Sentiment Analysis of Amazon Product Reviews by Supervised Machine Learning Models

Mohamad Faris bin Harunasir, Naveen Palanichamy *, Su-Cheng Haw, and Kok-Why Ng

Faculty of Computing and Informatics, Multimedia University, Cyberjaya, Malaysia;
Email: mhdfarisx@gmail.com (M.F.B.H.), sucheng@mmu.edu.my (S.-C.H.), kwng@mmu.edu.my (K.-W.N.)
*Correspondence: p.naveen@mmu.edu.my (N.P.)

***Abstract*—In recent times, e-commerce has grown expeditiously. As a result, online shopping and online product reviews are increasing, which makes it nearly impossible for companies to analyze them. In addition, ratings with high star ratings are often ignored, which may contain dissatisfied reviews that should be taken into account. Therefore, techniques are required for companies to extract information from the reviews and ratings, which helps them to analyze the data and make accurate decisions. The objective of this paper is to compare supervised Machine Learning (ML) classification approaches on Amazon product reviews to determine which method offers the most reliable sentiment analysis results. The product reviews are pre-processed and the extracted sentiments are labelled as either positive or negative sentiments. The sentiments are analysed using Multinomial Naive Bayes (MNB), Random Forest (RF), Long-Short Term Memory (LSTM) and Convolutional Neural Network (CNN). The feature extraction techniques Term Frequency-Inverse Document Frequency Transformer (TF-IDF(T)) and TF-IDF Vectorizer (TF-IDF(V)) were used for ML models, MNB and RF. The performance of the models was evaluated using confusion matrix, Receiver Operating Characteristic (ROC), and Area under the Curve (AUC). The LSTM provided an accuracy of 97% and outperformed other models.**

***Keywords*—Amazon, sentiment analysis, product review, feature extraction, machine learning**

I. Introduction

E-commerce is where people buy and sell products and services or transfer funds and data on the Internet [1]. Amazon is one of the biggest e-commerce and most influential brands worldwide that provides a marketplace for buying and selling goods and services. Many reviews are publicly available because customers write comments and feedback on various products and services [2]. For products in high demand and with many reviews, it is nearly impossible and challenging for a company to read and analyse all the product reviews [3]. In general, customer reviews with four or five stars are considered good reviews, while reviews with one or two stars are not considered good reviews. Despite leaving a high rating such as a four or five-star rating and good reviews, some leave feedback or complaints all in one review section as an indication of dissatisfaction towards the products. The extraction of reviews serves as a means to obtain in-depth insights regarding the issues that customers encounter; thus, ultimately facilitating the enhancement of businesses' merchandise offerings. A multitude of scholarly works focus on sentiment analysis and the utilization of Machine Learning (ML) models with regards to customer reviews. Notably, these works often neglect negative commentary contained within reviews consisting of four or five-star ratings. This method will help companies identify their products' problems and either improve or provide the right products that users are looking for based on user reviews and ratings. This research paper aims to do sentiment analysis on Amazon product reviews using ML algorithms with a feature extraction technique and Deep Learning (DL) algorithms, a part of ML. Afterwards, the aforementioned models were evaluated using the confusion matrix, the Receiver Operating Characteristic (ROC), and the Area under the Curve (AUC). The goal of this evaluation was to scrutinize and determine which model exhibited optimal performance.

The paper's arrangement is as follows: Section II is about the literature review. Section III describes the proposed framework. Section IV is a summary of the results. Finally, section V is a conclusion.

II. Literature Review

This section contains two categories. The first category discusses the feature extraction techniques used for ML algorithms by other researchers. The second category is about the state-of-art ML and DL models.

A. *Feature Extraction*

Feature extraction is the process of extracting and generating features suitable for model building, increasing learning speed. The feature extraction techniques in [3] were Term frequency, TF-IDF, Global Vectors (GloVe) and word2vec. TF-IDF uses word counts as frequencies to determine the relevance of words to a given document. GloVe indicates the probability that two words will co-occur, and word2vec learns meaningful relationships and encodes them into vector similarities. The feature extraction techniques vectorise the "Ready Document" from pre-processing. Each

Manuscript received July 30, 2022; revised October 14, 2022; accepted January 3, 2023; published August 28, 2023.

doi: 10.12720/jait.14.4.857-862

technique's outcome creates a matrix representing all documents in the dataset as vectors, based on the built up of the classification models. AlQahtani [4] used extraction techniques such as Bag-of-Words, which represents the occurrence of words in the document, TF-IDF, and GloVe for Machine Learning (ML) and Deep Learning (DL) algorithms. Ahuja *et al.* [5] used feature extraction techniques such as TF-IDF and n-grams in the ML models. TF-IDF Vectorizer (TF-IDF(V)) was implemented in [6, 7] to calculate the word count frequencies in the document by weighting the number of words.

