Webology (ISSN: 1735-188X)
Volume 18, Number 6, 2021

Big Data Analytics For Sentiment Analysis From The Reviews And Feedbacks Of E-Commerce Portal

Chetan Kumar Soni[1] and Atul D. Newase[2]

[1]Department of Computer Application, Dr.A.P.J. Abdul Kalam University Indore, India.

[2]Department of Computer Application, Dr.A.P.J. Abdul Kalam University Indore, India.

Abstract—Recently, big data analytics has emerged as boon in the field of computing technology which is capable of dealing with the very huge amount of data of varying velocity and variety in a very efficient way. The parallel distributed computing architecture with high computing capabilities has made it possible to access, process and analyze this big data. This paperpresents an implementation of this big data analytics technology to analyze the text data received from reviews and the feedbacks of the users of various e commerce companies and identify the sentiments of the users about the specific product or service. The concept of natural language processing is used in this work to identify the orientation of the users. The sentiments are classified into four categories in this work as positive, negative, neutral and unsure. A dictionary is prepared on the basis of keywords projecting the sentiments and the emotions of the users which is preprocessed as per the requirement. The novelty of this work lies into the deep learning algorithm presented in this work which classifies the emotions of the users into the above mentioned four categories and presents a complete framework for the stakeholders to take the necessary action. A neural network based deep learning algorithm is presented in this work and the implementation is done through Hadoop. As compared to the previously proposed algorithms, this model presents a larger range of emotions over a huge database. It also provides a better accuracy and efficiency.

Keywords-Sentiment analysis, Emotion mining, Deep learning, Neural networks,Big data analytics, Hadoop.

I. INTRODUCTION

Teaching The advent of e-commerce over the last decade and changed the complete paradigm of the business environment. Instead of going to the shops personally, the users can directly access the details of the products they want to buy, compare them of different ecommerce websites, and order them. The online shopping has saved so much of time, money and energy of the users as well as the vendors. The delivery experience of the users is also improving day by day as the online shopping is getting popular among the last mile also. The users can provide their feedbacks and reviews about the products and the services they have received on the shopping portals. These reviews also help the other users to assess the quality of the product of services. These reviews also help the vendors to improve their services and enhance their market base. The expectations and the needs of the users can also be understood from these feedbacks.

However, as the number of users and the service providers are increasing day by day, it is not an easy task to analyze the response of each and every user about each and every product. The amount of data generated through these feedbacks is too huge for a person to evaluate and come to a conclusion manually. Considering the business base of very big giants of the online market like, Wallmart, Amazon, Flipkart, Reliance, etc, who have millions of users, it is near to impossible to handle this huge amount of data. The speed at which this data is generated also make it a near to impossible challenge to handle manually. The problem becomes more severe, when the heterogeneous nature of the reviews came into account. The users are from very different backgrounds, having varying expectations and range of products is also huge. Incorporating all these challenges associated with the amount of data with a huge velocity and variety, the assessment of the data is a big challenge for the service providers.

The potential of this platform has brought the users as well as the businesses to showcase their products and services to the users. The users also have the flexibility to preset their expectations from the service providers and the companies. This environment has added a new dimension to the business models and the respective revenue generation. Millions of users on e-commerce express their thoughts and opinion about the purchase experience and the quality of the products they have bought. This expression has reflected the attitude of the users towards any concern of the world. The expression is mostly in the form of text messages and the continuous accumulation of these messages results into a huge amount of data with so much variety and randomness. The cumulative effect of the response of the millions of people possesses a great capacity to drive the orientation of world commercial policies. Even the human behavior can also easily be predicted and analyzed through this data. The extensive use of e-commerce has generated massive amount of data and the major portion of this data is in the form of text. For example, the amount of data generated through e-commerce technologies like Amazon, Flipkart, etc has reached to 20ZB and more than half of this data is in the form of text. The careful analysis and study of these text messages may give a clear idea about the sentiments and opinion about each and every aspect of business. The analysis of this huge amount of data and predicting the behavior of users has emerged as a biggest challenge over the last decade.

The field of analyzing the opinion of users through the text messages is known as opinion mining. It includes the classification of sentiments of the users into different polarities like positive, negative or neutral. The parameters resembling the characteristics of the data governs the complete results of the analysis. However, the huge amount of data poses a severe challenge in the field of sentiment analysis. It needs a completely different framework for the analysis which can process the huge amount of data with high randomness and speed in a smaller amount of time with high accuracy.

