Logistic Regression Method for Sarcasm Detection of Text Data

Bipin Gupta
M.Tech Scholar
gbipin002@gmail.com
RIMT University
Mandi Gobindgarh, Punjab

Ankur Gupta
Assistant Professor
ankurgupta@rimt.ac.in
RIMT University
Mandi Gobindgarh, Punjab

ABSTRACT
The prediction analysis is approach which can predict future possibilities. This research work is based on the sarcasm detection from the text data. In the previous time SVM classification is applied for the sarcasm detection. The SVM classifier classifies data based on the hyper plane which give low accuracy. To improve accuracy for sarcasm detection logistic regression is applied in this work. The existing and proposed techniques are implemented in python and results are analyzed in terms of accuracy, execution time. The proposed approach has high accuracy and low execution time as compared to SVM classifier for sarcasm detection

Keywords
SVM, Logistic Regression, Sarcasm detection

1. INTRODUCTION
With the increase in amount of data available in different applications, many large sized databases have been designed to store them. People are using new methods to perform operations and the inputs and outputs are stored with time on these large sized databases. Data mining is a KDD based technology that examines and reveals the important information from the raw data collected in large databases [1]. Cluster analysis is known as an important method for data mining. Either the dissemination of data index can be understood or the calculation values can be performed over the available data through pre-processing using this extracted data. The data mining approach is applied for differentiating the new, legitimate, reasonable and potentially valuable designs. This technology aims to extract the prescient data that is available in huge databases for examining the data available in the information distribution centers. It also helps in recognizing the valuable and pertinent data from databases [2]. Cognitive science is another technology that studies the human brain with the application of data mining technologies. Huge sized data is collected and placed in the databases and measures are used to extract only the verified information. So, it is important to expand the need of productive and compelling examination strategies. The examination of important information is done to perform a group investigation through which a client can view the common structure available in the collected data [3]. Therefore, in data mining, several new improved calculations and methods have been proposed over the years. 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 [4]. In the applications like social media analysis or commercial product reviews, opinion mining is performed. 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 [5]. 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 [6]. Each of the sentiment denotes a different category since every sentiment analysis can be considered as a separate task of classification process. 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. Different kinds of classifiers are applied to perform twitter sentiment classification through text classification. There are lexicographical resources included here [7]. The collection of seeds of sentiment words and their orientation is the initial step. The sets can be expanded by finding their antonyms and synonyms. The iteration process will stop after no words are left to be analyzed. Machine Learning Approaches are used to solve the problems associated with the classification of sentences. These approaches are broadly categorized among supervised and unsupervised learning approaches [8]. To perform text classification that can be applied for twitter sentiment classification, various classifiers have been designed. Maximum Entropy Algorithm is applied for providing extra semantic, syntactic features that are used with huge flexibility. To perform classification, the massive edge is provided by SVM classifier. The tweet and hyperplane are compared to separate the tweets and the hyper plane approach is used to do so. For performing classification the linear kernel is used through which the gap among two classes can be maintained. Today, several ensemble classifiers are available [9]. By ensuring that all the embedded features of the base classifiers are used to the fullest, the best classification is provided here. The base classifiers that can be used here are SVM, Naïve Bayes and Maximum entropy. The generation of ensemble classifier is done using the voting rule. Based on the output of huge part of classifiers, the classifier is classified.

2. LITERATURE REVIEW
Dan Cao, et.al (2016) stated that Automatic Text Summarization approach intended to make a condensed adaptation of documents. This version should be able to cover all important contents and common information. In this study, all features that utilized metrics and thought of complicated network for scoring sentences were reviewed [10]. In this

study, tested outcomes on individual module and mixture of various presented were analyzed. DUe 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.

RasimAlguliyev, 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 [11]. 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. Initially, the score of each sentence was evaluated and further most representative sentences were selected from the text 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.

NarendraAndhale, et.al (2016) stated that the procedure used for the generation of compressed structure of text document was known as text summarization. The wide-ranging analysis of both approaches was presented within the text summarization in this study [12]. In this study, numerous extractive and abstractive sorts of summarization techniques were 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.

Rupal Bhargava, et.al (2017) stated 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 [13]. In this study, a novel technique was presented for analyzing various languages in order to discover sentiments in these languages and performed sentiment analysis. The proposed technique implemented different machine learning methodologies for content inspection. In order to deal with various languages, Machine translation was utilized in this mechanism. Therefore, it was advantageous to retrieve important text occurring within this text for reducing further processing. Thus, the presented structure utilized text summarization procedure for extracting significant elements of text. These parts were then utilized to examine sentiments regarding some specific matter and its features.

ArchanaN.Gulati, et.al (2017) proposed a new method for multiple documents and extractive text summarization by considering this issue [14]. A condenser or summarizer for Hindi language was fabricated by considering it the most common language of India. Input to the system was applied in form of News editorials regarding sports and political affairs from Hindi newspapers which were available online. Eleven important aspects of the text had been utilized for retrieval procedure using Fuzzy inference system. A standard accuracy of about 73% over numerous Hindi documents 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 Precision, Recall and F-score values.