Alsubae *et al.* [8] used TF-IDF(T), a variation of TF-IDF, that measures how important a word is to a document in a corpus. The related work about feature extraction techniques shows that TF-IDF performed better than other feature extraction techniques for ML algorithms.

B. Supervised Machine Learning

ML is one of the applications of Artificial Intelligence (AI) where systems can learn and improve from experience without having to be explicitly programmed. DL, a subset of ML based on artificial neural networks, mimics human brain behaviour.

Naive Bayes (NB) is a classification method that assumes the presence of a particular feature is independent of other features. Aljuhani *et al.* [3] used NB on balanced and unbalanced datasets. They observed that using Bigrams of TF-IDF and NB achieved an accuracy of 85.82% when used on unbalanced datasets. In contrast, using Trigrams of TF-IDF with the NB algorithm achieved an accuracy of 74.90% when used on balanced datasets. In their research, Xiao *et al.* [9] used NB and Logistic Regression as Machine Learning (ML) models. The NB model achieved accuracies of 67.50%, 79.41%, and 85.07% on the shopping reviews dataset, Weibo reviews dataset, and a combination of online shopping and Weibo datasets, respectively. Multinomial Naive Bayes (MNB) and Support Vector Machine (SVM) were the models used in [10] to analyse the data gathered from the customer's feedback. MNB [11] is analyzed sentiments on Amazon reviews.

Random Forest (RF) builds an ensemble of decision trees, adding more randomness while growing the trees. AlQahtani [4] specifically implemented RF and NB, and utilized the bagging technique to train the machine learning model, which is a combination of several learning models that improve the overall outcome. Through this research, the RF algorithm achieved the best performance with GloVe, resulting in an accuracy of 90%. Karthika *et al.* [12] performed sentiment analysis on a dataset from Kaggle using SVM and RF. Aribowo *et al.* [13] and Zhang *et al.* [14] used RF with other ML models.

Convolutional Neural Network (CNN) comprises neurons with ascertainable weights and biases. Each neuron receives multiple inputs, and the weighted sum of those inputs is passed through an activation function to produce the output. Aljuhani *et al.* [3] mentioned that they used CNN algorithms. CNN combined with word2vec achieved an accuracy of 92.73% for unbalanced datasets. With balanced datasets, the combination of CNN and word2vec resulted in the best accuracy of 79.60%. Paredes-Valverde *et al.* [15] utilised CNN to classify tweets into positive and negative classes. The building of CNN specifies concatenated word vectors of the text to be used as the input. This approach has shown promising results with 88.85% precision, recall of 88.8% and F-measure of 88.7% for the positive class. Conversely, the negative class achieved a precision of 88.8%, with a recall of 88.4% and an F-measure of 88.6%.

Long-Short Term Memory (LSTM) controls how the information in a sequence of data enters, stores and leaves the network using a series of gates such as forget gate, input gate and output gate. As AlQahtani [4] posited, using LSTM algorithms in conjunction with fine-tuned GloVe embedding resulted in peak performance at a notable 93% accuracy level. Xiao *et al.* [9] chose to use LSTM as the algorithmic model in their research for sentiment analysis. The LSTM model was successful, with a recorded accuracy of 85.48% for the online shopping reviews dataset, 69.66% for the Weibo reviews dataset, and 89.85% for a fusion of both online shopping and Weibo datasets. Bodapati *et al.* [16] and Shamal *et al.* [17] used the LSTM model and achieved better results for sentiment analysis. According to Güner *et al.* [11], the LSTM model significantly outperformed other models for binary classification when the sentiment analysis outcome is binary.

