class c, calculate

$$P(X|C_i) = \prod_{k=1}^{n} P(x_k|C_i) = P(x_1|C_i) * P(x_2|C_i) * \ldots\ldots\ldots * P(x_n|C_i)$$

d. To predict the most likely class label, then you can just pick the class c with the highest scoring value. To calculate the probability distribution, use equation:

For testing purpose, cross validation method is used. Cross-validation, it is also referred to as rotation estimation, is a model validating technique to assess how the results of a statistical analysis will simplify to an independent data set [17]. The main goal of cross validation is to define a dataset to "test" the data model in the training period i.e. in the validation dataset, so that to reduce the problem of over fitting, giving an insight on how the model will simplify to an independent dataset.

Two types of cross validation methods are there:

i) Exhaustive: These are the methods which do learning and testing on all possible ways so as to divide the original sample into two parts i.e. (i) training and (ii) a validation set. It includes two type:
    a) Leave-p-out cross validation
    b) Leave one out cross validation.
ii) Non-exhaustive: These are the methods that don't learning and testing on all possible of splitting the original sample. It also includes two types:
    a) K-fold cross validation
    b) 2-fold cross validation

IV. ANALYTICAL APPLICATION:

It provides valuable things from text mining so that it can provide information that helps in improving decision and processes. It includes following ways such as sentiment analysis, document imaging, fraud analysis etc.

V. PROPOSED APPROACH

1. Generating Dataset

Datasets were collected from Online Review Dataset. The online review dataset consists of around 800 user's reviews archived on the IMDB (Internet Movie Database) portal. And for, Twitter dataset around 1000 review

were collected and each review were formatted according to. arff file where review text and class label are only two attributes.

2. Collection of raw data and then apply filtering techniques to make that raw data into structured format. For doing the classification, Text pre-processing and feature extraction is a preliminary phase. Pre-processing involves 3 steps:

a) 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.
b) Removal of stop words: Stop words are the words that contain little information so needed to be removed. As by removing them, performance increases.
c) 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.

3. Applying the Correlation based feature selection algorithm on collected data. CFS is correlation based feature selection. Feature selection method try to pick up subsets of features (small) that are relevant to the target concept.

4. Classifying the clustered data by using Balanced Random Forest

5. Analyse the performance parameters like FP rate, TP rate, Recall, Precision of existing algorithm and new proposed algorithm then Compare the results of both.

a) Accuracy: It is measured as the proportion of correctly classified instances to the total number of instances being evaluated.
b) Precision: It is defined as ratio of the number of correctly labelled as positive to the total number that has been classified as positive.
c) Recall:It is defined as the ratio of the number of correctly labelled as positive to the total number that are truly positive.
d) F-measure: It is referred as the harmonic mean of precision and recall. It helps to give score needed to balance between precision and recall.