Manisha Gupta, et.al (2016) proposed a new scheme for summarizing Hindi text document in this study [15]. This approach was based on several linguistic rules. For generating smaller amount of words from the real document, dead wood words and phrases were eliminated from the actual text as well. The performance of resented approach was 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 proposed approach reduced the text size of information up to 60% - 70 %. The extractive summary provided by user was generated by proposed approach.

3. RESEARCH METHODOLOGY

A. Data Preprocessing

In this research, three main preprocessing methods such as Stemming, error correction and stop word removal are performed. In stemming procedure, recognition of root of a word is the fundamental job. The main objective of this technique is to eliminate included suffixes and amount of words. This scheme ensures that the system will consume minimum time and memory. As all reviewers do not utilize analogous grammatical rules, punctuation and spellings, therefore, an error correction method should be developed here. Due to these errors, the context is recognized in different way and this generates the need of error correction. In order to minimize the complexity of the text, the stop words are 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.

B. Lexical Analysis of Sentences

Either a positive or a negative sentiment is included in a subjective sentence. Nevertheless, some questions or sentences written by clients may not comprise any sentiments inside them. These sentences are identified as objective sentences. These types of sentences can be eliminated to lessen the overall size of review. Generally, A query mainly is generated through the inclusion of words like where and who. These 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.

C. 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.

D Define Scarstic and non Scarstic data

The words representing a particular feature can be retrieved using Stanford parser. The parser 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. 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. Define Logistic regression is chosen as a classification model in this research work. Logistic regression is selected because sentiment analysis 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 is provided in training set. The variable x represents the score of an opinion word and the score whereas y represents the positivity or negativity of word. A score of the opinion word relevant to a feature in the review is applied as input to Logistic regression classification model.