As a conclusion, NB and RF performed better than other supervised machine learning algorithms for sentiment analysis. MNB is useful for sentiment analysis. On the other hand, LSTM and CNN are the best compared to other deep learning models for sentiment analysis.

III. Research Methodology

This section explains the steps involved in comparing the ML models based on the sentiment analysis of Amazon product reviews, as shown in Fig. 1.

A. Data Obtained

The dataset used in this project is the Amazon product reviews from Kaggle.com. The dataset contains over 34,000 reviews from customers on Amazon products such as electronic products, home furniture and other products. In addition, the dataset included customer reviews, product ratings, and much more. There are 21 features available in the dataset, including product information, the star rating of the products, customers' reviews and other features.

B. Data Pre-processing

The dataset needs pre-processing before feeding to the ML algorithms as it is a textual dataset and to achieve higher accuracy. The general flow of pre-processing, as shown in Fig. 2. Firstly, it is imperative that the dataset undergoes a rigorous process of data cleaning, wherein extraneous features and null values are removed. Subsequently, all reviews included within the dataset must be converted to lowercase. This should be followed by tokenization, whereby the sentences within the reviews are parsed into individual words, or tokens, based on spatial segmentation. Finally, in order to eliminate stop words that do not contribute significantly to the overall meaning of the reviews, such as “for,” “a,” “and,” and other similar terms, they must be systematically removed. After that, the punctuation marks in the reviews are removed. Punctuation in this context refers to full stops, exclamation marks, question marks, commas, and other marks. Lastly, the reviews will undergo a lemmatisation process. Lemmatization of a word is the process of returning words to their roots by eliminating prefixes and suffixes. Fig. 3 displays how the review before getting pre-processed and after pre-processing.

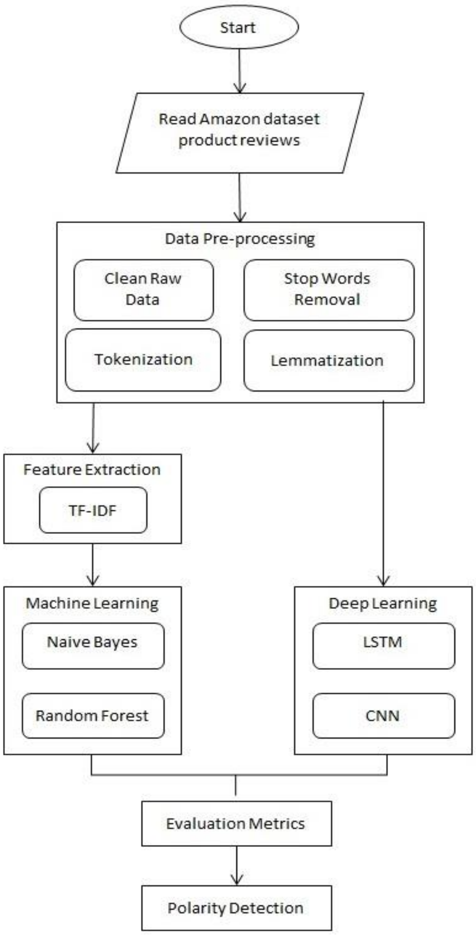

Figure 1. General flow of the experiment.

C. Sentiment Analysis

The reviews are from the Amazon product reviews dataset. Next, extract four or five-star rating reviews, as they are “positive” due to the high rating number. After extraction, search for negative words in the list of all words from the reviews. Negative words in four or five-star rating reviews are of “Negative” sentiment; otherwise, as “Positive” sentiment. As shown in Fig. 4, 91.58% of reviews were “positive” while another 8.42% were “Negative”.

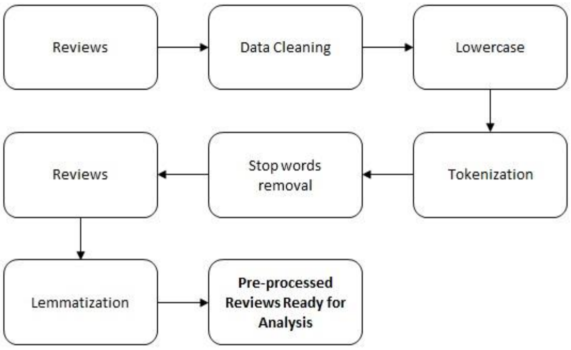

Figure 2. Pre-processing.

