International Information and Engineering Technology Association
Advancing the World of Information and Engineering

A Combined Approach of Sentimental Analysis Using Machine Learning Techniques

Ketan Gupta[1] , Nasmin Jiwani[1] , Neda Afreen[2]*

[1] Department of Information Technology, University of The Cumberlands, Williamsburg, KY 40769, USA
[2] Department of Mathematics and Computer Science, University of Cagliari, Cagliari 09124, Italy

Corresponding Author Email: n.afreen@studenti.unica.it

ABSTRACT

Sentiment analysis is a vital area of current research. The area of sentiment analysis is extensively used for observing text data and identifying the sentiment element. Every day, e-commerce sites produce a massive amount of text information from customer's comments, reviews, tweets, and feedbacks. One of the most recent technological advances in web development is the emergence of social networking websites. It aids in communication and knowledge gathering. Aspect - based evaluation of this information can help businesses to gain a greater understanding of their consumers' expectations and then shape their plans accordingly. It is difficult to convey the exact sentiment of a review. In this study, we demonstrated an approach that focuses on sentimental aspects of the item's characteristics. Consumer reviews on Amazon and IMDB have been presented and evaluated. We obtained the dataset from the UCI repository, where each analysis's opinion rates are first observed. To get meaningful information from datasets, and to eliminate noise, the pre-processing operations are performed by the system such as tokenization, punctuation, whitespace, special character, and stop-word removal. For the purpose of accurately representing the preprocessed data, feature selection methods such as word frequency-inverse document frequency are utilized (TF–IDF). The customer reviews from three datasets Amazon, Yelp, and IMDB is merged and classification is performed using classifiers such as Naïve Bayes, Random Forest, K-Nearest Neighbor (KNN), and Support Vector Machine (SVM). In last, we provide some insight into the future text classification work.

1. INTRODUCTION

Sentimental analysis is among the machine learning processing techniques that aid in the identification of sentiments, allowing entrepreneurs to access information about their clients' opinions via various online media like surveys, social media, and reviews on e-commerce website. People are dealing products by using various e-commerce websites as the world's commercial sector has almost completely transitioned to an online platform. As a result, reviewing items or commodities before purchasing is a common scenario. Customers are also more likely to purchase something or go for movies based on reviews these days [1]. As a result, nowadays, examining the information from such customer reviews to transform the data is an important area. Sentimental analysis applications in industry range from predictions of future market movements on the basis of sentiment displated in blogs and news, to analyze customer satisfaction and dissatisfaction based on posts of social media and reviews. This also serves as the foundation for several other application fields such as recommender systems. The sophistication of sentiment analysis involves the removal of noisy data from the original dataset, the selection of suitable features for representation, and the selection of an appropriate classifier. Sentiment analysis gained popularity during the early 2000s. Scholars have shown a strong interest in this field [2].

As part of sentiment analysis, two terms can be investigated:

'Polarity' and 'Subjectivity.' Subjectivity is about beliefs of individual, their opinions, or personal sentiment, whereas polarity basically refers to feelings expressed as negative, positive, or neutral. Sentiment analysis encompasses working at the different levels of sentence, sub-sentence, and document. On different areas, different forms of sentiment analysis can be carried out, such as fine-grained sentiment analysis by operating on polarities ranging from very negative to very positive, intent-based or emotion detection. To perform sentence analysis, a conventional method based on a Lexicon can be used, as well as a machine learning-based technique too. Both methods have some advantages and disadvantages.

The aim of this study is to identify and classify positive and negative customer feedback on various products and movie reviews and use machine learning model to classify them [3]. According to an analysis conducted on Amazon previous year, more than 80% of online customers valued reviews more than the specific recommendations. Any online product with a huge number of positive feedback makes a strong comment on the item's credibility, and without reviews on the other hand, causes potential customers to be doubtful. Simply put, more reviews appear more credible [4]. People value other people's opinions and experiences, and reviewing a product is the only way to understand what others think of it [5]. Opinions gathered from consumer experience with particular products or topics directly affect future purchases made by customers. Likewise, negative reviews frequently result in a loss of sales. In this approch, we merged two dataset from amazon and

IMDB movie review. After preprocessing, perform feature extraction using TF-IDF approach and finally applied machine learning classifier to get the performance result.

The remainder of the paper is structured as follows: in Section 2, we provide a quick overview of the literature on sentiment analysis. Section 3 describes the methodologies involved. Section 4 contains comprehensive experimental analysis. Finally, Section 5 concludes with a discussion of future work.