Big Data analytics has emerged as a great opportunity for business prospect and in research field. It is becoming popular and gaining more and more attention because of its capability to deal with huge amount of data in a smaller amount of time and with a great level of accuracy as per the need.This paper presents an intelligent sentiment analyzing framework on the basis of the reviews and the feedbacks of the users on e-commerce website. This work classifies the sentiments in the classes of positive, negative, neutral and unsure to present an emotion mining framework. Neural network has been designed in this work to

augment the intelligence in the proposed model. The dataset used to train the sentiment analysis system is derived through the e-commerce websites and portals. The major contribution of this research work is the implementation of big data analytics for the processing of huge amount of data to extract the emotions. The impact of the outcome of this system has the potential to improve the business framework and customer'sexpectations many folds. Proposed model presents a personalized and customized recommendation to specific stakeholder over a particular time period. The common patterns of the preferences of the stakeholders in the commercial framework have been identified in this work to utilize them for the effective and impactful analysis.

The paper is organized as follows: section II deals with the review of the existing techniques of sentiment analysis in various fields.The mathematical framework for the neural network used in the system is given in section III. The proposed bag data analytics based sentiment analysis system is discussed in sectionIV.Section V discusses the effectiveness of the proposed strategy through the analysis of the performance parameters while section VI concludes the paper.

II. RELATED WORK

The potential of information which can be retrieved from the text reviews and feedbacks of the users have drawn the attention of many researchers over the last decade. Various aspects of data analytics have been addressed by the researchers for sentiment analysis through the data received in the form of text, images or videos. They have also explored the feasibility of applying different statistical algorithms and machine learning frameworks on this huge amount of data with large diversity and velocity. Some research works have been reviewed here to present a rationale behind this research work. The detailed analysis is as follows:

Divya Sehgal et al. [1] (2006) proposed sentimental analysis framework for the social media data in their papertitled"Sentiment Analysis of Big Data Applications using Twitter Data with the Help of HADOOP Framework". They have used the text data of twitter and performed the big data analytics for the emotional and sentimental value of the users associated with issues discussed on social media. The similar framework of bigdata was implemented on Geosocial Networks for safety and relief analysis in the situation of any disaster by M. Mazhar Rathore et. al. [2] (2017). The healthcare data and the reviews from the citizens were collected for the analysis of the preparation and readiness of the concerned authorities. They also evaluated the quality of response in the disaster through the sentiment analysis.

C. Fellbaum [3](2005) described about resources and variations in the field of verbal communication. They presented the concept of Wordnet which resembles to the group of words which are very familiar and popular among the people. These words are linked with each other semantically. The derived relationship also described the specialization and generalization in these words. They have encompassed various aspects like psychology, language pathology, anthropology, linguistics, cognitive science and medicine & computer science.

B. Liu [4] introduced the idea of opinion Lexicon which consists of annotations which are positive and are manually chosen. The classification of these opinion lexicons was performed by the researchers on the basis of various verbal, grammatical and spelling characteristics. The frequency of use of these

words was also taken into consideration for the analysis. The author categorized the textual information gathered from various sources into two classes: facts and opinions.

H. Ji et al. [5] proposed a big data framework for the analysis of verbal data on the basis of and referred them as contextual words contextonyms. Considering the frequency of occurrence of words, words are linked as edges with representing networks of word node. The corresponding relationship is established between the group of similar semantics with contextonyms and the sub-graphs are derived for each node. A strong semantics are summarized by plotting graph and linking them with the composition of words with context. Strong bonding is indicated in a word by connecting nodes. Chuanming Yu [6] extended the work of sentiment analysis using the modern classification strategywhere they have used SVM model experimentally for four data sets. A detailed dataset was prepared and the feature extraction was performed using classifier which extracted the maximum Entropy. The features are selected in such a way that the statistical characteristics of the data are achieved in a desirable manner. During the analytical study of the proposed work, it was found that the SVM based classification is better than the conventional techniques.

