MACHINE LEARNING APPROACH FOR SENTIMENT ANALYSIS OF TWITTER DATA

Meetali [1], Amandeep Kaur Sohal[2] and Mandeep Kaur[3]
[1] Student , [2]Assistant Professor and [3]Assistant Professor
Computer Science and Engineering Department
Guru Nanak Dev Engineering college, Ludhiana-141006,Punjab,India

Abstract: Sentiment analysis is the process of computationally identifying and categorizing opinions expressed in a piece of text, especially in order to determine whether the writer's attitude towards a particular topic, product, is positive, negative, or neutral. Sentiment analysis is done to get public opinions and also used for marketing of products to make business strategies on basis of these opinions. It can be used to identify the follower's attitude towards a brand and its products through the use of variables such as context, product reviews, tones, emotions. Different kinds of sentiments that people have towards amazon products or their services are analysed using Amazon data. The best source available to collect the sentiments is internet. Twitter is social networking platform that provides information regarding sentiment of the user's w.r.t to their purchase of products. In this research, the KNN-LSTM mechanism is applied instead of WDE-LSTM to predict the more accurate sentiments. Python simulator is used to implement the proposed techniques. Around 94.5% of accuracy is achieved approximately. The proposed algorithm also compared in terms of execution times and it shows that the proposed algorithm outperforms in terms of execution time.

Keywords: Sentiment analysis, SVM, KNN

I. INTRODUCTION

This technology extracts the patterns from data by performing few tasks on the dataset [1]. The various tasks performed here are categorized among predictive and descriptive data mining tasks. The former are responsible for performing predictions based on the existing dataset and the latter help in understanding the characteristic properties of dataset. The applications using data mining examine the data using various parameters which include clustering, pattern analysis, classification and association. It is known as the process that examines the features of new object and assigns them to a predefined class. For characterizing the classification task, well-defined classes are used along with a training set that includes pre-classified instances [2]. This method generates a model through which the unclassified data is identified and classified into classes.

Estimation and classification are the other names commonly used for this approach which are much alike. In case if data mining access the phone line that is being used for accessing internet, it is not possible to perform a recheck to ensure that correct classification is performed. Mainly due to the presence of incomplete knowledge, there exists an uncertainty that the classification is correct or not. In real time scenarios, relevant actions are being performed [3]. It is possible that the phone is being used or not to dial the local ISP primarily. For the credit card transaction is the possible that the fraudulent is seen or not. Sufficient efforts need to be done for performing checking. Predictive tasks perform differently since the classification of records can be done [4] based on few predicted future behaviors or estimated future values. An approach also known as opinion mining in which the opinions of people related to particular services are categorized is known as sentiment analysis. Based on the emotions and attitudes of certain event or object, the opinions and perspectives of humans are analyzed through sentiment analysis. In the applications like social media analysis or commercial product reviews, opinion mining is performed [5].

For creating the recommender systems, semantic analysis is considered as a valuable technique. On the e-commerce and social networking websites, several online reviews and comments are mentioned by users. These sources help in understanding the opinions of users in an effective manner [6]. For checking if the reviews of users about the products are positive, negative or neutral, the sentiment analysis is performed. These reviews help in defining the important or popularity of products in the competitive market. For a specific event, the opinions, feelings, thoughts and emotions are different for every human being. Each of the sentiment denotes a different category since every sentiment analysis can be considered as a separate task of classification process [7]. Since it deals with the human and computer language interaction, the AI and computer science play an important role in NLP. Due to the huge changes arising in market level of competition, more research needs to be done in sentiment analysis such that effective outcomes can be achieved.

The opinions of speaker are determined by the sentiment analysis otherwise commonly known as opinion mining [8].Here, an appropriate review related to a product or service is provided through this method [9]. It analyzes all the opinions and information provided only related to the product under review. For analyzing the opinions of an individual user, the data collected from various users is improved. For posting their opinions or using blog posts, several social networking platforms are provided to the users online. The platforms like Instagram, Google, Twitter, and Facebook are some of the commonly known sites [10].