2. RELATED WORK

A great deal of research has been conducted in the field of text classification and sentiment analysis. One of the fundamental problems in sentiment analysis is to categorize sentiment score. The task is to determine whether a fraction of text is negative or positive. The aspect or entity level, the document level, and the sentence level are the three different forms of sentiment polarity classification. The entity type is concerned with what is the opinion of people about their likes and dislikes. The document level perceives the valence, i.e., negative or positive sentiment, for the whole text, whereas the sentiment categorization for each sentence comes under sentence level.

Many scholars from all over the world have conducted studies using semi-supervised, unsupervised, and supervised machine learning techniques. A sentiment analysis methodology was suggested by Bhatt et al. [6] for evaluation of iPhone5 reviews on amazon. The technique combines several pre-processing methods for reducing noise in data such as punctuation, numbers, and HTML tags. The features are identified with the help of part-of-speech (POS) tagger, and a rule-based procedure is used for reviews categorization. Haddi et al. [7] investigate the impact of text pre-processing on reviews of online movie. To remove noisy data, many preprocessing procedures such as stemming, html tag removal, and data cleaning are performed. The chi-square technique for feature selection is used to eliminate the irrelevant features. The SVM is applied on reviews to categorize them into negative or positive. Wassan et al. [8] proposed a new approach which focuses on sentimental components of the item's characteristics. Customer reviews on Amazon have been presented and evaluated. They derived dataset from data world center, and detect opinion rates and extract meaningful information such as negativity or positivity, and performs pre-processing activities on datasets such as tokenization, stone-coating, stop-words deletion, and boxing. They focus to examine the data's aspect level to understand consumer preferences and continuing to develop their behavior accordingly.

Srujan et al. [9] first eliminate noise from the reviews of amazon book by using various preprocessing techniques such as URL removal and HTML tag, whitespace, punctuation, stemming, and special character removal. Term frequency-inverse document frequency(tf-idf) are utilized for feature selection to represent the preprocessed data. They compare the accuracy of various classifiers such as K-Nearest Neighbor (KNN), Decision Tree (DT), Support Vector Machine (SVM), Random Forest (RF), and Naive Bayes (NB) and the time taken by each classifier and also find the sentimental score of different books. Shrestha and Nasoz [10] assess the consistency of reviews from Amazon.com to their correlating

ratings using sentiment analysis. Sometimes there is a discrepancy between the review and the rating what consumers submit on e-commerce websites. They employed deep learning sentiment analysis on product review from amazon.com to identify reviews with mismatched ratings. Using paragraph vector, the product reviews were transitioned to vectors, which were later used to train a recurrent neural network with gated recurrent unit. They consider both the product information and the semantic relationship of the review text. Using the trained model, they created a web service application that predicts the rating score for a submitted review and provides feedback to the reviewer if there is a difference in the submitted rating score and the predicted rating score. Nandal et al. [11] proposed a novel approach for aspect level sentiment detection, which emphasis on the features of item. The task was trialled on Amazon customer reviews, at which aspect terms were identified first for each review. The preprocessing operations were performed by the system on the dataset such as tokenization, stemming, stop-word removal, and casing, to retrieve meaningful insights before assigning a rank to its classification in positivity or negativity. Chen et al. [12] perform review embedding on the IMDB and Yelp datasets using a one-layer convolutional neural network (CNN). That CNN generates 300-dimensional vectors from reviews of varying length. They padded shorter reviews with zero vectors to accommodate varying lengths of reviews. Filters with widths of 3 and 5 performed one-dimensional convolution on the word embeddings, yielding multiple feature maps. By using max-overtime pooling in the pooling layer, only useful features were captured. Multiple filter outputs were then concatenated to form a 300-dimensional vector. To train the network over K-classes, the Softmax function was used as an activation function. The Softmax function was used as an activation function to train the network over K-classes. They then employed a Recurrent Neural Network (RNN) to learn the temporal information and capture both user and product information, reporting cutting-edge results on the Yelp and IMDB datasets. Their research is based on the assumption that an item which gets positive feedback initially is more likely to get positive feedback in the future, and vice versa.

We, in this work, presents a method for elimination of noise and feature extraction using TF-IDF from amazon and IMDB datasets by combining them and performs comparative analysis on the accuracy of various classifiers and their sentiment scores. The methodology and classifiers used in our experiment are discussed in the following section.

