Understanding the Voice of Customers in Amazon Reviews: Comparison of Machine Learning models

Sunkeswaram Sreeja[1], Sri Vaishnavi Koduri[2], L. Chandra Sekhar Reddy[3], M. Parameswar[4]
[1,2]*Artificial Intelligence and Data Science, CMR College of Engineering and Technology, Hyderabad*
[3,4]*Computer Science and Engineering (Data Science), CMR College of Engineering and Technology, Hyderabad*

***Abstract*—Sentiment analysis, text analysis, and stemming are the core research aspects of contemporary NLP. They expand a researcher's toolkit in the form of additional approaches and tools to process unstructured data, generating objective insights. These methods help computers "understand" human speech, which is a beneficial skill due to people's incredible ability to be subjective while sharing their thoughts. The internet is filled with content reflecting various subjective approaches while various people share data considering their personal points of view. It is difficult for people to find necessary information and discover the truth, especially about products, because companies do not know their customers fully. While processing product reviews helps one understand the public's sentiment about a commodity, one should summarize all positive, neutral, and negative reviews due to their significant amounts. This work can be extended to cover more product review websites and look at more complex natural language processing features in the future. To ensure precision, the system retains only the words present in the dataset, filtering out any extraneous terms that do not contribute to the analysis.**

***Index Terms*—Natural Language Processing, Stemming, Term Frequency-Inverse Document Frequency, Bag-of-Words, feature-based Learning.**

I. INTRODUCTION

The use of Machine Learning (ML) has grown remarkably in recent years it has a strong impact on both research and industry. NLP (Natural Language Processing) is one of the main domains in ML to help computing aficionado understand human language (even slang/per-say). Data analysis is a technique of exploring structured, unstructured, or semi-structured data systematically to extract meaningful insights. One of the most useful applications of NLP is indefinite comprehension, also called opinion mining, which detects the view tone of textual material and classifies it as bad, good, or neutral. This step is especially helpful for companies, since it allows them to analyze customer comments and analyze public perception of their products or services [1]. Analyzing customer reviews from e-commerce platforms such as Amazon provides insights into consumer preferences, aiding companies in improving product quality and overall customer experience. With the help of sentiment analysis, companies can improve their marketing, product offerings, and customer satisfaction, which can therefore increase revenue and growth in stock value [2].

Since customers frequently leave detailed product reviews on quality, quantity, brand perception, and user experience, this research examines the sentiments in reviews of Amazon products, including electronics, jewelry, toys, car accessories, watches, gardening tools, personal care items, farm equipment, games, gourmet food, and healthcare products. However, manually analyzing such large volumes of data is impractical [3]. To handle this challenge, various machine learning techniques allow large datasets to be processed and classified efficiently. This study investigates the feasibility of determining whether product reviews are positive, negative, or neutral using various ML-based sentiment analysis techniques. This work using TF-IDF and BoW and classification techniques is dedicated to the advancement of automated sentiment analysis [4].

II. RELATED WORKS

The ways in which sentiment analysis techniques for product reviews have been investigated have become a subject of numerous studies with the adoption of varying machine learning methods and frameworks. A very good method of investigation on product reviews should be mentioned, performed by Pankaj et al. [5], concerning the smartphone product reviews classified as positive, negative, or neutral. The authors tackled an interesting problem in sentiment polarity categorization using Amazon product review data from August to December 2018. In particular, over 500 reviews were investigated spanning almost four product categories: mobiles, computers, flash drives, and electronics.

Sunny Kumar et al. explored how the R language and Rhadoop can be applied to sentiment analysis of social media data. They carried out a performance and architectural comparison of these frameworks which shows how tools for big data can improve the fabricity of sentiment classification techniques [6].

Zeenia Singla et al. conducted statistical sentiment analysis with a focus on consumer product reviews on mobile products. Their findings proved useful for both consumers and product designers, as their analysis revealed the top three most preferred brands—Apple, BLU, and Samsung—based on consumer sentiment trends [7].

Tanjim Ul Haque et al. have proposed a supervised learning model for sentiment classification, in which they used a combination of two feature extraction techniques. Using a collection of models, they checked out how feature extraction worked in the model. Sentiment analysis methods are used well beyond e-commerce [8].

Wankhade et al. and Tan et al. performed a similar analysis, contrasting several sentiment classification algorithms in respect to their performances using large-scale text data [9, 10].

These studies deal with the importance of feature selection and optimization of a model in the boost of accuracy in sentiment analysis. This work builds on previously conducted ones wherein different machine-learning models were used for sentiment analysis of Amazon product reviews to improve classification accuracy and scalability.

III. SENTIMENT ANALYSIS: TECHNIQUES AND BENEFITS

A basic component of sentiment analysis is the reliance on opinions, which opens more space for detailed analysis of opinions on products, services, or experiences. This will afford businesses insight into customer sentiments [11]. This process of sub-task involves the detection of polarity that classifies the opinions on the basis of positive, negative, or neutral. In this context, the term "opinion analysis" is often interchangeably referred to as sentiment analysis [12]. Sentiment analysis is in essence a text classification approach to decipher if a consumer review is aligned with a feeling favorable, unfavorable, or neutral toward the item. Such reviews play a major role in e-commerce related to product development, consumer satisfaction, and hybridization of approaches for engaging users [13]. Interpretation of various products in the market trends can allow the manufacturers and retailers and, in turn, the businesses that apply sentiment analysis to redesign their marketing strategies as per consumer expectations [14].

