A Feature Selection Based Hybrid Classification Using Naive Bayes And Decision Table For Sentiment Analysis In Map Reduce Environment

[1]Behjat U Nisa, [2]Er. Kiran Gupta

[1]M.Tech Scholar, Department of CSE, Swami Devi Dyal Institute of Engineering and Technology, Haryana, India.

[2]Assistant Professor, Department of CSE, Swami Devi Dyal Institute of Engineering and Technology, Haryana, India.

Abstract- Data is created from a couple of resources and it is ever-growing. Content development inside the Internet these days has made a massive extent of facts on hand. This data is exhibited in numerous codecs, for instance, posts, information articles, remarks, and surveys. Particularly within the car, hardware and movie segments, clients have composed audits approximately objects or their functions. By amassing and examining these surveys, new clients find out others' opinion approximately various highlights of the item. In this research feature selection with mutual information and then optimized features is fed to hybrid category set of rules comprising of Naïve bayes primarily based on possibility and Decision Table primarily based on decision rules is applied. Experimental effects demonstrate that the proposed technique outperforms the prevailing technique by numerous parameters.

Keywords: Naïve Bayes, Decision Table, Big data, Map reduce

I. INTRODUCTION

Research in Big Data and investigation offers chances to apply the proof based way to deal with basic leadership in numerous spaces. The different types of investigation plus approaches utilized in Big Data, for example, AI, can be utilized to filter through the tremendous measure of traffic data to separate valuable and empower the transportation position to take preventive activities and make fitting decisions [1]. The development of big data is the result of ongoing innovation advancements. A few creators characterized "Big Data" as gigantic measure of data originating from heterogeneous sources at an extremely fast, that it isn't workable for conventional apparatuses and procedures to investigate and separate an incentive from it "Big Data" alludes to the broad hurl of data having differed as well as complex structure that can't be overseen by customary data dealing with strategies and methods [2].

Sentimental analysis might be particular as the characterization of a book or report into a positive or a negative gathering by passing judgment on the undertone contained in the content. A positive assessment indicating content is given a positive name though a negative mark offers a negative input any target estimation would be relegated a nonpartisan label [3]. People and machines must be recognized by assumptions or feelings. A few analysts are attempting to build systems to make sentiments in machines. In equal, others are additionally attempting to consequently remove specific news, items or some other required part of life. At present, assumption analysis through normal language preparing is the most testing activity being broadly looked into by the educational network. The escalated utilization of

web based life is likewise affecting conclusions of individuals for or against explicit points including government, radicalism, training or money related approaches, association, and so forth. Along these lines, to perceive assumptions behind the posts via web-based networking media gatherings, there is a definitive need of an efficient, viable, and effective system. For sentimental analysis, abstract dictionary and AI based techniques are used [4].

II. LITERATURE REVIEW

In [5], In this work, authors begin a gated consistent neural network with sentimental kindred (GRNN-SR)[1]to incarcerate the sentimental relations' information from sentiment modifier environment as well as mock-up their belongings in texts. At each time measure, GRNN-SR separately encodes the information of sentiment polarity furthermore sentiment modifier context. The new sentiment inputs are personalized multiplicatively by the prior resolute sentiment modifier background before they are updated into current states of sentiment polarity, which is more effective than the approach of traditional GRNNs. The new grades demonstrate that our 9representation not simply can detain sentimental relations but as well is a growth over state-of-the-art gated standard neural network baselines.

In [6], a twitter API is used to fetch data from twitter. Tweets are pre-processed like usernames, urls, special symbols, hashtags are removed and emoticons are converted into words. Then Naïve-Bayes algorithm is applied for classification, which uses a wordnet dictionary. In Map phase, polarity of tokens is generated to check overall polarity of a tweet. In reduce phase, categorisation of the polarity into specific class is done. Converting emoticons into equivalent words increases the efficiency of the system.

