

Sentiment Analysis of Amazon Data Using KNN-LSTM Algorithm

Neha Rajput¹, Mrs. Shivani Chauhan²

¹Research Scholar, ²Assistant Professor, Department of Computer

^{1,2}Bhagwant Institute of Technology, Muzaffarnagar UP 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. This research work is based on the sentiment analysis of product reviews of Amazon data. To apply sentiment analysis the technique of feature extraction and classification is applied. For the sentiment analysis in the previous work, the WDE-LSTM technique is applied and which is replaced with the KNN-LSTM technique. The existing and proposed techniques are implemented in python and simulation results shows that accuracy of the proposed technique is approx 94.56 percent. The simulation results show that execution time of the proposed algorithm is less as compared to existing algorithm.

Keywords- WDE-LSTM, KNN-LSTM, Sentiment Analysis, Amazon

I. INTRODUCTION

There has been large amount of data generated with time due to the increase in level of knowledge within various applications. There has been increase in sizes of databases along with the need of storing the inputs and outputs with the increase in new methods being used by people over time. The KDD approach in which the information can be examined and revelations can be calculated is known as data mining [1]. An important strategy of database mining is known as cluster analysis. The extracted information can either be used to understand the dissemination of informational index or in the form of pre-processing to perform values calculations over the data available. In order to differentiate the legitimate, novel, potential valuable as well as reasonable designs that exist within the information, data mining approach is used. A new approach in which the covered prescient data present in huge databases is extracted such that the important data from the information distribution centre can be examined is known as data mining. From the database, the pertinent and valuable data is identified in this approach [2]. The process in which the patters are extracted from the data by performing certain tasks on the dataset is known as data mining. The descriptive data mining tasks and predictive data mining tasks are the two

different tasks that are performed in data mining. The characteristic properties of dataset are understood with the help of descriptive data mining tasks. The predictions are performed on the basis of available dataset to perform predictive data mining tasks. For examining the data several parameters are utilized by data mining applications. These parameters include association, classification, clustering as well as the sequence/pattern analysis. All these parameters have their own actions to perform and provide their own contribution to examine the data. The process in which the features of a newly present object are examined and further assigned to the predefined class is known as classification [3]. The well-defined classes and a training set that includes pre-classified examples are used to characterize the classification task. A model that can be used to un-classify the data such that it can be classified is required to be generated here. Any kind of prediction can be considered as the classification or estimation. There is very minimal difference here. One cannot go back to check if there is correct classification or not if the phone line that is utilized to access internet is utilized by data mining. A rule through which particular association relationships present amongst the set of objects present in the database can be implied is known as the association rule [4]. A method, in which diverse groups are segmented into a number of similar subgroups or clusters, is known as clustering. The breaking down of data into such a form that it can be 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. Sentiment Analysis is also known as the opinion mining [5]. It uses the NLP in order to categorize the opinions of people about the products or the reviews. 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 [6]. 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. User's opinion is a major criterion for the improvement of the quality of services rendered and enhancement of the deliverables. Blogs, review sites, data and micro blogs provide a good understanding of the reception level of the products and services. 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 [7]. 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. The combination of the both these approaches lead to increase in the performance and accuracy of the sentiment analysis.

II. LITERATURE REVIEW

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 [8]. 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.

ArchanaN.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. Thinking about the above issue a novel procedure for multi document, extractive text summarization is proposed [9]. 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.

Akshi Kumar, et.al (2017) proposed that there are various different algorithms that are used in the text summarization [10]. In this paper author has analyze and compare the performance of three different algorithms. Firstly, the different text summarization techniques explained. Extraction based techniques are used to extract important or keywords to be included in the summary. Abstraction based techniques generates its own sentences for text summary. Three algorithms are explained and implemented in python language. The ROUGE-1 is used to evaluate the effectiveness in extracted keywords. The result of all algorithms compares with the handwritten summaries to evaluate the performance of the algorithms at the end the performance of TextRank Algorithm is much better than other algorithms.