II. LITERATURE SURVEY

The Literature survey reveals that sentiment analysis had been a major investigational domain in the last few years. The most of the researches conducted in this area were mainly focused on English language. Various methods and machine learning techniques have been used for automatic text summarization like kernel tree ,naïve bayes have been used ,binary and ternary classification for unigram model. Each of the methods and their studies along with some of the limitations have been described below.

Cao D. et al. (2016) stated that Automatic-Text Summarization approach intended to make a condensed adaptation of documents. All the important contents as well as common information should be covered using this version. In this study, all features that utilized metrics and thought of complicated network in order to score the sentences have been reviewed [20]. Also, the tested outcomes on individual module and mixture of various presented were analyzed.2002 data sets were utilized to evaluate quantitative and qualitative features. Shortened ways were identified as amazing for text summarization. With respect to the quality of produced summary, these ways attained maximum grades. An additional significance was the detecting those that featured mixtures with same assets of network and specified unbelievable effect on chosen sentences. It was identified that Sentence correlation among sentences became a necessary element in the retrieval of fine abstracts.

Alguliyev R. et al. (2016) stated that a good example of sentence scoring and selection procedure was text summarization. Due to this approach, massive text documents had been generated in the web and e-government. The quantity of these documents increased exponentially over the time. This massive growth in the amount of text documents had made reading process difficult for users. This also created problem in the retrieval of useful information from these documents. Thus, ATS approach had become an imperative discovery process due to the increase of text documents. Thus, this approaches grabbed attention of various researchers in the past few years [21]. In this study, main attention was given to extractive text summarization. In this approach, a summary was created with the help of scoring and selection of sentences in the source text. The score of each and every sentence has been evaluated initially in view of the fact that semantic resemblance among selected sentences would be low. In order to score sentences, one more formula was established. The proposed approach depicted accomplishments for finding equilibrium between coverage and repetition in an abstract. In this study, a human learning optimization algorithm was utilized to handle optimization problem.

Andhale N. et al. (2016) stated that the methodology used for the producing compressed structure of text documents was known as text summarization. This compacted document maintained important information and frequent significance of source text. An important technique with relevant information which has been recognized from enormous documents was called automatic text summarization approach. The study also highlights about the wide-ranging analysis of both approaches which were presented using text summarization [22]. In this study, numerous extractive & abstractive sorts of summarization approaches have been analyzed. An effective summary was to be generated by summarization approach in minimum time slot. This summary had less redundancy and included well-formed sentences. High-quality results were attained with the help of extractive and abstractive methodologies. These results could be utilized further by the users. The testing for hybridization was analyzed in this study for generating helpful, fine condensed and understandable abstracts.

Bhargava R. et al. (2017) predicts that S.A had been a major investigational domain in the last few years. The most of the researches conducted in this area were mainly focused on English language [23]. This study also predicts , a novel approach has been presented for analyzing various languages in order to discover sentiments in these languages and performed S.A. The proposed technique implemented various M.L.A methodologies for content inspection. Machine translation has been utilized in this mechanism to deal with different kinds of languages. In order to find sentiments in content, content was processed following the machine translation. The introduction of blogs, forums and online surveys resulted in the ample amount of online text. This sentiment regarding a specific topic or an object could be analyzed using this text. Therefore, it was advantageous to retrieve important text occurring within this text for reducing further processing. Thus, the presented structure utilized text summarization procedure in order to extract the significant kind of elements of text. On the other hand these parts has been utilized to examine the sentiments regarding some particular matter as well as its features.

