ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 8, Issue 4, pp: (1-6), Month: October - December 2020, Available at: www.researchpublish.com

Sentiment Analysis Using Hybrid Classification

VIKAS THAKUR[1], POONAM CHOUDHARY[2]

[12]Sirda Institute of Engineering Technology, Sundernagar, Mandi, H.P. India

Abstract: **Sentiment can be described in the form of any type of approach, thought or verdict which results because of the occurrence of certain emotions. This approach is also known as opinion extraction. In this approach, emotions of different peoples with respect to meticulous rudiments are investigated. For the attainment of opinion related data, social media platforms are the best origins. The hybrid classification is designed in this work which is the combination of KNN and random forest. The KNN classifier extract features of the dataset and random forest will classify data. The approach of hybrid classification is applied in this research work for the sentiment analysis. The performance of the proposed model is tested in terms of accuracy and execution time**

Keywords: **ISP, SVM classifier, KNN classifier, Data Mining, Sentiments analysis architecture, random forest.**

1. INTRODUCTION

The predictive and descriptive are the two categories of data mining tasks that are performed using data mining technology. The predictive data mining tasks are accountable to perform predictions on the basis of existing dataset while the descriptive data mining tasks help to understand the feature assets of the dataset. The data having different parameters is analyzed by the applications using data mining. These parameters include clustering, pattern analysis, classification and association. Every parameter should perform individual action. This approach is also known as estimation and classification. Both of these terms have a lot of similarity. If data mining access the telephone line that is being utilized to access internet, in this case, a recheck cannot be done to guarantee accurate classification. An ambiguity about the correctness of classification occurs generally due to the existence of incomplete information. The applicable actions are being performed in the real time situations. The phone may or may not be used to call the local ISP initially. It is possible in case of credit card transaction to see whether the fraud has happened or not. In order to do verification, enough attempts should be made. It is possible to perform predictive tasks in a different way as the records can be classified on the basis of few predicted future activities or probable future values. Wait and watch is just the way to validate the precision of classification using forecasting [6]. A kind of normal communication dispensation for knowing the opinion of customers for a meticulous object is known as sentiment or emotion analysis. The other name of emotion analysis is opinion or belief mending. This analysis develops a scheme for collection and examining views about a certain object appeared in social media posts, evaluation, tweets or remarks. The technique of emotion investigation can be beneficial in various ways. This analysis shows its presence from computer discipline to administration discipline and public science because of its worthiness in public and industries. In recent years, manufacturing actions adjoining emotion study are also flourished. Various novel industries have been developed. A lot of huge businesses encompass self relies domestic ability. Emotion scrutiny schemes have established their claims in approximately each industry and public region [8].Opinion study may be depicted as a procedure which includes computerized extraction of sentiments, estimation, vision and feeling from Natural Language Processing in the form of content, language, chirp, record and so on. During Opinion investigation, the tweets are mainly categorized in three categories. These categories are “optimistic”, “pessimistic” and “unbiased”. This analysis can also be understood in the form of prejudice investigation, belief extraction and assessment mining.

Following are the implementation techniques of sentiment analysis architecture:

a) Pre-processing of the datasets: Some tweet involves a lot of sentiments about the information articulated in dissimilar traditions by dissimilar clients. Twitter data sample utilized in this study is tagged into two sections viz. pessimistic and affirmative division and thus the emotion scrutiny of the information becomes simple to scrutinize the consequence of different characteristics. The unprocessed information having division is extremely vulnerable to discrepancy and superfluous [11].

b) Feature Extraction: The preprocessed data sample includes numerous characteristic belongings. In the characteristic withdrawal technique, we mine the features from the developed data sample. Later these are utilized for the computation of optimistic and pessimistic polarity in a phrase helpful for formatting the estimation of the persons using replicas such as unigram, bigram etc. Machine learning methods need representation of the key features of content or papers for dispensation. These input characteristics are measured as characteristic vectors which are utilized for the categorization job.

c) Training: Managed learning is a significant system for resolving categorization issues. The training of classifier formulates it easier for prospect forecasting for unidentified information.

d) Classifiers: Following are the classifiers used for the implementation of sentiment analysis.

- **Naive Bayes:** Naïve Bayes is a probabilistic classifier relied on Bayes' Theorem. By investigative a set of papers, it can learn the prototype. An evaluation is made between the topic stuff of the paper and a specified deposit of language, for the attainment of an accurate group of categorization. NLTK comes with all wherewithal's for opinion study such as characteristic withdrawal.
- **Maximum Entropy [ME]:** This form is considered as exponential. In this Classifier, development of any assumptions related to provisional sovereignty among characteristics is not mending. This Classifier desires additional time for training in comparison with Naive Bayes because of development issues. This scheme can manage overlapping characteristic also and then decides the replica with utmost entropy [12].
- **Support Vector Machine:** This classifier is typically utilized for prototype acknowledgment and information investigation. This was made-up by Vladimir Vapnik. In this classifier, categorization is implemented by the erection of an N-Dimensional hyper plane, which is capable of separating information into detached groups. Mainly two vectors of a meticulous dimension are applied in the form of inputs and after this categorization is implemented.

2. LITERATURE REVIEW

Jianqiang, et.al (2018) suggested the use of deep convolution neural system for the categorization of twitter data sentiments [14]. In this technique, sentiments characteristics vector of t tweeter data utilized emotion lexicon and n-gram characteristics. In the presented approach, already trained statement enclosed characteristics were developed with the help of GloVe statement attitude divisional characteristics. The characteristics of twitter sentiments were given as input to deep intricacy neural system. The conceptual data was confined with the help of persistent arrangement. With the help of a complicated neural system, the demonstration of content was constructed. Almost five data samples were used for the validation of investigational outcomes.

Ankit, et.al (2018) projected a novel approach for the classification of twitter data sentiments and this approach was named as ensemble classification approach [15]. A number of conventionally utilized twitter emotion investigating approaches were considered for calculating the valuation of proposed approach but the proposed ensemble classification approach was declared best. A number of base learners were utilized for the representation of the proposed approach. The proposed approach of ensemble classification showed better results in comparison with stand alone approaches. For the observation of clients' beliefs about their goods, the presented system was quite appropriate for the corporations.

Das, et.al (2018) stated that stream-based setting by using the incremental active learning approach, gave capability to algorithm for choosing new training data from a data stream for hand-labeling [16]. Stream based active learning in financial domain could be helpful to both sentiment analysis and the active learning research area. With the use of RNNs Long Short –Term Memory, this experiment also proved helpful for feasibility study through batch processing. To analyze the sentiments and the current stock trends, a hybrid model could also be developed. This model would improve the reliability of prediction. In future for analyzing the stock data, addition of machine learning algorithms can be done. Some other methods of data ingestion like data ingestion through Apache Flume or NodelJS can also be used in future.

Alzahrani, et.al (2018) proposed a novel approach of hybrid internet of things system utilizing calculative syntax realm. For the generation of genuine tweets, API of twitter was utilized [17]. For the investigation of twitter sentiments and beliefs creation, an internet of things system utilizing Raspberry Pi set up was implemented. For conducting the experiments, Arabic tweets information samples and Naïve Bayes approaches were used. This classifier showed considerable precision on the used data sample for the classification of twitter data into optimistic or pessimistic.

Symeonidis, et.al (2018) proposed that various pre-processing techniques evaluated on their resulting classification accuracy and the number of features they developed [18]. The obtained results indicated that some techniques like lemmatization, removing numbers and replacing contractions improved precision while other techniques kike removing punctuations did not. To investigate the interactions between the techniques when they were employed in a pipeline manner, an ablation and combination study was done. The outcomes of these techniques clearly indicated the importance of techniques like replacing numbers and replacing repetitions of punctuations.

Tasoulis, et.al (2018) proposed a practical mechanism relied on open source technique for the recognition of genuine sentimental changes [19]. The proposed approach was mega proficient in terms of memory utilization as well as in the case of calculation rate. For the accomplishment of this work, tweets were gathered reiteratively in actual time and also discarded instantly after their utilization. For the classification of sentiments, Lexicon technique was utilized. Also suitably controlled graphical representation was utilized for the attainment of alter recognition.

Patil, et.al (2017) analyzed that the Micro blogging became a very important part of everyone's life in present scenario [20]. A number of internet users shared their feelings on different blogging sites like face book, twitter. Various tools were used for sentiment analysis of data by using some of the tweeted data as input and got respective scores as output. The earlier developed unigram model utilized as the gauge model and gave 4% of general report. This method used two classification methods, one is two-way classification methods and the other is three-way classification methods.

3. RESEARCH METHODOLOY

This research work is related to sentiment analysis of the twitter data. Following are the various steps of the research methodology: -

e) Pre-processing of the datasets: Some tweet involves a lot of sentiments about the information articulated in dissimilar traditions by dissimilar clients. Twitter data sample utilized in this study is tagged into two sections viz. pessimistic and affirmative division and thus the emotion scrutiny of the information becomes simple to scrutinize the consequence of different characteristics. The unprocessed information having division is extremely vulnerable to discrepancy and superfluous [4].

f) Feature Extraction: The preprocessed data sample includes numerous characteristic belongings. In the characteristic withdrawal technique, we mine the features from the developed data sample. Later these are utilized for the computation of optimistic and pessimistic polarity in a phrase helpful for formatting the estimation of the persons using replicas such as unigram, bigram etc. Machine learning methods need representation of the key features of content or papers for dispensation. These input characteristics are measured as characteristic vectors which are utilized for the categorization job.

g) Training: Managed learning is a significant system for resolving categorization issues. The training of classifier formulates it easier for prospect forecasting for unidentified information. The approach of KNN classifier is applied which can extract the features of the dataset. The KNN classifier approach can apply k-mean approach means it can define the centriod points and eculidian distance is calculated from these points. The points which have similarity will be classified will be specified into one class.

h) Classifiers: A classifier named random forest is chosen for this approach. Because, emotion study is a dual categorization and a large amount of data samples are present for execution, therefore random forest classifier is selected in this study. A manually generated training set is utilized for training the classification; a physically produce training sample is used here. An X: Y relation is provided inside the training sample where x represents the score of an estimation text and y is used for the representation of optimistic or pessimistic word and gives them score accordingly. The score of t estimation statement associated with characteristic inside the appraisal is applied as key to random forest approach.

4. RESULT AND DISCUSSION

This section shows the results of the SVM and Hybrid classifiers for the sentiment analysis. The performance of the both classifiers are analyzed in terms of accuracy and execution time. The results are analyzed by varying the test and training set rations. The proposed model is implemented in python using anaconda.

ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 8, Issue 4, pp: (1-6), Month: October - December 2020, Available at: www.researchpublish.com

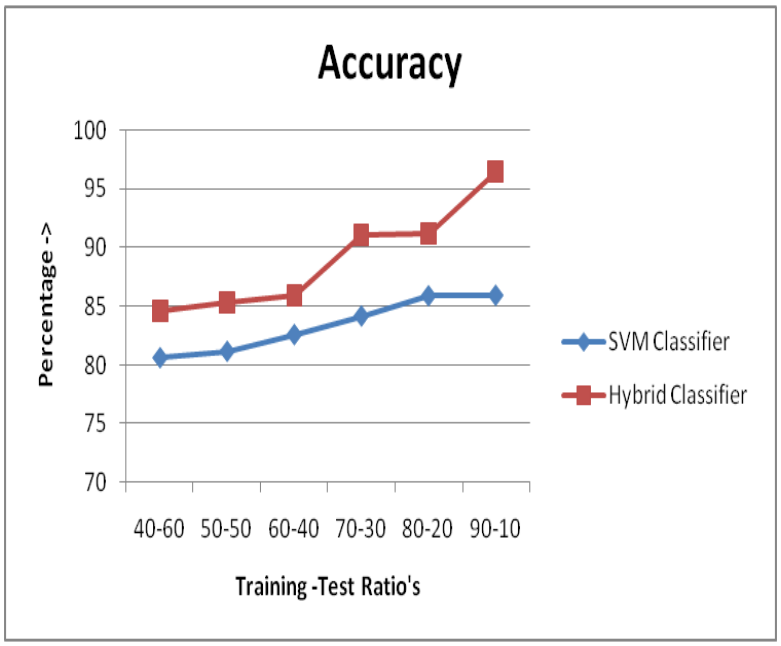

Fig 1: Accuracy Analysis

As shown in figure 1, the accuracy of SVM classifier is compared with the hybrid classifier. The accuracy of hybrid classifier is high as compared to SVM on different set of rations of training and test

Table 1: Accuracy Analysis

Training, Test Ratio	SVM Classifier	Hybrid Classifier
40-60	80.63	84.63
50-50	81.17	85.32
60-40	82.59	85.92
70-30	84.17	91.08
80-20	85.88	91.15
90-10	85.92	96.43

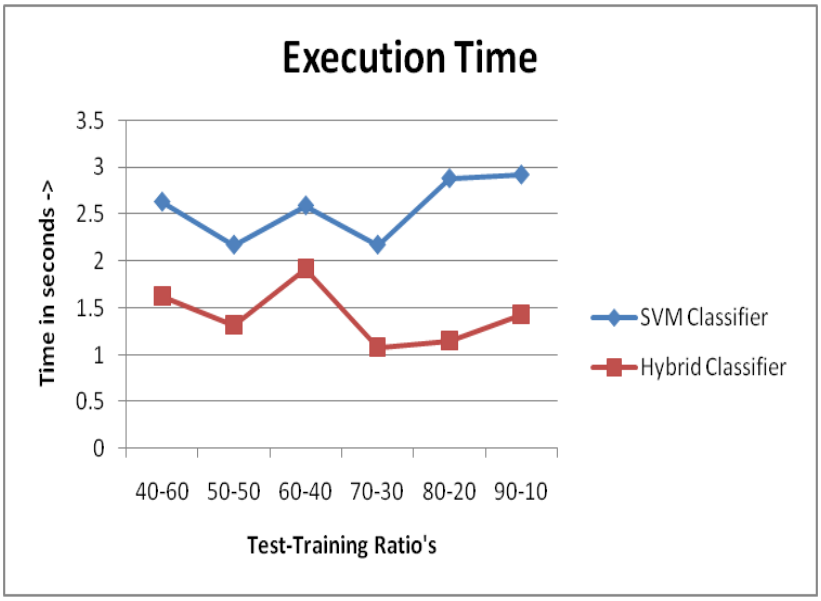

Fig 2: Execution Time

As shown in figure 2, the execution time of SVM classifier for sentiment analysis is compared with hybrid classifier. It is analyzed that execution time of hybrid classifier is low as compared to SVM classifier

Table 2: Execution Time

Training, Test Ratio	SVM Classifier	Hybrid Classifier
40-60	2.63	1.63
50-50	2.17	1.32
60-40	2.59	1.92
70-30	2.17	1.08
80-20	2.88	1.15
90-10	2.92	1.43

5. CONCLUSION

The sentiment analysis is the approach which is applied to analyze the sentiment of the users. This research work is related to analyze sentiments of the twitter data. The sentiment analysis has three steps which are pre-processing, feature extraction and classification. The classification approach plays important role in analyzing sentiments from the data. In the previous work, approach of SVM classification is applied for the sentiment analysis. The hybrid classifier is the combination of KNN and random forest. The hybrid classifier is applied on the place of SVM for the sentiment analysis which increase accuracy and reduce execution time.