In [7], authors aim to provide a well and effective solution to distribute political party's tickets during the election. The sole parameter of the current distribution of political party's tickets is based on the money power and person's references from their superior leaders. The purpose is to highlight the discrepancy between the real candidate who is predicted to win the election based on their popularity with other parameters and those who have only references with money power. The work choose the deserving candidate by analyzing parameters such as social work, criminal records, educational qualification, and his/her popularity on social media (twitter). It is clearly stated that current distribution of political party's ticket is not suitable. This is why eligible candidate lose their election and corrupt opponent wins.

In [8], A hybrid sentimental thing recognition model (HSERM) has been calculated. Using 100 million together communication from Twitter, the hashtag is supposed as the label for sentimental classification. In this moment, features as emoji furthermore N-grams have been found in addition to confidential the composed topic communication into four dissimilar sentiment parts based on the circumplex sentimental representation. At last, machine learning methodology are utilized to categorize the sentimental data set, in addition to an 89 % exact result has been gained. Additional, entities that are behind emotions could be achieved with the assist of SENNA deep learning representation.

In [9], authors have proposed a unique approach by performing abstract-level sentimental analysis of user review by n-gram classification and POS tagging. This classification is then used as entropy for machine learning algorithm. In this work , leverages upon the proposed methodology with promising outcomes and improved accuracy by evaluating the data with the help of two algorithms, MaxEnt model and Naïve Bayes classifier, after analysing few algorithms including SVM and random forest. It is clearly observed that the combined methodology of performing MaxEnt and Naïve Bayes together with the help of n-gram classification and POS tagging provides estimation of ratings with highest accuracy.

In [10], authors proposed a method for text sentimental analysis based on dimension reduction of Chi-square statistic (CHI) multi-grams mixture in this paper. It can not only effectively improve the effect of feature extraction, but also precisely determine the feature dimensions, which is different from the traditional methods using experience value. Experimental results show that the proposed method outperforms the exiting methods and the highest accuracy rate reached 94.85%. Moreover, it is proved that this method is universal for the subjective and objective classification as well as the different length of text classification reviews.

In [11], SoftMax based attitude detection algorithm is proposed to identify the user nature efficiently. The reliability and the accuracy of the sentiment prediction can be substantially increased with taking user attitude into account. The proposed algorithm is evaluated on tweets fetched from micro blogging website twitter.

In [12], research splits related work in two other classification: the first one works with detecting the term orientation and the other works with detecting the term subjectivity. These divisions only refer to research study on the term/word level classification, and not document- level classification. The purpose of this research is the sentiment evaluation which refers to get the sentiment polarity (positive, negative, or neutral) of a text reviews data and evaluate the sentiment score of the text review. Essentially a text review is split into single sentences (“sentence-based”) and words (“words-based”) or very short texts from a single source.

In [13], authors are concerned with conducting entity-level sentiment analysis. Firstly, a manually labelled dataset containing 3,000 issue comments selected from 231,732 issue comments collected from 10 open source projects in GitHub is built. Then the design and develop SentiSW, an entity-level sentiment analysis tool consisting of sentiment classification and entity recognition, which can classify issue comments into <;sentiment, entity> tuples is done. Authors evaluated the sentiment classification using ten-fold cross validation, and it achieves 68.71% mean precision, 63.98% mean recall and 77.19% accuracy, which is significantly higher than existing tools. We evaluate the entity recognition by manually annotation and it achieves a 75.15% accuracy.

In [14], for enhancing the performance of sentiment prediction, an individual prediction model must be established for each image category. However, collecting much ground truth sentiment data is one of the obstacles encountered by studies on this field. Here, authors proposed an

approach that acquires a training data set for category classification and predicting sentiments from images. Using this approach, authors collect a training data set and establish a predictor for sentiments from images. Firstly, estimation of the image category from a given image is done and then the prediction of the sentiment as coordinates on the arousal–valence space using the predictor of an estimated category.

This paper [15] provides predictive analysis on demonetization data using support vector machine approach (PAD-SVM). The proposed PAD-SVM system involved three stages including pre-processing stage, descriptive analysis stage, and prescriptive analysis. The pre-processing stage involves cleaning the obtained data, performing missing value treatment and splitting the necessary data from the tweets. The descriptive analysis stage involves finding the most influential people regarding this subject and performing analytical functionalities. Semantic analysis also is performed to find the sentiment values of the users and to find the compound polarity of each tweet. Predictive analysis is performed to view the current mind-set of people and the society reacts to the issue in the current time. This analysis is performed to find out the overall view point of the society and their view may change in the near-future in regarding to the scheme of demonetization as well.