Gulati A. et al. (2017) highlights that actual text has been reduced in text summary process for text summarization by selecting important information [24]. In past decade, due to the expansion of internet, massive volume of data had been generated and accessible though internet. The demand of a general idea regarding some specific subject from various online accessible information causes raised the need of text summarization. In this study, a new method for multiple documents and extractive text summarization was presented by considering this issue. As hindi is common language in India so summarizer or condenser has been fabricated for this language. Sports as well as political affairs input has been applied to the system in the form of news editorials which were available online. Eleven important aspects of the text had been utilized for retrieval procedure using Fuzzy inference system. The standard accuracy about 73% was achieved by the proposed approach. The system produced summary was similar to the human produced summary up to some extent. The system generated summary showed good values of various parameters like Recall, Precision & F1-score.

Gupta M. et al. (2016) state that an important role was played by automatic text summarization in document processing scheme and information extraction scheme. An imperative branch of NLP was to generate the summary of a text document. Automatic building of these summaries was quite helpful in various situations [25]. A longer text was compressed into smaller form in text summarization procedure. In this process, the information of text document had been maintained. Abstract of a larger text saved reading time due to lesser number of lines. This abstract however contained all significant information of actual text paper. A new scheme for summarizing Hindi text document was proposed in this study. This approach was based on several linguistic rules. For generating smaller amount of words from real document, as well as from phrases & dead-wood words has been eliminated from the actual text. The performance of resented approach had been tested on numerous Hindi inputs. The accuracy of proposed approach was obtained as amount of lines retrieved from actual document holding significant information of the actual text. The size of information text has been reduced using proposed approach within the range of 60% to 70 %. The extractive summary provided by user was generated by proposed approach.

III. RESEARCH METHODOLOGY

Following are the various research gaps :-

1. The sentiment analysis is the approach which can analyze sentiments of the text data with the classification. The classification methods used in the previous study give low accuracy which needs to be improved.

2. The technique which are proposed for the feature extraction in the previous study are not so efficient for the feature extraction which needs to improved.

Sentiment analysis approach has been implemented on various micro blogging sites. The attributes of input data have been extracted using pattern matching algorithm in sentiment analysis method. Classification methodologies have been applied to detect sarcasm. N-gram technique is utilized for feature extraction from micro blogging sites. Pattern-matching algorithm has been implemented with neural networks. The features have been classified using WDE-LSTM classification model. In this work, major issue is classification. The complexity is increased at steady rate after the implementation of WDE-LSTM classification model which in turn increases execution time. K-Nearest Neighbor (KNN) approach is proposed to implement for the elimination of both classification and regression analytical issues.

Following are the various phases of the research methodology:-

A. Data-set

In this study, two sorts of datasets are generated in manual way. Among these two sets, one set is used for training while other is used for testing. Within the training set, a relationship X: Y exist. Variable X represents the score of feasible opinion word while y represents the positivity or negativity of score. The testing set is created by collecting online reviews. A review regarding the positivity or negativity of testing set is labeled manually w.r.t positive and negative sentiments, the reviews will be divided. These reviews will be included after the completion of training. The system is tested using those reviews which are collected from the test set. The polarity of this test set is identified earlier. The generated output of system determines the accuracy of the system

B. Data Preprocessing

In this research, three preprocessing methods such as error correction, Stemming as well as stop word removal are being performed. In stemming procedure, recognition of root of a word is the fundamental job. The main goal of this approach is to eliminate included suffixes & amount of words. It includes following things like all URL (e.g. www.xyz.com), hash- tags (e.g. #topic) as well as targets (@username) are being removed. Upper-case letters are to be converted to lower-case letters. All the texts are to be broken down into tokens. This process is called tokenization. For example “this is very amazing mobile phone” is broken down into individual tokens such as “this”, “is”, “very”, “amazing”, “mobile”, “phone”. With spaces individual tokens are indentified. Stop words like articles, conjunctions, prepositions as well as pronouns are removed. This scheme ensures that the system will be consuming minimum amount of time & memory. As all reviewers do not utilize analogous grammatical rules, punctuation as well as spellings, therefore, an error correction method should be developed here. Due to these various types of errors, the context has been recognized in different ways which is able to generate the need as well as requirement of error correction. To reduce the complexity of the text, the stop-words should be removed to lessen the complexity of text. The removal of some words can affect the core reference of resolution for example “it” which should be eliminated.

