iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** XII **Month of publication:** December 2021

DOI: https://doi.org/10.22214/ijraset.2021.39404

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue XII Dec 2021- Available at www.ijraset.com

An Enhanced Approach for Sentiment Analysis Using Association Rule Mining

Abhishek Sharma[1], Anita Ganpati[2]
[1]*M.Tech Scholar,* [2]*Professor, Department of Computer Science, Himachal Pradesh University, Shimla*

Abstract: ***In today's world social networking platforms like Facebook, YouTube, twitter etc. are a great source of communication for internet users and loaded with large number of emotions, views and opinions of the people. Sentiment analysis is the study of attitudes, emotions and opinions of the people and is also known as opinion mining. Sentiment analysis is used to find the opinion i.e. negative or positive about a particular subject. In this paper an Enhanced sentiment analysis approach is presented by using the Association rule mining i.e. Apriori and machine learning approach such as Support Vector Machine. The Enhanced approach is compared with the baseline approach, on accuracy, precision, recall, and F1-score measures. The Enhanced approach for sentiment analysis is implemented using the R programming language. The Enhanced approach shows better performance in comparison to the baseline approach.***
Keyword: Sentiment Analysis, Opinion Mining, Support Vector Machine, Association Rule Mining, Machine Learning

I. INTRODUCTION

Any review, opinion given by any person, through which the attitude, thoughts and feelings can be told is known as sentiment. Sentiment analysis is the analysis of the data attained from user reviews, comments, news reports and microblogging sites. "Sentiment analysis, is also called opinion mining, is the area of study that analyse people's review, sentiment, feelings, emotions, and attitudes towards entities and their attributes conveyed in written content" [1]. Many related names like opinion mining, sentiment mining, review mining, opinion extraction and emotion analysis are comes under the umbrella of the sentiment analysis. Sentiment analysis is very helpful in many applications like from identifying people's opinion, to monitor the mental health of a patient based on the patient's posts on social media platforms. Common application areas of the sentiment analysis are Government intelligence, Business intelligence, Healthcare and medical domain and recommendation system etc.

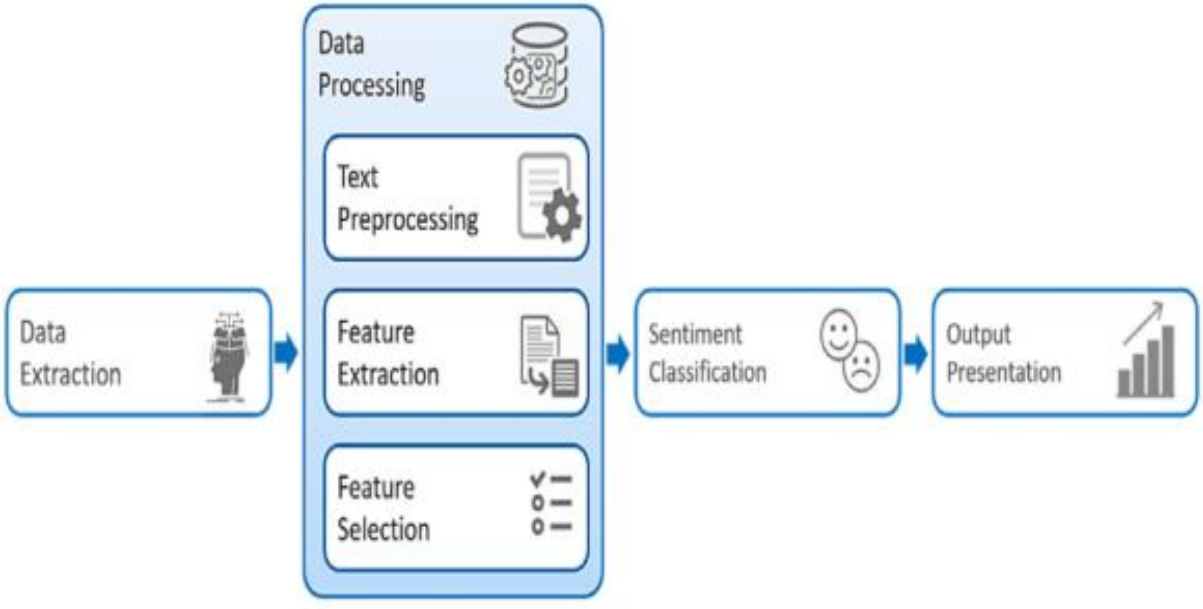


Figure 1: The process of sentiment analysis [2]