3. METHODOLOGY FOR THE SENTIMENTAL ANALYSIS

For our analysis, we will use the Amazon product review, IMDB movie review, and yelp review dataset. Techniques include preprocessing, classification, and representation. Pre-processing approaches include data cleaning, number and punctuation, stop word removal, and the HTML/URL tag removal. The TF-IDF representation model is used to represent the preprocessed text. To categorize the dataset into negative and positive classes, classifiers such as Naive Bayes (NB), Random Forest (RF), Support Vector Machine (SVM), and K-Nearest Neighbours (KNN) [13] are used and their performances are compared. The overall work procedure is shown in Figure 1.
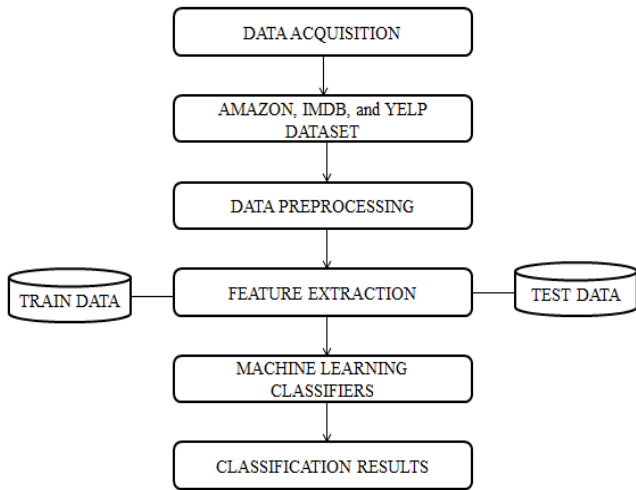
Figure 1. Work flow of sentimental analysis

3.1 Data acquisition

We obtained our three dataset amazon product review, IMDB movie review, and yelp review [14] from UCI repository. We merge all the three reviews in single dataset. As all these reviews are based on a 5-star scale rating, 3 star ratings are regarded as neutral, which means they are neither negative nor positive. So we remove reviews with a three-star rating from required dataset and proceed to the next step, 1 is considered as positive rating and 0 is considered as negative rating [15]. The scree shot of small random samples is shown in Figure 2 after combining Amazon, IMDB, and Yelp dataset.

	Review	Sentiment
0	Wow... Loved this place.	1
1	Crust is not good.	0
2	Not tasty and the texture was just nasty.	0
3	Stopped by during the late May bank holiday of...	1
4	The selection on the menu was great and so wer...	1
...	...	...
2742	I never walked out of a movie faster.	0
2743	I just got bored watching Jessice Lange take h...	0
2744	Unfortunately, any virtue in this film's produ...	0
2745	In a word, it is embarrassing.	0
2746	Exceptionally bad!	0

Figure 2. Screenshot of combined Amazon, IMDB, and yelp review dataset

3.2 Data pre-processing

The process of converting raw data into a logical format is known as data preprocessing. Data preprocessing method is used to improve data quality. It includes various operations such as tokenization, punctuation, whitespace, special character, and stop-word removal [16].

Tokenization- It is the method of breaking down a string pattern into separate components such as phrases, words, symbols, keywords, and other components defined as tokens. Phrases, Individual words, or even entire sentences can serve as tokens. Some characters, such as punctuation marks, are deleted during the tokenization process. Tokens are used as input for various processes such as text mining and parsing.

Eliminating Stop Words- Stop words are objects in a sentence that are not needed in any segment of text mining. To enhance the analysis accuracy, we normally ignore such words. There are various stop words based on the country, language, and so on. There are numerous stop words in English such as 'or', 'each', 'both', 'when', 'whereupon', etc.

POS tagging- Parts of Speech (POS) tagging refers to the method of allocating one of the parts of speech to a specific word. It is commonly known as POS tagging. Verbs, adverbs, nouns, pronouns, adjectives, conjunctions, and their subclasses are all examples of parts of speech. Parts of Speech tagger, also known as POS tagger, is a program that performs this function [17].

3.3 Feature extraction

The significant step in sentiment classification is representation. In general, noise is present in raw data and must be filtered using many pre-processing methods. The data that has been preprocessed is next converted into a term document matrix, also known as a TDM, which determines the frequency of each individual word. The TF-IDF and bag of words are two of the feature extraction methods that are supported by the TDM. When multiplied together, TF and IDF make up a word's TF-IDF score; this score may be found by looking at the product of these two elements. The review dataset's terms that are found most frequently are given more weight in the TF score than other words. The least common words in the dataset are given more importance thanks to the IDF scaling factor. This score is lower for uncommon and common words than it is for other types of words. We are able to get rid of them if we ignore terms that have poor TF-IDF values [18].