Shahnawaz, et.al, (2017) presented sentiment analysis is the process that helps to provide idea to the customer to identify the product or service is satisfactory or not before the customer buys it. Public opinions on different types of social media are the major concern of the scientific communities and business world to gather and extract public views [11]. Inadequacy accuracy, inability to perform well in different domain and performance are the main issues in the current techniques. Author concluded, by using semi-supervised and unsupervised learning based models, it will be easy minimize lack of labeled data if sufficient amount of unlabeled data is available.

N. Moratanch, et.al (2017) proposed that the text summarization has various techniques that are classified as the extractive and abstractive approaches of summarization of the text [12]. In this paper the author has presented the comprehensive review of extraction based text summarization techniques. In this paper the author provides survey on extractive summarization approach. Then different methodologies, the advantages are presented in the paper. The author also includes various evaluation methods, challenges and future research direction in the paper.

Pierre Ficamos, et.al (2017) proposed that the social media become more popular nowadays; the more researches have been focusing on automatic processing and extracting the sentiment information from the large data [13]. In this paper the author proposed a feature extraction method that relays on Part Of Speech (POS) tags. That helps in selection of the unigram and bigram features. The paper focuses on the sentiment analysis of the Chinese social media. The grammatical relations between the different words are used in

construction of the bigram and unigram features. The experiment shows that the proposed method provides the better results with the Naïve Bayes.

III. RESEARCH METHODOLOGY

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.

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.

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.

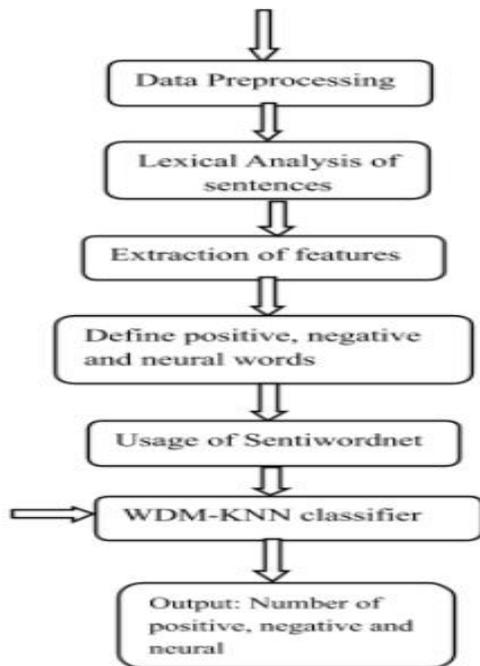


Fig.1: Proposed Flowchart

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

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.

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

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.

G. WDE-K-Nearest Neighbor Classifier

In order to use a classifier within this approach, WDE-KNN is selected. Since, sentiment analysis is a binary classification and there are huge datasets which can be executed, WDE-KNN is chosen here. A manually generated training set is utilized for training the classifier here.

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.

IV. EXPERIMENTAL RESULTS

The proposed work is implemented in MATLAB and the results are analyzed in terms of accuracy and execution time.

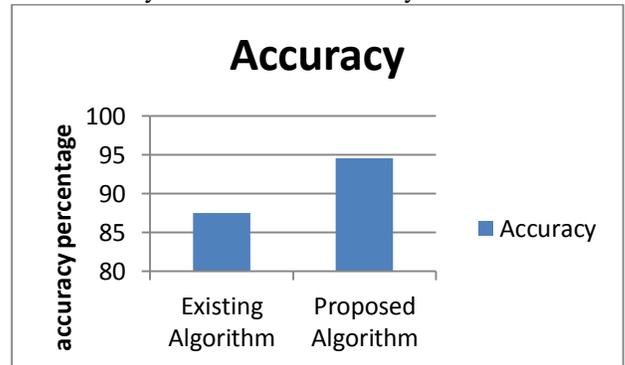


Fig.2: Accuracy Comparison

As shown in figure 2, the accuracy of the proposed and existing algorithm is compared for the performance analysis. It is analyzed that accuracy of existing WDE-LSTM algorithm is less as compared to proposed WDE-KNN algorithm.