In this paper [16], the authors propose an adaptable sentiment analysis approach that analyses social media posts and extracts user's opinion in real-time. The proposed approach consists of first constructing a dynamic dictionary of words' polarity based on a selected set of hashtags related to a given topic, then, classifying the tweets under several classes by introducing new features that strongly fine-tune the polarity degree of a post. To validate the approach, they classified the tweets related to the 2016 US election. The results of prototype tests have performed a good accuracy in detecting positive and negative classes and their sub-classes.

III. PROPOSED WORK

1. Collection of data: Data can be collected from Twitter Data source using Twitter API.

Twitter has more than 200 million months to month dynamic clients which results in billions of tweets every week. One more imperative cause of using tweet data is that tweets are mostly in text, while on others, there are usually images, videos etc.

2. Pre-Processing and Filtering: The collected dataset from Twitter is in the form of raw data that needs to be filtered in order to do classification on the data. For the Filtration of the raw data; various URL's, hashtags, punctuation marks, stop words and Digital words needs to be removed.

3. Feature Optimization using Mutual Information (MI): Mutual Information (MI) is a crucial technique to determine the relation or mutual dependence of a random variable with the other. The mutual information concept is associated with the attribute's entropy.

4. Classification using Map Reduce Platform: Decision table will store the conditional probabilities for the naïve bayes algorithm. Attributes are chosen and based on them; information is stored in the Decision table. During prediction making, decision table will make use of this model. There is a probability for every entry in

the table. Naïve Bayes makes use of conditional probability and bayes theorem and it assumes that features are independent. The overall class probability is estimated by combining the estimated probability. The goal of the proposed approach is to construct the model which will predict labels, so only training data needs to be tested.

In proposed approach, Decision table represent the conditional probability table for naïve bayes. Each point in the search, the algorithm estimates the values by isolating the qualities of attributes into two disjoint subsets: one for each of naïve Bayes and decision table. Initially all attributes are modelled by the decision table. At each step, forward selection search is used, and the attributes selected from this are modelled by naive Bayes and the remainder by the decision table.

IV: RESULTS IN HADOOP PLATFORM

Figure 4.1: Tokenizing the dataset

Loading the data from Hadoop source and then tokenizing the data using string to word vector filter. Tokenization includes converting -the string data into word vector filter and removing the stop words, converting the string data into numerical data etc.

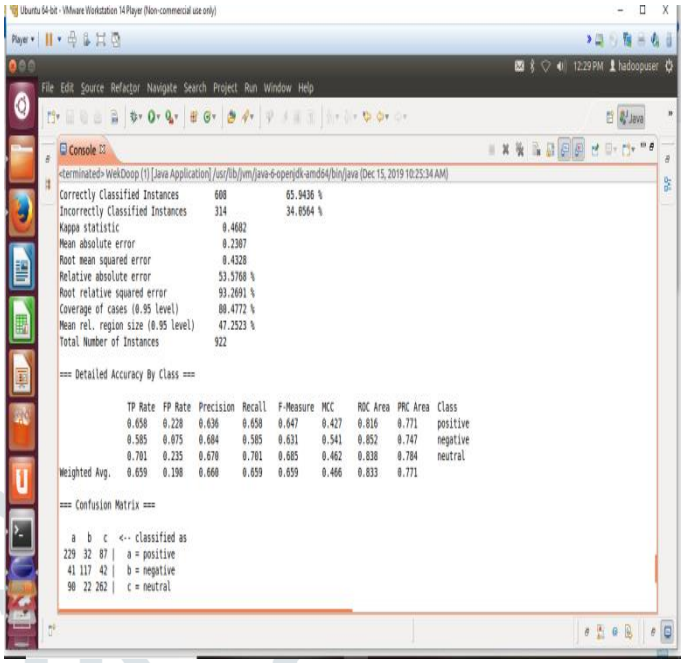

Figure 4.2: Evaluation results of Exiting Technique

Above figure shows the results of the existing naïve bayes technique.