Some of the general steps of the sentiment analysis process that can be used to analyze text for sentiments or opinions is shown in Figure 1.

A. Sentiment Analysis Approaches

Sentiment analysis is a lively and fast-growing research field and it can be used in many domains. Figure 2 outlines the different approaches used for performing sentiment analysis.

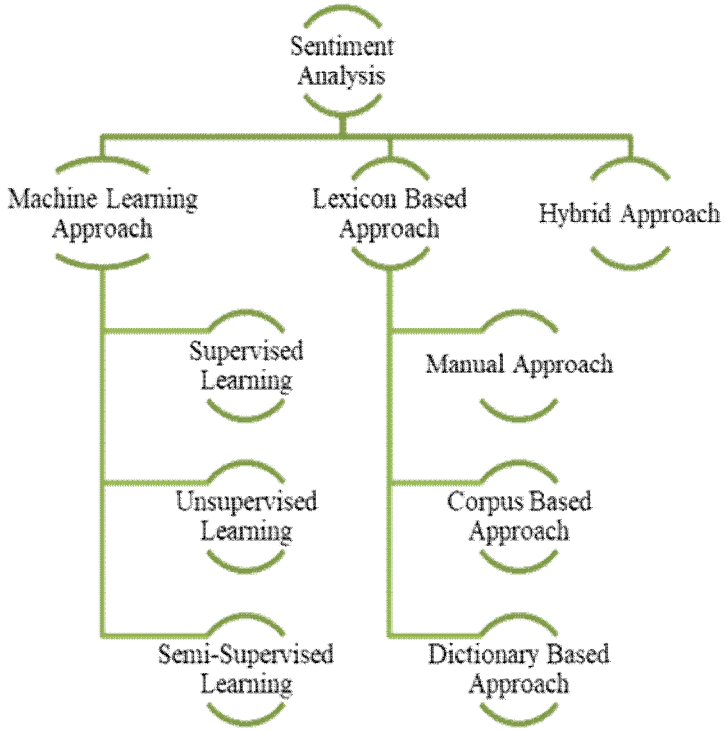


Figure 2: Classification of Sentiment Analysis Approaches.

1) *Machine Learning Approach:* Machine learning approach is one of the most widely used approach for sentiment analysis and it relies on the use of algorithms like Support Vector Machine (SVM), Naïve Bayes (NB), Random forest (RF) etc. It uses the classification techniques in order to classify the text into negative or positive sentiment polarity [14].
2) *Lexicon Based Approach:* Lexicon based approach is also known as knowledge based approach, is one of the two most widely used approaches for sentiment analysis and uses a built-in dictionary to find the polarity of the comments, reviews. Lexicon based approach use sentiment dictionary that represents a list of phases and words to match them with the data and to express the negative or positive sentiment. Some of the popular lexicons for sentiment analysis are: Bing Liu's lexicon, AFINN lexicon, VADER lexicon etc. [15].
3) *Hybrid Approach*: Hybrid approach combines lexicon approach and the machine learning approach and has the capability to improve the overall performance of the sentiment analysis. Hybrid approach combine the speed of the lexicon approach with the mobility of the machine learning approach [2].

B. Association rule Mining

In association rule mining the term association rule represents the connection between two or more items that are frequently occurring in the database. Association rules helps the market analyst to find the connection between the items that are most commonly bought by the customers. Like eggs and milk are purchased together. Some of the most widely used algorithms for finding the association rules are Frequent Pattern (FP)-Growth, Equivalent Class Transformation (Eclat) and Apriori algorithms. Apriori algorithm is one of the easiest and most widely used algorithm for discovering the association rules.