C. Lexical Analysis of Sentences

Either a positive or a negative sentiment is included in a subjective sentence. Nevertheless, some of the questions as well as sentences written by clients may or may not comprise any sentiments within them. These sentences have been identified as Objective-sentences. These types of sentences should be eliminated to lessen the overall size of review. Generally, A query should be generated through the inclusion of words such as who & where. Such kinds of words do not have any sentiment within them. This sort of sentence is eliminated from the data as well. The standard expressions included in python do not identify these queries.

D. Extraction of Features

In sentiment analysis, one main issue occurs during the extraction of features from data. The features of a commodity are represented using a noun. For recognizing and extracting all nouns, POS tagging is applied. This is done to identify all features. Extremely infrequent features should be eliminated here. After eliminating rarely present features, the list of frequently occurring features can be created. The N-gram algorithm is implemented for feature extraction and post tags the sentences.

E. Define Positive, Negative and Neutral Words

The words which represents a particular kind of feature is being retrieved using Stanford parser. Parser generally collects grammatical-reliance occurring among sentence’s words and gives them as output. The dependencies should be considered in further steps for identifying opinion word for features that have been collected from the final step [14]. The direct recognition of opinion words for some specific features is called direct dependency. In this step, transitive dependencies should be included along with direct dependencies.

F .Senti-Word-Net

Senti-word-net has been generated particularly in various kinds of opinion-mining applications. For each & every word within the Senti-word-net , three kinds of pertinent polarities may occur. These polarities are identified as positivity-negativity as well as subjectivity e.g. 125

is the total score for the word "high" in the Senti-Word-Net. Though, the word "high" cannot be looked as positive in the sentences i.e. "cost is high". Actually, this sentence has negative meaning. Therefore, such situations should also be considered.

G. WDE-K-Nearest Neighbor Classifier

WDE-KNN has been selected as a classification model in this research work. WDE-KNN is selected because of S.A as it is a binary classification. With the help of this classifier, massive datasets can be executed. In order to train classifier, a manually produce training set is used. An X: Y relation has been provided in training set, where variable x illustrates the score of the opinion word and y illustrates the score of positivity and negativity [15]. A score of the opinion word relevant to a feature in the review is applied as input to WDE-KNN classification model.

H. Extraction of Feature Wise Opinion

In all reviews, features should be taken into account for extracting opinion related to a special feature. The ratio of total amount of reviews comprising positive sentiments to the total amount of reviews available is measured for a particular feature. The ratio of total amount of reviews within which a negative sentiment interrelated to a feature is given to the total number of reviews present is computed as the ultimate negative score for special feature.

IV. RESULT AND DISCUSSION

Python is a high-level programming language which comprises dynamic semantics. This language is generated within the data structures. This tool can easily be used by The Rapid Application Development due to the integration of data structures with dynamic typing and binding. In this tool, the previously accessible components get interrelated using scripting As this language is extremely simple and easy to learn. This also minimizes the maintenance cost of program. Results are visualized using various parameters which consists of original-values and predicted-data values, which can further sub-categorized such as true-positive , false-negative, true-negative, false–positive. Python comprises of various inbuilt functions which are used to determine results.

Fig 1: Classification Report Plotting

As shown in fig.1, the classification parameters like precision, recall and f-measure is calculated. The precision, recall and f-measure is calculated for each class like positive, negative. The generation of report is given in the form of numerical-scores with a color-coded-boxes which specifies the values of different parameters. Different values , are visualized in some particular range of values of classification-report. The values range has been lying in-between the 0.4 -0.9. The values of each parameter are plotted in the form of figure.