Raisa Varghese and Jayasree M [7] has presented abstracted the advantages of Senti-WordNet with the resolution of co-reference and dependency. The relationship between the words in a context and specific domain are used in their work. The sentimental value associated with the words which is known as Senti-WordNet was considered in this work. A large dictionary was prepared to tag the sentiment related to the respective word and was been used to train a support vector machine (SVM) based machine learning algorithm.Amit Gupte et al. [8] presented a comparative analysis between the various classification and estimation algorithmswhere they have performed the sentiment analysis using Random Forest technique and compared the results with Maximum Entropy, Boosted tree and Naive Bayes approaches. The challenges related to the training and testing accuracy of the conventional machine learning algorithms were addressed in this work and they are been overcome using the random forest algorithm.

Suchita V Wawre et al. [9] have also presented a comparative study of Support Vector Machine (SVM) based classification with the Naive Bayes classifierdepending upon the characteristics of the data and the statistical analysis. The features are processed and engineered in a meaningful manner so as to present a best suited classification model. The superiority of the proposed SVM based sentiment analysis framework was also proved through the comparative analysis with the Naïve Bayes classifier.

Alessia D'Andrea et al. [10] presented a hybrid approach for the sentiment analysis for the lexicon data by incorporating the potential of machine learning algorithms with the conventional feature extraction and processing algorithms. The authors in this paper have put their efforts in the feature engineering as well to give the best finishing to the features. This has improved the training performance of the machine learning model as the variance of the features has been improved through the feature engineering. This paper has shown a hybrid approach of feature extraction, feature processing along with the machine learning algorithms. A comparative analysis of various approaches for the sentiment analysis was performed in this work along with the different tools available for the same. The combined strategy as resulted into an improved and accurate classification performance.

Hailong Zhang et al. [11] presented a detailed survey to address the lexicon classification for sentiment analysis using cross-domain and cross lingual method. This work has extended the scope of implementation of machine learning based sentiment analysis over the larger range over multiple lingual and domains. The authors have encompassed the complete spectrum of text analysis by discussing the different techniques available for the same.

Rousseau et al. [12] presented a detailed study of different evidences in management and organizational scienceand derived a scientific framework for the synthesis of classification algorithm. The sentiment analysis was done on the organizational and management domain. The authors have undertaken several case studies in the management and organizational system and considered the respective data. The data has been used for the analysis of the sentiments of the stakeholders associated with the organization along with their managerial responsibilities.

III. NEURAL NETWORK FRAMEWORK

Neural network has gained a lot of attention over the last decade due to its superior learning capabilities and ability to solve the classification problems. It presents an optimal classification solution in terms of complexity, speed and accuracy. A typical configuration of ANN consists of the most basic building block known as artificial neuron or perceptron which is represented by synapses, an adder, and an activation function. Each neuron is connected through the synapses associated with its own weights which resembles to the strength of the respective input link. All these weighted outputs of neurons are added through the adder component. Activation function or squashing function refers to the learning function. A typical activation function for linearly separable classes is represented by

$$\phi(x) = \begin{cases} 1, & x \geq 0 \\ -1, & x < 0 \end{cases}$$

The scaled version of single layer perceptron is known as multi-layer perceptron which enhances the learning capabilities of the conventional network. It is achieved through the nonlinear decision region by correctly classifying a set of inputs within an input space. The commonly used activation function for non-linearly separable classes is sigmoid function because of its simple derivative. The sigmoid function is mathematically defined as

$$\phi(x) = \frac{1}{1+e^{-x}}$$

A typical architecture of multi-layer perceptron comprises of an input layer, one or more hidden layers and an output layer. However, the learning capabilities and training accuracy is determined through the number of neurons and number of layers. The training of the neural network is performed through a known training set along with the respective target values. The results from the output layer are then compared with the desired response to derive the error signal. This error signal is then back propagated through the neural network, against the direction of the synaptic connections. This algorithm of using the error signal to tune the weights of the neural network is known as backpropagation learning algorithm represented as $\Delta w_{ji}(n) = \eta \delta_j y_i(n)$

where η is the learning rate, and $y_i(n)$ is the neuron output value.

If j is in the output layer,

$$\delta_j(n) = (d_j(n) - y_j(n)).\frac{b}{a}.(a - y_j(n))(a + y_j(n)),$$

or if j is in a hidden layer,

$$\delta_j(n) = \frac{b}{a}.(a - y_j(n)).(a + y_j(n)).\sum_k \delta_k(n).w_{kj}(n),$$

where d_j is the desired response, and a and b are the scaling values from the neuron activation function. Training continues until the weights of the neural network produce outputs that converge. Convergence is defined by an average error signal, ε_{av} reaching a threshold.

$$e_j(n) = d_j(n) - y_j(n)$$

$$\varepsilon(n) = \frac{1}{2}.\sum_{j \in c} e_j^2(n)$$

where c is the set of all neurons in the output layer.

$$\varepsilon_{av} = \frac{1}{N}.\sum_{n=1}^{N} \varepsilon(n)$$

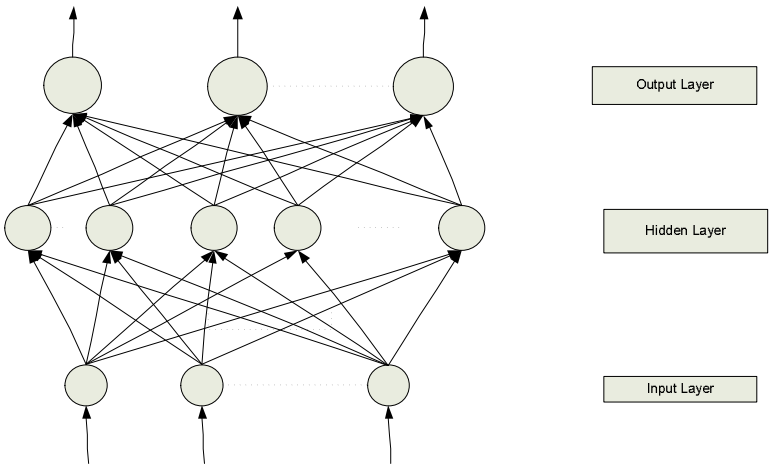

Fig.1. Three layer Architecture of Neural Network

IV. PROPOSED METHODOLOGY

The challenges and the drawbacks associated with the statistics based or probabilistic sentiment analysis has provided the required motivation to the researchers to implement some intelligent solutions. The diversity associated with the text data collected from various sources enhance the complexity level of the problem many fold. The problems with text data like unstructured format, different languages, spelling mistakes, abbreviations, emojis, etc are very difficult to deal with in real time application. The estimation result of any classification algorithm which is not very accurate may also result into some hazardous outcomes in real world as it has the direct impact on the organizational and functional decision making of many organizations.As depicted in the fig. 4.1, the process of sentiment analysis is broadly dived into

three phases: Text data preprocessing through the annotated sentiment defining words and dictionary, Machine learning based classification algorithm derivation and the estimation for a real time system. With an objective of estimating the sentiments associated with the feedbacks and reviews of the users on any e-commerce company for some specific product or service, an intelligent emotion mining framework is implemented in this work. The proposed method is implemented over a big data environment due to the size of the text data collected from the various sources. The large size and diversity of the data has added a complete novel dimension to the sentiment analysis process. The analytics scenario required to deal with big data is composed of a large memory, bandwidth and high computation and processing capabilities in real time. The distributed framework utilized is Hadoop which is an open source computing and processing of large datasets in distributed environment.

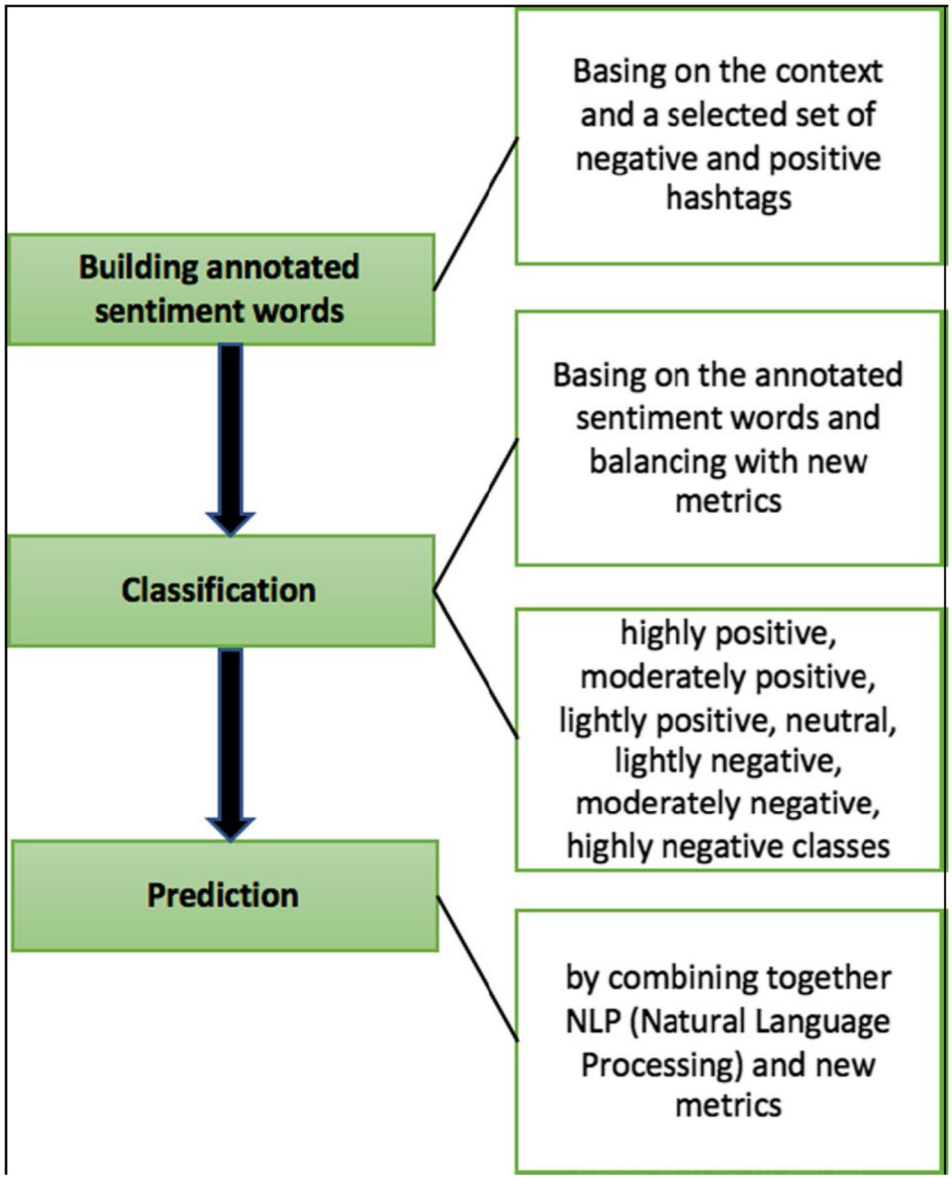

Fig 2. Proposed Sentiment Analyzing Framework

The processing algorithm of Spark framework is used for the implementation of complete sentiment analysis on the real time data in this work. The text data used in this work for the sentiment analysis is received from the feedbacks and reviews of various users. Incorporating all the products of e-commerce Company and the user's input on them collectively generates a huge amount of data with large diversity and speed. This data is preprocessed in this work to develop the dictionary for the text analysis in the form of sentiments associated with them.

During the implementation of the classification phase of sentiment analysis model also, a great level of processing is required for the training and testing data so as to present it to the machine learning algorithm. This processing of data in terms of tokenization, filtering, stemming, etc is performed on the Hadoop platform through Spark. It also has performed the machine learning algorithm to train, test and validate the data. It in turn is used to estimate the sentiments of the users under the categories of positive, negative, neutral and unsure. Hadoop can provide the results of sentiment analysis classification only, which can be manipulated by the decision server. The server checks the authenticity of the decision making on the basis of its comparison with some real time scenario. The cumulative classification results of each batch in Hadoop environment is then used to identify the overall sentiment of the user towards any product or service.

V. EXPERIMENT ANALYSIS

The system specifications used in this work to implement the proposed sentiment analysis using the text data are consist of cluster of 3 servers for the prototype implementation. These servers are having two intel Xeon E5530quad core CPU with 2.4 GHz processor, 24 Go DDR3 RAM and 1 TB hard disk who run on 64 bit Linux operating system. Data gathering is done using the distributed streaming platform, Apache Kafka, which offers a distributed and replicated service using the publish- subscribe messages. Integrated Stream API libraries are utilized in this work to build the applications for stream processing. This large sized data is collected and stored in Hadoop Distributed File System (HDFS).

The performance of the proposed classification approach is evaluated through the automatically generated dynamic dictionary. A randomly selected subset of reviews is considered with 600 reviews of 50 sentiment classes. The reviews are carefully inspected and sentiments associated with them is identified and labeled manually as positive, negative, neutral and unsure for each candidate. The same data is then taken through the proposed framework after the preprocessing including the tokenization, stemming, filtering, etc. This step is taken care by a tool named as Tree Tagger which provides the facility to handle negation, Usernames, hashtags, twitter mentions, URLs, intensifiers, etc. Neural Network is then applied on this data to estimate the sentiments using the posterior probability of the class over the bag of words.

The performance of the classification algorithm in this work is evaluated using the accuracy. While considering accuracy and macro F-measure, it is observed that classification using our proposed method achieves a good accuracy (90.21%)

Result analysis of the complete work is described in this section which is demonstrated through dataset size in MB and computation time for that dataset is in seconds

Below table shows the proposed work explanation which is demonstrated in table.

Table 5.1

Dataset Size (MB)	Computation Time (sec)
1	128
10	178
50	756
100	1568

Figure 5.1: Proposed work graph

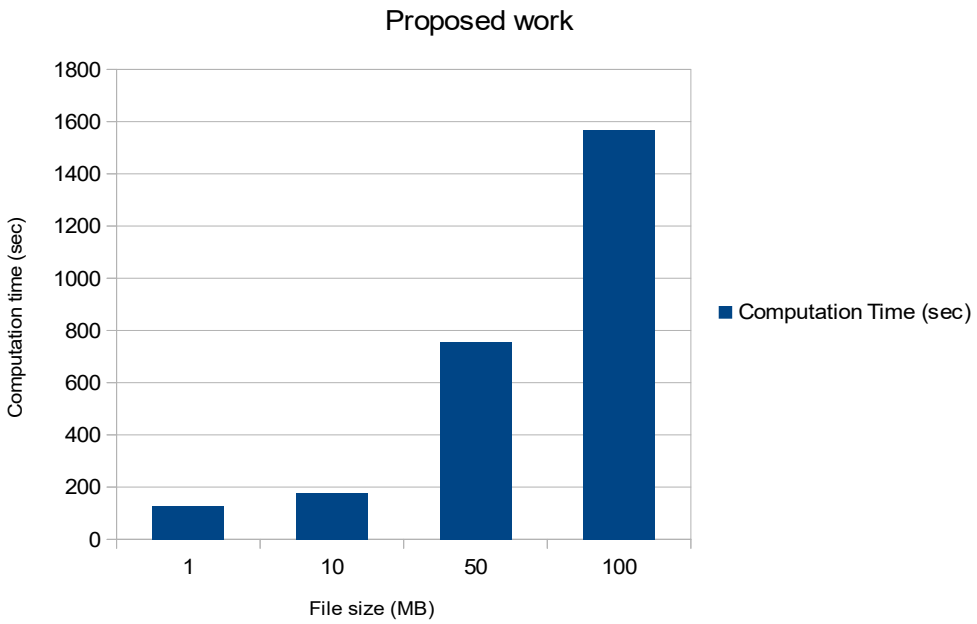

Table 5.2: Existing Work table

Dataset Size (MB)	Computation Time (sec)
1	100
10	419
50	1720
100	4120

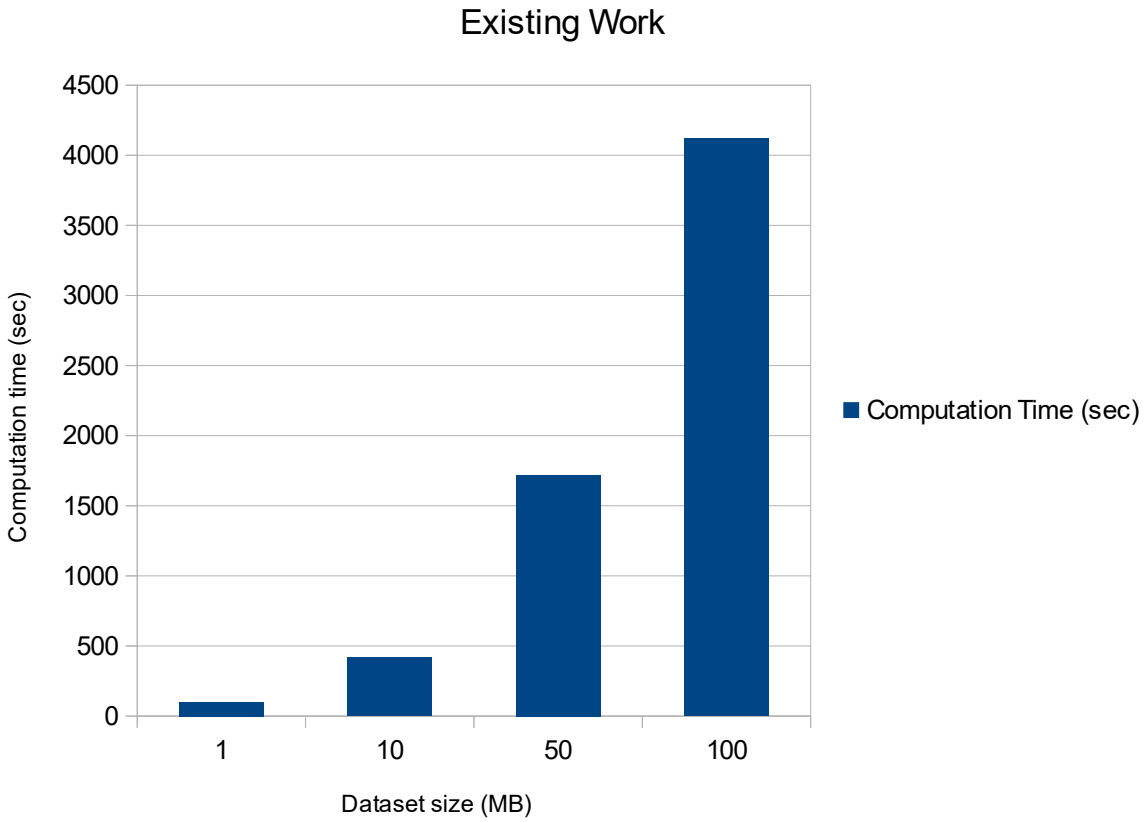

Figure 5.2: Existing work graph

Table 5.3: Comparison Table

Dataset Size (MB)	Proposed Work	Existing Work
1	128	100
10	178	419
50	756	1720
100	1568	4120

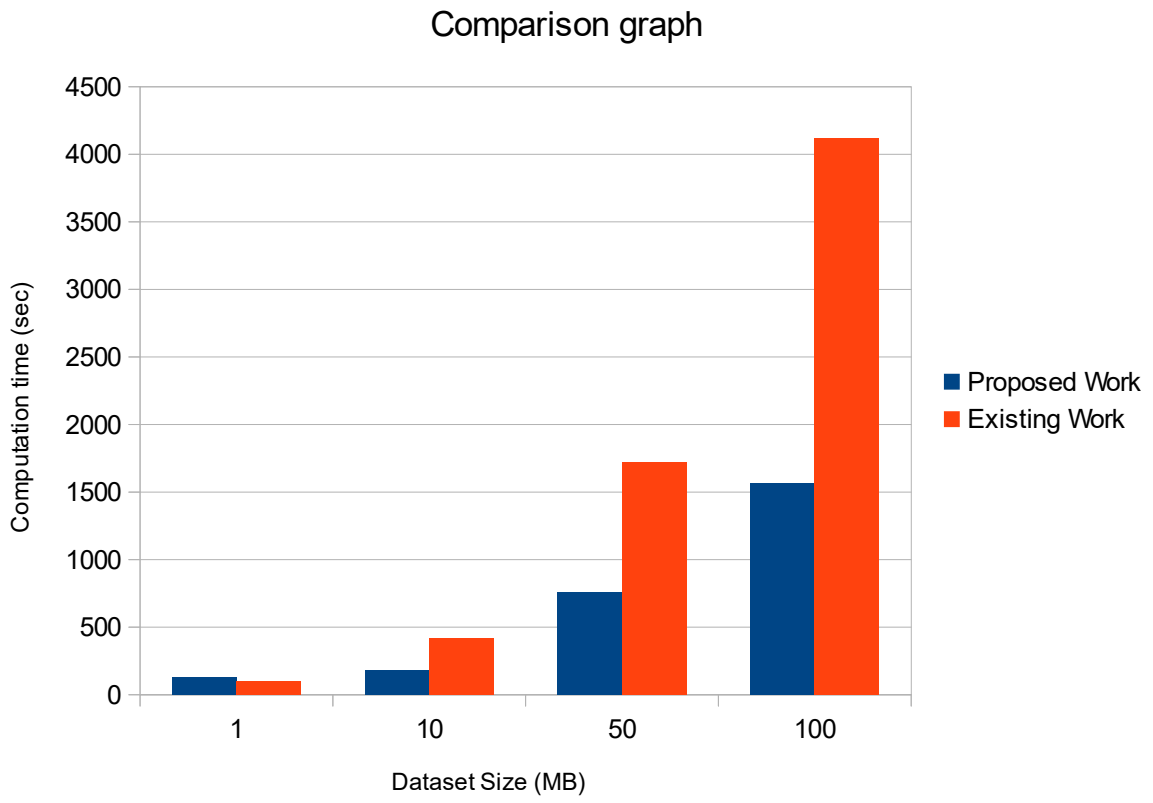

Figure 5.3: Comparison graph

The figures shown above represent the effectiveness of the proposed sentiment analysis framework using the Neural network algorithm. The accuracy and the computation time represents that the proposed technique has performed in the desired manner.

VI. CONCLUSION

A sentiment analysis model using big data analytics has been presented in this paper using the neural network framework to deal with the diverse distribution of the attributes. The analysis is proposed on the basis of various reviews and feedbacks of the users on the e-commerce company websites which directly and indirectly governs the characteristics of the business. The sentiments like positive, negative, neutral and unsure have been identified in this work to evaluate the emotions of the users towards any product. The performance of the proposed technique is evaluated in terms of parameters like accuracy, precision and recall. It is also compared with the other conventional techniques and found to be performing better than those techniques.

REFERENCES

[1] Divya Sehgal and Ambuj Kumar Agarwal, “Sentiment Analysis of Big Data Applications using Twitter Data with the Help of HADOOP Framework”. 5th International Conference on System Modeling & Advancement in Research Trends, 2016, IEEE.

[2] M. Mazhar Rathore, Anand Paul, Awais Ahmad, “Big Data Analytics of Geosocial Media for Planning and Real-Time Decisions”. ICC SAC Symposium Big Data Networking Track, 2017 IEEE.

[3] C. Fellbaum, “Wordnet and wordnets,” Encyclopedia of Language and Linguistics, pp. 665–670, 2005.

[4] B. Liu, “Sentiment analysis and subjectivity,” Handbook of Natural Language Processing, pp. 627–666, 2010.

[5] H. Ji, S. Ploux, and E. Wehrli, “Lexical knowledge representation with contextonyms,” in Proceedings of MT Summit IX, New Orleans, USA. Association for Machine Translation in the Americas, 2003.