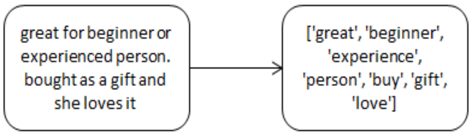
great for beginner or experienced person. bought as a gift and she loves it
['great', 'beginner', 'experience', 'person', 'buy', 'gift', 'love']

Figure 3. Before and after pre-processing.

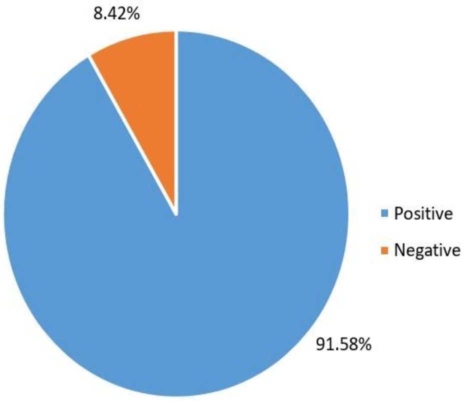

Figure 4. Percentage of positive and negative sentiment analysis.

D. Feature Extraction

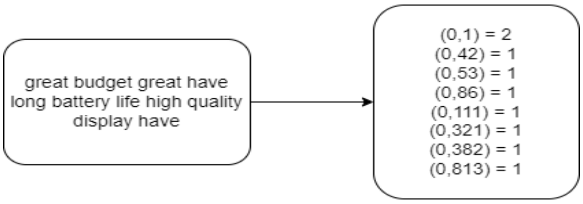

Figure 5 Before and after using TF-IDF (T).

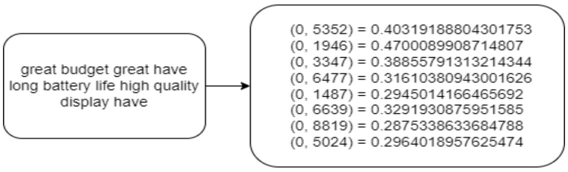

Figure 6 Before and after using TF-IDF (V).

In this project, the feature extraction technique used for the ML models are TF-IDF (T) and TF-IDF (V). The input for TF-IDF was the pre-processed reviews, and the output is a word index with TF-IDF values of a word having an index. Figs. 5 and 6 show the output for TF-IDF (T) and TF-IDF (V) after putting a cleaned review into it. However, feature extraction for DL is not required as it extracts features by the network while training.

E. *Classification Models*

The present study examines four classification models, namely MNB, RF, LSTM, and CNN. Following the labelling of reviews, the dataset is partitioned into two sets: training and testing. The training set accounts for 80% of the data, and the testing set accounts for 20%. This partitioning is applied across all four classification models.

1) Multinomial Naïve Bayes: The MNB algorithm uses Bayesian learning to guess the tags of a text using Bayes' theorem. MNB calculates the probability of each independent tag for a given sample. The output of this algorithm is the tag with the highest probability. The MNB used a default set of parameters. First, the value of the alpha parameter is 1.0. Next, fit_prior is True, and class_prior is None, which is a default value.

2) Random Forest: For the parameter of RF, the number of trees is 100, which is a default number since a higher number of trees increases the algorithm's performance. The max depth is 2. The criterion used was "entropy" instead of the default criterion "gini". Aznar [18] mentioned that "entropy" criterion results are better than "gini" even though it is less computationally expensive.

3) Long-Short Term Memory: The LSTM model for this project will have an embedding layer which creates the input layer. The input used is the total number of words collected from each review. The embedding dimension value is 32, and the input length is 17, which is the maximum length of the longest review in the dataset. The number of neurons for the LSTM layer is 64, with a dropout of 0.2. The expected output for this model is one because it is a binary classification, and the activation function is sigmoid.

4) Convolutional Neural Network: CNN's embedding layer is similar to the LSTM model, where the length of all words from the reviews is the input. The embedding dimension is 32, with a dropout layer of 0.2. The CNN layer has 16 neurons, a kernel size of 3 × 3, and a sigmoid as its activation function. Next, adding a pooling layer reduced the feature maps' dimensions. In addition, the CNN layer also includes eight neurons with sigmoid activation functions and a kernel size of 3 × 3. It is then necessary to add another layer, known as the flatten layer, to alter the shape of the data. Finally, to classify the output from the convolutional layers, a dropout layer with a rate of 0.2 and a dense layer with the Sigmoid activation function were used.