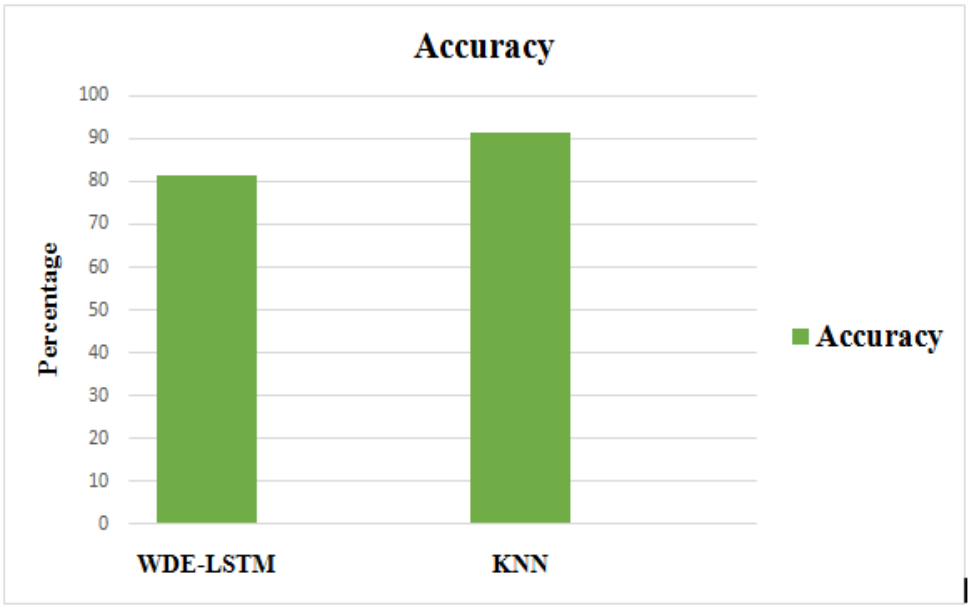

Fig. 2: Accuracy Analysis

As shown in fig.2, the accuracy of WDE-LSTM is compared with KNN. It is analyzed that performance on basis of accuracy of KNN classifier is high as compared to WDE-LSTM for the sentiment analysis. Accuracy is basically defined as the number of points correctly classified to total number of points (Eq.1) Accuracy is of great measure if values available in datasets are systematic of false positive and false negative. Accuracy of WDE-LSTM is about 81.51 % whereas KNN has been achieved upto 91% which is considered to be better.

$$\text{Accuracy} = \frac{Number\ of\ points\ correctly\ classified}{Total\ Number\ of\ points} * 100 \qquad \text{(Eq.1)}$$

