Sentiment Analysis of Twitter Data using KNN Classification Technique

Anjume Shakir[1] Jyoti Arora[2]
[1]M.Tech Scholar [2]Assistant Professor
[1,2]Desh Bhagat University, Punjab, India

Abstract— The data mining is the approach which can extract the useful information from the large amount of data. The classification is the approach of data mining which can classify whole data into finite classes. This research is based on the sentiment analysis in which sentiments of the input data are analyzed. In the existing work, SVM classifier is applied for the classification. In the proposed work, SVM will be replaced with the KNN which increase accuracy of classification. The proposed algorithm will be implemented in anaconda and results show improvement in accuracy and reduction in execution time.

Key words: SVM, KNN, Anaconda, Sentiment Analysis

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. Within the big data analysis, there are three major categorizations in which the data can be differentiated. They are structured, semi-structured and unstructured [1].There is majorly the text and multimedia type of data present within this category. The data present in the e-mail messages, word documents, images, presentations, videos and various other documents is all included within this category. Natural language Processing (NLP) is an application of the computational linguistics [2]. NPL is an important tool in area of artificial intelligence as it helps in interaction of robots in human natural language with humans. With the advancement in the technology, the internet is accessible through various devices like smart phones, smart watches and within the reach of common people. The communication of users through the social media has been exponentially increased [3]. A considerable piece of information exchange happens as online conversations like Internet Relay Chats (IRC), Facebook and Twitter streams. Among them, we concentrate on conversations from the online support gatherings which plan to examine and resolve user-related issues. 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. 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 [4]. The user gives the text reviews like online reviews, comments or the feedbacks on the social media sites, e-commerce websites. This text is an important source of user's opinions. 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. The resources used are lexicographical, and the lexical method is to collect the seeds of the sentiment words and their orientation to find their antonyms and synonyms to expand their set. When no more words left this iteration stops. The issue related to the sentences classification has been solved with the help of machine learning approach as it totally based on the algorithms [5]. A text classifier is trained on a human labeled training dataset. Supervised leaning approach and unsupervised learning approach are the two utilized approaches. Supervised learning is defined as the bulk of labeled training document and is divided further in two parts such as Naïve Bayes algorithm and Maximum Entropy Classifier. There are different sorts of classifiers that are generally utilized for text classification which can be likewise utilized for twitter sentiment classification. Naive Bayes provides good result in spite of having low Naïve Bayes Classification probability [6].

II. LITERATURE REVIEW

Dan Cao, et.al proposed that an Automatic Text Summarization is an important research area in the domain of information systems. In extractive text summarization, sentences are scored on a few of features. A large number of features network based have been proposed by researchers in the past literatures. This paper reviews every one of the features that utilization metrics and idea of complex network for scoring sentences [7]. Rasim Alguliyev, et.al proposed that text summarization is represented as a sentence scoring and selection process in this paper. The process is displayed as a multi-objective optimization issue. 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 [8]. 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. Narendra Andhale, et.al 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 [9]. There are various extractive and abstractive types of summarization methods which are studied in this paper. 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 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. The strategy leverages diverse techniques of machine learning to dissect the text. Machine translation is

utilized as a part of the system to give the component of dealing with various languages. After the machine translation, text is processed for finding the sentiments in the text [10]. Archana N.Gulati, et.al 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 thinking about the above issue a novel procedure for multi document, extractive text summarization is proposed [11]. 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 proposed by author that automatic summarization assumes an important part in document processing system and information recovery system. In this paper we present a novel approach for text summarization of Hindi text document based on some linguistic principles [12]. 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. Tools and Techniques

Since the data is taken from the twitter, the algorithm will work for any data in text format after the relevant preprocessing techniques. We have used the KNN algorithm for this research as it is a supervised machine learning algorithms and yields good results. The implementation of the code was done in Anaconda which supports lot of inbuilt packages for machine learning like numpy, sci-py etc. Anaconda supports lot of tools, for this research Spyder was used which is the python development tool.

IV. Research Methodology

The summarization technique includes following steps:

1) Dataset inherited: - The data which is given as input will be taken from the twitter. The data will be downloaded using the twitter API
2) Data Pre-Processing:- In the second phase, the data which is taken as input will be pre-processed means the un-wanted data will be removed.
3) Analyzing features of the Dataset :- The dataset which is pre-processed and on that data algorithm of n-gram is applied for the feature extraction
4) Chat Summarization:- In the last step, the rating to each word is given on the basis of their occurrences and the words with maximum rating is considered as most important words that are included in the final chat summary and others are removed.

The pattern based algorithm is the algorithm which generates patterns of the input data. The input data will be divided into certain phases and these phases are generated using the N-gram algorithm which will make combinations with the others words in the dataset. The weight is assigned to each word, character in the chat for generation of final chat summary. The patters are generated using N-Gram algorithm. Figure 1 shows the methodology followed for research.