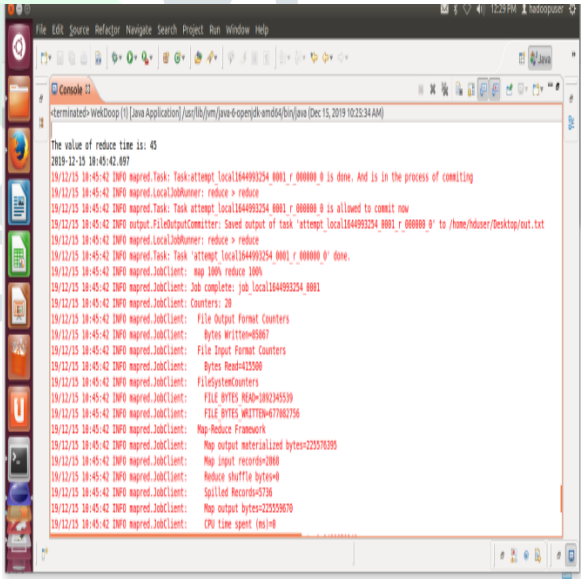

Figure 4.3: Showing the Hadoop information of running of existing technique in hadoop

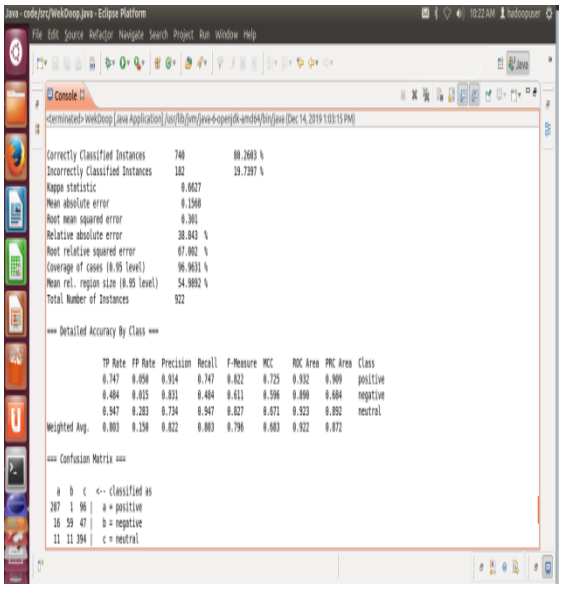

Figure 4.4: Evaluation results of proposed technique

Above figure shows the results of the proposed technique. It gives 740 correctly classified instances, 182 incorrectly classified instances. Kappa statistics of proposed technique is 0.6627, mean absolute error is 0.1568, root mean squared error is 0.301, relative absolute error is 38.843, root relative squared error is 67.002, TP rate is 0.803, FP rate is 0.150, precision is 0.822, recall is 0.803, f-measure is 0.796.

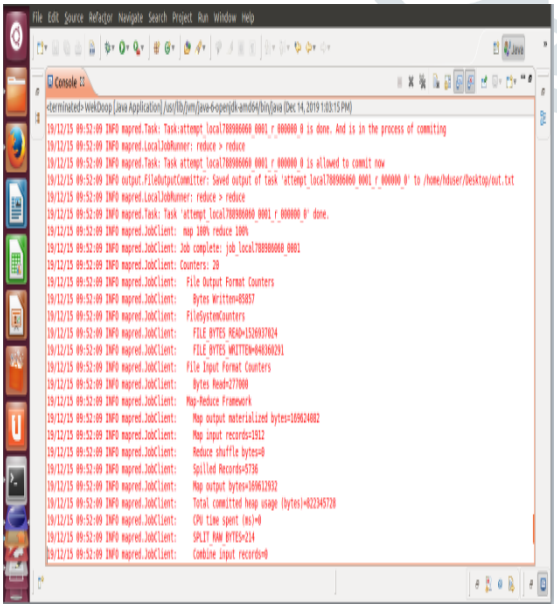

Figure 4.5: Showing the Hadoop information of running of existing technique in Hadoop

