Hybrid Classification Technique for Sentiment Analysis of the Twitter Data

Jaanu Sharma, Vinayak Khajuria, Dilbag Singh

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. Twitter may be recognized as a social media platform which is socially accessible to numerous followers. When these followers post some message on twitter, then this is recognized as tweet. The sentiment of twitter data can be analyzed with the feature extraction and classification approach. 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 : Sentiment analysis, KNN, SVM, random forest

I. INTRODUCTION

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 [1]. 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”.

Manuscript published on 30 August 2019.
***Correspondence Author(s)**
Jaanu Sharma , M.E CSE from Chandigarh university, India.
Vinayak Khajuria, Assistant Prof. in Chandigarh University, Gharuan, Mohali, Punjab, India.
Dilbag Singh, assistant professor at Chandigarh University, Gharuan, Mohali, Punjab, India

This analysis can also be understood in the form of prejudice investigation, belief extraction and assessment mining.In sentiment analysis, some areas are considered very important and they are:

1. Opinion Cataloging: In this approach, whole reports are categorized in accordance with the estimation about some product.

2.Attribute relied Opinion Categorization: This categorization takes into consideration, certain beliefs about some product

3.Sentiment Characterization: The work of sentiment characterization is distinct from the conventional content characterization. In this task only the product characteristics are extracted on the basis of which the customer has articulated his views.

1.1. Challenges in Sentiment Analysis

Opinion study is an extremely difficult job. Difficulties faced during the task of sentiment analysis are given below:

1. Recognizing Slanted Fraction of Content**:** Prejudiced sections symbolize emotion-carrying text. The similar statement may be utilized in the form of slanted in single case, or purposeful in another case.
Identification of the slanted parts of content can be more challenging because of this.

2. Region Reliance**:** Identical idiom or axiom may have dissimilar sense in dissimilar areas.

3. Mockery Recognition**:** Mocking phrases articulate pessimistic estimation concerning an objective by means of optimistic expressions in different method.

4. Dissatisfied Terminology**:** In several phrases merely some section of a sentence concludes the entire divergence of the report.

5. Unambiguous Contradiction of Opinion**:** Emotions may be annulled through various means like for opposing something, use normal words like no, not, never and so on. Identification of such kind of contradictions is complicated.

6. Sort Reliance**:** Conversation arrangement scrutiny is necessary for opinion extraction.

7. Object Acknowledgment: Separation of content from a certain object is essential because it investigates the emotion about object.

8. Development of a Classifier for Slanted vs. Intentional Tweets: Existing study effort concentrates mainly on classification of optimistic vs. pessimistic properly. The tweets must also be classified on the basis of emotions vs. no emotions directly [2].

9. Management of Assessment**.** Container of vocabulary representation cannot manage all evaluation in a satisfactory approach.

10. Utilization of Opinion Study on Facebook Communication: Only some small researches are carried out in Facebook for the investigation of emotions because of certain limitations.

11. Internationalization**:** Existing study effort concentrates chiefly on English text but on the same time it is also true that twitter has many followers all across the world.

II. LITERATURE SURVEY

Jianqiang, et.al (2018) suggested the use of deep convolution neural system for the categorization of twitter data sentiments [7]. 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. The tested results depicted that the presented approach performed well in comparison with several other approaches. Thus, it was concluded that the depth complicated neural system exploiting previously trained statement trajectories showed superior performance. Ankit, et.al (2018) projected a novel approach for the classification of twitter data sentiments and this approach was named as ensemble classification approach [8]. 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 standalone approaches. For the observation of clients' beliefs about their goods, the presented system was quite appropriate for the corporations. The presented system was applicable for the clients also because with the help of public beliefs they could choose the better goods. In future, major area of consideration will be the learning of unbiased tweets because of their neutral nature.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 [9]. 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 NodeJS 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 [10]. 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. The classifier was trained with the help of gold standard data sample and thereby a precision rate of 0.992 was achieved. The tested results displayed the efficiency and viability of the proposed approach. In future, some other classifiers will also train with the help of different data samples. Improvement in the proposed approach will also be performed in near future for showcasing the prevailing tweeter emotions. Symeonidis, et.al (2018) proposed that various pre-processing techniques evaluated on their resulting classification accuracy and the number of features they developed [11]. 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 [12]. 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. It was also believed that for the determination of fraudulent reports, the projected approach prompted an impending huge scale threads' scrutinizing. The tested results depicted that with the help of proposed methodology significant attitude alterations athwart hash tag life span could be detected. In future, the researchers will analyze the methodologies which are used for improving the sturdiness of alter recognition approaches.

