Hybrid Classification Technique for Sentiment Analysis of Twitter Data

Ramneet[1], Amandeep Kaur[2], G.N. Verma[3]
[12]*Department of Computer Science & Engineering, Sri Sukhmani Institute of Engineering & Technology, I.K.*
[3]*Gujral Punjab Technical University, Jalandhar, India.*

Abstract- Any kind of attitude, through or judgment that occurs due to any feeling is known as a sentiment which is also known as opinion mining. The sentiments of individuals towards particular elements are analyzed in this approach. To gather sentiment information, web or internet is the best known source. A platform that is accessed socially by various users to post their views is known as Twitter. The messages that are posted by these users are known as tweets. The properties of Tweets are highly unique due to which new challenges have raised. In comparison to several other domains, the sentiment analysis requires higher analysis studies. In the form of data source, the tweets are used from Twitter. The twitter public API is given for the extraction of tweets on a large scale. The "twitteroauth" version of the public API by Williams (2012) is utilized in this situation. It is possible to run this version on local host or web servers and the implementation is done on PHP. For performing feature analysis in sentiment analysis, the N-gram approach is used here.

Keywords- SVM, KNN, Twitter, Analysis, Sentiments

I. INTRODUCTION

The breaking down of data into such a form that it can useful to other users in the form of important knowledge is known as data analytics. The real scenario of the user's work can be understood better with the help of data analytics process. Better decisions can be made with the help of this process. In order to discover the useful information from the present data, various actions such as inspection, cleansing, transformation as well as modeling are performed which are collectively known as the data analytics process [1]. There are numerous facets as well as approaches present within the process of data examination. Within the different domains, numerous techniques are proposed with separate names. A specific data investigation process in which the modeling is performed, and useful information is extracted such that is can be utilized in predictive manners is known as data mining process. However, the analysis that is performed based on aggregations which is completely dependent on business information is known as business intelligence process. The process through which the complete document is broken down such that the individual components can be investigated separately is known as investigation. Within the big data analysis, there are three major categorizations in which the data can be differentiated. The data that can be stored within the database SQL in a tabular form which includes rows and columns is known as structured type of data [2]. The structured data is highly organized. The information that is not present within the relational database but still has some authoritative properties such that it can be analyzed in an easy way is known as the semi-structured type of data. Around 80% of the total amount of data present today is considered to be the unstructured type of data. The unstructured data does not have any specific structured. There is majorly the text and multimedia type of data present within this category. The human language is converted into computer language with the help of this tool. In the area of Sentiment analysis NLP has been primarily used. NPL is an important tool in area of artificial intelligence as it helps in interaction of robots in human natural language with humans. Sentiment Analysis is also known as the opinion mining. It uses the NLP in order to categorize the opinions of people about the products or the reviews [3]. Sentiment analysis deals with opinions and perspective of human related to emotions and attitude about some occurrence or the event. Opinion mining is most useful in various fields like commercial product reviews, social media analysis and movie reviews etc. the semantic analysis is a valuable technique in creation of recommender systems. The user gives the text reviews like online reviews, comments or the feedbacks on the social media sites, e-commerce websites. The opinions of users are known in better way with the help of this source. The sentiment analysis is done to check the positive, negative and neutral opinion of users about products to check its popularity or importance in the market [4]. Every human being has their different opinion, feelings, thoughts, and emotion for a particular event this can be known with the help of sentiment analysis. A new category is represented by each sentiment, as sentiment analysis can be interpreted as a task of a classification. An important role is played by the computer science and artificial intelligence in the field of natural language processing as it deals with the human and computer language interaction. It is required to do more research in the field of the sentiment analysis as marketing level competition is change drastically. For the text classification, there are various types of classifiers have been utilized that also utilized for the twitter sentiment classification [5]. The resources used are lexicographical,

