Sentiment Analysis on Tweets using Machine Learning

Richa Garg[1], Dr. Ochin Sharma[2]
[1]Research Scholar, [2]Assistant Professor
[12]Manav Rachna International Institute of Research and Studies, Faridabad, India

Abstract- The approach that is used to predict the polarity of any content as being positive, negative or neutral is called sentiment analysis. This research work designed a new mechanism for this study using the previous study as its base. The classification and feature extraction techniques are combined together to design this proposed technique. To perform feature extraction, n-gram algorithm is applied. Further, the input data is categorized among positive, negative and neutral using KNN classifier. Certain performance parameters such as recall, accuracy and precision are calculated for validating the proposed system. It is seen through the experimental results that in comparison to the existing approach that uses SVM classifier, the performance of proposed approach is better.

Keywords- SVM, KNN, Naïve Bayes, Fake Profile Detection

I. INTRODUCTION

A study through which important knowledge is extracted from the raw data by breaking it down is called data analytics. This process can help in understanding the actual scenario of the user's work. It can help in making better decisions. The data analytic process includes certain actions like cleansing, inspection, modeling and transformation which collectively help in discovering important information present within the data [1]. In the data examination process, several facets and approaches are designed. With separate names, numerous techniques have been designed in individually separate domains. To extract important information such that it can be used in predictive forms a particular data investigation method has been proposed which is named as data mining [2]. However, business intelligence process is known as the analysis which is performed on the basis of aggregations that are completely dependent on the business information. Investigation is known as the process in which multiple components are generated from one complete document for proper investigative study [3]. To ensure that the data can be used at the time of generating decisions, the raw data is converted into useful form. The collection of data from numerous sources and then analyzing it helps in answering several questions such that the hypothesis can be tested or any theory can be disapproved. For data investigation a procedure is followed. Further, different techniques are executed for result interpretation and also for proper data arrangement. Thus, a highly precise data analysis can be performed and when applying proper measurements, data dissecting can be performed. There are three major categorizations in which data is divided in the big data analysis. Structured data is stored in rows and columns of a table which is included in the database SQL [4]. It is highly organized and also has a relation key through which mapping the pre-designed fields is very easy. Semi-structured data has some authoritative properties even if it is not available in the relational database. It is easier to analyze this kind of data. A special organizational format is generated by organizing this data [5]. The unstructured type of data comprises of around 80% of the total amount of existing data. There is no particular structure of this kind of data. This category mainly includes text and multimedia kind of data. An application that includes computational linguistics is called Natural Language Processing (NLP). The text is interpreted and analyzed through NLP. The area of Computer Science and Artificial Intelligence which helps in interacting and interpreting computer and human natural language is known as NLP. For providing appropriate review about the product, the complete information and opinion about the product are collected and categorized. For the analysis of opinions of individual users, several improvements of the collected data are done [6]. The opinions of users can be posted through blog posts with the help of various social networking platforms. The manners in which users express their views and opinions are changed using the social network sites among which few popularly known sites are Google, Instagram, Twitter and Facebook. All the reviews of clients related to the products and services can be achieved here. The recovery of textual information technique is processed, searched and analyzed by the available accurate data. Developing new applications becomes challenging since the blogs and social sites include huge amount of data [7]. For tweeting the ongoing messages in twitter, micro blogging and social networking sites are includes. For creating new challenges and shaping various domains and methods included in sentiment analysis, various unique properties included in tweets are used. Several kinds of approaches are used to perform text classification which further helps in performing twitter sentimental classification. There are lexicographical resources are used in these approaches. The parts of sentiment words and their orientation are used to expand their set by finding their antonyms and synonyms are collected in the initial method [8]. The machine learning based approach is used to solve the various problems related to sentence classification. On a human labeled training dataset, a text classifier is trained. There are two commonly known approaches used here which are supervised and

unsupervised learning approaches. Hybrid approach is generated by combining the elements of lexicon-based as well as machine learning techniques to perform sentiment analysis. For determining the semantics available in sophisticated manner, these approaches are used as semantics networks and ontology. To perform text classification, various kinds of classifiers are used [9]. They help in performing twitter sentiment classification in highly efficient manner. Even though there is low Naïve Bayes classification probability, efficient results are achieved by Naïve Bayes algorithm. This algorithm is based on Bayes theorem and is a supervised machine learning approach. To perform classification, a huge edge is achieved through SVM classifier. The hyper plane technique is used to separate the tweets based on the different between tweet and hyper plane.