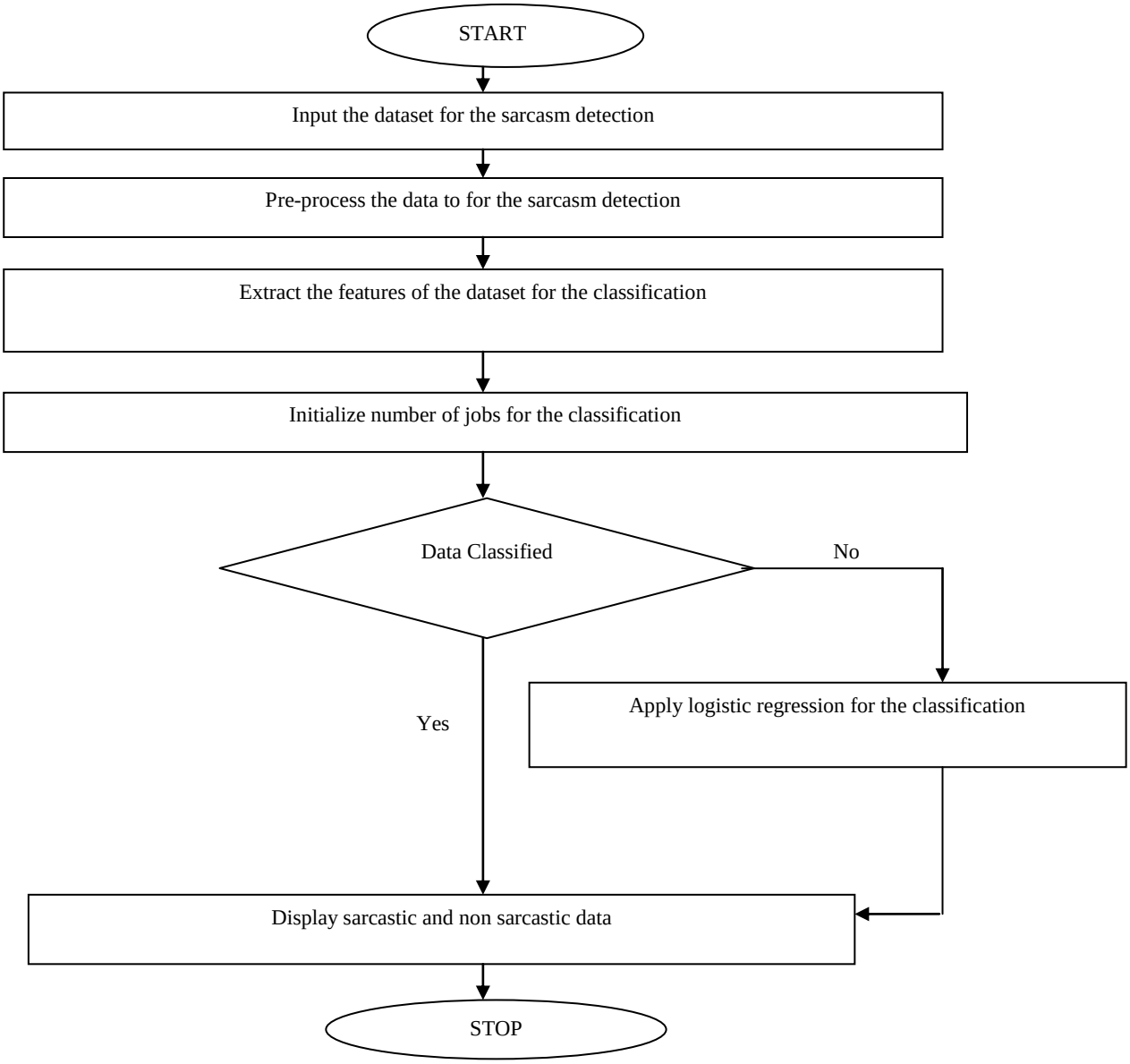

Figure 1: Proposed Flowchart

4. RESULT AND DISCUSSION

Python is a high-level programming language. This language 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, therefore, it can be read easily. This also minimizes the maintenance cost of program.

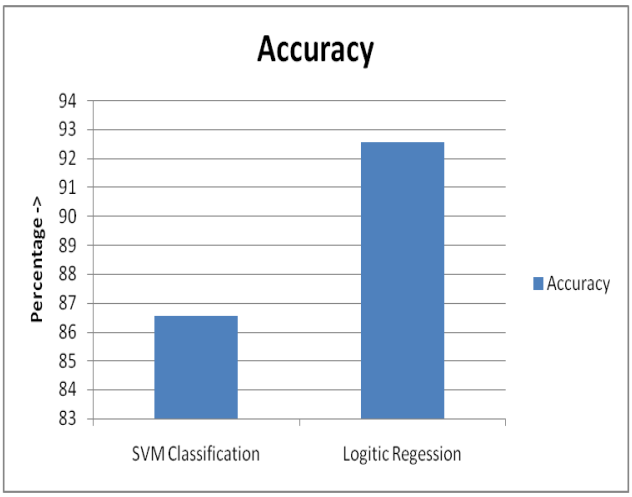

Figure 2: Accuracy Analysis

As shown in figure 2, the accuracy of the SVM classifier and logistic regression is compared for the sarcasm detection. The logistic regression given high accuracy for the sarcasm detection as compared to SVM classifier

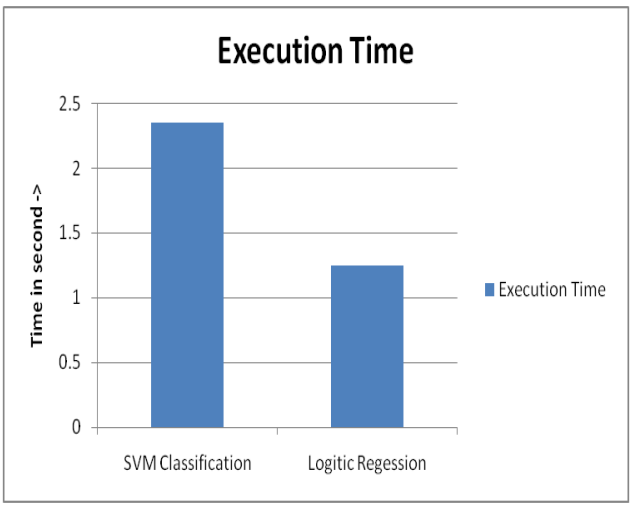

Figure 3: Execution Time

As shown in figure 3, the execution time of the SVM classifier and logistic regression is compared for the performance analysis. The execution of logistic regression is low as compared to SVM classifier.

5. CONCLUSION

In this paper, it is concluded that sarcasm detection is the challenge of the prediction analysis. To detect sarcasm from the twitter classification approach can be applied on input data. To detect sarcasm from the data in the previous work approach of SVM is applied. To detect sarcasm accurately approach logistic regression is applied in this work. When the logistic regression is applied on the data it gives accuracy about 93.5 percent.

6. REFERENCES

[1] Tharindu Weerasooriya, Nandula Perera, S.R. Liyanage, "A method to extract essential keywords from tweet using NLP", 2016 l16th 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 logic in Hybrid Intelligent Systems", 2009, HIS'09, Ninth International Conference on, vol. 1, IEEE, pp. 142-146.

[4] L. Suanmali, N. Salim, and M. S. Binwahlan, "Fuzzy logic based method for improving text summarization", 2009.

[5] X. W. Meng Wang and C. Xu, "An approach to concept oriented text summarization", Proceedings of ISClTS05, IEEE international conference, China, pp-1290-1293, 2005.

[6] M. G. Ozsoy, F. N. Alpaslan, and 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] 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

[11] 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

[12] Narendra Andhale, L.A. Bewoor, "An Overview of Text Summarization Techniques", IEEE, vol. 9, iss. 8, pp- 97-110, 2016

[13] Rupal Bhargavaand Yashvardhan Sharma, "MSATS: Multilingual Sentiment Analysis via Text Summarization", IEEE, vol. 9, iss. 8, pp- 97-110, 2017

[14] 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.

[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.