[6] Chuanming Yu, “Mining Product Features from Free-Text Customer Reviews: An SVM-based Approach”, 2009, Nanjing, China. ICISE 2009 December 26-28,

[7] Raisa Varghese and Jayasree M, “Aspect Based Sentiment Analysis using Support Vector Machine Classifier”, International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2013.

[8] Amit Gupte, Sourabh Joshi, Pratik Gadgul, Akshay Kadam, “Comparative Study of Classification Algorithms used in Sentiment Analysis”, International Journal of Computer Science and Information Technologies, Vol. 5 (5) , 2014, 6261-6264.

[9] Suchita V Wawre, Sachin N Deshmukh, “Sentiment Classification using Machine Learning Techniques”, International Journal of Science and Research (IJSR) Volume 5 Issue 4, April 2016.

[10] Alessia D’Andrea, Fernando Ferri, Patrizia Grifoni, Tiziana Guzzo, “Approaches, Tools and Applications for Sentiment Analysis Implementation”,International Journal of Computer Applications (0975– 8887) Volume 125 – No.3, September 2015.

[11] Hailong Zhang, Wenyan Gan, Bo Jiang, “Machine Learning and Lexicon based Methods for Sentiment Classification: A Survey”, 11th Web Information System and Application Conference, 2014

[12] D. M. Rousseau, J. Manning, and D. Denyer, “11 evidence in management and organizational science: Assembling the field's full weight of scientific knowledge through syntheses," Acad. Manage. Ann., vol. 2, no. 1, pp. 475_515, 2008.

[13] I. Korkontzelos, A. Nikfarjam, M. Shardlow, A. Sarker, S. Ananiadou, and G. H. Gonzalez, ``Analysis of the effect of sentiment analysis on extracting adverse drug reactions from tweets and forum posts," J. Biomed. Inform.,vol. 62, pp. 148_158, Aug. 2016.

[14] E. H.-J. Kim, Y. K. Jeong, Y. Kim, K. Y. Kang, and M. Song, “Topic-based content and sentiment analysis of Ebola virus on Twitter and in the news,” J. Inf. Sci., vol. 42, no. 6, pp. 763_781, 2016.