For any association rule to be valuable it must satisfy three measure:

1) *Support:* Number of times item A and item B are occurring together divided by the total number of transactions.
2) *Confidence:* It is the ratio of Number of times item A and item B are occurring together to the number of times item A is occurring.
3) *Lift:* It is the ratio of confidence of rule (A →B) to the number of times item B is occurring [3].

II. LITERATURE REVIEW

Thakare Ketan Lalji et al. [4] presented hybrid approach for the sentiment analysis of the twitter data. They first detect the polarity of the words by using MPQA (Multi-Perspective Question Answering) lexicon dictionary and then apply the results to machine learning algorithms i.e. SVM. They use accuracy measure to evaluate the performance for features like Unigram, Bigram, Trigram for different size of training data. Marouane Birjali et al. [2] presented a survey on sentiment analysis and also discussed their approaches and challenges. They conclude that because of simplicity, algorithms of supervised machine learning are mostly preferred for sentiment analysis. Indrajeet Kaur Chhabra et al. [5] proposed a hybrid approach in which SentiWordNet lexicon is used to discover the polarity of the words and then linear SVM classifier is used to the classify the reviews. Dipti Sharma et al. [6] presented a review of various methods that are used for the sentiment analysis, and conclude that for sentiment classification NB and SVM algorithms are most frequently used. Rajkumar S. Jagdale et al. [7] performed the sentiment analysis of the amazon product reviews by using NB and SVM and conclude that for camera reviews SVM achieves accuracy of 93%. Binita Verma et al. [8] proposed a model for predicting the sentiments of movie reviews using machine learning approach. They used Logistic regression and SVM classifiers in which SVM achieves maximum accuracy of 91%. S. A. El Rahman et al. [9] proposed a model for sentiment analysis of the twitter data by combining the unsupervised and supervised models. For extracting tweets on KFC and McDonald's they use R programming language and conclude that maximum entropy achieves highest accuracy than NB, SVM and other algorithms. Jyotsna Anthal et al. [10] performed the sentiment analysis on tweets, for Ola and Uber using R Programming language. For comparison of accuracies NB and SVM algorithms is used and conclude that in both cases NB is the dominating algorithm for classifying the Ola & Uber datasets. Nitika Nigam et al. [11] performed the twitter sentiment analysis using lexicon approach and classify the tweets into negative or positive using external dictionary. A. Mukwazvure et al. [12] proposed a hybrid approach for analysing the sentiments of the news comments. They used AFINN lexicon to assign polarity to the comments and then uses SVM and (KNN) K-Nearest Neighbour classifiers and conclude that SVM outperforms KNN for news comments.

III. RESEARCH METHODOLOGY

The study was based on observation of several resources in which firstly, various websites, surveys and research papers are studied related to sentiment analysis and Secondly, an Enhanced approach using association rule mining is proposed which is implemented using R programming language. The dataset used is STS-Gold dataset [13] [19], which was taken from Kaggle repository [20]. This dataset was used to compare the Enhanced approach using association rule mining, with the baseline approach i.e. svm alone used for sentiment analysis.

A. Tool and parameter used for evaluation

The RStudio an open source software is used, which is an IDE for R programming language and makes R easier to use. To evaluate and compare the performance of the two approaches, Accuracy, precision, recall, and F1-score measure is used, which are most commonly used evaluation metrics for sentiment analysis [2].

B. Proposed Enhanced Approach

The Enhanced approach for sentiment analysis is a combination of, association rule and machine learning approaches, hence utilizing the best characteristics of two in one. For this purpose, we choose support vector machine, along with the simplest and widely used association rule mining algorithm i.e. Apriori algorithm. This Enhanced approach is compared with the baseline machine learning based approach i.e. support vector machine alone. The Enhanced approach follows the four basic steps namely data pre-processing, generating association rule by using Apriori algorithm, extracting tweets for each rule and last step is using machine learning classifier such as SVM to classify the text and produce results.