initial method is to collect the seeds of the sentiment words and their orientation to find their antonyms and synonyms to expand their set. The issue related to the sentences classification has been solved with the help of machine learning approach as it totally based on the algorithms. Supervised leaning approach and unsupervised learning approach are the two utilized approaches. The combination of both the elements from lexicon-based techniques and machine learning in the sentiment analysis is known as the Hybrid approaches. These are used as the semantics networks and ontologies in order to determine the semantics that are present in the sophisticated manner [6]. The combination of the both these approaches lead to increase in the performance and accuracy of the sentiment analysis. The application of this algorithm used as the advantage as it provides the extra semantic, syntactic feature and very much flexible in use. SVM classifier has the massive edge, for the classification. For the separation of the tweets a comparison difference between the tweet and hyper plane has been done with the help of hyper plane technique. There are different kinds of Ensemble classifiers are present [7]. In order to do the best classification, this classier try to make full utilizes of all the embedded features of the base classifiers. Naive Bayes, Maximum entropy and SVM are the utilized base classifier here. With the help of voting rule, an ensemble classifier is created. The classification of this classifier is done on the basis of the output of the greater part of classifiers.

II. LITERATURE REVIEW

Dan Cao, et.al (2016) proposed that an Automatic Text Summarization intends to make a compressed version of documents, which should cover all the significant contents and general information. This paper reviews every one of the features that utilization metrics and idea of complex network for scoring sentences [8]. The experiment results on single component and combinations of different features we proposed are discussed. Another contribution was the discovery of results that features combinations with a similar kind property of network indicated incredible influence to choose sentences. About sentence relationship between sentences which turned out to be an essential element in the extraction of good rundowns, cause may concern about the structure of text document it inferred well.

Rasim Alguliyev, et.al (2016) proposed that text summarization is represented as a sentence scoring and selection process in this paper. As a result of the large amounts of text documents are created in the web and e-government and their volume increments exponentially along years. In result, expanding the volume of text documents has made troublesome for clients to read and extract helpful information from them [9]. This paper is centered on the extractive text summarization where a summary is generated by scoring and choosing the sentences in the source text. At first it assesses the score of each sentence and afterward chooses the most representative sentences from the text by considering that semantic similarity between chose sentences will be low. For scoring the sentences another formula is introduced. The proposed show endeavors to find balance amongst coverage and redundancy in a summary. For taking care of the optimization issue a human learning optimization algorithm is used.

Narendra Andhale, et.al (2016) presented the process in which the condensed type of document can be generated that can help in recording the significant information and provides importance to the source text is known as text summarization. An important method through which the related information can be identified from huge documents is known as automatic text summarization method. The comprehensive survey of both of the techniques present within the text summarization is presented in this paper [10]. An effective summary is to be generated by summarization method which has less redundancy and involves correct sentences which are grammatically correct. Good results are achieved within the extractive and abstractive methods which can be utilized further by the users. The testing for hybridization is studied within this paper which helps in generating the information which is compressed and readable by the users.

Rupal Bhargava, et.al (2017) proposed the fundamental use of the Sentiment Analysis has been a sharp research area for recent years. In any case, a significant part of the exploration that has been done supports English dialect as it were. This paper proposes a strategy utilizing which one can break down various languages to find sentiments in them and perform sentiment analysis. After the machine translation, text is processed for finding the sentiments in the text [11]. With the coming of blogs, forums and online reviews there is substantial text present on internet that can be utilized to break down the sentiment about a specific subject or an object. Thus to reduce the processing it is beneficial to extract the important text present in it. So the system proposed utilizes text summarization process to extract important parts of text and after that utilizations it to examine the sentiments about the specific subject and its aspects. Experiment demonstrates that proposed strategy can deliver promising results.

Archana N.Gulati, et.al (2017) proposed a text summary is a reduction of original text to condensed text by choosing what is important in the source. Over a period of years, the World Wide Web has expanded with the goal that tremendous measure of data is created and accessible on the web [12]. Additionally, considering the normal dialect in India being Hindi, a summarizer for a similar dialect is assembled. News articles on games and governmental issues from online Hindi newspapers were utilized as contribution to the system. Fluffy inference motor was utilized for the extraction process utilizing eleven important features of the text. The system accomplishes an average precision of 73% over multiple

Hindi documents. The summary generated by the system is discovered near summary generated by humans. The Precision, Recall and F-score values demonstrates good accuracy of summary generated by the system.