II. LITERATURE REVIEW

Rupal Bhargava, et.al (2017) proposed approach used different machine learning techniques for analyzing text. In the system, Machine translation was used to deal with different features of different types of languages [10]. In order to find sentiments within the text, text was processed after machine translation process. Substantial text was obtained on the internet with the introduction of blogs, forums and online analysis. The extraction of important text from this Substantial text was proved helpful in the reduction of processing. Therefore, in this study, text summarization process was used for extracting significant portions of text. These portions were utilized to scrutinize sentiments about the specific subject and its sides. The tested outcomes demonstrated that the proposed approach showed good performance.

Archana N.Gulati, et.al (2017) stated that a text summary was identified as the reduction of real text. In text summary, the text was summarized by choosing important text within the source. In the last few years, with the growth of World Wide Web, large volume of information was generated and presented through internet [11]. Text summarization was needed to get the essence of a specific topic from several sources of information existing online. In this study, a novel method named as extractive text summarization was proposed for various documents. The proposed approach achieved a standard accuracy of 73% over numerous Hindi documents. The system made summary was very close to the human made summary. The accuracy of summary system generated summary was shown in terms of Precision, Recall and F-score values.

Akshi Kumar, et.al (2017) performed the comparison of three keyword extraction techniques. These approaches were utilized in the automatic text summarization systems on the basis of text demonstration and summary creation factors [12]. These approaches were compared on a universal data suite of physically generated articles. The comparison of these algorithms was performed on time scale basis and on the basis of their efficiency in the extraction of keywords. The outcomes were achieved and compared with the scores of manually created summaries. These summaries were used as benchmark for comparison. The algorithms by now reached the score achieved by human summaries. In future, the more attention will be given to the enhancement of summary accuracy by using these algorithms together with the help of machine learning and swarm based techniques.

Shahnawaz, et.al, (2017) stated that sentiment analysis procedure was used for identifying opinion or beliefs articulated in the opinioned information to recognize the feelings of writer with respect to a specific topic [13]. The curiosity of the technical society and trade world was increasing gradually in gathering, processing and extracting of information from the public reviews presented on different social media platforms. Incapability to show good performance in different fields and scarce correctness were the major issues of existing methods. These issues rose due to inadequate labeled information, inability to handle difficult sentences that required surplus sentiment words and simple assessments. It was identified that semi-supervised and unsupervised learning based models could be used to solve the issue of labeled information scarcity.

N. Moratanch, et.al (2017) stated that the text summarization system contained several methods. These methods were categorized as the extractive and abstractive methods [14]. In this study, a wide-ranging analysis of extraction based text summarization methods was given. In this study, a review on extractive summarization methods was presented by classifying these methods into supervised learning and unsupervised learning approaches. On the basis of different techniques, the benefits of these approaches were presented in this study. A number of assessment techniques, concerns and future researches were included in this review as well.

Manisha Gupta, et.al (2016) proposed a new technique for the text summarization of Hindi text document on the basis of a number of linguistic conventions [15]. In order to generate lesser number of words from the original text, dead wood words and phrases were detached from the real document as well. The proposed approach was examined on different Hindi sources of info and precision of the system in terms of number of lines retrieved from real text having significant information of the real text document**.** It was identified that the proposed approach reduced the text size of information up to 60% - 70 %. It was also recognized that system generated the extractive summary provided by the client. This denoted that it did not produce text summary based on the principle of the text semantics.

III. RESEARCH METHODOLOGY

The figure 1, shows the construction of the proposed system which is based on N-gram and KNN classification model.