The Enhanced Sentiment analysis approach shown in Figure 3 is summarized as follows:

1) *Data Pre-processing:* The opinion sentences contain URL, annotation "@", hashtags "#", numbers, stop words, which does not contain any sentiment. Removing them is very important to enhance the performance and it easy to perform operation on cleaner data. It involves removing RT, @usernames, twitter hashtags, external web links, Unnecessary Space, stop words, punctuation, number, replace internet slang, replace word elongation and change to lower case.
2) *Association Rule Mining:* Association rules are generated by using support and confidence measure after reading the tweets in the basket format. By using Apriori algorithm, rules are generated. The generated rules are then filtered by using lift measure by the keeping the value of lift>1 [16]. Now removing the redundant rules from the filtered rules.

3) *Extracting Tweets for Each Rule:* From the non-redundant rules extract the tweets for each rule and then pass the data obtained, based on the non-redundant rules, to the machine learning classifier.

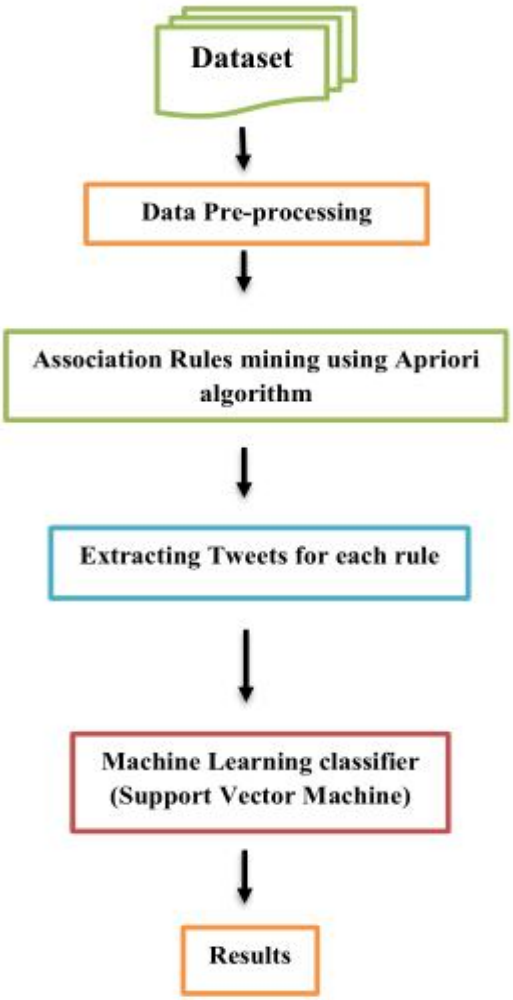


Figure 3: Enhanced approach for sentiment analysis

4) *Sentiment Classification:* Finally, supervised Machine Learning classifiers such as SVM is trained on the data obtained, based on the non-redundant rule. One of the most widely used classifier for sentiment analysis i.e. support vector machine were used to evaluate the performance of two approaches. Accuracy, precision, recall, and F1-score measure is used to evaluate the performance of Enhanced and baseline approaches.

IV. RESULT AND ANALYSIS

The paper aim to develop an Enhanced approach for sentiment analysis using association rules mining. The Enhanced approach for sentiment analysis using association rule mining is implemented and compared for performance along with the baseline approach. The Enhanced approach was implemented using the R programming language. The tool utilized is RStudio, which makes R easier to use. The Enhanced approach using association rule mining is designed such as association rules are discovered using Apriori algorithm where each word is considered as an item and then extracting the tweets for each rule. Association rules whose lift value is >1 is only selected as in those rules the antecedent and consequent are positively related. The high confidence value against the low support value will be selected as it generates a cohesive and acceptable number of rules [18]. Percentage split test method is used in which dataset is split into two parts based on the famous Pareto principle [17] where 80% is training data and 20% is testing data. The results are plotted using MS Excel 2016.