REFERENCES

[1] O. Rambow, L. Shrestha, J. Chen, and C. Lauridsen. Summarizing email threads. In Proceedings of HLT-NAACL 2004: Short Papers, pages 105–108, 2004.

[2] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. Information Processing and Management, 24:513–523, 1988.

[3] O. Sandu. Domain Adaptation for Summarizing Conversations. PhD thesis, Department of Computer Science, The University Of British Columbia, Vancouver, Canada, 2011.

[4] S. Teufel and M. Moens. Summarizing scientific articles: Experiments with relevance and rhetorical status. Computational Linguistics, 28:409–445, 2002.

[5] J. Ulrich, G. Murray, and G. Carenini. A publicly available annotated corpus for supervised email summarization. In AAAI08 EMAIL Workshop, Chicago, USA, 2008. AAAI.

[6] D. C. Uthus and D. W. Aha. Plans toward automated chat summarization. In Meeting of the Association for Computational Linguistics, pages 1–7, 2011.

[7] C. Whitelaw, B. Hutchinson, G. Chung, and G. Ellis. Using the web for language independent spellchecking and autocorrection.In Empirical Methods in Natural Language Processing, pages 890–899, 2009

[8] Kiruthika M, Sanjana Woonna, “Sentiment analysis of twitter data”, 2016, International journal of innovations in engineering and technology.

[9] Varsha Sahayak, Vijaya Shete, Apashabi Pathan, “Sentiment Analysis on Twitter Data”, 2015, International Journal of Innovative Research in Advanced Engineering (IJIRAE)

[10] Apoorv Agarwal Boyi Xie Ilia Vovsha Owen Rambow Rebecca Passonneau, “Sentiment Analysis of Twitter Data”, 2011, Conference of the European Chapter of the ACL

[11] Chandan Arora, Dr. Rachna, “SENTIMENT ANALYSIS ON TWITTER DATA”, 2017, International Research Journal of Engineering and Technology (IRJET)

[12] Vishal A. Kharde, S.S. Sonawane, “Sentiment Analysis of Twitter Data: A Survey of Techniques”, 2016, International Journal of Computer Applications (0975 – 8887)

[13] Onam Bharti, Mrs. Monika Malhotra, “SENTIMENT ANALYSISON TWITTER DATA”, 2016, International Journal of Computer Science and Mobile Computing, Vol.5 Issue.6,

[14] Zhao Jianqiang, Gui Xiaolin, “Deep Convolution Neural Networks for Twitter Sentiment Analysis”, 2018, IEEE

[15] Ankit, Nabizath Saleena, "An Ensemble Classification System for Twitter Sentiment Analysis", 2018, International Conference on Computational Intelligence and Data Science

[16] Sushree Das, Ranjan Kumar Behera, Mukesh Kumar, Santanu Kumar Rath, "Real Time Sentiment Analysis of Twitter Streaming Data For Stock Prediction",2018,International Conference on Computational Intelligence and Data Science

[17] Salha M. Alzahrani, "Development of IoT Mining Machine for Twitter Sentiment Analysis", 2018, 15th Learning and Technology Conference (L&T)

[18] Symeon Symeonidis , Dimitrios Effrosynidis , Avi Arampatzis, " A comparative evaluation of pre-processing techniques and their interactions for twitter sentiment analysis",2018, Expert Systems With Applications

[19] Sotiris K. Tasoulis, Aristidis G. Vrahatis, Spiros V.Georgakopoulos, Vassilis P. Plagianakos, "Real Time Sentiment Change Detection of Twitter Data Streams",2018, Innovations in Intelligent Systems and Applications (INISTA)

[20] Rashmi H Patil, Siddu P Algur," Sentiment Analysis by Identifying the Speaker's Polarity in Twitter Data", 2017, International Conference on Electrical, Electronics, Communication, Computer and Optimization Techniques (ICEECCOT)