A. Dataset

Mainly two categories of information samples are produced here manually. One data sample is utilized for training and other dataset is utilized for testing. Inside the training sample, X: Y association is presented. X is used for the representation of possible estimation remark grade and Y is utilized for the estimation of positive or negative grade. The testing set is produced after the attainment of remarks available on different social media sites. For the identification of optimistic or pessimistic test sample, a remark is tagged physically. After the completion of training, the appraisals will be alienated on the basis of optimistic and pessimistic opinions. With the help of appraisal from the test sample, gathered earlier with known polarity, the testing process of method is performed. The accurateness of the arrangement may be resolute on the base of outcomes produce by the scheme.

B. Data Preprocessing

Mainly three kind of pre processing methodologies named Stemming, error correction and stop word removal are implemented in this research work. The essential job of stemming process is the detection of a root of a statement. The main objective of this technique is the removal of suffixes and number of terminology concerned. With the help of this approach, time and memory consumption of the scheme can be reduced up to large extent. The development of error correction system is necessary because alike grammatical regulations, punctuation as well as spellings are not used by all the assessors. The situation may be unstated in dissimilar because of these errors and therefore some kind of rectification is needed. The stop words are removed for the minimization of text complication. The nucleus orientation of the declaration may get effect because of the removal of some terminology such as "it" which must be evaded.

C. Lexical Analysis of Sentences

The sentence which comprises either an optimistic or a pessimistic opinion is known as subjective sentence. Though, there are some questions or phrase written by the followers which may not embrace any emotions inside them are called as objective sentences. In order to minimize the absolute size of the review, these sentences can be detached to reduce the absolute dimension of review. A query is mostly produced with the adaption of some words like "where" and "who", these words in the phrase do not give any opinion. Such kind of phrase is eliminated from the records as well. The usual terminology implicated inside python does not distinguish these queries.

D. Extraction of Features

During the extraction of characteristic from data sample, the main problem occurs inside the sentiment analysis. For the representation of the characteristics of a product, a noun is used always. POS tagging is used to identify and extract all the nouns for the recognition of all features. Extremely exceptional features are needed to be removed from here. After the elimination of rarely present features, a trail of frequently generated characteristics can be obtained. For feature extraction and for the post tagging of phrases, N-gram approach is utilized.

E. Define Positive, Negative and Neutral Words

The words representing a particular feature can be extracted with the help of Stanford parser approach. The parser gathers the grammatical reliance's present amid the words used in sentences and applies it as output. The reliance will be considered in the next few stages, for the identification of features of opinion words gathered from the last stage. The straight reliance is considered a straight recognition of view lexis for specific features. The indulgence of transitive reliance is also needed in association with straight reliance's inside this stage.

F .SentiWordNet

The Sentiwordnet is generated particularly in the opinion mining applications. Mainly three related polarities exist for every word inside the Sentiwordnet and they are recognized as positivity, negativity and subjectivity. For example within the SentiWordNet, 125 is the complete grade for word 'high'. Also the word high may not be used as positive word in some sentences like "cost is high". This sentence represents a negative approach in fact. Therefore these kinds of situations also measured carefully.

G. K-Nearest Neighbor Classifier

A classifier named KNN is chosen for this approach. Because, sentiment analysis is a binary classification and a large amount of data samples are present for execution, therefore KNN classifier is selected in this study. A physically generated training used to train the classification model. An X: Y relation is provided inside the training suite where x represents the score of an opinion word while y is used for the representation of positive or negative word and gives them score accordingly [15]. The score of opinion word relevant to the feature in the review is applied as input to the KNN classification model.

H. Extraction of Feature Wise Opinion

For the extraction of opinion associated with a specific characteristic, entire remarks which involve characteristic must be utilized. For the attainment of a particular characteristic, the remarks containing optimistic opinions rationed with complete existing remarks are computed. For

the evaluation of negative score of a specific characteristic, the ratio of whole number of appraisal inside a negative emotion connected to a feature is applied to the whole figure of appraisal present is considered.