Table 1: Results of Enhanced and Baseline approach

Approaches	Accuracy	Precision	Recall	F1-score
Enhanced	0.8276	0.8678	0.8824	0.8750
Baseline	0.7753	0.8264	0.8530	0.8395

The computed values for the Enhanced approach and the Baseline approach is shown in Table 1. In Figure 4 vertical axis represents the different value and horizontal axis shows the different parameters. From the results it is evident that the Enhanced approach using association rule shows better results than the baseline machine learning based approach for accuracy, precision, recall and F1-score measure. The Enhanced approach achieves an accuracy of 82.76% while baseline approach achieves 77.53% accuracy.

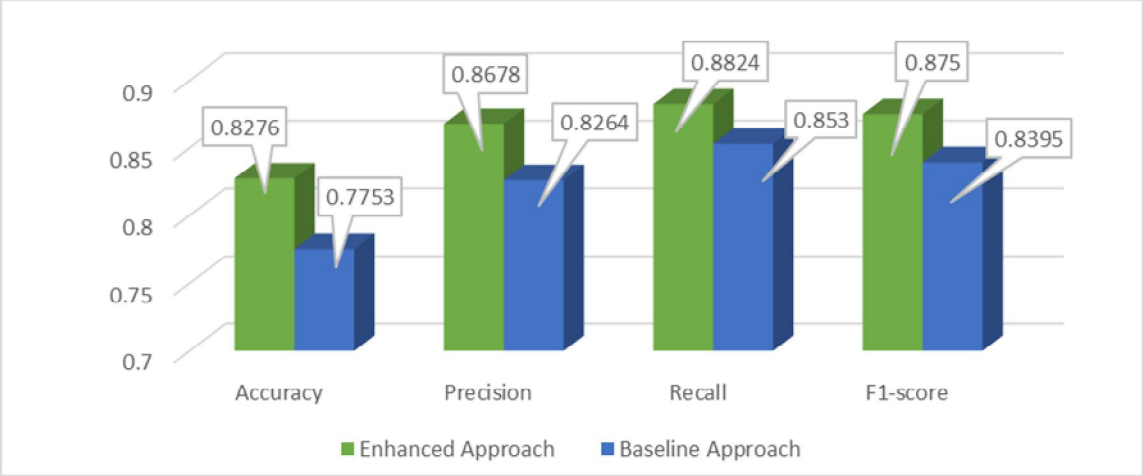


Figure 4: Comparison of Enhanced and Baseline approach

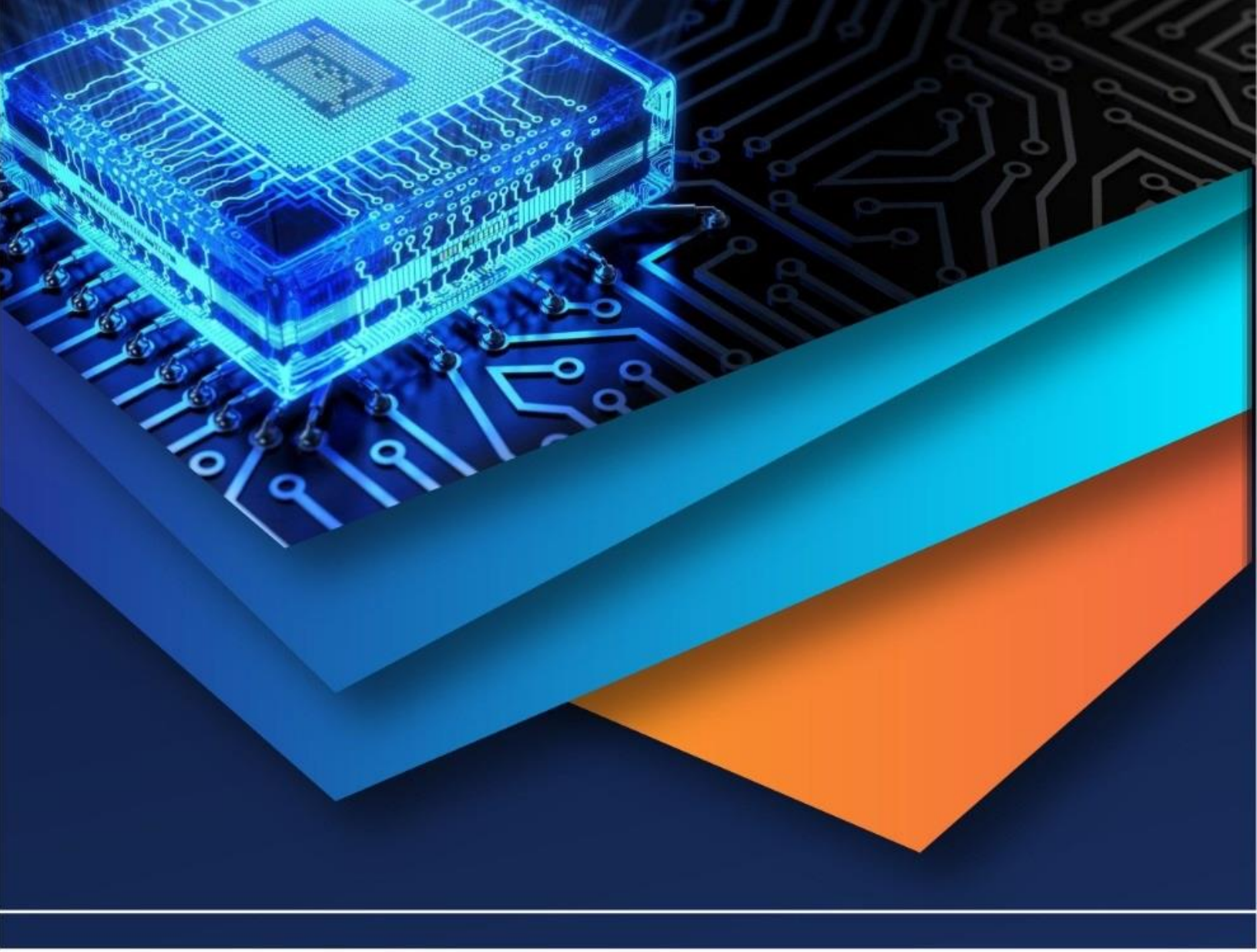
V. CONCLUSION AND FUTURE SCOPE

In sentiment analysis, the sentiments imply perception or opinion of person. In sentiment analysis the word sentiment usually represents the different feeling like anger, sadness, trust, fear, surprise, anticipation, disgust and joy. Sentiment analysis helps in understanding the perception of the people toward any particular topic, problem etc. In this paper an Enhanced approach for sentiment analysis using association rule mining is proposed and implemented using the R programming language. This Enhanced approach is compared with baseline approach of SVM alone. The comparison of approaches was done on accuracy, precision, recall and F1-score measure. The results indicate that the Enhanced approach outperforms the Baseline approach and shows an increased accuracy. In Future the Enhanced approach may be tested by using different machine learning algorithms.