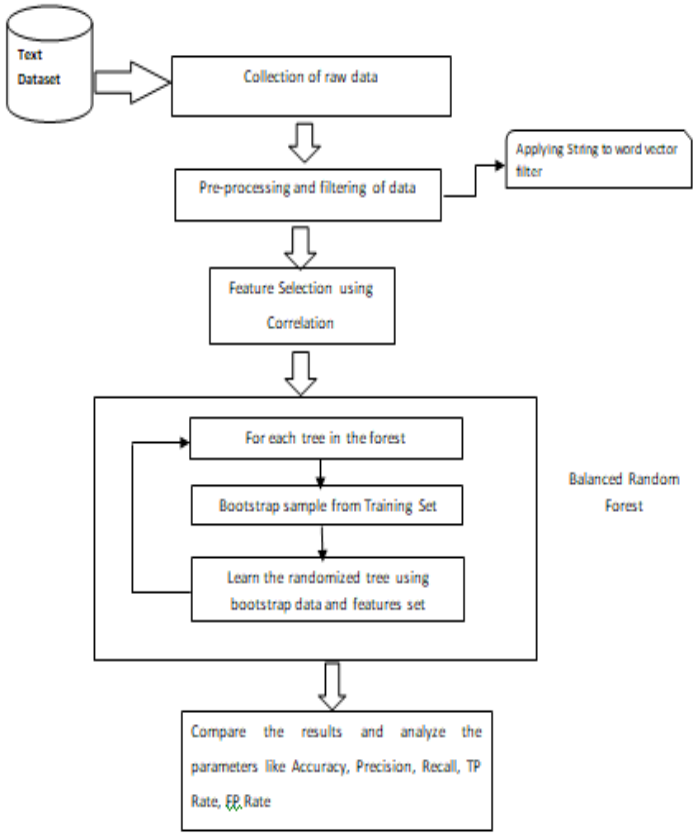

Fig 1. Flowchart of Proposed Methodology

VI. RESULTS AND DISCUSSION

The proposed methodologies implemented with the help of Weka and Net beansIDE8.0.Weka is the library that provides the simulation environment of data mining and also provide primary classes for evaluating the classification models.

Table 1: IG with CFS, Balanced RF and CFS with Balanced RF comparison on Movies dataset

Parameters /Algorithms	IG with CFS	Balanced RF	CFS with Balanced RF
Accuracy	74.717	78.4906	91.5723
Precision	0.724	0.753	0.915
Recall	0.747	0.785	0.916
F-Measure	0.733	0.76	0.913
TP Rate	0.747	0.785	0.916
FP Rate	0.219	0.221	0.08
Kappa Statistics	0.5134	0.5697	0.8361

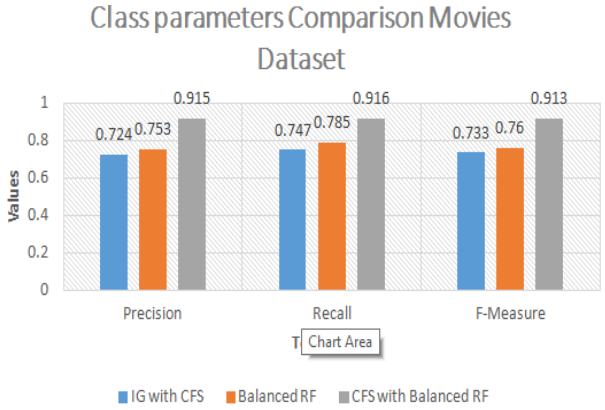

Fig 2. Class parameters comparison of movies dataset

Table 2: IG with CFS, Balanced RF and CFS with Balanced RF comparison on various News articles **dataset**

Parameters /Algorithms	IG with CFS	Balanced RF	CFS with Balanced RF
Accuracy	75.2094	76.7169	79.397
Precision	0.824	0.825	0.841
Recall	0.761	0.767	0.794
F-Measure	0.733	0.776	0.801
TP Rate	0.752	0.767	0.794
FP Rate	0.104	0.095	0.087
Kappa Statistics	0.6266	0.6495	0.6812

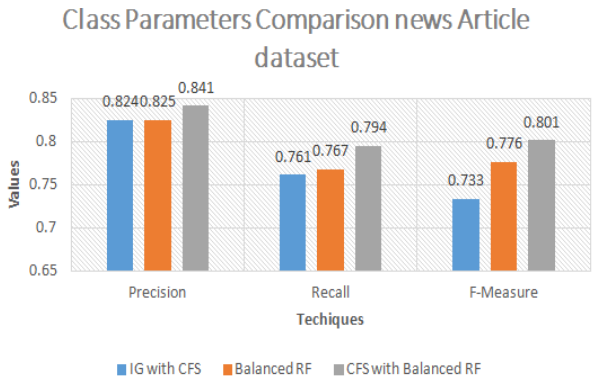

Fig 3. Class parameters comparison of new articles dataset

Table 2: IG with CFS, Balanced RF and CFS with Balanced RF comparison on various Tweets dataset

Parameters /Algorithms	IG with CFS	Balanced RF	CFS with Balanced RF
Accuracy	72.9927	74.717	83.9416
Precision	0.693	0.724	0.835

