Proceedings of the Third International Conference on Computing Methodologies and Communication (ICCMC 2019)
IEEE Xplore Part Number: CFP19K25-ART; ISBN: 978-1-5386-7808-4

Implementation of Sentiment Classification of Movie Reviews by Supervised Machine Learning Approaches

Tejaswini M. Untawale[1], Prof. G. Choudhari[2]

[1]PG Scholar, Department of Computer Technology, Yeshwantrao Chavan College of Engineering, Nagpur, Maharashtra, India.

[2]Assistant Professor, Department of Computer Technology, Yeshwantrao Chavan College of Engineering, Nagpur, Maharashtra, India.

Abstract— **Entertainment is crucial part of human life entertainments like songs, music, drama and movies etc. So for watching good movies most of the peoples generally prefer theater. If movie is not good then we feel nervous and we think wasted our money and time for watching bad movie so people prefer to go to movie by reading reviews and rating for that movie on various apps like IMDb, flixter and voice etc. But by reading one or two reviews we cannot say movie is good or bad because different peoples have different opinions some peoples like action or some like thrill or romance so people gives reviews based on their area of interest. So we cannot predict movie for that we proposed movie reviews based on sentiment analysis and classification algorithms such as Naïve bays and Random forest (RF). Sentiment analysis generally utilized to identify the sentiment of huge amount of text. We compare naïve bayes and RF machine learning techniques for measuring negative, positive and neutral reviews.**

Index Terms: Sentiment analysis, Naïve Bayes algorithm, Random Forest algorithm.

I. INTRODUCTION

The enhancement in the field of web technology has changed the manner by which individuals can express their perspectives. Individuals rely on this user perspective information for analyzing the items for online shopping or while booking film tickets for watching movies in theaters. The users are interfacing together through posts, Facebook, tweets on twitter etc. The measure of information is huge to the point that it is troublesome for a typical human to examine and come to conclusion. Sentiment analysis is extensively arranged in the two kinds initial one is an information based methodology and the other classification techniques. First methodology requires an expansive database of predefined feelings and a proficient information portrayal for recognizing sentiments. Then again the Machine learning approach makes utilization of a datasets and a test informational collection to build up a classifier. It is preferably more straightforward over Knowledge base methodology. Since the improvement of calculations a few difficulties were looked in the field of Sentiment analysis. The first is that a sentiment word can be sure or negative contingent on the circumstance. The second test is that individuals don't in every case express conclusions similarly. Sentiment mining comprehends the connection between literary audits and the outcomes of those reviews.

Sentimental analysis can be utilized to differentiate customers and followers depends on their attitude towards a specific brand or a movie or a product with the help of reviews. One can identify whether the product review is positive or negative or whether the user email is satisfied or not.

Feature Extraction categorized into four types Syntactic Feature, Semantic Feature, Link based Feature, Stylistic Feature. The most commonly utilized features are the first two features. Syntactic feature utilizes word tags, patterns, phrases and punctuations. On the other hand, Semantic feature works on the relationship between words, signs and symbols. Linguistic semantics can be utilized to know the human expression through language accurately.

Classification is also known as “Supervised learning”. Linear Classifiers: Logistic Regression/Naive Bayes Classifier, Support Vector Machines, Decision Trees, Random Forest, Neural Networks are classification algorithms in Machine Learning

The section I explains the Introduction of movie review using classification method such as NB and RF. Section II presents the literature review of existing systems and Section III present proposed system implementation details Section IV presents experimental analysis, results and discussion of proposed system. Section V concludes our proposed system. While at the end list of references paper are presented.

II. LITERATURE REVIEW

Sentiment analysis has been done for movie Review, Twitter and Gold dataset utilizing optimized SVM. Author used two machine leaning techniques for analyzing sentiments that is SVM and Naïve bayes. Also author utilized lexicon approach to convert structured review into numerical score value.

Here Liu, B [2] discussed sentimental analysis applications and its problems also presented types of sentimental analysis, also two relevant and important concepts of subjectivity and emotion were also introduced, which are highly related to but not equivalent to opinion.

Based on textual review Mouthami, K., Devi, K.N. and Bhaskaran [5] did classification and sentimental analysis. First they used mining techniques to extract the features from huge data and sentimental analysis utilized to analyze the opinions or emotions of users from text data this analysis is widely adopted in CRM. Analyzing sentiment using Multi-theme document is very difficult and the accuracy in the classification is less. So proposed a new algorithm called Sentiment Fuzzy Classification algorithm with parts of speech tags is utilized to improve the classification accuracy on of Movies reviews dataset.