TF-IDF- Information can be extracted using a method known as TF-IDF, which takes into account both term frequency (TF) and inverse document frequency (IDF) (IDF). There is a distinct TF mark and IDF mark assigned to each phrase and word. A term's total TF and IDF product scores are added together to determine its TF*IDF weight. To put it another way, the phrase is considered more rare when its TF*IDF score (weight) is higher, and vice versa. The TF of a term is a representation of how frequently it is used. The IDF of a word is an estimation of its importance all across the text [19].

When words in content have a high weight of TF*IDF, the information will always be among the leading search outcomes, allowing any of us to:

1. Quit worrying about the use of stop-words,
2. Look for words with low competition and high search volume.

3.4 Experimentation

Following the transformation of the training and testing datasets into individual term document matrices (TDM), the frequency of occurrence of each word is analyzed. The TDM is then fed into classifiers like SVM, KNN, Naive Bayes, and Random Forest [20]. The algorithm for proposed work is given

below.

Proposed work algorithm
Input: Labelled data
Output: Classifiers accuracy;
Recall, Precision, F-1Measure for positive and negativevalues
1. Load negative and positive (0 & 1) labeled data
2. Preprocessing of data that has been labeled
3. for each Value {V1...Vn} in the data that is tagged
4. Feature extraction(Vi)
5. validate into train and test dataset
6. classifier.train()
7. calculate accuracy
8.Compare result (precision, recall, accuracy, f1 measure) of different machine learning classifiers such as, SVM, KNN, NaiveBayes, and Random Forest
9. End

4. RESULTS AND DISCUSSION

This section contains a detailed description of evaluation metrics, and a discussion of the results obtained.

4.1 Evaluation metrics

Evaluation metrics are critical for measuring performance of classification. Accuracy is the commonly used metric for performance checking. A classifier's accuracy on a specified test dataset is the proportion of those datasets classified correctly by the classifier. And, as the accuracy measure is not sufficient in the text mining analysis to make a valid judgment, we also implement some other metrics to evaluate classification performance [21]. F-measure, recall, and precision are three fundamental measures which are generally used. Before delving into various measures, there are a few terms we should become acquainted with:

- TP (True Positive) shows the number of data points that were classified correctly.
- FP (False Positive) shows the number of misclassified correct data.
- FN (False Negative) shows the number of incorrect data that has been classified as positive.
- TN (True Negative) represents the number of incorrect data classified as negative.

Accuracy- Accuracy indicates how oftenly the classifier appears to make correct prediction. The accuracy ratio is calculated by dividing the number of correct predictions to the total number of predictions.

$$Accuracy = \frac{Correct\ Prediction}{Total\ prediction} \qquad (1)$$

Recall- Recall calculates a classifier's sensitivity. More recall means lesser false negatives. The recall ratio is the number of correctly classified occurrences divided by the total number of predicted occurrances. This can be illustrated as-

$$Recall(R) = \frac{TP}{TP+FN} \qquad (2)$$

Precision- The exactness of a classifier can be measured by its precision. A low level of precision will result in a greater number of false positives, whereas a high level of precision

will result in fewer of these errors. The term "precision" (P) is defined as-

$$Precision(P) = \frac{TP}{TP+FP} \qquad (3)$$

F-Measure- A single metric called F-measure (the weighted harmonic mean of recall and precision) is formed when precision and recall are combined. It is defined as follows-