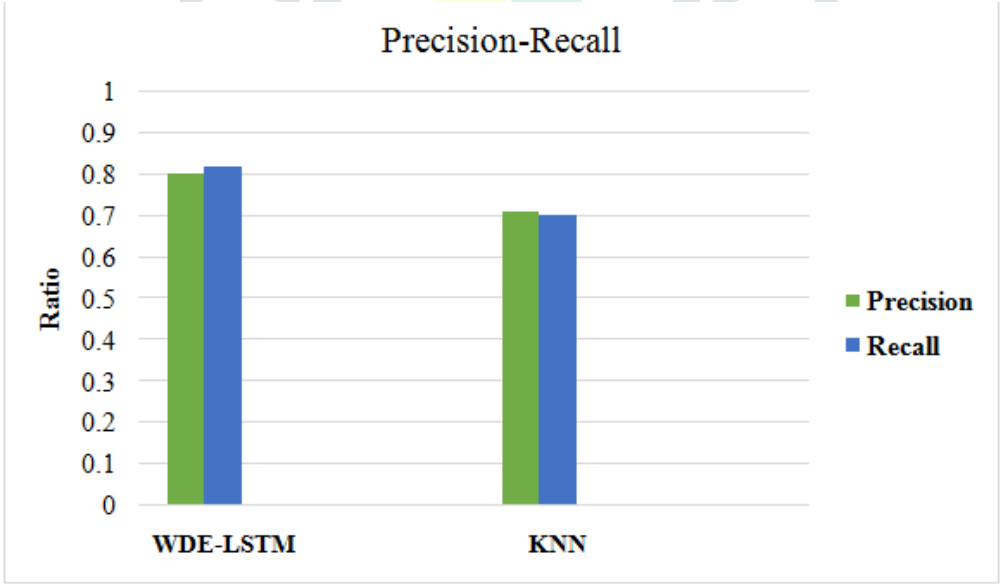

Fig 3: Precision-Recall Analysis

As shown in fig.3, the precision-recall of the WDE-LSTM is compared with the KNN classifier. Values of the given parameters in K-Nearest-neighbour (KNN) algorithm vary as compared to SVM algorithm on the basis of the performance- analysis. Precision (also called positive- predictive value) is the fraction of relevant- instances among the retrieved-instances (Eq.2) .It is basically a kind of measure which intends a classifier to avoid tagging a value as positive which actually depicts some negative values

$$Precision = \frac{TruePositive}{TruePositive+FalsePositive} \quad \text{(Eq.2)}$$

Recall is the fraction of relevant-instances that have been retrieved over the total amount of relevant-instances (Eq..3). Recall is also called as sensitivity or true positive rate. In other way its like only finding out the positive outcomes. The values of the precision and recall is shown in the Fig.3

$$Recall = \frac{TruePositive}{TruePositive+FalseNegative} \quad \text{(Eq.3)}$$

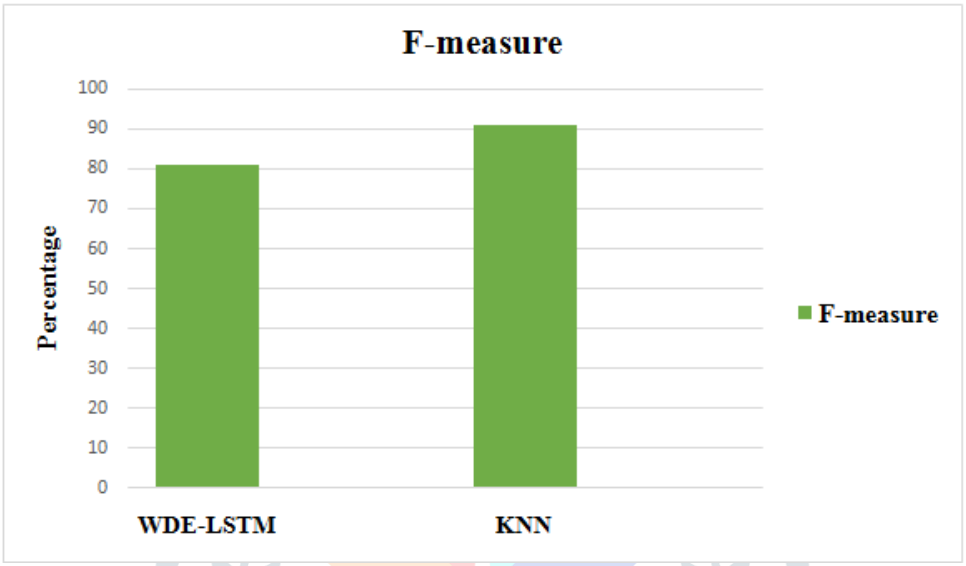

Fig.4: F-measure Analysis

As shown in fig.4, the F-measure of WDE-LSTM is compared with KNN. It is analyzed that F-measure of KNN classifier is high as compared to WDE-LSTM for the sentiment analysis F1 score is weighted average of precision and recall. This score takes both false-positives as well as false-negatives .F1-score is more useful than accuracy if values have uneven class distribution. It is mostly required when there is need to make a comparison between two classifiers as it takes into consideration the re-call as well as precision, which are used to find out the kind of harmonic-mean shown in (Eq.4)

$$\text{F1 score} = 2 * \frac{(\text{Recall} * \text{Precision}) * 100}{(\text{Recall} + \text{Precision})} \quad \text{(Eq.4)}$$