Here Kanakaraj, M. and Guddeti [6] analyzing the mood, emotions of the society on a particular news from Twitter posts. The key point of this is to enhance the accuracy of classification by including Natural Language Processing Techniques (NLP) especially semantics and Word Sense Disambiguation.

Author Chaovalit, P. and Zhou, L [7] examines movie review mining utilizing two methodologies that is machine leaning and semantic analysis. The methodologies are adjusted to movie review area for correlation. The outcomes demonstrate that our outcomes are equivalent to or surprisingly better than past discoveries. We likewise find that film audit mining is a more difficult application than numerous different sorts of review mining. The difficulties of film review mining lie in that authentic data is constantly blended with genuine audit information and amusing words are utilized recorded as a hard copy film review.

Text analysis important type are Sentiment analysis or opinion mining that aims to support decision making by extracting and analyzing opinion oriented text, finding positive and negative opinions, and estimating how positively or negatively an entity regarded. Most of the users express their political or religious views on Twitter so tweets become valuable sources of individual express. Tweets data can be efficiently utilized to infer people's opinions for social studies. Author proposes a Tweets Sentiment Analysis Model (TSAM) [8] that can spot the social interest and general people's opinions in a social event.

Gautam, G. and Yadav, D [9] analyze opinions or sentiments of the twitter data using machine learning approaches and semantic analysis. machine learning approaches such as Naive Bayes, Maximum entropy and SVM utilized to analyze the twits on twitter and lastly measured the performance of classifier in terms of recall, precision and accuracy.

By using unigram feature extraction technique twitter dataset is analyzed and utilized. Then after that, different machine learning techniques [10] trains the dataset with feature and then the semantic analysis gives a large set of similarity and synonyms which gives the polarity of the content. WordNet enhances the accuracy. As a part of preprocessing they have cleared ambiguous information and not required blank spaces. After preprocessing, this preprocessed data is converted into numerical vector using TF-IDF and Count Vectorizer. SVM and NB classifiers are utilized to classify numerical vector [11].

Zhang et al. [15] utilized rule based semantic analysis to classify the sentiment.

III. SYSTEM ARCHITECTURE

A. System Architecture

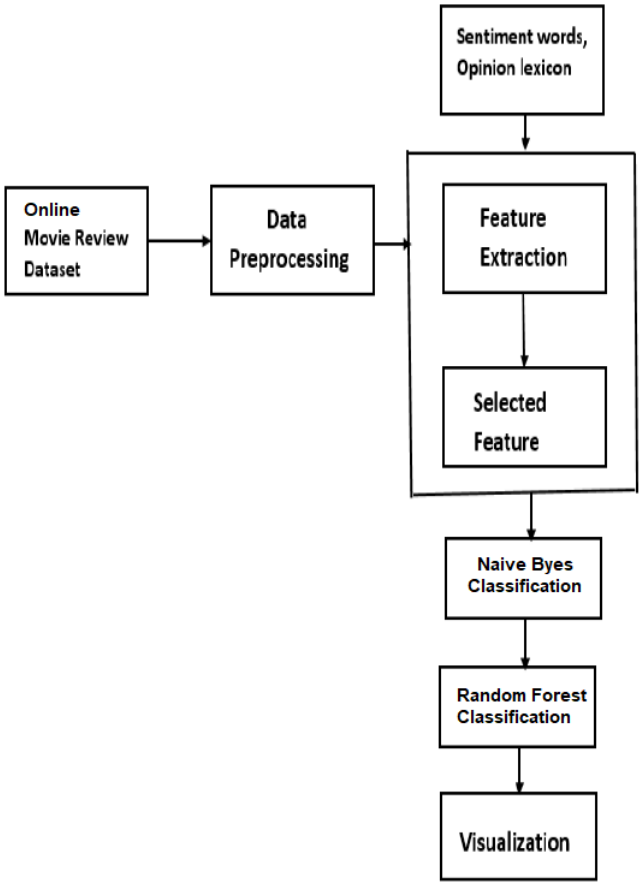

Fig 1. System Architecture

Initially we get online or offline dataset in the form of movie list. Then we select particular movie for review. Then we add review record. After that preprocessing is performed for feature extraction and selection Data needs to be cleaned

before it is processed. It includes removal of URL's, punctuations, symbols, emoticons, stemming and stop words. Then we classify that data using classifiers such as NB and RF. NB gives result as a status it takes first 20 reviews and classify them strongly positive, strongly negative, weekly positive, weekly negative and neutral review and by taking the average we get final result in terms of positive, negative and neutral. Then we use RF classification for sentiment analysis for that taken Polarity count i.e total count of positive, negative comments for analyzing total comments. Based on that training file is generated with id, positive/negative word count, polarity count and class as 0, 1, and 2. 0 indicates positive, 1 negative and 2 is neutral then RF classification performed to predict the class of reviews based on positive count, negative count and neutral count.