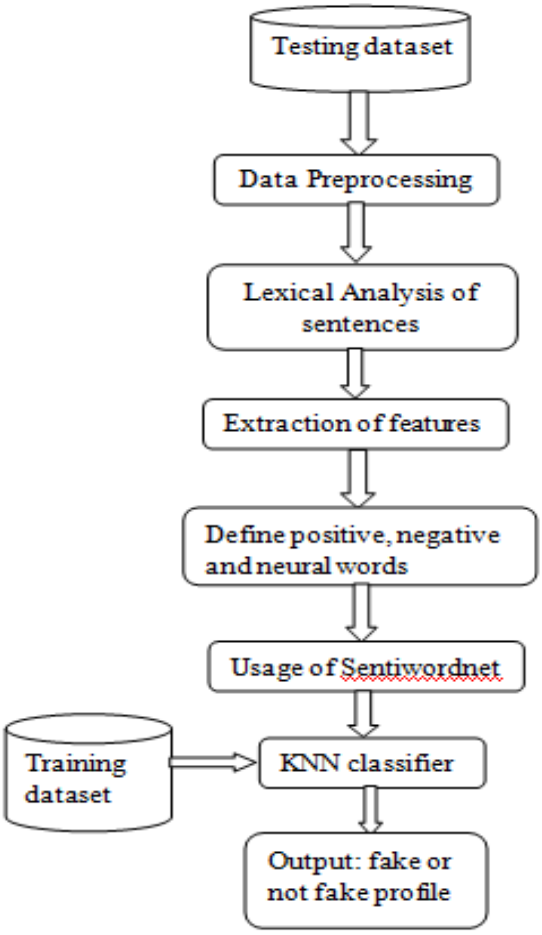

Fig.1: Proposed System Architecture

IV. EXPERIMENTAL RESULTS

The proposed research work is implemented in Python and the results are evaluated by comparing the results of proposed and existing techniques in terms of different performance parameters.

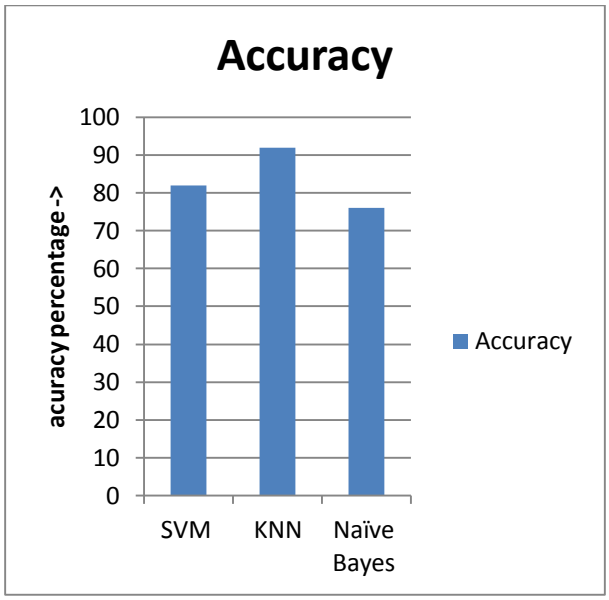

Fig.2: Accuracy Comparison

As shown in figure 2, the accuracy of the three classifiers are compared for the performance of analysis. The three classifiers are SVM, KNN and Naïve bayes for the performance analysis.