$$F - measure = \frac{2PR}{P+R} \qquad (4)$$

4.2 Experimental results

In the study, we consider two classes of problems, negative and positive. We have taken reviews of three different sites, namely amazon, IMDB, and Yelp. Each of these review includes the user's rating and text reviews. Based on their ratings, these reviews are divided into two categories: negative and positive. Reviews with ratings of 4 or 5 are considered positive, while reviews with ratings of 1 and 2 are considered negative. The set is divided in 80:20 proportions for training and testing. In this experiment, various machine learning algorithms like Support Vector Machine Classifier (SVC), K Nearest Neighbor (KNN), Naive Bayes, and Random Forest are used. Several other types of preprocessing operations, such as data cleaning, the elimination of whitespace, and stemming, are carried out on the dataset. The train set and the test set are both transformed into individual term document matrices (TDM), and then the frequency of occurrence of each word is counted. The TF-IDF is utilized so that the data can be represented. After that, the TDM is used as an input for a number of different classifiers. The confusion matrix of the given classifiers is shown below in Figure 3.
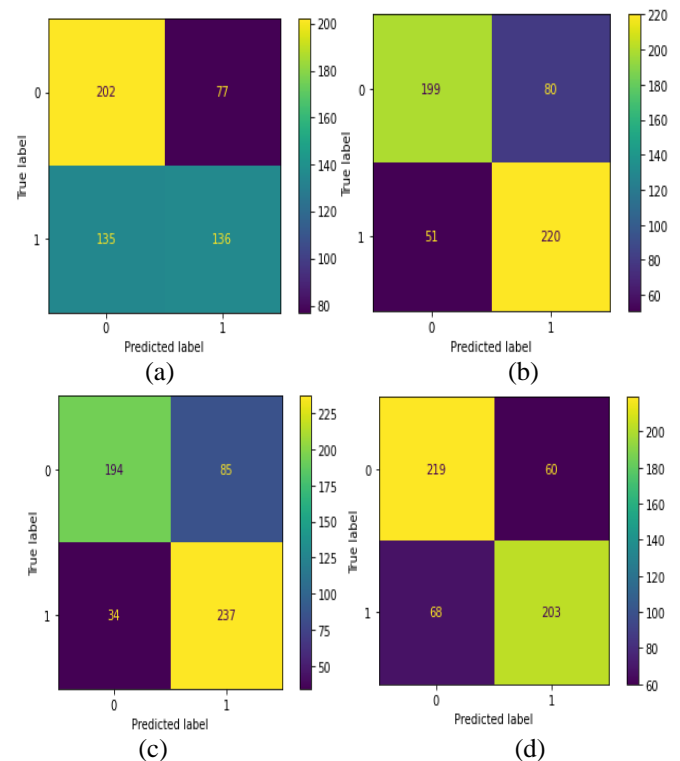
Figure 3. Confusion matrix of (a) KNN, (b) SVM, (c) Naive Bayes, and (d) Random Forest

Experimental result of different classifier is shown in Table 1. From Table 1, We can notice that Random Forest outperforms the other methods in the two-class classification problem. The highest accuracy obtained from these results was 78.96%. The TDM is made up of feature vector values known as points. The categorization accuracy significantly improved by constructing an ensemble of trees and allowing them to vote on the most popular class.

Table 1. Experimental result of classifiers on combined dataset

Dataset	Classifier	Accuracy	Precision	Recall	F1 score
AMAZON+IMDB+ YELP	support Vector machine	0.76	0.76	0.76	0.76
	Random Forest	0.78	0.79	0.78	0.78
	Multinomial Naïve Bayes	0.77	0.77	0.77	0.77
	KNN	0.61	0.62	0.61	0.61

5. CONCLUSIONS AND FUTURE SCOPE

Analysis of customer sentiment is an indispensable tool for any online retailer to fully comprehend their customers' reactions. Reviews must be analyzed in order to develop recommendation systems that reach consumers according to their preferences. In this work, we demonstrated different machine learning model for a large unlabeled review dataset from amazon, IMDB, and yelp. With different machine learning model, we used feature extraction technique. We highlighted the fundamental theory behind the model, as well as the approaches that we took in our research and the performance metric for a large-scale experiment. Random Forest classifier is able to achieve over 78% accuracy.

For better representation, the research scope can be expanded by embedding different techniques for feature selection such as mutual information (MI), information gain, and chi-squared test. To improve accuracy, we can combine hybrid classifiers such as SVM with other methods. A good recommendation methodology can be developed by considering the emotions elicited by customer reviews.

REFERENCES

[1] Alsaeedi, A., Khan, M.Z. (2019). A study on sentiment analysis techniques of Twitter data. International Journal of Advanced Computer Science and Applications, 10(2): 361-374. http://dx.doi.org/10.14569/IJACSA.2019.0100248

[2] Vijayarani, S., Ilamathi, M.J., Nithya, M. (2015). Preprocessing techniques for text mining-an overview. International Journal of Computer Science & Communication Networks, 5(1): 7-16.