Manisha Gupta, et.al (2016) proposed by author that automatic summarization assumes an important part in document processing system and information recovery system. Era of summary of a text document is an important piece of NLP [13]. Dead wood words and phrases are likewise removed from the original document to generate the lesser number of words from the original text. Proposed system is tested on different Hindi sources of info and accuracy of the system in type of number of lines extracted from original text containing important information of the original text document. Info text size can be decreased to 60% - 70 % with the assistance of proposed system. System generates the extractive summary given by the client i.e. it doesn't generate the summary of the text on the premise of the semantics of the text.

III. RESEARCH METHODOLOGY

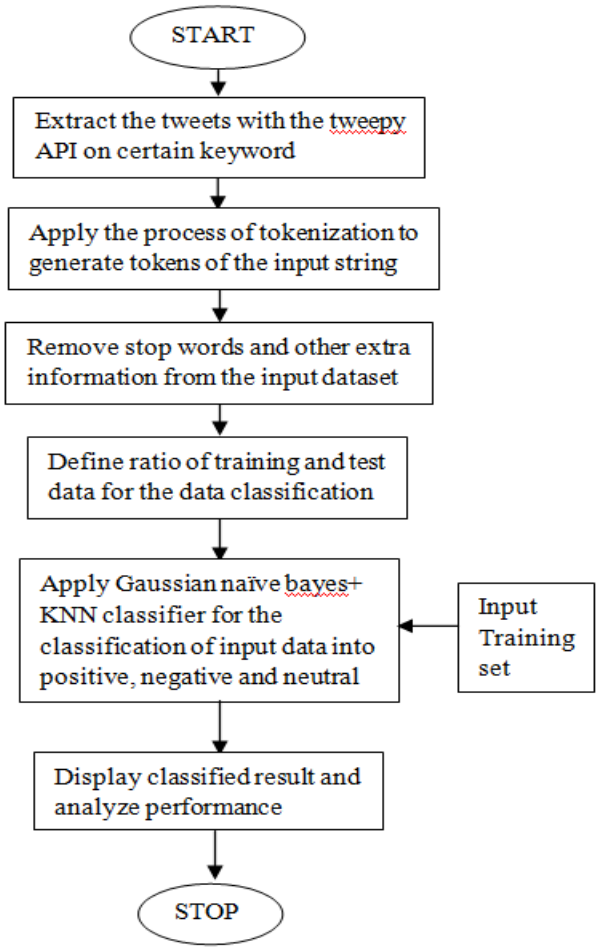

Fig. 1: Proposed System Architecture

As shown in figure 1, the systematic approach for proposed technique is explained in which both N-gram and KNN classifiers are utilized.

A. Dataset: Two types of datasets are generated manually here amongst which one is used for training and another is used for testing. X:Y is the relation present within the training set. The score of probable opinion word is represented by X here and the representation whether the score is positive or negative is done by Y. By gathering reviews from the e-commerce sites, the testing set is generated. A review whether the testing set is positive or negative is manually tagged. The reviews will be separated on the basis of positive and negative sentiments they include once the training is completed. With the help of reviews that are gathered from the test set whose polarity is known previously, the system is tested. The accuracy of the system can be determined on the basis of output that is generated by the system.

B. Data Preprocessing: Stemming, error correction and stop word removal are the three main preprocessing techniques which are performed here. The identification of root of a word is the basic task within stemming process. The elimination of suffixes and number of words involved is the major aim of this method. It also ensures that the time as well as memory utilized by the system is saved up to maximum. Since, similar grammatical rules, punctuation as well as spellings are not utilized by all the reviewers; there is a need to develop error correction mechanism. The context is understood in different manner due to such mistakes and thus, correction needs to be done here. In order to minimize the complexity of the text, the stop words are eliminated. The core reference of the resolution might get effected due to elimination of some words such as "it" which should be avoided.

C. Lexical Analysis of Sentences: A subjective sentence is known as one which includes either a positive or a negative sentiment. However, there are some queries or sentences written by the users which might not include any sentiments within them and thus are known as the objective sentences. In order to minimize the complete size of the review, such sentences can be removed. A question mainly is generated by including words such as where and who which a sentence which also does not provide any sentiments. This type of sentence also is removed from the data. The regular expressions involved within python do not recognize these questions.