III. RESEARCH METHODOLOGY

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

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 [4].

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.

c)

d)

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

e) These input characteristics are measured as characteristic vectors which are utilized for the categorization job.

f) 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 centroid points and Euclidian distance is calculated from these points. The points which have similarity will be classified will be specified into one class.

g) 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.

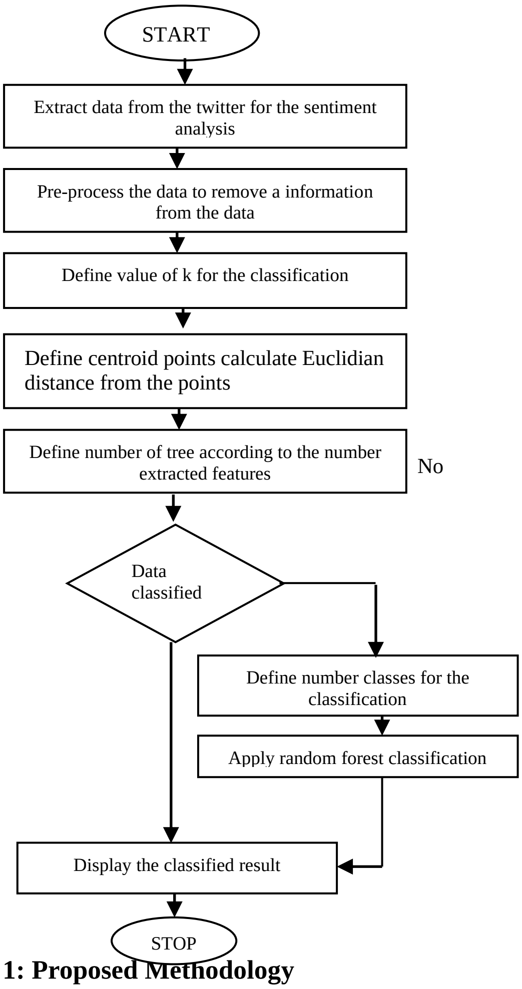
START
Extract data from the twitter for the sentiment analysis
Pre-process the data to remove a information from the data
Define value of k for the classification
Define centroid points calculate Euclidian distance from the points
Define number of tree according to the number extracted features
No
Data classified
Define number classes for the classification
Apply random forest classification
Display the classified result
STOP