F. *Model Performance Evaluation*

After implementing the models, the confusion matrix is used to analyze the model's performance. The confusion matrix includes the following components. First, True Positive (TP), in which actual and predicted values are positive. Second, the predicted values are negative, whereas the actual values are positive in True Negative (TN) and vice versa in the component False Positive (FP). Lastly, False Negative (FN), in which both actual and predicted values are negative. These components are essential to calculate the accuracy, precision, recall and F1-score. Table I shows the formulas to calculate confusion metrics [19].

Next, the Area under the Curve (AUC) [20] is calculated through a thorough analysis of the True Positive (TP) and True Negative (TN) rates. AUC indicates the level of distinction between negative and positive classes. In conclusion, the models' performance is effectively demonstrated by plotting the Receiver Operating Characteristic (ROC) curve, which exhibits the false positive rate (FPR) on the X-axis and the True Positive Rate (TPR) on the Y-axis.

G. *Polarity Detection*

After evaluating each model and identifying the model with the best performance, that model is ideal for detecting the polarity of a new review. The review must be pre-processed first before feeding it into the model. The threshold for the review to be considered as "positive" is more than 0.5; else, it would be "negative".

IV. Results And Discussions

This section discussed the results achieved by the models MNB, RF, LSTM and CNN from the input of cleaned Amazon product reviews dataset into these models. The models were of python version 3.7 using Google Colab. First, section A discusses the performance of the MNB and RF models using features with TF-IDF (T) and TF-IDF (V). Next, sections B and C present the DL models' results and the overall findings.

A. *Results of MNB and RF with TF-IDF*

The MNB and RF models used TF-IDF (T) and TF-IDF (V) as feature extraction techniques. Table II displays the evaluation results of MNB and RF with different methods of using TF-IDF (T and V). MNB with TF-IDF (T) achieved better with an accuracy of 96%, 98% precision, a recall of 0.98, and an F1-score of 0.9. In contrast, the results of RF with Transformer and Vectorizer were identical. It achieved an accuracy of 91%, precision of 91%, recall of 1.00, and F1-score of 0.95.

B. *Results of LSTM and CNN*

Table III presents the evaluation results of LSTM and CNN. The LSTM model obtained 97% accuracy, 97% precision, a recall of 0.99, and an F1-score of 0.98. For the CNN model, it achieved the evaluation results of an accuracy of 95%, precision of 96%, 0.99 of recall, and F1-score of 0.97. It is evident that LSTM, the most preferred model in the field of DL, outperformed CNN.

TABLE I. Formula for Confusion Metrics [19]

Metrics	**Formulas**
Accuracy (ACC)	TP + TN / TP + TN + FP + FN
Precision (PR)	TP / TP + FP
Recall (RC)	TP / TP + FN
F1-score (F1)	2(TP) / 2(TP + FP + FN)

TABLE II. THE RESULTS OF MNB AND RF

Models	ACC	PR	RC	F1
TF-IDF(T) & MNB	0.96	0.98	0.98	0.98
TF-IDF(V) & MNB	0.92	0.92	1.00	0.96
TF-IDF(T) & RF	0.91	0.91	1.00	0.95
TF-IDF(V) & RF	0.91	0.91	1.00	0.95

TABLE III. THE RESULTS OF LSTM AND CNN

Models	ACC	PR	RC	F1
LSTM	0.97	0.97	0.99	0.98
CNN	0.95	0.96	0.99	0.97

C. *Overall Findings*

Fig. 7 compares the evaluation results of all the ML models with TF-IDF (T) and DL models. MNB with TF-IDF (T) of ML outperformed RF using TF-IDF (T) with an accuracy of 96%, whereas in the DL models, LSTM is the best model with an accuracy of 97% compared to CNN. Additionally, using the ROC curve and AUC value, the performance of each of these models can be determined with certainty. AUC shows how well the positive and negative classes are differentiated by considering the TP and TN rates. The ROC curve is a graph which shows the performance of the models. Fig. 8 shows the ROC curves and AUC for the models. Each model yields a result that is close to one, indicating a lower FPR, higher TPR and reasonable threshold. The LSTM model outperforms the MNB with the TF-IDF (T) model, indicating that the LSTM model is the best performer.