IV. RESULT AND DISCUSSIONS

A. Experimental Setup

I. All the experimental cases are implemented in Java in congestion with Netbeans tools and MySql as backend, algorithms and strategies, and the competing classification approach along with various feature extraction technique, and run in environment with System having configuration of Intel Core i5-6200U, 2.30 GHz Windows 10 (64 bit) machine with 8GB of RAM

II. Dataset Description: The data was taken from web, current movie reviews taken from different sites like Times of India, Rotten Tomatos, etc.

B. Results Comparison

This section presents the performance of the NB and RF algorithms. Fig 2 Shows Memory Comparison of NB and RF algorithms for various Threshold. X-axis shows Algorithm & Y-axis shows Memory in bytes. NB require more Memory than RF.

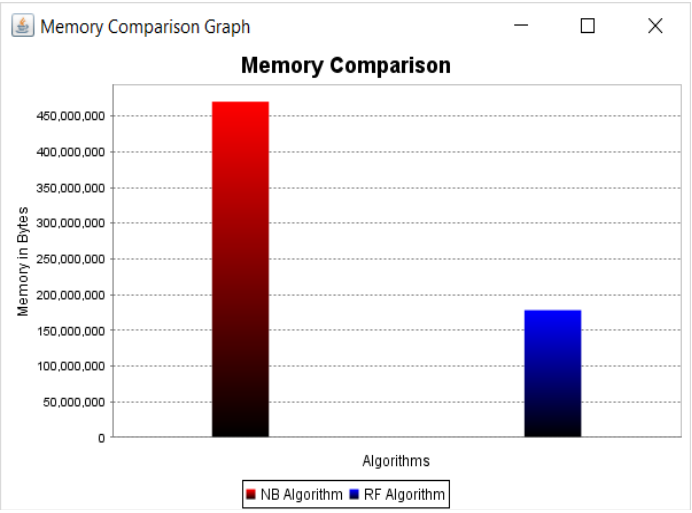

Fig. 2: Memory Comparison Graph

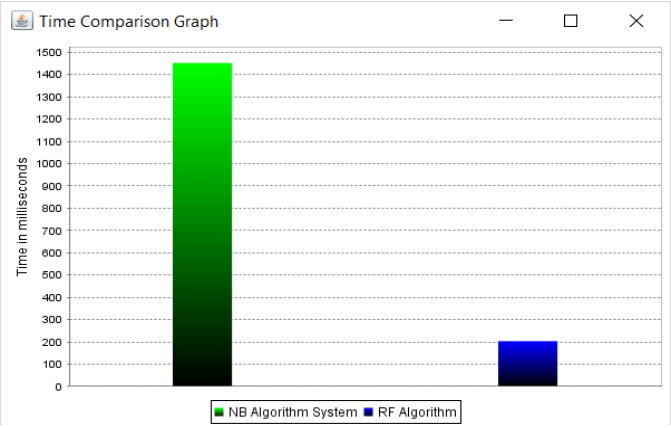

Fig. 3: Time Comparison Graph

Fig 3 shows the Time comparison of NB and RF algorithms for various size. The X-axis shows algorithms and Y- axis shows Time in ms. The RF takes less time than NB.

V. CONCLUSION

In proposed system NB and RF classification techniques and sentimental analysis are utilized that will provide interested movie reviews to web user. Generally sentimental analysis is termed as an opinion mining. It is utilized to identify user's emotions, mood, interest and behavior by using text pattern data. Classification techniques such as NB and RF utilized for feature selection and extraction, NB will not work properly in terms of execution time and it requires more memory. RF takes less time and less memory to execute than NB. Then we compare NB and RF on the basis of time and memory, results show that Random Forest Algorithm is better than Naive Bayes Algorithm in terms of time and memory to recommend the good movie to users.