Fig 1: Proposed Methodology

IV. 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

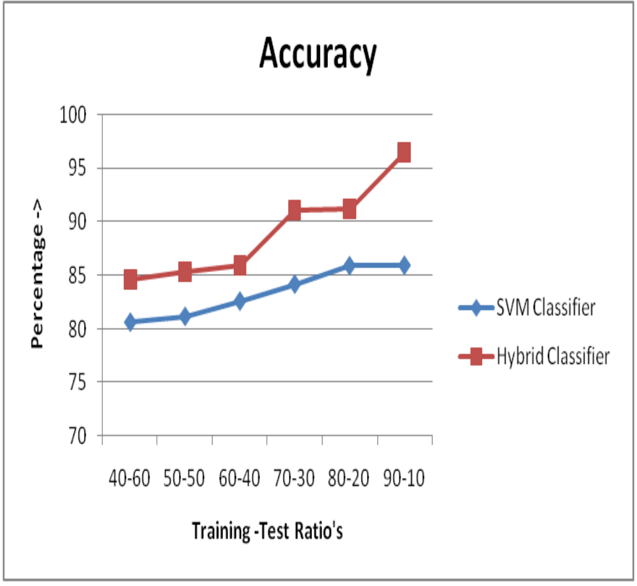

Fig 1: Accuracy Analysis

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

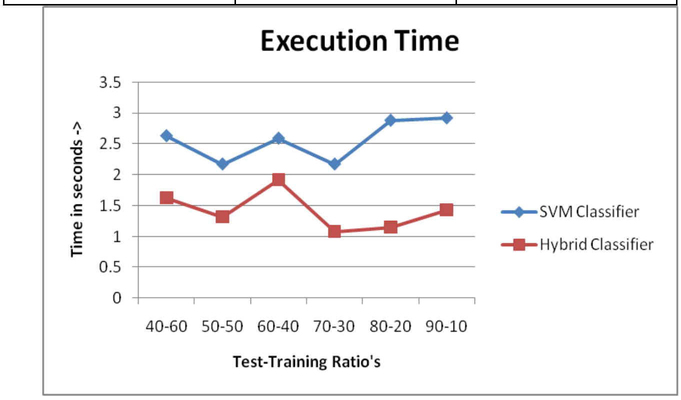

Fig 2: Execution Time

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

V. 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. The hybrid classifier optimizes results up to 8 percent for the sentiment analysis. In future hybrid classification model can be designed for the sentiment analysis.

REFERENCES

1. Kiruthika M, Sanjana Woonna, “Sentiment analysis of twitter data”, 2016, International journal of innovations in engineering and technology.
2. Varsha Sahayak, Vijaya Shete, Apashabi Pathan, “Sentiment Analysis on Twitter Data”, 2015, International Journal of Innovative Research in Advanced Engineering (IJIRAE)
3. Apoorv Agarwal Boyi Xie Ilia Vovsha Owen Rambow Rebecca Passonneau, “Sentiment Analysis of Twitter Data”, 2011, Conference of the European Chapter of the ACL
4. Chandan Arora, Dr. Rachna, “SENTIMENT ANALYSIS ON TWITTER DATA”, 2017, International Research Journal of Engineering and Technology (IRJET)
5. Vishal A. Kharde, S.S. Sonawane, “Sentiment Analysis of Twitter Data: A Survey of Techniques”, 2016, International Journal of Computer Applications (0975 – 8887)
6. Onam Bharti, Mrs. Monika Malhotra, “SENTIMENT ANALYSISON TWITTER DATA”, 2016, International Journal of Computer Science and Mobile Computing, Vol.5 Issue.6,
7. Zhao Jianqiang, Gui Xiaolin, “Deep Convolution Neural Networks for Twitter Sentiment Analysis”, 2018, IEEE
8. Ankit, Nabizath Saleena, “An Ensemble Classification System for Twitter Sentiment Analysis”, 2018, International Conference on Computational Intelligence and Data Science
9. 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
10. Salha M. Alzahrani, “Development of IoT Mining Machine for Twitter Sentiment Analysis”, 2018, 15th Learning and Technology Conference (L&T)
11. 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
12. 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)

AUTHORS PROFILE

Jaanu sharma is pursuing M.E CSE from Chandigarh university. He has done his B. Tech from Baba Ghulam Shah Badshah University Rajouri. Currently, he is working as Teaching Assistant at Chandigarh University, Gharuan, Mohali, Punjab, India.

Vinayak Khajuria completed M.E CSE from Chandigarh University. Done his B.Tech in IT from B.G.S.B. University, Rajouri. Currently, working as Assistant Prof. in Chandigarh University, Gharuan, Mohali, Punjab, India.

Dilbag Singh received his PhD degree from Department of Computer Science and Engineering, Thapar Institute of Engineering and Technology, Patiala. He has done his Master in Technology (Computer Science and Engineering) from Guru Nanak Dev University, Amritsar, Punjab, India (2012). Currently, he is working as an assistant professor at Chandigarh University, Gharuan, Mohali, Punjab, India. He has published more than 27 research papers in well-known reputed SCI indexed journals and international conferences. His research interest includes Wireless sensor networks, Digital image processing and Meta-heuristic techniques.