D. Extraction of Features: The major issue arises within the sentiment analysis while extractive the features from data. A noun is always utilized in order to represent the features of a product. POS tagging is utilized in order to recognize and extract all the nouns such that all the features can be recognized. There is a need to eliminate the features that are very rare. A list of features that occur very frequently can be generated after the rarely present features are eliminated. The

N-gram algorithm is applied which can extract the features and also post tag the sentences.

E. Define Positive, Negative and Neutral Words: With the help of Stanford parser, the words that represent a specific feature can be extracted. The grammatical dependencies present amongst the words present in the sentences will be gathered by the parser and given as output [13]. In order to identify the opinion word for features that have been gathered from the last step, the dependencies have to be looked upon in further steps [14]. The direct dependency is referred to as the direct identification of opinion words for particular features. There is also a need to include the transitive dependencies along with direct dependencies within this step.

F. SentiWordNet: Within the opinion mining applications, the Sentiwordnet is generated especially. There are 3 relevant polarities present for each word within the Sentiwordnet which are positivity, negativity and subjectivity. For instance, 125 is the total score for the word "high" within the SentiWordNet. However, the word high cannot be considered as positive within the sentences such as "cost is high". In fact, there is negative meaning represented by this sentence. Thus, such situations need to be considered here as well.

G. K-Nearest Neighbor Classifier and naïve bayes: In order to use a classifier within this approach, KNN is selected. Since, sentiment analysis is a binary classification and there are huge datasets which can be executed, KNN is chosen here. A manually generated training set is utilized for training the classifier here. There is X:Y relation provided within the training set in which the score of an opinion word is represented by x and the score whether the word is positive or negative is represented by y [15]. A score of the opinion word related to a feature within the review is given as input to KNN classifier. Naïve Bayes is a statistical classifier which accepts no dependency between attributes. This classifier calculation utilizes conditional independence; means it expects that an attribute value on a given class is independent of the values of different attributes. The advantage of utilizing credulous bayes is that one can work with the Naïve Bayes model without utilizing any Bayesian methods

H. Extraction of Feature Wise Opinion: All the reviews that include that feature are to be considered in order to extract the opinion relevant to a particular feature. For a specific feature, the ratio of total number of reviews that have positive sentiments to the total number of reviews available is calculated. The ratio of total number of reviews within which a negative sentiment related to a feature is given to the total number of reviews present is calculated as the eventual negative score for particular feature.

IV. EXPERIMENTAL RESULTS

The proposed approach is implemented in Python and the results are evaluated by making comparisons against proposed and existing techniques as shown below.

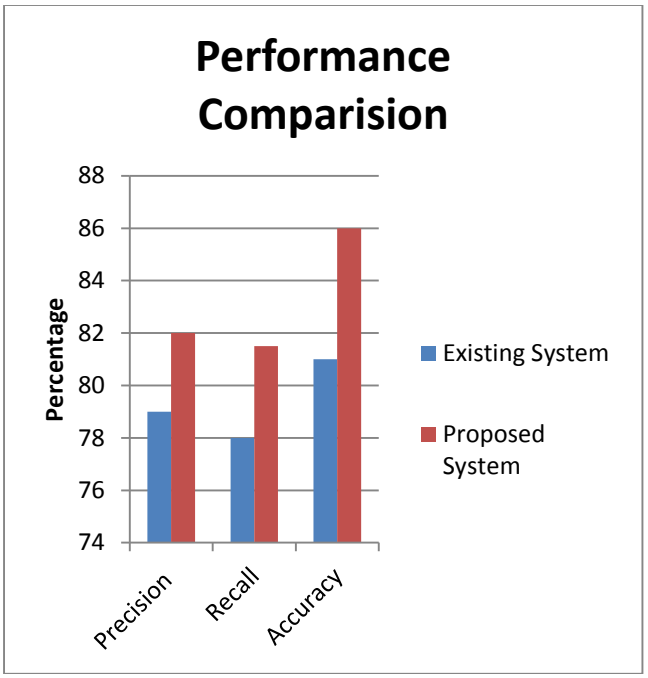

Fig2: Performance analysis

As shown in figure 2, The value of precision of proposed system is approx 82 % on the other hand the precision value of existing system in which SVM is used is approx 79 percent. The recall of proposed system is 81.5 percent where recall value of existing system in which SVM is used is upto 78 percent. The accuracy of proposed system is achieved upto 86 percent where accuracy of existing system is approx 81 percent of positive, negative and neural tweets classification.