REFERENCES

[1] Bing Liu, “Sentiment Analysis: Mining Opinions, Sentiments, and Emotions”, Cambridge University Press, 2015.
[2] Marouane Birjali, Mohammed Kasri and Abderrahim Beni-Hssane, “A comprehensive survey on sentiment analysis: Approaches, challenges and trends”, Knowledge-Based Systems, Volume 226, 17 August 2021.
[3] Abhishek Sharma and Anita Ganpati, “Association rule mining algorithms: A Comparative review”, International Research Journal of Engineering and Technology, Volume: 08 Issue: 11, Nov 2021.
[4] Thakare Ketan Lalji and Sachin N. Deshmukh, “Twitter Sentiment Analysis Using Hybrid Approach”, International Research Journal of Engineering and Technology, Volume: 03 Issue: 06, 2016.
[5] Indrajeet Kaur Chhabra and Gend Lal Prajapati, “Sentiment Analysis of Amazon Canon Camera Review using Hybrid Method”, International Journal of Computer Applications, Volume 182 – No.5, July 2018.
[6] Dipti Sharma, Munish Sabharwal, Vinay Goyal and Mohit Vij, “Sentiment Analysis Techniques for Social Media Data: A Review”, First International Conference on Sustainable Technologies for Computational Intelligence, Advances in Intelligent Systems and Computing, Volume 1045. Springer, Singapore, pp 75-90.
[7] Rajkumar S. Jagdale, Vishal S. Shirsat and Sachin N. Deshmukh, “Sentiment Analysis on Product Reviews Using Machine Learning Techniques”, Cognitive Informatics and Soft Computing, pp. 639-647, Springer, Singapore, 2019.
[8] Binita Verma and Ramjeevan Singh Thakur, “Predicting Sentiment from Movie Reviews Using Machine Learning Approach”, International Journal of Multidisciplinary Research in Science, Engineering and Technology, Volume 2, Issue 11, November 2019.

[9] S. A. El Rahman, F. A. AlOtaibi and W. A. AlShehri, "Sentiment Analysis of Twitter Data," International Conference on Computer and Information Sciences, pp. 1-4. IEEE, 2019.
[10] Jyotsna Anthal, Anand Upadhyay, Yash Indulkar and Abhijit Patil, "Twitter Sentimental Analysis & Algorithm Comparison for Uber & Ola Using 'R'", International Journal of Future Generation Communication and Networking Vol. 13, No. 1s, 2020, pp. 352—358.
[11] Nitika Nigam and Divakar Yadav, "Lexicon-based approach to sentiment analysis of tweets using R Language", International Conference on Advances in Computing and Data Sciences, pp. 154-164. Springer, Singapore, 2018.
[12] A. Mukwazvure and K. P. Supreethi, "A hybrid approach to sentiment analysis of news comments", 4th International Conference on Reliability, Infocom Technologies and Optimization, (Trends and Future Directions), pp. 1-6. IEEE, 2015.
[13] Hassan Saif, Miriam Fernandez, Yulan He and Harith Alani, "Evaluation Datasets for Twitter Sentiment Analysis A survey and a new dataset, the STS-Gold", 1st ESSEM Workshop, Turin, Italy 2013.
[14] Dipanjan Sarkar, Raghav Bali, and Tushar Sharma, "Practical Machine Learning with Python", Apress,2018.
[15] Alessia D'Andrea, Fernando Ferri, Patrizia Grifoni, and Tiziana Guzzo. "Approaches, tools and applications for sentiment analysis implementation", International Journal of Computer Applications 125, no. 3 2015.
[16] K. Rajeswari, "Feature Selection by Mining Optimized Association Rules based on Apriori Algorithm", International Journal of Computer Applications, Volume 119 – No.20, June 2015
[17] Shahzad Qaiser, Nooraini Yusoff and Farzana Kabir Ahmad, Ramsha Ali, "Sentiment Analysis of Impact of Technology on Employment from Text on Twitter", International Journal of Interactive Mobile Technologies 14, no. 7, 2020.
[18] J. A. Diaz-Garcia, M. D. Ruiz and M. J. Martin-Bautista, "Non-Query-Based Pattern Mining and Sentiment Analysis for Massive Microblogging Online Texts", IEEE Access, vol. 8, pp. 78166-78182, 2020.
[19] Akrivi Krouska, Christos Troussas and Maria Virvou, "Comparative Evaluation of Algorithms for Sentiment Analysis over Social Networking Service", Journal of Universal Computer Science, vol. 23, no. 8, 2017.
[20] https://www.kaggle.com/divyansh22/stsgold-dataset, Accessed on : 04/12/2021 at 07:035 P.M.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY