Fuzzy Systems and Soft Computing
ISSN : 1819-4362

SENTIMENT ANALYSIS USING MACHINE LEARNING ALGORITHMS FOR MUSIC INSTRUMENTS REVIEWS DATASET

K. Jayabharathi Assistant Professor Department of MCA Gurunanak College, Velachery, Chennai, India, jayabharathikannan4@gmail.com;
P.C. Sridevi Research Scholar PG and Research Department of Computer Science, Dwaraka Doss Goverdhan Doss Vaishnav College, Arumbakkam, Chennai, India sridevipc@gmail.com;
T. Velmurugan Associate Professor PG and Research Department of Computer Science, Dwaraka Doss Goverdhan Doss Vaishnav College, Arumbakkam, Chennai, India
velmurugan_dgvc@yahoo.co.in

Abstract:
In every consumer-facing company, ratings and comments have great significance. The majority of the data used for this is generated by reviews and feedback from different social networking sites. This research examines many machine learning methods on data set to comprehend consumer's perceptions about an item. A comparison analysis is conducted to discuss the effectiveness of these algorithms on various text categorization datasets. The goal of this research is to use sentiment analysis to determine which classifier best fits consumer perception. Therefore, to achieve this goal, datasets are first subjected to text preprocessing techniques, and then the processed data is subjected to feature extraction techniques. After that, machine learning methods are used for classification and clustering, respectively. Additionally, these methods are assessed, and the findings show that, for the music instruments review dataset, machine learning algorithms—particularly Random Forest (RF) perform better than Support Vector Machine (SVM) and Decision Tree (DT) algorithms.

Keywords:
Data mining; Music instrument review dataset; Random Forest Algorithm; Support Vector Machine Algorithm; Decision Tree Algorithm

I. INTRODUCTON

Sentiment analysis, also called opinion mining or emotion AI, is the systematic identification, extraction, quantification, and evaluation of affective states and subjective data using natural language processing, text analysis, computational linguistics, as well as biometrics. In marketing, customer service, and clinical medical applications, sentiment analysis is widely used to voice of the customer materials such as reviews and survey replies, internet and social media, and healthcare materials. Sentiment analysis is the analysis of the positive or negative attitude of the customer in text [1]. Businesses may track online chats to use contextual analysis of identifiable data to better understand the social attitude of their customers [2].

Customers are sharing their ideas and thoughts about the company more freely than ever before, and sentiment analysis has grown into a powerful tool for monitoring and understanding online conversations. By automatically analyzing survey and social media feedback and reviews, you may learn what makes a consumer pleased or dissatisfied [3]. We may also use this information to enhance your brand by tailoring your products and services to your customers' demands. The data utilized in sentiment analysis on product reviews comes from customer comments or reviews on a specific product. It may be used to monitor customer perceptions of the product and determine if they are generally positive or negative [4]. This kind of information may be used to make decisions about the product, such as whether it should be advertised or withdrawn. It may also help organizations determine which areas or facets of the product want improvement. Customizing the consumer experience can also benefit from the information found in customer evaluations, since many of them provide specific details about the customers' experiences [5].

II. LITERATURE REVIEW

A. Krouska et al. conducted research and released a paper on the use of text preparation techniques to sentiment analysis using Twitter datasets [6]. The results of the experiments have shown

that the feature selection stage increases the sentiment analysis accuracy of each machine learning system. Z. Jianqiang [7] suggested utilizing five Twitter datasets to compare the results of different text preparation methods. Enhancing the classification accuracy performance of each dataset is the aim of this research. G. Angiani et al. [8] examined the several text processing techniques that influence accuracy performance. While it takes a long time, the data cleaning procedure is very important and demands more attention than previous processes.

E. Haddi et al. presented the experimental results of the text preparing phase to predict sentiment data from online movie reviews using SVM [9]. The data transformation and filtering processes were integrated in terms of performance and accuracy. According to D. Torunolu et al.'s [10] investigation, there are three text processing procedures that affect the accuracy of Turkish text analysis: stemming, stop word filtering, and word weighting. Text preparation techniques were used in M. Mhatre et al.'s [11] examination of the "bag of Words Meets Nags of popcorn" dataset. Combining three text processing strategies—managing slang, eliminating stop words, and lemmatization—resulted in the greatest accuracy.

The LeBERT sentiment classification approach, presented by Mutinda et al. in [12], combines CNN, BERT, N-grams, and a sentiment lexicon to overcome these drawbacks. Words chosen from a portion of the input text are vectorized by the model using sentiment lexicon, N-grams, and BERT. After mapping features, CNN, a deep neural network classifier, provides an output sentiment category. Three publicly available datasets are used to evaluate the proposed model: Yelp restaurant reviews, Amazon retail reviews, and Imbd movie reviews. S. Vijayarani et al. [13] provided a detailed description of text mining and text processing techniques for locating knowledge in text material that users of social media platforms publish. They discussed the TF/IDF method, stemming, and stop word removal as part of the core text preparation step. Each group's stemming algorithms were also provided, along with the advantages and disadvantages of each stage. This paper illustrated every stage of text preparation, which was essential for doing sentiment analysis or text mining.

III. DESCRIPTION OF DATA SET

The present research examines preprocessing methods for sentiment analysis using Python modules, utilizing the Musical Instruments Review dataset. The input data is provided as a.csv file. The dataset consists of 9 columns and 10261 rows. reviewerName, reviewerID, asin, and helpful reviewText, general, synopsis, and unixReviewTime, reviewTime, and dtype: int64 are the input dataset's properties. Figure 1 displays a sample of the dataset for reviews of musical instruments before preprocessing.

	reviewerID	asin	reviewerName	helpful	reviewText	overall	summary	unixReviewTime	reviewTime
0	A2IBPI20UZIR0U	1.4E+09	cassandra tu "Yeah, well, that's just like, u...	[0, 0]	Not much to write about here, but it does exac...	5	good	1.4E+09	02 28, 2014
1	A14VAT5EAX3D9S	1.4E+09	Jake	[13, 14]	The product does exactly as it should and is q...	5	Jake	1.4E+09	03 16, 2013

2	A195EZSQDW3E21	1.4E+09	Rick Bennette "Rick Bennette"	[1, 1]	The primary job of this device is to block the...	5	It Does The Job Well	1.4E+09	08 28, 2013
3	A2C00NNG1ZQQG2	1.4E+09	RustyBill "Sunday Rocker"	[0, 0]	Nice windscreen protects my MXL mic and preven...	5	GOOD WINDSCREEN FOR THE MONEY	1.4E+09	02 14, 2014

Figure 1: Sample of Musical instruments reviews dataset

IV. METHODOLOGY

This section discusses the problem definition for this research work. Organizing the many forms of unorganized information in a customer review is the main challenge for data mining activities. It is necessary to comprehend the patterns and important phrases in the customer review to analysis the sentiment [14]. The dataset is preprocessed to remove redundant data, missing data, and unnecessary features. The cleaned music instruments reviews dataset is then used in the sentiment analysis process to determine whether or not the review impacted individuals would be identified using Random Forest, Support Vector Machine and Decision Tree Algorithms. The.csv file format for the music instrument dataset can be obtained from the repository. The dataset is named musicinstrumentsreviews.csv in the file.

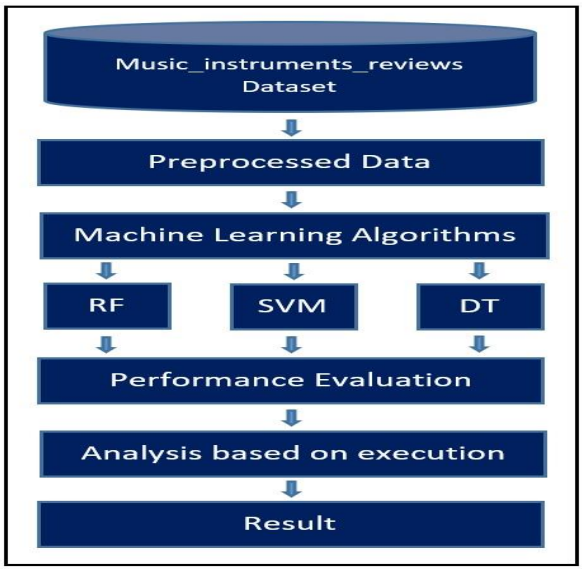

Figure 2: Research Methodology

According to the sentiment among the various input data forms, three distinct machine learning algorithms Random Forest, Support Vector Machine, and Decision Tree Algorithms are used to do the research. Based on how the data points are separated from one another, each emotion has been categorized and arranged [15]. A range of colors are used to display each categorized data point, and the execution time is expressed in seconds. Figure 2 shows the overall methodology of this research work.

A. Performance Metrices

The dataset's performance is shown by performance metrics. The presentation of the suggested system was assessed using the following criteria: F-measure, Precision, and Recall. Conventional count values, such as True Positive (Tp), True Negative (Tn), False Positive (Fp), and False Negative (Fn), are utilized here.

$$Accuracy = TP + TN \ / \ TP + TN + FP + FN \quad (4)$$

$$Sensitivity = TP \ / \ TP + FP \quad (5)$$

$$Specificity = TN \ / \ TN + FP \quad (6)$$

These metrics are carefully utilized to evaluate the algorithms' performance when compared to the analysis's evaluation of the data that was selected set.

B. Preprocessing

Data preparation is the process of converting raw data into a readable format. It is also a vital step in data mining because we cannot work with raw data. Prior to applying deep learning or data mining techniques, the data's dependability should be evaluated [16]. Duplicate entries, missing data, & noisy data in the same format will be removed from the database during preparation. Data transformation, data reduction, data integration, and data purification are the four main tasks of data preprocessing.

C. Machine Learning Algorithms

Supervised learning, unsupervised learning and Reinforcement learning are the three basic categories into which machine learning algorithms may be generally divided. Labeled training data—that is, input-output pairs—is what supervised learning algorithms need to run. Learning from the training data and predicting outputs for fresh, unknown data is the main objective of these algorithms [17] [18]. Frequently utilized supervised learning methods comprise Logistic Regression for binary classification issues, Support Vector Machines (SVM) for both classification and regression applications, and Linear Regression for continuous value prediction[19]. A few more noteworthy algorithms in this area include neural networks, especially the deep learning varieties such as Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs), as well as Random Forests, Decision Trees, and k-Nearest Neighbors (k-NN).

C.1 Random Forest

There are several steps involved in using the random forest method for sentiment analysis on a dataset of reviews of musical instruments. Gathering a dataset with reviews and sentiment labels (positive, negative, or neutral) is the first step [20]. You may have to manually label a portion of the dataset if sentiment labels are absent, or you may need to construct sentiment labels using a trained model. Subsequently, preprocess the text data by lemmatizing or stemming, tokenizing, deleting stop words, and cleaning. Transform the text into numerical representations by applying word embeddings or TF-IDF algorithms. Divide the preprocessed data into training and testing sets using an 80-20 or 70-30 ratio, if applicable [21].

Utilizing the training data, train the random forest classifier. The RandomForest Classifier from the sklearn.ensemble package in Python may be used for this. Utilizing TfidfVectorizer, vectorize the text data, fit the classifier to the training set of data, and then forecast the test set's attitudes [22]. Utilizing measures like accuracy, sensitivity and specificity all of which can be acquired via scikit-learn's classification_report evaluate the model's performance. Utilize the model to foresee the sentiment of fresh reviews after it has been trained and evaluated. Because of its robustness and capacity to manage intricate data and connections, the random forest algorithm is a great option in sentiment analysis in music instrument reviews, yielding insightful user comments [23].

C.2 Support Vector Machine

A structured method is necessary to achieve reliable sentiment classification when using the support vector machine (SVM) algorithm for sentiment analysis on a dataset of reviews of musical instruments. First, compile a dataset of reviews of musical instruments together with sentiment labels (neutral, negative, or positive). Use a pre-trained model or manually label a subset if labels are absent. Preprocess the text data by tokenizing, eliminating stop words, cleaning (removing HTML elements, special characters, and punctuation), and lemmatizing or stemming to guarantee consistency[24]. Use methods like TF-IDF to convert the written material into numerical representations that accurately

reflect the context and meaning of each word. utilizing an 80-20 or 70-30 ratio, divide the dataset into training and testing sets, utilizing the larger fraction for training [25].

Utilizing the SVC class from scikit-learn and vectorizing the text input with TF-IDF, train the SVM classifier. Using accuracy, precision, recall, and F1-score metrics gleaned from scikit-learn's classification_report, assess the trained model on the test set. Use grid search or random search techniques to fine-tune hyperparameters like the kernel type (linear, polynomial, radial basis function), regularization parameter (C), and gamma parameter (for non-linear kernels) in order to increase performance [26]. After the model is refined, use it to forecast the tone of fresh reviews. The SVM technique is a strong option for sentiment analysis in music instrument reviews because of its capacity to identify the best hyperplane for class separation, yielding insightful user comments [27].

C.3 Decision Tree

In order to achieve correct sentiment classification, there are many crucial measures to follow when applying the decision tree algorithm for sentiment analysis on a dataset of reviews of musical instruments. First, gather a dataset comprising reviews of musical instruments and sentiment labels (neutral, negative, or positive). You might have to apply a sentiment analysis model that has already been trained or label a subset by hand if sentiment labels are not accessible [28]. Tokenize the text data into individual words, remove stop words, clean the text data (removing HTML elements, special characters, and punctuation), and conduct stemming or lemmatization to check consistency. Utilize methods such as TF-IDF to convert the cleaned text into numerical representations that accurately reflect the significance and context of each word. Using an 80-20 or 70-30 ratio, divide the dataset into training and testing sets, with the bigger set being used to train the model [29].

Utilizing the DecisionTreeClassifier from scikit-learn and vectorizing the text input with TF-IDF, train the decision tree classifier [30]. Utilizing measures like accuracy, precision, recall, and F1-score—all of which can be found in scikit-learn's classification_report—evaluate the trained model on the test set. Optimize performance by adjusting hyperparameters such as the tree's maximum depth, the number of samples needed to split a node, and the criterion (such as entropy or Gini impurity) used to determine the quality of a split [31] [32]. After the model is refined, use it to forecast the tone of fresh reviews. Sentiment analysis of music instrument evaluations can benefit from the interpretability and robustness of the decision tree technique, which allows for the extraction of insightful user input.

V. RESULTS AND DISCUSSIONS

This research uses machine learning methods, to analyze raw data in order to identify emotive product reviews for musical instruments and to assess the effectiveness of these algorithms, and the outcomes are tracked and compared. Analysis has been done in a few prediction parameters. The following outcomes are the product of the procedure.

Table 1: Comparative Performance of all algorithms

Algorithms	Accuracy	Sensitivity	Specificity
Random Forest	81.85	72.81	77.06
Support Vector Machine	75.16	66.66	70.91
Decision Tree	71.68	61.36	66.52

Table 3 shows the performance analysis of Random Forest, Support Vector Machine and Decision Tree algorithms. The three approaches are evaluated for effectiveness using the following metrics: f-measure, recall, accuracy, and precision.

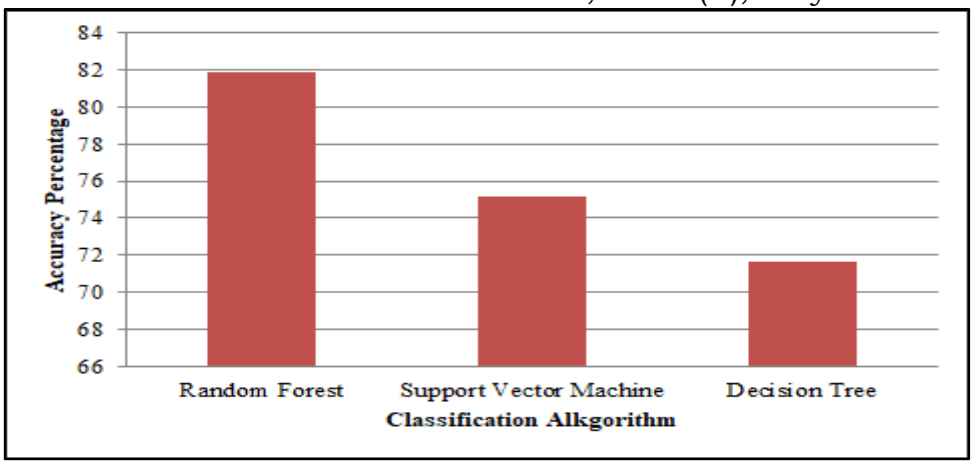

Figure 3: Accuracy of all algorithms

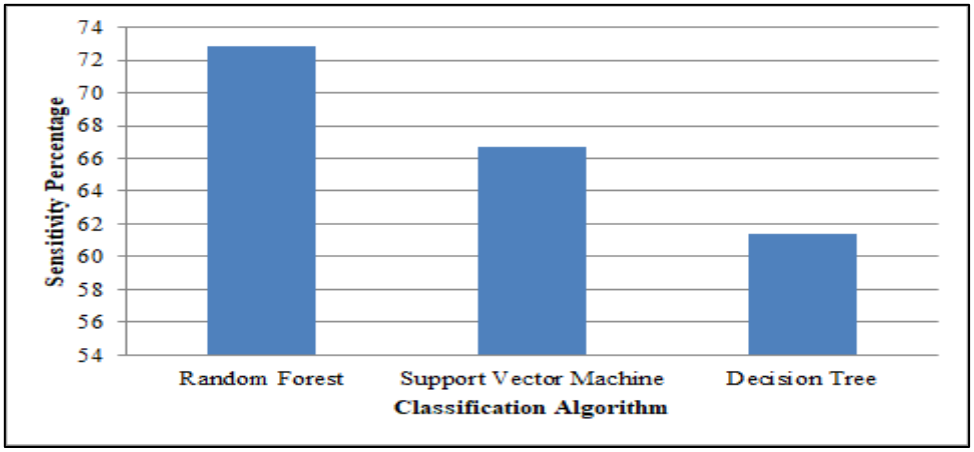

Figure 4: Sensitivity of all algorithms

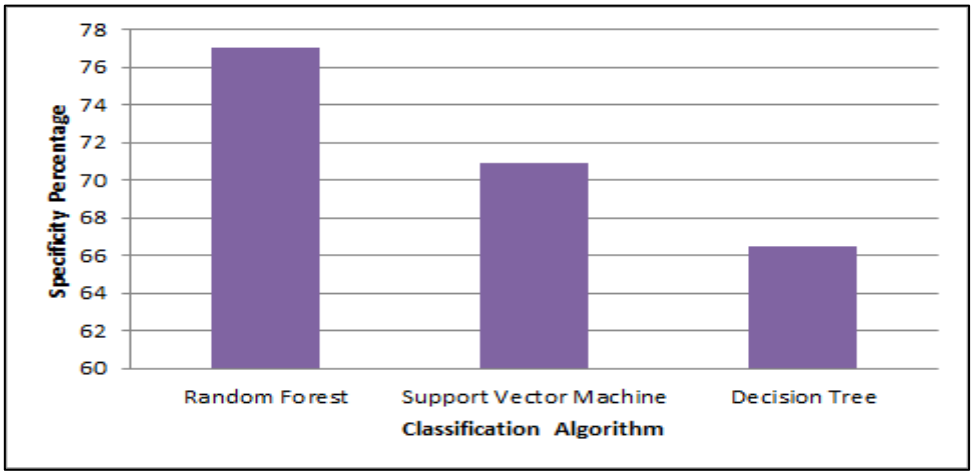

Figure 5: Specificity of all algorithms

Figure 3,4, and 5 depicts the performance analysis of all three methods, with the Random Forest algorithm achieving the precision value of 81.85%, recall value of 72.81%, and F-measure value of 77.06%. The Support Vector Machine algorithm achieves 75.16% precision, 66.66% recall, and 70.91% f-measure value. The Decision Tree Algorithm achieves a precision value of 71.68%, a recall value of 61.36%, and an F-measure value of 66.52%. Based on the results, it is evident that the Random Forest algorithm outperforms the other two existing approaches.

Table 2: Average Computational time and Memory utilization of algorithms

Algorithms	Execution time (sec)	Memory utilization (kb)
Random Forest	45.41	15.69
Support Vector Machine	54.03	22.42
Decision Tree	66.85	29.67

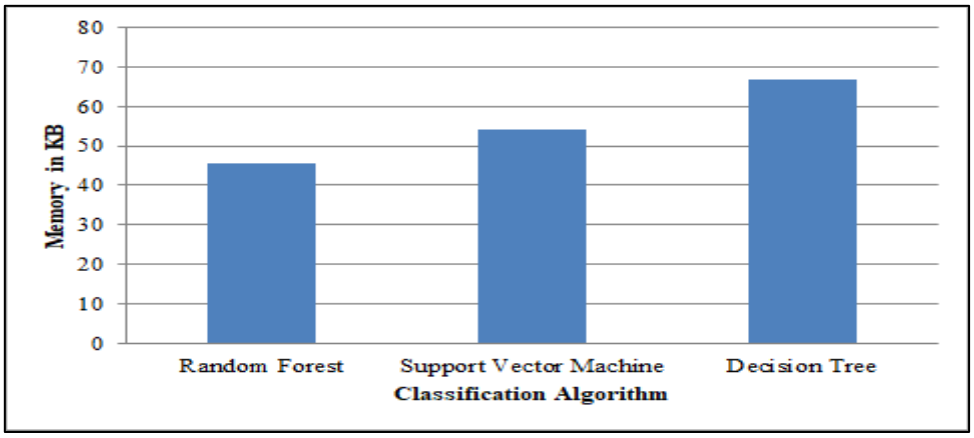

Figure 6: Time Complexity of algorithms

Table 2 displays the execution time and memory usage for Random Forest, Support Vector Machine and Decision Tree algorithms. Figure 6 shows the graphical representation of the execution time taken by the resultant dataset of the three classification algorithms. Figure 7 shows the graphical representation of the memory space occupied by the resultant dataset of the three classification algorithms.

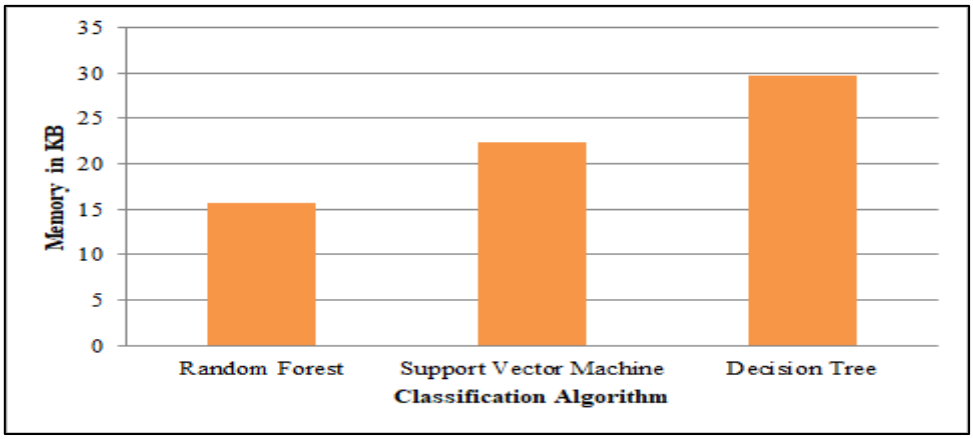

Figure 7: Space Complexity of algorithms

Figure 6 shows that the time to compute the Random Forest algorithm is very less when compared to Support Vector Machine and Decision Tree algorithms. Figure 7 shows that the memory space occupied by Random Forest algorithm is also relatively less when compared to Support Vector Machine and Decision Tree algorithms for the selected dataset.

VI. CONCLUSION

The research compares the results range along with various analyses based on emotions from the dataset to examine the performance of the algorithms. Sentiment analysis is a rapidly developing field in text mining and computational linguistics; writings are categorized using this method based on the feelings they convey. This article covers sample methodologies for the three main components of a typical sentiment analysis model: data preparation, review analysis, and sentiment categorization. Future research should look into novel classification models that can account for the ordered declares property in rating inference, as well as advanced techniques for collecting viewpoint and product attributes. Applications that capitalize on the findings of sentiment analysis will also likely based on surface. According to our findings, the Random Forest method performed better than the Support Vector Machine and Decision Tree algorithms in terms of accuracy, recall, and f-measure. It also required less time and space.

REFERENCES

[1] Lee, Dong-yub, Jae-Choon Jo, and Heui-Seok Lim, "User sentiment analysis on Amazon fashion product review using word embedding", *Journal of the Korea Convergence Society* 8, no. 4, pp.1-8, 2017.

[2] Shah, Arkesha, "Sentiment analysis of product reviews using supervised learning", *Reliability: Theory & Applications* 16, no. SI 1 (60), pp. 243-253, 2021.
[3] Salmony, Monir Yahya Ali, and Arman Rasool Faridi, "Supervised Sentiment Analysis on Amazon Product Reviews: A survey", IEEE, *2nd International Conference on Intelligent Engineering and Management (ICIEM)*, pp. 132-138, 2021.
[4] Hawlader, Mohibullah, Arjan Ghosh, Zaoyad Khan Raad, Wali Ahad Chowdhury, Md Sazzad Hossain Shehan, and Faisal Bin Ashraf ,"Amazon product reviews: Sentiment analysis using supervised learning algorithms", IEEE, *International Conference on Electronics, Communications and Information Technology (ICECIT)*, pp. 1-6, 2021.
[5] Haseeb, Abdul, Rabia Taseen, Marryam Sani, and Qamas Gul Khan, "Sentiment Analysis on Amazon Product Reviews using Text Analysis and Natural Language Processing Methods", *International Conference on Engineering, Natural and Social Sciences*, vol. 1, pp. 446-452, 2023.
[6] Krouska, C. Troussas and M. Virvou, "The effect of preprocessing techniques on Twitter sentiment analysis," Proc. 7th International Conference on Information, Intelligence, Systems & Applications (IISA), pp. 1-5, 2016, doi:10.1109/IISA.2016.7785373.
[7] Martis, Eesha, Rutuja Deo, Sejal Rastogi, Keshav Chhaparia, and Ameyaa Biwalkar, "A Proposed System for Understanding the Consumer Opinion of a Product Using Sentiment Analysis", *Sentiment Analysis and Deep Learning: Proceedings of ICSADL 2022*, pp. 555-568, 2023.
[8] Mutinda, James, Waweru Mwangi, and George Okeyo, "Sentiment analysis of text reviews using lexicon-enhanced bert embedding (LeBERT) model with convolutional neural network", *Applied Sciences* 13, no. 3, pp.1445, 2023.
[9] Solairaj, A., G. Sugitha, and G. Kavitha. "Enhanced Elman spike neural network based sentiment analysis of online product recommendation", *Applied Soft Computing 132*, pp. 109789, 2023.
[10] Z. Jianqiang, "Pre-processing Boosting Twitter Sentiment Analysis," Proc. IEEE International Conference on Smart City/SocialCom /SustainCom (SmartCity), pp. 748-753, 2015, doi: 10.1109/SmartCity.2015.158.
[11] G. Angiani and et al., "A Comparison between Preprocessing Techniques for Sentiment Analysis in Tewtter," Proc. 2nd International Workshop on Knowledge Discovery on the WEB (KDWEB 2016), Volume. 1748, 2016.
[12] Mutinda, James, Waweru Mwangi, and George Okeyo, "Sentiment analysis of text reviews using lexicon-enhanced bert embedding (LeBERT) model with convolutional neural network", Applied Sciences, Volume 13(3), pp.1445, 2023.
[13] S. Vijayarani, J Ilamathi and Nithya, "Preprocessing Techniques for Text Mining-An Overview," International Journal of Computer Science & Communication Networks, Volume 5, pp. 7-16, 2015.
[14] E. Haddi, X. Liu and Y. Shi, "The role of text pre-processing in sentiment analysis," Procedia Computer Science, Volume 17, pp. 26-32, 2013, doi:org/10.1016/ j.procs.2013. 05.005.
[15] D. Torunoğlu, E. Çakirman, M. C. Ganiz, S. Akyokuş and M. Z. Gürbüz, "Analysis of preprocessing methods on classification of Turkish texts," Proc. International Symposium on Innovations in Intelligent Systems and Applications, pp. 112-117, 2011, doi:10.1109/INISTA.2011.5946084.
[16] M. Mhatre, D. Phondekar, P. Kadam, A. Chawathe and K. Ghag, "Dimensionality reduction for sentiment analysis using preprocessing techniques," Proc. International Conference on Computing Methodologies and Communication (ICCMC), pp. 16-21, 2017, doi:10.1109/ ICCMC.2017.8282676.
[17] Diekson, Ziedhan Alifio, Muhammad Rivyan Bagas Prakoso, Muhammad Savio Qalby Putra, Muhammad Shaden Al Fadel Syaputra, Said Achmad, and Rhio Sutoyo, "Sentiment analysis for customer review: Case study of Traveloka", *Procedia Computer Science* 216, pp. 682-690, 2023.
[18] Li, Xianyong, Jiabo Zhang, Yajun Du, Jian Zhu, Yongquan Fan, and Xiaoliang Chen, "A novel deep learning-based sentiment analysis method enhanced with Emojis in microblog social networks", *Enterprise Information Systems* 17, no. 5, pp. 2037160, 2023.
[19] D'souza, Stephina Rodney, and Kavita Sonawane, "Sentiment analysis based on multiple reviews by using machine learning approaches", IEEE, *International Conference on Computing Methodologies and Communication (ICCMC)*, pp. 188-193, 2019.

[20] Pujari, Chetana, and Nisha P. Shetty, "Comparison of classification techniques for feature-oriented sentiment analysis of product review data", *Data Engineering and Intelligent Computing*, Springer, pp. 149-158, 2018.
[21] Fang, Xing, and Justin Zhan, "Sentiment analysis using product review data", *Journal of Big Data 2*, Vol. 1, pp. 1-14, 2015.
[22] Al Amrani, Yassine, Mohamed Lazaar, and Kamal Eddine El Kadiri, "Random forest and support vector machine based hybrid approach to sentiment analysis", *Procedia Computer Science*, Vol. 127, pp. 511-520, 2018.
[23] Başarslana, Muhammet Sinan, and Fatih Kayaalpb, "Sentiment Analysis with Machine Learning Methods on Social Media", Vol. 5, 2015.
[24] Wang, Hanshi, Lizhen Liu, Wei Song, and Jingli Lu, "Feature-based Sentiment Analysis Approach for Product Reviews", *Journal of Softw. 9*, Vol. 2, pp. 274-279, 2014.
[25] Latha.U, and T. Vemurugan, " Analyzing agricultural text data using classification algorithms.", *International conference on small and medium business*, pp. 339-344, 2016.
[26] Alsolamy, Afnan Atiah, Muazzam Ahmed Siddiqui, and Imtiaz Hussain Khan, "A Corpus Based Approach to Build Arabic Sentiment Lexicon", *International Journal of Information Engineering & Electronic Business*, Vol. 11, No. 6, 2019.
[27] Kumar, Akshi, and Teeja Mary Sebastian, "Sentiment analysis on twitter", *International Journal of Computer Science Issues (IJCSI)*, Vol. 9, No. 4, pp. 372, 2012.
[28] Darwich, Mohammad, Shahrul Azman Mohd, Nazlia Omar, and Nurul Aida Osman, "Corpus-Based Techniques for Sentiment Lexicon Generation: A Review", *Journal of Digital Information Management*, Vol.17, No. 5, pp. 296, 2019.
[29] Chathuranga, P. D. T., S. A. S. Lorensuhewa, and M. A. L. Kalyani, "Sinhala sentiment analysis using corpus-based sentiment lexicon", IEEE, *19th international conference on advances in ICT for emerging regions (ICTer)*, Vol. 250, pp. 1-7, 2019.
[30] Abdulla, Nawaf A., Nizar A. Ahmed, Mohammed A. Shehab, and Mahmoud Al-Ayyoub, "Arabic sentiment analysis: Lexicon-based and corpus-based", IEEE, *Jordan conference on applied electrical engineering and computing technologies (AEECT)*, pp. 1-6, 2013.
[31] Cordeiro, Cheryl Marie, "A corpus-based approach to understanding market access in fisheries and aquaculture: a systematic literature review", *International Journal of Economics and Management Engineering*, Vol. 13, no. 10, pp. 1324-1333, 2019.
[32] Singh, Jaspreet, Gurvinder Singh, and Rajinder Singh, "Optimization of sentiment analysis using machine learning classifiers", *Human-centric Computing and information Sciences*, Vol. 7, no.1, pp.1-12, 2017.