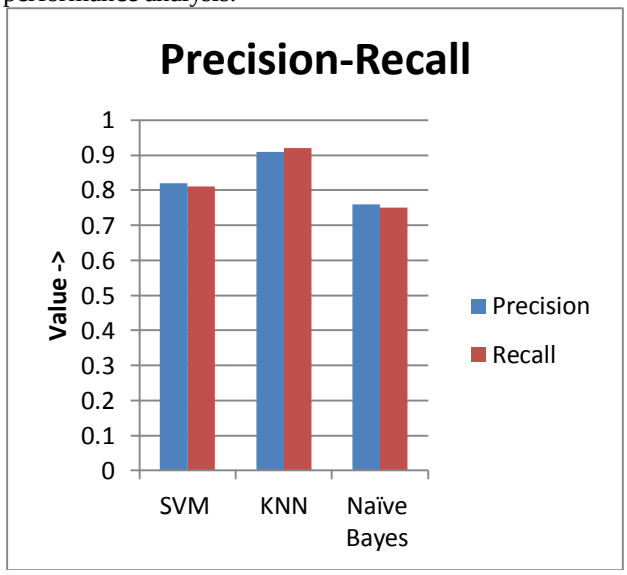

Fig.3: Precision-Recall Comparison

As shown in figure 3, the precision-recall of the three classifiers are compared for the performance of analysis. The three classifiers are SVM, KNN and Naïve bayes for the performance analysis

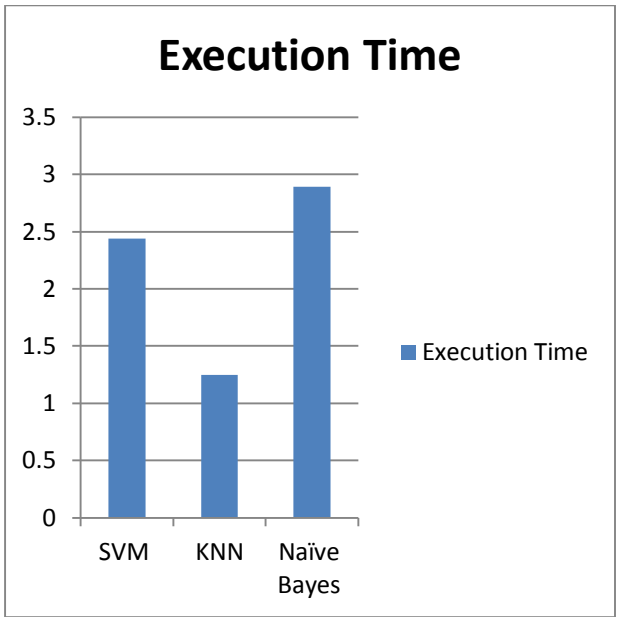

Fig.4: Execution Time Comparison

As shown in figure 4, the execution time of three classifiers is compared for the performance of analysis. The three classifiers are SVM, KNN and Naïve bayes for the performance analysis

V. CONCLUSION

The emotions and attitudes of individuals on certain events are handled through sentiment analysis. In different applications like reviewing products, analyzing social media content or reviewing the movies, opinion mining is used commonly. This research is based on combining the KNN and n-gram classifiers for performing sentiment analysis. Several techniques have been designed for performing sentiment analysis over the past years. The previously designed technique that used SVM classifier for classifying the tweets as positive, negative and neutral was used as base to design the new approach. N-gram and KNN classifiers were used in the new approach such that the extraction of input was done using N-gram and data was categorized based on their polarity using the KNN classifier. The proposed and existing approaches are compared with each other for performance evaluations. It is seen that around 7% of improvement in sentiment analysis is achieved by implementing the proposed approach.

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]. K. Kaviya, C. Roshini, V. Vaidhehi, J. Dhalia Sweetlin. Sentiment for Restaurant Rating. 2017 IEEE International Conference on Smart Technologies and Management for Computing, Controls, Energy and Material (ICSTM).

[9]. Devika M D, Sunitha C, Amal Ganesh "Sentiment Analysis:A Comparative Study On Different Approaches",Procedia Computer Science, vol.87 , pp. 44-49,2016

[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]. Akshi Kumar, Aditi Sharma, Sidhant Sharma,Shashwat Kashyap. Performance Analysis of Keyword Extraction Algorithms Assessing Extractive Text Summarization. International Conference on Computer, Communication, and Electronics (Comptelix), 2017.

[13]. Shahnawaz, Parmanand Astya "Sentiment Analysis: Approaches and Open Issues" International Conference on Computing, Communication and Automation, vol. 9, pp. 1-5, 2017

[14]. N. Moratanch, S. Chitrakala. A Survey on Extractive Text Summarization.IEEE International Conference on Computer, Communication and Signal Processing (ICCCSP), vol. 8, pp. 1-4, 2017.

[15]. 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.