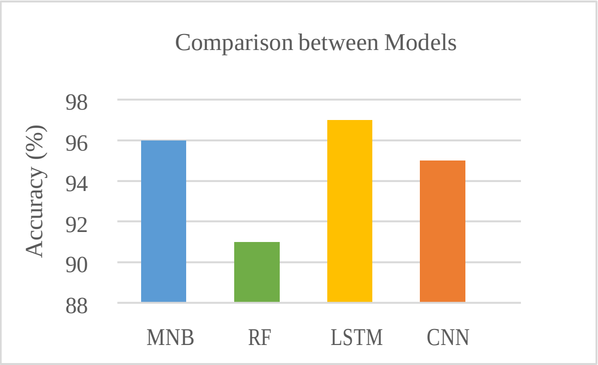

Figure 7. Results of the models.

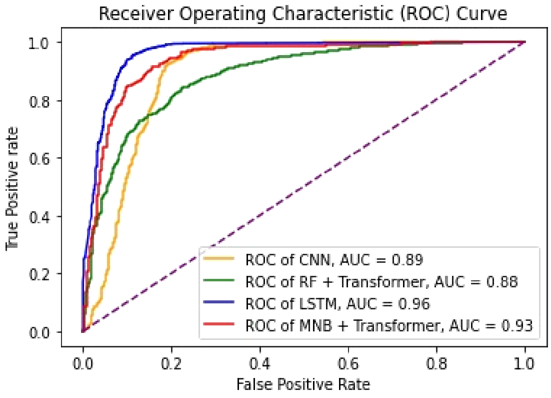

Figure 8. ROC curves for the models.

D. *Discussions*

In this paper, different ML and DL algorithms were implemented to perform sentiment analysis on the Amazon dataset. A sentiment analysis was performed where 8.42% of the four or five-star rating reviews contained negative words. These reviews were input to the models to evaluate which model has the best performance. It was observable that the LSTM model provided the best accuracy of 97%, when compared with other MNB, RF and CNN models. The LSTM model was selected for the polarity detection. Suppose a new raw review with four or five-star ratings that contain negative words becomes the input for the polarity detection. In that case, the review will be grouped into “Negative” despite having a high rating. The high accuracy obtained from the LSTM model may benefit companies and organisations to perform sentiment analysis on their product reviews where the result is highly reliable. Companies and businesses may utilise this model with high confidence to understand their customer better.

V. CONCLUSION

This paper implemented several models, such as MNB, RF, LSTM, and CNN, to perform sentiment analysis. The feature extraction techniques used were TF-IDF (T) and TF-IDF (V) for ML algorithms. They were evaluated based on confusion metrics, AUC and ROC to know the best performance model. The best model in this experiment is the LSTM model, with an accuracy of 97%. BERT DL model may be utilised for future work with another feature extraction, such as GloVe and word2vec, to see how it improves the accuracy of the models.

CONFLICT OF INTEREST

The authors declare no conflict of interest

AUTHOR CONTRIBUTIONS

The presented concept was developed by Faris, Naveen, Haw, and Ng. Faris conducted the experiment and wrote the paper. The project was supervised and received constructive feedback from Naveen, Haw, and Ng. All authors had approved the final version.

REFERENCES

[1] W. L. Koe and N. A. Sakir, “The motivation to adopt e-commerce among Malaysian entrepreneurs,” *Organisations and Markets in Emerging Economies*, vol. 11, no. 1, pp. 189–202, 2020.

[2] J. P. Singh, S. Irani, N. P. Rana, *et al.*, “Predicting the “helpfulness” of online consumer reviews,” *Journal of Business Research*, vol. 70, pp. 346–355, 2017.

[3] S. A. Aljuhani and N. S. Alghamdi, “A comparison of sentiment analysis methods on Amazon reviews of mobile phones,” *International Journal of Advanced Computer Science and Applications*, vol. 10, pp. 608–617, 2019.