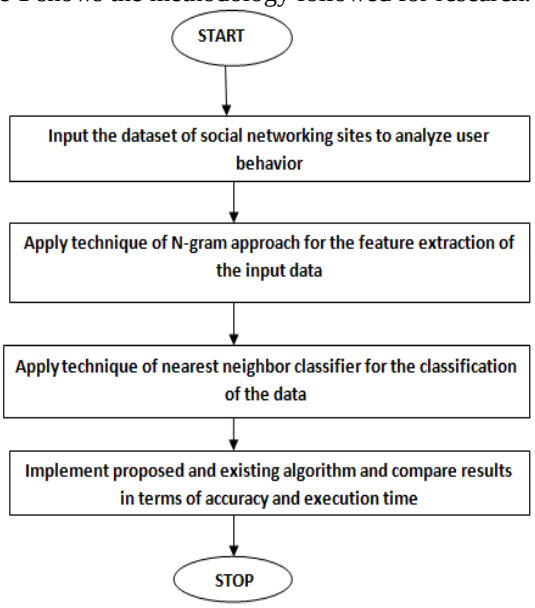

Fig. 1: Research Methodology

V. Experimental Results

The Data set was taken from the twitter and preprocessed and then KNN algorithm was applied on it. The algorithm has shown good results in classification of the sentiments and the accuracy is better than the existing system. This algorithm was applied to binary, ternary and multi-class sentiment analysis. The figure 2 shows the comparison of accuracies of the existing and proposed system.

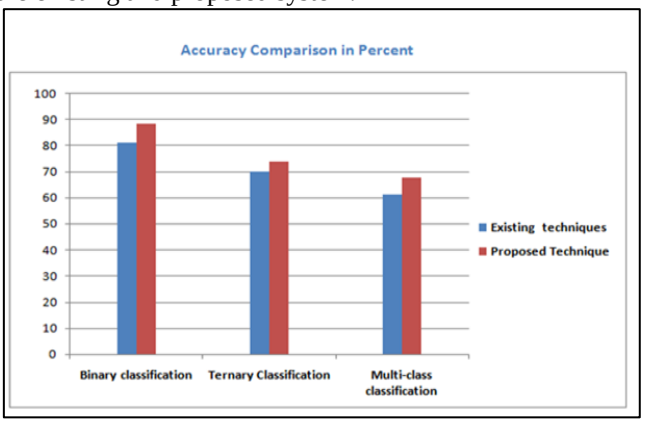

Fig. 2: Comparison of Accuracy of Proposed and Existing System

VI. Conclusion

In this work, the movie dataset of around 6000 tweets was taken and preprocessed and sentiment analysis approach was applied to classify the sentiments related to these tweets. The accuracy of binary, ternary and multi-class sentiment analysis was improved by using the KNN algorithm. The execution time using KNN is also less than the other existing techniques.

VII. FUTURE SCOPE

The size of the dataset can be increased and the new techniques can be applied to classify the sentiments. There is always scope for improvement therefore there is room for improving the accuracy and execution time. The hybrid approach can be used to make the sentimental analysis more effective.

REFERENCES

[1] Diakopoulos, N., Shamma, D.," Characterizing debate performance via aggregated twitter sentiment", 2010, Proc. 28th International Conference on Human Factors in Computing Systems, ACM, vol. 4, pp.45
[2] Gangemi, A., Presutti, V., Reforgiato Recupero, D.," Frame-based detection of opinion holders and topics: A model and a tool", 2014, IEEE Computational Intelligence Magazine, Volume1, Issue 13, pg no.19
[3] Go, A., Bhayani, R., Huang, L.," Twitter sentiment classification using distant supervision", 2009, vol46 issu 16, pp. 59, CS224N Project Report, Stanford
[4] Speriosu, M., Sudan, N., Upadhyay, S., Baldridge, J.," Twitter polarity classification with label propagation over lexical links and the follower graph", 2011, Proceedings of the EMNLP First workshop on Unsupervised Learning in NLP, Edinburgh, Scotland volume 30, issue 32, page no. 456-475
[5] Thelwall, M., Buckley, K., Paltoglou, G.," Sentiment strength detection for the social web", 2012, J. American Society for Information Science and Technology, volume 63 issue 1, pp- 163–173
[6] Thet, T.T., Na, J.C., Khoo, C.S., Shakthikumar, S.," Sentiment analysis of movie reviews on discussion boards using a linguistic approach", 2009, Proc. the 1st International CIKMWorkshop on Topic-sentiment Analysis for Mass Opinion, vol 6 issue 2, pp- 52
[7] 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.
[8] 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.
[9] Narendra Andhale, L.A. Bewoor. An Overview of Text Summarization Techniques. IEEE, vol. 9, iss. 8, pp- 97-110, 2016.
[10] Rupal Bhargavaand Yashvardhan Sharma. MSATS: Multilingual Sentiment Analysis via Text Summarization, IEEE, vol. 9, iss. 8, pp- 97-110, 2017
[11] 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.
[12] 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.