[3] Gupta, K., Jiwani, N., Whig, P. (2023). Effectiveness of machine learning in detecting early-stage leukemia. International Conference on Innovative Computing and Communications. 471: 461-472. https://doi.org/10.1007/978-981-19-2535-1_34

[4] Prabowo, R., Thelwall, M. (2009). Sentiment analysis: A combined approach. Journal of Informetrics, 3(2): 143-157. https://doi.org/10.1016/j.joi.2009.01.003

[5] Miao, Q., Li, Q., Dai, R. (2009). AMAZING: A sentiment mining and retrieval system. Expert Systems with Applications, 36(3): 7192-7198. https://doi.org/10.1016/j.eswa.2008.09.035

[6] Bhatt, A., Patel, A., Chheda, H., Gawande, K. (2015). Amazon review classification and sentiment analysis. International Journal of Computer Science and Information Technologies, 6(6): 5107-5110.

[7] Haddi, E., Liu, X., Shi, Y. (2013). The role of text pre-processing in sentiment analysis. Procedia Computer Science, 17: 26-32. https://doi.org/10.1016/j.procs.2013.05.005

[8] Wassan, S., Chen, X., Shen, T., Waqar, M., Jhanjhi, N.Z. (2021). Amazon product sentiment analysis using machine learning techniques. Revista Argentina de Clínica Psicológica, 30(1): 695-703. http://dx.doi.org/10.24205/03276716.2020.2065

[9] Srujan, K.S, Nikhil, S.S, Rao, R., Kedage, K., Harish, B.S., Kumar, H.M.K. (2018). Classification of Amazon Book Reviews Based on Sentiment Analysis. Information Systems Design and Intelligent Applications, 672: 401-411. http://dx.doi.org/10.1007/978-981-10-7512-4_40

[10] Shrestha, N., Nasoz, F. (2019). Deep learning sentiment analysis of amazon. com reviews and ratings. arXiv preprint arXiv:1904.04096. https://doi.org/10.5121/ijscai.2019.8101

[11] Nandal, N., Tanwar, R., Pruthi, J. (2020). Machine learning based aspect level sentiment analysis for Amazon products. Spatial Information Research, 28: 601-607. https://doi.org/10.1007/s41324-020-00320-2

[12] Chen, W., Lin, C., Tai, Y.S. (2015). Text-based rating predictions on amazon health & personal care product review. Computer Science. https://doi.org/10.1145/1235

[13] Jiwani, N., Gupta, K., Whig, P. (2023). Assessing permeability prediction of BBB in the central nervous system using ML. International Conference on Innovative Computing and Communications. Lecture Notes in Networks and Systems, Springer, Singapore, pp. 449-459. https://doi.org/10.1007/978-981-19-2535-1_33

[14] Wang, M., Qiu, R. (2015). Text mining for yelp dataset challenge. Computer Science, pp. 1-5.

[15] Tan, W., Wang, X., Xu, X. (2018). Sentiment analysis for Amazon reviews. In International Conference, pp. 1-5.

[16] Shaikh, T., Deshpande, D. (2016). Feature selection methods in sentiment analysis and sentiment classification of amazon product reviews. International Journal of Computer Trends and Technology (IJCTT), 36(4).

[17] Jagdale, R.S., Shirsat, V.S., Deshmukh, S.N. (2019). Sentiment analysis on product reviews using machine learning techniques. Cognitive Informatics and Soft Computing. 768: 639-647. https://doi.org/10.1007/978-981-13-0617-4_61

[18] Trstenjak, B., Mikac, S., Donko, D. (2014). KNN with TF-IDF based framework for text categorization. Procedia Engineering, 69: 1356-1364. https://doi.org/10.1016/j.proeng.2014.03.129

[19] Haque, T.U., Saber, N.N., Shah, F.M. (2018). Sentiment analysis on large scale Amazon product reviews. In 2018 IEEE International Conference On Innovative Research and Development (ICIRD), Bangkok, Thailand, pp. 1-6. https://doi.org/10.1109/ICIRD.2018.8376299

[20] Rintyarna, B.S., Sarno, R., Fatichah, C. (2020).

Enhancing the performance of sentiment analysis task on product reviews by handling both local and global context. International Journal of Information and Decision Sciences, 12(1): 75-101. https://doi.org/10.1504/IJIDS.2020.104992

[21] Das, S., Jabirullah, M., Afreen, N., Prabhakara Rao, A., Gayatri Sarman, K.V.S.H. (2023). Gamma band: A bio-marker to detect epileptic seizures. Smart Technologies for Power and Green Energy. 443: 335-364. https://doi.org/10.1007/978-981-19-2764-5_29