V. CONCLUSION

The tweets reflect sentiment of the people regarding various topics. Companies will benefit immensely by getting this sentiment data. So we need to classify the data based on sentiment. One classification algorithm is the Naïve Bayes Algorithm. But it assumes that features are independent. Some decision rules can be combined with the probability in order to make more correct predictions. Also, to more optimize the features of the data, mutual information feature selection is used to optimize features first then classification using naïve bayes and decision table is done for the prediction of social issues tweets collected from twitter in Hadoop. In future, tweets on some different issues can be taken into consideration for more analysis of sentiments.

REFERENCES

1. Neilson, A., Indratmo, Daniel, B., & Tjandra, S. (2019). Systematic Review of the Literature on Big Data in the Transportation Domain: Concepts and Applications. Big Data Research. doi:10.1016/j.bdr.2019.03.001

2. Loubna Rabhia, Noureddine Falihb, Abdlekbir Afraitesa, Belaid Bouikhalene" Big Data Approach and its applications in Various Fields: Review", International Workshop on Applying Data Mining Techniques to E-Learning and Pedagogical Approaches (ADMEPA) August 19-21, 2019, Halifax, Canada.

3. Kaur, V. D. (2018). Sentimental Analysis of Book Reviews using Unsupervised Semantic Orientation and Supervised Machine Learning Approaches. 2018 Second International

Conference on Green Computing and Internet of Things (ICGCIoT).

4. Muhammad Asif, Atiab Ishtiaq, Haseeb Ahmad, Hanan Aljuaid, Jalal Shah, Sentiment Analysis of Extremism in Social Media from Textual Information, Telematics and Informatics, 10 January 2020.

5. Chaotao Chen, Run Zhuo, Jiangtao Ren, Gated recurrent neural network with sentimental relations for sentiment classification, Information Sciences Volume 502, October 2019, Pages 268-278.

6. Huma Pandey and Shikha Pandey, "Sentiment Analysis on Twitter Data-set using Naive Bayes Algorithm", IEEE, 2nd International Conference on Applied and Theoretical Computing and Communication Technology, 2016, pp. 416-419.

7. Aniruddha Prabhu, B. P., Ashwini, B. P., Anwar Khan, T., & Das, A. (2019). Predicting Election Result with Sentimental Analysis Using Twitter Data for Candidate Selection. Lecture Notes in Networks and Systems.

8. Hibo Wang, Xiaohui Cui, Lu Gao, Qi Yin, Lei Ke & Shurong Zhang , A hybrid model of sentimental entity recognition on mobile social media, EURASIP Journal on Wireless Communications and Networking**,** volume 2016,

9. Shah, A., Kothari, K., Thakkar, U., & Khara, S. (2019). User Review Classification and Star Rating Prediction by Sentimental Analysis and Machine Learning Classifiers. Advances in Intelligent Systems and Computing.

10. Fulian Yin, Yanyan Wang,Jianbo Liu, A Text Sentimental Analysis Method Based on Dimension Reduction of CHI Multi-gram Features Mixture, The International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery, 07 November 2019.

11. Khanna N, B., Moses J, S., & M, N. (2018). SoftMax based User Attitude Detection Algorithm for Sentimental Analysis. Procedia Computer Science, 125, 313–320.

12. Godbole, Srinivasaiah, Skiena, "Large-Scale Sentiment Analysis for News and Blogs", 2007.

13. Jin Ding, Hailong Sun; Xu Wang; Xudong Liu, Entity-Level Sentiment Analysis of Issue Comments, 2018.

14. Sanghyun Seo, Dongwann Kang, Study on predicting sentiment from images using categorical and sentimental keyword-based image retrieval, The Journal of Supercomputing, September 2016.

15. Kannan, N., Sivasubramanian, S., Kaliappan, M. et al. Cluster Comput (2019) 22(Suppl 6): 14709.

16. El Alaoui, I., Gahi, Y., Messoussi, R. et al. J Big Data (2018) 5: 12