[4] A. S. AlQahtani, “Product sentiment analysis for Amazon reviews,” *International Journal of Computer Science & Information Technology (IJCSIT)*, vol. 13, pp. 15–30, 2021.

[5] R. Ahuja, A. Chug, S. Kohli, *et al.*, “The impact of features extraction on the sentiment analysis,” *Procedia Computer Science*, pp. 341–348, 2019.

[6] A. A. Wadhe and S. S. Suratkar, "Tourist place reviews sentiment classification using machine learning techniques," in *Proc. the 2020 International Conf. on Industry 4.0 Technology (I4Tech)*, Pune, 2020, pp. 1–6.

[7] U. Parida, M. Nayak, and A. K. Nayak, "News text categorization using Random Forest and Naïve Bayes," in *Proc. the 1st Odisha International Conf. on Electrical Power Engineering, Communication and Computing Technology (ODICON)*, Odisha, 2021, pp. 1–4.

[8] S. M. Alsubaie, K. M. Almutairi, N. A. Alnuaim, *et al.*, "Automatic semantic sentiment analysis on twitter tweets using machine learning: A comparative study," *Journal of Theoretical and Applied Information Technology*, vol. 97, no. 23, pp. 3497–3508, 2019.

[9] S. Xiao, H. Wang, Z. Ling, L. Wang, and Z. Tang, "Sentiment analysis for product reviews based on deep learning," in *Proc. 2020 the Second International Conf. on Artificial Intelligence Technologies and Application (ICAITA)*, Dalian, 2020, 012103.

[10] M. S. Elli, Y.-F. Wang, *et al.* (2015). Amazon reviews business analytics with sentiment analysis. [Online]. Available: https://docplayer.net/151407565-Amazon-reviews-business-analytics-with-sentiment-analysis.html

[11] L. Güner, E. Coyne, and J. Smit, "Sentiment analysis for amazon.com reviews," *Big Data in Media Technology (DM2583) KTH Royal Institute of Technology*, 9, 2019.

[12] R. Khan, F. Rustam, K. Kanwal, *et al.*, "US based COVID-19 tweets sentiment analysis using TextBlob and supervised machine learning algorithms," in *Proc. 2021 International Conference on Artificial Intelligence (ICAI)*, Islamabad, 2021, pp. 1–8.

[13] A. S. Aribowo, H. Basiron, N. S. Herman, *et al.*, "An evaluation of pre-processing steps and tree-based ensemble machine learning for analysing sentiment on Indonesian YouTube comments," *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 9, pp. 7078–7086, 2020.

[14] X. Zhang, H. Saleh, E. M. G. Younis, *et al.*, "Predicting coronavirus pandemic in real-time using machine learning and big data streaming system," *Complexity*, vol. 2020, pp. 1–10, 2020.

[15] M. A. Paredes-Valverde, R. Colomo-Palacios, M. D. P. Salas-Zárate, *et al.*, "Sentiment analysis in Spanish for improvement of products and services: A deep learning approach," *Scientific Programming*, vol. 2017, pp. 1–6, 2017.

[16] J. D. Bodapati, N. Veeranjaneyulu, and S. Shaik, "Sentiment analysis from movie reviews using LSTMs," *Ingénierie des Systèmes d Inf.*, vol. 24, pp. 125–129, 2019.

[17] A. J. Shamal, R. G. H. Pemathilake, S. P. Karunathilake, *et al.*, "Sentiment analysis using Token2Vec and LSTMs: User review analyzing module," in *Proc. 2018 18th International Conference on Advances in ICT for Emerging Regions (ICTer)*, 2018, pp. 48–53.

[18] P. Aznar. (2020). Decision trees: Gini vs entropy. [Online]. Available: https://quantdare.com/decision-trees-gini-vs-entropy/

[19] B. Gaye and A. Wulamu, "Sentiment analysis of text classification algorithms using confusion matrix," in *Proc. International 2019 Cyberspace Congress, CyberDI and CyberLife*, Beijing, 2019, pp. 231–241.

[20] M. S. Satu, M. I. Khan, M. Mahmud, *et al.*, "TClustVID: A novel machine learning classification model to investigate topics and sentiment in COVID-19 tweets," *Knowledge-Based Systems*, vol. 226, 107126, 2021.