SENTIMENT ANALYSIS USING AMAZON DATA FOR WDE-KNN ALGORITHM

[1]Kanika Sharma, [2]**Akanksha Sambyal**

[1]Research Scholar, kanikasagar004@gmail.com

[2]Assistant Professor, Er.akankshasambyal@gmail.com

Department Of Computer Science & Engineering, Sri Sai University Palampur (HP)

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 analysis is the technique which is applied on analyses sentiment of the input data. The sentiments of individuals towards particular elements are analyzed in this approach. To gather sentiment information, web or internet is the best known source. In comparison to several other domains, the sentiment analysis requires higher analysis studies. This research work is based on the sentiment analysis of product reviews of Amazon data [7]. To apply sentiment analysis the technique of feature extraction and classification is applied. The subjective view of a product review reflects the opinions expressed by opinion words, while the objective view is constructed by the remaining text features. For the sentiment analysis in the previous work, the WDE-LSTM technique is applied and which is replaced with the WDE-KNN technique. In the proposed method every product review has two views: subjective view and objective view. The existing and proposed techniques are implemented in python and simulation results shows that accuracy of the proposed technique is better than previous. The simulation results shows that execution time of the proposed method is less as compared to existing method.

KEYWORDS:

WDE-LSTM, WDE-KNN, Sentiment Analysis, Amazon, NLP.

I. Introduction

Sentiment analysis is a type of data mining that assess the inclination of people's opinion through NLP. It uses the NLP in order to categorize the opinions of people about the products or the reviews. Sentiment analysis sometimes known as opinion mining. Sentiment analysis is a technique which helps to grasp or understand the behavior of user and speaker. It deals with opinions and perspective of human related to emotions and attitude about some occurrence or the event [1]. In social media sites it helps to determine whether the reviews, blog-posts, news, articles is positive, negative and neutral. Sentiment analysis helps merchants for improving their products and services. The two important tasks involved in Opinion Mining and Sentiment Analysis are 1) Opinion Extraction: extracting the opinionated phrases, in proper context, from free text and (2) Sentiment classification: classifying opinionated phrases based on sentiment orientation. It utilizes various machine

learning techniques such as SVM, Naïve Bayes, Character Based N-gram model etc. [17] for sentiment classification. It also help future customers to make decision such as they buy products or not. In this technique of sentiment analysis the features of the input data are extracted using pattern matching algorithm and for the sarcasm detection, classification techniques are applied .Sentiment Analysis is also known as the opinion mining [2]. Opinion mining or sentiments analysis is most useful in various fields like commercial, product reviews, social media analysis and movie reviews etc. [3]. 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. Different machine learning classifiers are also used to predict offline results using sentiment analysis.

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. Blogs, review sites and data give a good understanding of the acceptance level of the products and services. The issue related to the sentences classification has been solved with the help of machine learning approach as it totally based on the algorithms [4]. Supervised leaning approach and unsupervised learning approach are the two utilized approaches. In the proposed approach every customer review has two views: subjective view and objective view. The subjective view of a customer review reflects the opinions expressed by opinion words, while the objective view is constructed by the remaining text features [5].

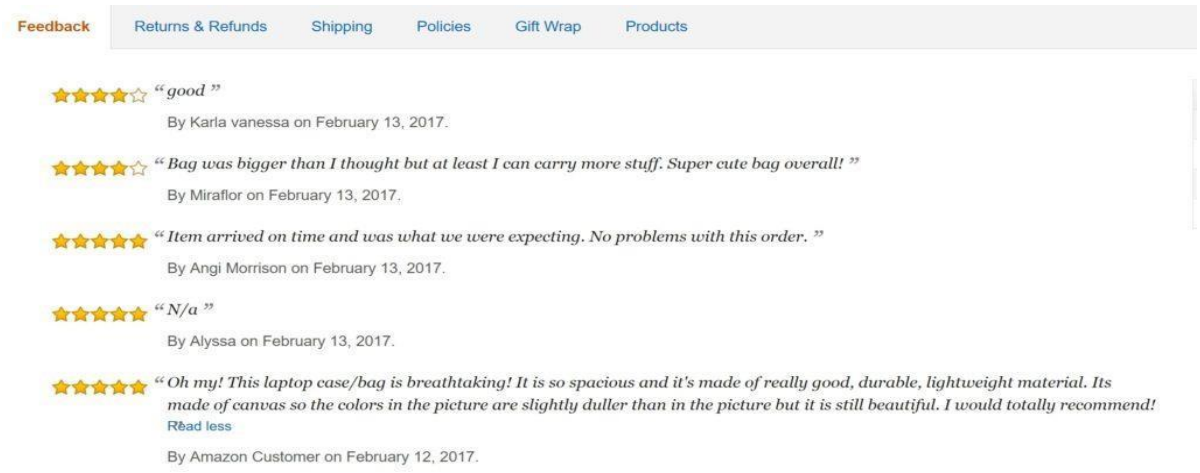

Figure 1: A Product Review Example

The above figure shows a product review which may include both subjective and objective information. Subjective information which indicates the opinions of opinion holders, while objective information indicates some objective facts. Mostly web users or peoples are likely to express their opinions using some opinion words. Opinion words such as "good" and "specious" in the sentence "It is so specious and it's made of really good." These opinion words are subjective. Subjective texts contains both positive and negative sentiment. Positive sentiment have good, super cute, spacious and negative sentiment has slightly dull. Figure 1 is an example of product review coming from Amazon [6] .Not every word is subjective but most of the words of the review give objective information. To simplify the problem, subjective information may contains extracted opinion words whereas the remaining text of reviews contain objective information. The data set which we have

used in our paper is amazon data. It contains labelled review sentences and unlabeled or weakly labelled sentences. The data collected from amazon is in three domain i.e. cell phones, digital cameras and laptops. The data set can be downloaded at https://www.dropbox.com/s/aji68llxmtcuu5l/data.zip [7].The reviews collected are sometimes known as subjective and objective reviews. The rest of the paper is organized as follows: next section shows the related work, section 3 shows research methodology, and experiment result shows in next 4section, section 5 contains conclusion and future work.

II. BACKGROUND STUDY

Opinion mining has become an interesting research area in the field of data mining. Sentiment classification is the major effort of opinion mining. Sentiment analysis sometimes known as opinion mining. Sentiment classification has bait much attention in recent years. The sentiments analysis is the technique which is applied on analyses sentiment of the input data. Following work had been done in this research area and still goes on. In "Sentiment Analysis: Approaches and Open Issues" Shah Nawaz, presented a sentiment analysis process that helps to provide idea to the customer to identify the product or service that is satisfactory for a customer or not before they buy 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 [8].Accuracy, inability to perform well in different domain and performance are the main issues in the current techniques. Semi-supervised and unsupervised learning based models are used that will help to be easily minimize lack of labeled data if sufficient amount of unlabeled data is available. Nowadays, the more researches have been focusing on automatic processing and extracting the sentiment information from the large data [9].

Pierre Ficamos, Proposed a feature extraction method that depends on Part Of Speech (POS) tags, that helps in selection of the unigram and bigram features. It mainly 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 result shows that the proposed approach provides the better results with the Naïve Bayes. We know Aspect level sentiment classification is a major task in the field of sentiment analysis. Lemieux, Jin Lin, Lina Wang, proposed "Deep Convolutional Neural Network based Approach for Aspect-based Sentiment Analysis" The main task is to extract aspects from the review text and then inferring the sentiment polarity (e.g. positive, negative) of the aspect [11].Deep convolutional neural networks (CNN) utilize layers with convolving filters that are applied to local features, and CNN models have demonstrated remarkable results for text classification and sentiment analysis [10].

ZiyuGuan1, LongChen1 Proposed a novel deep learning framework for Weakly-Supervised Deep Learning for Customer Review Sentiment Classification" [11] where customer reviews are an important form of opinionated content. Its goal is to identify each sentence's semantic orientation (e.g. positive or negative) of a review. It uses CNN architecture for sentence classification [12]. Experiment results on reviews collected from Amazon.com show that WDE is sufficient and outperforms baseline methods.

Another advancement is based on "Weakly-supervised Deep Embedding for Product Review Sentiment Analysis ". In this Wei Zhao, Ziyu Guan, test two low level network for modelling review sentence such as Convolutional feature extractor [13] and long- short- term –memory [14]. The experiment result shows that

WDE-LSTM perform better than WDE –CNN in terms of accuracy. Much work have been included for sentiment analysis on reviews in [15], we know sentiment analysis is a classification problem, where feature which contain sentiments and opinion should be identified before classification. Pang and, Lee [16] recommend to remove objective sentences by extracting subjective ones for feature selection.

III. PROPOSED WORK

- **Dataset**

The data set can be downloaded from Amazon product reviews. The product reviews contain data in 3 domains: digital cameras, cell phones and laptops. All unlabeled reviews were extracted from the Amazon data product dataset. For the labelled dataset, we crawled latest reviews in 2015 for random products in the above 12 categories, in order to be disjoint with the unlabeled data. We attempt to keep an equity between reviews with 4 & 5-stars and those with 1 & 2-stars.

- **Data Pre-processing**

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.

- **Lexical Analysis of Sentences**

It is the first phase of a compiler. It breaks sentence into a series of tokens, by removing comments, or whitespace in the source code. In sentiment analysis 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.

- **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.

- **Define Positive, Negative and Neutral Words**

Each word used to make a complete sentence having sense has its own importance. Although we have mentioned positive words as good ,better, brilliant etc. whereas negative sentence as bad , worst, etc. in sentiment analysis neutral word are the words having no positive as well as negative values. In sentiment analysis positive and negative values can be calculated through rating. For example a rating between 4 & 5 is considered as positive, a rating between 1&2 is considered as negative whereas rating 3 has neutral.

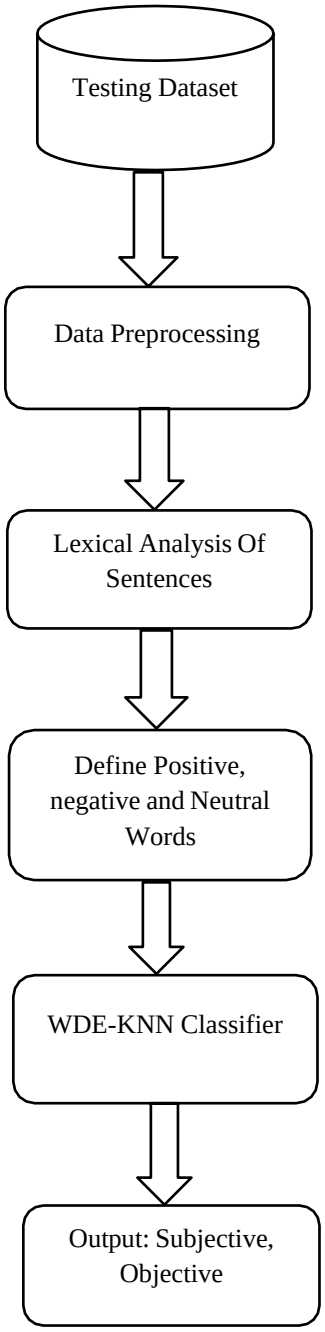

Figure 2: A Systematic Approach for Proposed Technique

- **WDE-K-Nearest Neighbour 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.KNN is the non-parametric algorithm used in case of classification and regression. In classification and regression, the input is consisting of K-nearest training

examples in the feature space and on the other hand, the output depends upon whether KNN belongs to regression category or classification category. The output belongs to the classification category when:

- An object is classified by the majority of the votes from its neighbor along with the objects being assigned to that class which is most common its K-nearest neighbor. If k=1, then the object directly belongs to the class of that single nearest neighbor.

The output belongs to the regression class when

- The output is the property value for the objects. This value is the average mean of all the nearest neighbors.

KNN is the simple algorithm in which all the cases are stores and classifies based on the similarity measures and has been used in statistical estimation and pattern recognition

Once the value of k is selected, the prediction can be made. For the regression, KNN prediction is the average of KNN outcomes

$$y = 1/k \sum_{i=1}^{k} y\mathrm{i}$$

Here y_i is the ith case and *y* denotes the prediction or outcome of the query point. A non-parametric approach that is used to perform classification, as well as regression, is known as K-NN classifier. The data is assumed to be present in the feature space by KNN. The data points tend to exist within a metric space here. Either within the scalars or multidimensional vectors, the data is available in the applications. A notion of distance is important here since the points are present in the feature space. Even though the Euclidean distance is the most commonly used algorithm to calculate the distance, it is not necessary to use only this method for calculation. In order to estimate the density at point x, a hypercube that is centered at x is placed and its value is increased until *k* neighbors are captured. Then, the following formula is applied to estimate the density:

$$p(x) = \frac{k/n}{V}$$

Here, the total numbers of data points available are denoted by n. The volume of the hypercube is represented by V.

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.

- **Output/ Extraction of Feature Wise Opinion**

The output comes in the form of subjective and objectives reviews**.** All the reviews that include 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. As shown in figure 1, the systematic approach for proposed technique is explained.

IV. EXPERIMENTAL RESULTS

The proposed work is implemented in Python and the results are analyzed in terms of accuracy and execution time. Different parameters are used to calculate the accuracy or execution time.

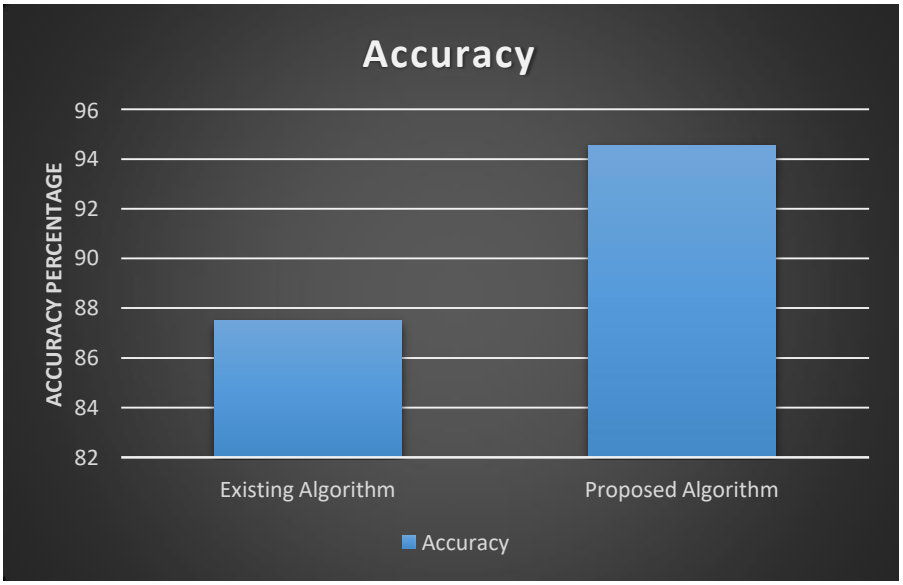

Figure 3: Accuracy Comparison

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.

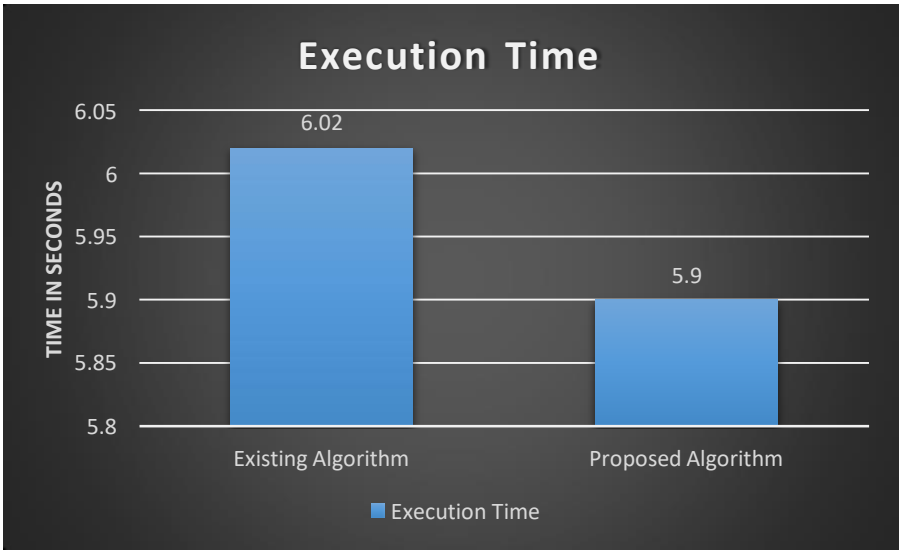

Figure 4: Execution Time Comparison

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.

Table 1: Execution Time and Accuracy Comparison

Method	Execution time	Accuracy
WDE-LSTM	6.02 sec	87.26
WDE-KNN	5.09 sec	94.12

As shown in table 1, the accuracy and execution time of proposed and existing algorithms are compared for the performance analysis. It is analyzed that proposed algorithm performs well in terms of accuracy and has less execution time as compared to existing algorithm.

V. CONCLUSION

The behavior of user is analyzed in this research work on the basis of analysis sentiments of amazon data. N-gram technique is applied 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. 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. 87.12 percent and when the technique wde-knn is applied it is increase up to approx.94 percent. The performance of wde-lstm and wde-knn technique is also compared in terms of execution time. It is analyzed that proposed technique performs well in terms of all parameters. In future wde is applied with other classifier to increase accuracy. Also different classifier are combined based on their polarity to increase accuracy. Our work falls on aspect level whereas another research work have been done using sentence level or document level.

ACKNOWLEDGEMENT

I am grateful and sincerely express my deep gratitude Er. Akanksha Sambyal (assistant professor, CSE Department, Sri Sai University Palampur (HP), India for her constant encouragement, motivation and supervision throughout my work. Her endless support and motivation helped me to put my ideas and my efforts in proper direction.

REFERENCES

[1] X. Ding, B. Liu, and P. S. Yu. A Holistic Lexicon-Based Approach to Opinion Mining. In WSDM, pages 231–240, 2008.

[2] NarendraAndhale, L.A. Bewoor. An Overview of Text Summarization Techniques.2016, IEEE.

[3] Chhaya Chauhan, SmritiSehgal, "Sentiment analysis on product Reviews" International Conference on Computing, Communication and Automation, vol. 10, pp. 1-8, 2017

[4] Mika V. Mäntylä1 Daniel Graziotin, Miikka Kuutila, "The Evolution of Sentiment Analysis - A Review of Research Topics, Venues, and Top Cited Papers", Volume 27, February 2018, Pages 16-32, ISSN, 2018.

[5] Suke Li, Yiheyuan, "Sentiment Classification using Subjective and Objective Views" International Journal of Computer Applications (0975 – 8887) Volume 80 – No7, October 2013.

[6] http://blog.refundsmanager.com/wp-content/uploads/2017/02/seller-feedback-example-1024x458.jpg

[7] https://www.dropbox.com/s/aji68llxmtcuu5l/data.zip.

[8] Shah Nawaz, ParmanandAstya "Sentiment Analysis: Approaches and Open Issues" International Conference on Computing, Communication and Automation, vol. 9, pp. 1-5, 2017

[9] Yan Liu, Pierre Ficamos, "Naive Bayes and Maximum Entropy approach to sentiment analysis: Capturing domain-specific data in Weibo", IEEE International Conference on Big Data and Smart Computing (Big Comp). vol. 8, pp. 1-4, 13-16 Feb. 2017

[10] Y. Kim "Convolutional Neural Networks for Sentence Classification" Proceedings of the 2014 Conference on EMNLP October 25-29 2014, 1746–1751. Retrieved from http://emnlp2014.org/papers/pdf/EMNLP2014181.

[11] Lemieux, Jin Lin, Lina Wang, Chunyong Yin, Jin Wang" Deep Convolutional Neural Network based Approach for Aspect-based Sentiment Analysis".

[12] Baotian Hu, Zhengdong Lu, Hang Li, and Qingcai Chen "Convolutional neural network architectures for matching natural language sentences" Corer, abs/1503.03244, 2015. URL http://arxiv.org/abs/1503.03244.

[13] S. Hoch Reiter and J. Schmidhuber. Long short-term memory. Neural computation, 9(8):1735–1780, 1997.

[14] Alex Graves and Jürgen Schmidhuber"Frame wise phoneme classification with bidirectional lstm and other neural network architectures "Volume 18. Issues 5-6, August 2005, Pages 602-610.

[15] Bing Liu"Sentiment Analysis and Opinion Mining" Morgan & Claypool Publishers, May 2012.VOL.5, No.1, Pages 1-167.

[16] Pang B, Lee L (2004) A sentimental education: sentiments analysis using subjectivity summarization based on minimum cuts In: Proceeding of the 43d annual meeting on association for computational linguistics, ACI '04. Association for computational Linguistics, Stroudsburg, PA, USA 2014.

[17] 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.