Recall	0.73	0.747	0.839
F-Measure	0.697	0.733	0.836
TP Rate	0.73	0.747	0.839
FP Rate	0.469	0.219	0.208
Kappa Statistics	0.2963	0.5134	0.6464

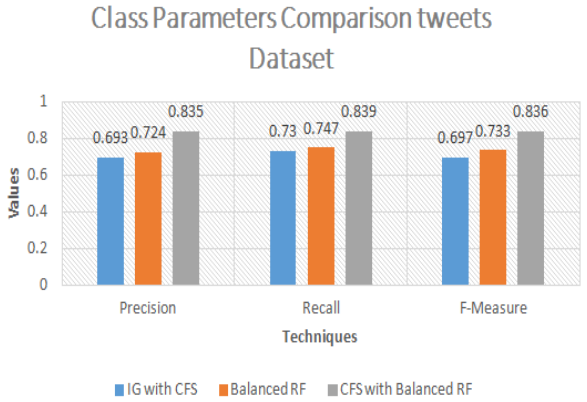

Fig 4. Showing the class parameters comparison of tweets dataset

VI. CONCLUSION

Social media monitoring has been growing very rapidly so there is a need for various organizations to analyse customer behaviour or attitude of particular product or any movie review. So, the concepts of sentiment analysis have been introduced. Text analytics and sentiment analysis can help organization to derive valuable business insights. Attitude can be calculated based on polarity check. Sentiment analysis refers to a broad range of fields of natural language processing, computational linguistics, and text mining. Sentiment classification of reviews and comments has merged as the most useful application in the area of sentiment analysis. Bag of words and feature based sentiment are the most popular approaches used by researchers to deal with sentiment analysis of opinions about products such as movies etc. In this, level sentiment analysis is considering three classes for sentiment polarity of each sentence (positive, neutral and negative). Each class prediction and classification is done by algorithm in terms of accuracy, precision, recall, TP rate, FP rate, F-measure, ROC area. Also, the comparison of proposed CFS based balanced random forest with the existing Information gain based random forest is done on the basis of accuracy or the correctly classified instances. Proposed technique performs better than IG based random forest with accuracy 91.57% in case of movies dataset and 83.94 % in case of tweets dataset.

VII. ACKNOWLEDGMENT

This research paper has been composed with the kind assistance, guidance and support of my department, computer Science who have helped me in this work. I would like to thank all the people whose encouragement and support has made the fulfilment of this work conceivable. I would like to thank my guide, S.Parminder Singh for the able guidance throughout the research work.

REFERENCES

[1] Asha S Manek, P Deepa Shenoy,M Chandra Mohan and Venugopal K R, "Aspect term extraction for sentiment analysis in large movie reviews using Gini Index feature selection method and SVM classifier", Springer, pg 135–154, Feb 04,2016.
[2] Tirath Prasad Sahu and Sanjeev Ahuja, "Sentiment Analysis of Movie Reviews: A study on Feature Selection & Classification Algorithms", IEEE Xplore , 28 July, 2016.
[3] Prerna Chikersal, Soujanya Poria, Erik Cambria (2015). "SeNTU: Sentiment Analysis of Tweets by Combining a Rule-based Classifier with Supervised Learning", Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval), 647–651.
[4] BogdonBatrinca, Philip C. Treleaven (2014) "Social media analytics: a survey of techniques, tools and platform",Department of Computer Science, Gower Street, London,UK published in Springer.
[5] Ana Mihanovic, HrvojeGabelica, ZivkoKrstic (2014) "Big Data and Sentiment Analysis using Knime: Online Reviews Vs. Social Media", MIPRO Opatija, Croatia
[6] HarunaIsah, Paul Trundle, Daniel Neagu. (2014)"Social media Analysis for product Safety using Text mining and Sentiment Analysis", Artificial Intelligence Research Group, University of Bradford, UK, IEEE.
[7] Mrs. R.Nithya, Dr. D.Maheshwari. (2014) "Sentiment Analysis on Unstructured Review", International Conference on Intelligent Computing Application, IEEE, pp. 367-371, March 2014.
[8] Lukasz Augustyaniak, Tomasz Kajdanowicz, PrzemyslawKazienko, MarcinKulisiewicz, WlodzimierzTuliglowicz, "An Approach to Sentiment Analysis of Movie Reviews: Lexicon Based vs. Classification", Springer, Vol. 8480, pp. 168-178, 2014.
[9] NehaNehra, "A Survey on Sentiment Analysis of Movie Reviews", International Journal of Innovative Research in Technology, Vol. 1, Issue 7, 2014.
[10] BasantAgarwal, Narmita Mittal, Erik Cambria. (2013) "Enhancing Sentiment Classification Performance using Bi-tagged