V. CONCLUSION

The behavior of user is analyzed in this research work on the basis of analysis sentiments of twitter data. N-gram technique is applied here for sentiment analysis through which the features of input data are analyzed. Further, the behavior of user is analyzed by applying classification technique. The complete input dataset will be divided into various segments using the N-gram approach. For analyzing the sentiments, each of these segments is analyzed individually. The classifier used for this analysis is logistic regression. There are several number of classes generated during data classification. To perform sentiment analysis, further, the KNN classifier is used instead of logistic regression classifier. As per the simulation results it is seen that in comparison to logistic regression, the accuracy of KNN classifier is better. Also, the results are seen better for KNN classifier in terms of execution time.

VI. REFERENCES

[1]. Tharindu Weerasooriya, Nandula Perera, S.R. Liyanage. A method to extract essential keywords from tweet using NLP.

2016 16th International Conference on Advances in ICT for Emerging Regions(ICTer).
[2]. Ibrahim A. Hameed. Using Natural language processing for designing socially intelligent robots. 2016 Joint International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob).
[3]. L. SUanmali, M. S. Binwahlan, and N. Salim. Sentence features fusion for text summarization using fuzzy logicin Hybrid Intelligent Systems. 2009, HIS'09, Ninth International Conference on, vol. 1, IEEE, 2009, pp. 142-146.
[4]. L. Suanmali, N. Salim, and M. S. Binwahlan. Fuzzy logic based method for improving text summarization. arXiv pre print arXiv:0906.4690, 2009.
[5]. X. W. Meng Wang and C. Xu. An approach to concept oriented text summarization, Proceedings of ISClTS05, IEEE internationalconference, China,1290-1293" 2005.
[6]. M. G. Ozsoy, F. N. Alpaslan, and 1. Cicekli. Text summarization using latent semantic analysis. Journal of Information Science, vol. 37, no. 4, pp. 405-417, 2011.
[7]. Adyan Marendra Ramadhani, Hong Soon Goo. Twitter Sentiment Analysis using Deep Learning Methods. 7th International Annual Engineering Seminar (InAES), Yogyakarta, Indonesia, 2017.
[8]. Dan Cao, Liutong Xu. Analysis of Complex Network Methods for Extractive Automatic Text Summarization.2016 2nd IEEE International Conference on Computer and Communications, vol. 9, iss. 8, pp- 97-110, 2016.
[9]. Rasim Alguliyev, Ramiz Aliguliyev, Nijat Isazade. A Sentence Selection Model and HLO Algorithm for Extractive Text Summarization, IEEE, vol. 9, iss. 8, pp- 97-110, 2016.
[10]. Narendra Andhale, L.A. Bewoor. An Overview of Text Summarization Techniques. IEEE, vol. 9, iss. 8, pp- 97-110, 2016.
[11]. Rupal Bhargavaand Yashvardhan Sharma. MSATS: Multilingual Sentiment Analysis via Text Summarization, IEEE, vol. 9, iss. 8, pp- 97-110, 2017
[12]. Archana N.Gulati, Dr.S.D.Sawarkar. A novel technique for multi-document Hindi text summarization. 2017 International Conference on Nascent Technologies in the Engineering Field (ICNTE-2017), vol. 8, pp. 1-4, 2017.
[13]. Manisha Gupta, Dr.Naresh Kumar Garg. Text Summarization of Hindi Documents using Rule Based Approach, International Conference on Micro-Electronics and Telecommunication Engineering, vol. 8, pp. 1-4, 2016.