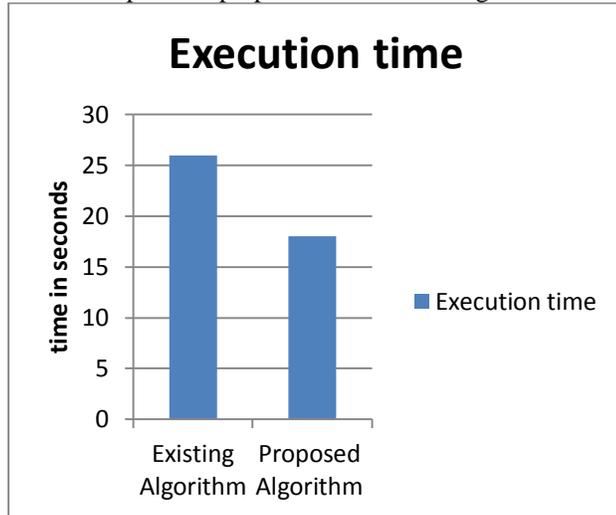


Fig.3: Execution Time Comparison

As shown in figure 3, the execution time of the proposed and existing algorithm is compared for the performance analysis. It is analyzed that execution time of existing WDE-LSTM algorithm is high as compared to proposed WDE-KNN algorithm.

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. In this research work, the technique of WDE-LSTM is compared with the WDE-KNN classification. It is analyzed that accuracy of WDE-LSTM technique is approx 83 percent and when the technique WDE-KNN is applied it is increase to 89 percent. The performance of WDE-LSTM and WDE-KNN technique is also compared in terms of accuracy and execution time. It is analyzed that proposed technique performs well in terms of all parameters.

VI. REFERENCES

- [1]. ShwetaRana, Archana Singh, "Comparative analysis of sentiment orientation using SVM and Naive Bayes techniques", 2016 2nd International Conference on Next Generation Computing Technologies (NGCT), vol. 8, pp. 1-4, 2016.
- [2]. HumaP arveen, Shikha Pandey, "Sentiment analysis on Twitter Data-set using Naive Bayes algorithm", 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), vol. 8, pp. 1-4, 2016.
- [3]. Wiraj Udara Wickramaarachchi, R. K. A. R. Kariapper, "An Approach to Get Overall Emotion from Comment Text towards a Certain Image Uploaded to Social Network Using Latent Semantic Analysis", 2017 2nd International Conference on Image, Vision and Computing, vol. 8, pp. 1-4, 2017.
- [4]. Ana Valdivia, M. Victoria Luz' and Francisco Herrera, "Neutrality in the Sentiment Analysis Problem Basedon Fuzzy Majority", vol. 7, pp. 1-6, 2017
- [5]. Chhaya Chauhan, SmritiSehgal, "Sentiment analysis on productReviews" International Conference on Computing, Communication and Automation, vol. 10, pp. 1-8, 2017
- [6]. PragyaJuneja, Uma Ojha "Casting Online Votes: To Predict Offline Results Using Sentiment Analysis by machine learning Classifiers" vol. 5, pp. 1-6, 2017
- [7]. Mika V. Mäntylä1 Daniel Graziotin, Miikka Kuuttila, "The Evolution of Sentiment Analysis - A Review of Research Topics, Venues, and Top Cited Papers", Volume 27, February 2018, Pages 16-32, ISSN, 2018.
- [8]. RupalBhargava and Yashvardhan Sharma, "MSATS: Multilingual Sentiment Analysis via Text Summarization", IEEE, vol. 9, iss. 8, pp- 97-110, 2017
- [9]. ArchanaN.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
- [10].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
- [11].Shahnawaz, ParmanandAstya "Sentiment Analysis: Approaches and Open Issues" International Conference on Computing, Communication and Automation, vol. 9, pp. 1-5, 2017
- [12].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
- [13].Pierre Ficamos;Yan Liu, WeiyiChenA., "Naive Bayes and Maximum Entropy approach to sentiment analysis: Capturing domain-specific data in Weibo", IEEE International Conference on Big Data and Smart Computing (BigComp). vol. 8, pp. 1-4, 2017