Phrases", 13th International Conference on DataMining Workshops, IEEE.
[11] Matthew Smith, Christian Szongott, Benjamin Henne, Gabriele von Voigt (2013) "Big Data Privacy Issues in Public Social Media", Distributed Computing & Security Group, Leibniz Universitat Hannover, Thailand, Germany IEEE.
[12] V.K. Singh, R.Piryani, A. Uddin, P.Waila. (2013) "Sentiment Analysis of Movie Reviews", Department of Computer Science, New Delhi, India, Published in IEEE.
[13] 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.
[14] 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.
[15] 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.
[16] NargizaBekmamedova, Graeme Shanks (2013) "Social Media Analytics and Business Value: A Theoretical Framework and Case Study", 2014 47th Hawaii International Conference on System Sciences (HICSS), pp. 3728-3737, January 2014.
[17] SimonaVinerean, IulianaCetina (2013) "The Effects of Social Media Marketing on Online Consumer Behavior", International Journal of Business and Management; Vol. 8, No. 14.
[18] V. S. Jagtap, KarishmaPawar. (2013) "Analysis of different approaches to Sentence-Level Sentiment Classification", International Journal of Scientific Engineering and Technology, Vol. 2, Issue 3, pp. 164-170, April 2013.
[19] Ya-Ting Chang, Shih-Wei Sun (2013) "A Real time Interactive Visualization System for Knowledge Transfer from Social Media in a Big Data",2013 9th International Conference on Information, Communications and Signal Processing (ICICS) ,IEEE.
[20] DeptiiD.Chaudhri, R.A. Deshmukh. (2012) "Feature –based Approach for Review mining Using Appraisal Words", Department of Post Graduate Computer Engineering, Pune, India, IEEE.
[21] Kristin Glass and Richard Colbaugh (2012) "Estimating the sentiment of social media content for security informatics applications", Institute for Complex Additive System Analysis, Socorro, USA, Springer Journal.
[22] SitaramAsur, Bernardo A.Huberma (2012) "Predicting the Future with Social Media",IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), Vol. 1, pp. 492-499, 2012..
[23] MostafaKaramibekr, Ali A. Ghorbani, "Sentiment Analysis of Social Issues", 2012 International Conference on Social Informatics, IEEE.
[24] Jiao Wu, Bin Zhang, WeihuaGao, Yi Hu, Jinsong Liu (2011),"Online Web Sentiment Analysis on Campus Network", 4th International Symposium on Computational Intelligence and Design, IEEE, Vol. 2, pp. 379-382, October 2011.
[25] Vivek Kumar Singh, Mousumi Mukherjee, Ghanshyam Kumar Mehta, "Combining a Content Filtering Heuristic and Sentiment Analysis for Movie Recommendations", Springer, 2011.
[26] LingyanJi, Hanxiao Shi, Mengli, MengxiaCai, PeiqiFeng. (2010) "Opinion Mining of Product reviews based on semantic role labeling", 5th International Conference on Computer Science and Education, IEEE, pp. 1450-1453, August 2010.
[27] Ms. K. Mouthami, Ms. K.Nirmala Devi, Dr. V.MuraliBhaskaran. (2010) "Sentiment Analysis and Classification based on Textual Reviews",Dept of CSE, Tamil Nadu, IEEE.