REFERENCES

[1] Bohemia M. Jadav, Vimalkumar B. Vaghela," Sentiment Analysis using Support Vector Machine based on Feature Selection and Semantic Analysis", *International Journal of Computer Applications (0975 – 8887) Volume 146 – No.13, July 2016*

[2] Liu, B., 2012. Sentiment analysis and opinion mining. Synthesis lectures on human language technologies, 5(1), pp.1-167.

[3] Y. Singh, P. K. Bhatia, and O. Sangwan, "A review of studies on machine learning techniques," International

Journal of Computer Science and Security, vol. 1, no. 1, pp. 70–84, 2007.

[4] Singh, Y., Bhatia, P.K., and Sangwan, O. (2007) A Review of Studies on Machine Learning Techniques. International Journal of Computer Science and Security, 1, 70-84.

[5] Mouthami, K., Devi, K.N. and Bhaskaran, V.M., 2013, February. Sentiment analysis and classification based on textual reviews. In

Information Communication and Embedded Systems (ICICES), 2013 International Conference on (pp. 271-276). IEEE.

[6] Kanakaraj, M. and Guddeti, R.M.R., 2015, February. Performance analysis of Ensemble methods on Twitter sentiment analysis using NLP techniques. In Semantic Computing (ICSC), 2015 IEEE International Conference on (pp. 169-170). IEEE.

[7] Chaovalit, P. and Zhou, L., 2005, January. Movie review mining: A comparison between supervised and unsupervised classification approaches. In System Sciences, 2005. HICSS'05. Proceedings of the 38th Annual Hawaii International Conference on (pp. 112c-112c). IEEE.

[8] Zhou, X., Tao, X., Yong, J. and Yang, Z., 2013, June. Sentiment analysis on tweets for social events. In Computer Supported Cooperative Work in Design (CSCWD), 2013 IEEE 17th International Conference on (pp. 557-562). IEEE.

[9] Gautam, G. and Yadav, D., 2014, August. Sentiment analysis of twitter data using machine learning approaches and semantic analysis. In Contemporary Computing (IC3), 2014 Seventh International Conference on (pp. 437-442). IEEE.

[10] Tripathy, A., Agrawal, A. and Rath, S.K., 2015. Classification of Sentimental Reviews Using Machine Learning Techniques. Procedia Computer Science,57, pp.821-829.

[11] Miller, G.A., 1995. WordNet: a lexical database for English. Communications of the ACM, 38(11), pp.39 41.

[12] Ronen Feldman," Techniques and Applications for Sentiment Analysis", Communications of the ACM, Vol. 56 No. 4, pp. 82-89, April 2013.

[13] Tsytsarau Mikalai, Palpanas Themis. Survey on mining subjective data on the web. Data Mining and Knowledge Discovery, Vol. 24, pp. 478–514, May 2012.

[14] Turney, P, "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews", ACL '02 Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Vol.21. No.4, pp. 417-424, July 2002.

[15] C. Zhang, D. Zeng, J. Li, F.-Y. Wang, and W. Zuo, "Sentiment Analysis of Chinese Documents: From Sentence to Document Level", Journal of the American Society for Information Science and Technology, vol. 60, No. 12, pp. 2474-2487, December 2009.

[16] Poirier, D., Bothorel, C., De Neef, E. G., &Boullé, M. Automating opinion analysis in film reviews: the case of statistic versus linguistic approach. In Affective Computing and Sentiment Analysis, Springer Netherlands,Vol. 45,pp. 125-140, July 2011.

[17] Ranjani Gandhi, V, Priya. N," Literature Survey on Data Mining and Statistical Report for Drugs Reviews", IJIRCCE, Vol. 3 Issue 3, pp. 1734-1739, March 2015.

[18] Richa Sharma, Shweta Nigam, Rekha Jain, "Opinion Mining of Movie Reviews at Document level", International Journal on Information Theory (IJIT), Vol.3, No.3, pp. 13-21, July 2014.

[19] B. Pang and L. Lee, "Opinion Mining and Sentiment Analysis", Foundations and Trends in Information Retrieval, vol. 2, nos. 1/2, pp. 1-135, January 2008.

[20] Liu, C. L., Hsaio, W. H., Lee, C. H., Lu, G. C., &Jou," E. Movie rating and review summarization in mobile environment", IEEE Transactions on, Systems, Man & Cybernetics: Part C - Applications & Reviews, Vol 42, pp. 397-407,2012.