Table.1 Performance Analysis

Parameter	**WDE-LSTM**	**KNN**
Accuracy	81.51	91.45
Precision	0.80	0.71
Recall	0.82	0.70
F-measure	0.81	0.91

V. CONCLUSION

In this research study, the behavior of user is analyzed on the basis of sentiment analysis of twitter data. N-gram method is implemented in this study for sentiment analysis for analyzing certain features of input data. Moreover, a classification technique is applied to analyze the behavior of client. The N-gram method divides whole input dataset into several sections. All segments are analyzed individually in order to analyze sentiments. In this research study, a classification model called logistic regression is used for analysis. During the classification of data, various classes are produced. A comparison between WDE-LSTM and WDE-KNN classifiers is performed in this study. It is analyzed that WDE-LSTM classifier shows less accurate results whereas WDE-KNN approach shows better and improved accuracy rate. Several matrices such as accuracy and execution time are used to compare the performance of WDE-LSTM and WDE-KNN approach. The proposed approach shows good performance on all matrices as per the inspection outcomes.

References

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

[2] G. Salton and C. Buckley, "TEXT RETRIEVAL," Information Processing and management , vol.24, no. 5, pp. 513-523, 1998.

[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, "Summarizing Scientific Articles: Experiments with Relevance and Rhetorical Status, "computational linguistics, vol. 28, pp. 409-445, 2002.

[5] J. Ulrich, G. Murray and G. Carenini, "A Publically Available Annotated Corpus for Supervised Email Summarization," Chicago, USA, pp. 77-82, 2008.

[6] D. C. Uthus and D. W. Aha, "Plans Toward Automated Chat Summarization", in Meeting of the association for computational Linguistics, no. Code 5514, pp. 1-7, 2011.

[7] C. Whitelaw, B. Hutchinson, G. Y. Chung and G. Ellis, "Using the web for Language Independent Spellchecking and Auto correction," in Empirical methods in Natural Language Processing, pp. 890-899, 2009.

[8] X. Wu, X. Zhu, G. Wu and W. Ding, " Data mining with Big data," IEEE Transactions on knowledge and data engineering, vol. 26, no. 1, pp. 1-26 , 2014.

[9] A. Dewan, "Prediction of Heart Disease Using a Hybrid Technique In Data Mining Classification," International Conference in Computing for Sustainable Global, pp. 704 706, 2015.

[10] L. Zhou and E. H. Hovy. Digesting virtual geek culture: The summarization of technical internet relay chats. In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, pp. 298-305, 2005.

[11] D. Cao and L. Xu, "Analysis of Complex Network Methods for Extractive Automatic Text Summarization," 2nd IEEE International Conference on Computer and Communications, vol. 9 , pp. 2749–2756, 2016.

[12] R. Alguliyev, "A Sentence Selection Model and HLO Algorithm for Extractive Text Summarization ," IEEE, vol. 9, pp. 97-110, 2016.s

[13] N. Andhale, "An Overview of Text summarization Techniques," IEEE, vol. 9, pp. 97-110, 2016

[14] R. Bhargava and Y. Sharma, "MSATS: Multilingual Sentiment Analysis via Text Summarization," IEEE, vol. 9, pp. 71-76, 2017.

[15] A. N. Gulati, "A novel Technique for multi document hindi text summarization," International Conference on Nascent Technologies in the Engineering Field (ICNT 2017), vol. 8, pp. 1-4, 2017.

[16] M. Gupta, N. Garg., "Text Summarization of Hindi Documents using Rule Based Approach", International Conference on Micro-Electronics and Telecommunication Engineering, vol. 8, pp. 1-4, 2016.