Sentiment analysis brings value to customers, helping them make informed choices when it comes to purchasing items, as well as to organizations, product manufacturers, and retailers. Companies can assess trends of sentiment to discern market phenomena and effect changes in service practices and user experience [15]. Sentiment analysis is applied in numerous fields which include brand reputation management, financial market prediction, and political discourse analysis [16].

A salting of various methodologies and techniques is applied to sentiment analysis with Natural Language Processing (NLP) and Machine Learning (ML) being widely researched. NLP techniques extract, process, and interpret data in the form of text, while within ML, models can classify and predict, which helps in increasing sentiment detection in accuracy and visibility [17]. The incorporation of deep learning algorithms-LSTM, GRU, and different variations based on the Transformer model-has provided further efficiency to sentiment classification in being robust and adaptable to systems with different data sources [18].

IV. SENTIMENT ANALYSIS STEPS ON PRODUCT REVIEW

This section presents the main steps of reviewing customer sentiment for classification. The process for sentiment analysis is a structured framework, which will be described below:

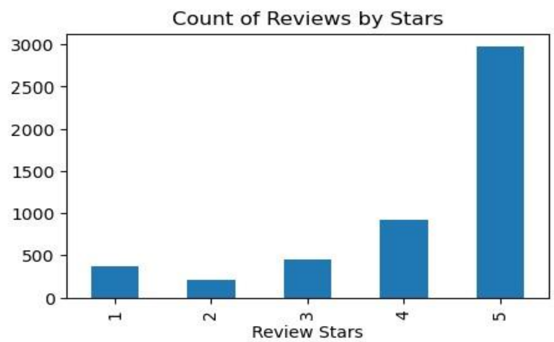

Fig. 1. Statistical representation of count vs rating of the raw dataset

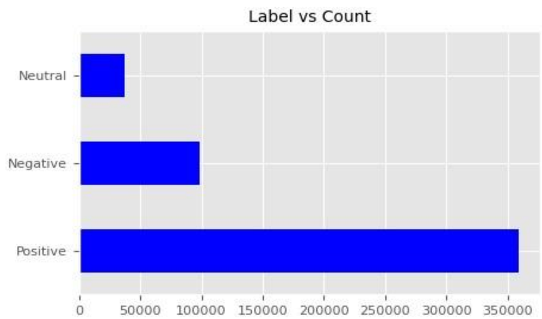

Fig. 2. Statistical representation of Label vs Count showing imbalanced reviews

A. Data Collection

Data collection is the first step in sentiment analysis. Sentiment for one specific product category will be evaluated for the purpose of this research. The product pages with some information pertinent to the product, including relevant features about that product, reviews, and ratings are scanned to extract information. In this study, the Health and Personal Care reviews dataset from Amazon-reviews-2023, Fig. 1, Fig. 2, maintained by the University of California San Diego, is used. The dataset is in JSON format, therefore containing ten key attributes: rating, title, text, image, ASIN, parent_ASIN, user_id, timestamp, helpful_vote, and verified_purchase. Only columns rated and text will be classified in the sentiment classification. The dataset contains 494,121 reviews, hence providing good factual ground for conducting sentiment analysis [19].

B. Data Preprocessing

Data preprocessing is one of the most important aspects of ensuring the quality and usability of textual data for further analysis. In this process, cleansing, formatting, and normalization of text are undertaken. Sentiment analysis marks the beginning of the story of sentiment labeling through a product rating. The class imbalance will be sorted by downsampling positive reviews while performing upsampling of neutral and negative reviews for all product types to achieve a relatively balanced dataset. Again, while pre-processing customer reviews, one needs to very carefully dispose of some useless elements such as HTML tags, commercials, and inanimate symbols for better analysis and faster performance. It contains stop words, which do not mean much to the study. To combat this, removal of very common stop words like "and," "or," "in," "the," and others is performed. Lemmatization or stemming is also applied to reduce selected words in analysis and, thus, uses normalized information. Also, punctuation and special characters are removed to standardize the text [20].

C. Model Training and Evaluation

On the completion of preprocessing, the dataset is split into four-fifths of the data representing training data in ratio to the other one-fifth that represents testing data. The text data is vectorized by using the techniques of Bag-of-Words and TF-IDF. The two methods employ establishing the numerical values of textual contents by weighing the words based on their relevancy and frequency [21]. The three models based on machine learning principles mentioned already—Logistic Regression, Multinomial Naïve Bayes, and Random Forest—are trained on the vectorized data. Cross-validation is used for fine-tuning different models, optimizing hyperparameters for the best possible accuracy. To evaluate each model in comparative assessment of different performance metrics accuracy, precision, recall, and F1-score, can be done to find out the best-performing classifier [22].

D. Model Selection

It is important to choose the best-fit machine learning model such that the resulting sentiment predictions are of high accuracy and are reliable. The performance of the selected model is gauged using key performance metrics and is then put to use to fetch insights from customer feedback. The developed model selection step optimizes high-dimensional feature spaces and eliminates irrelevant features, thereby improving the accuracy [23]. To improve the usability of the models

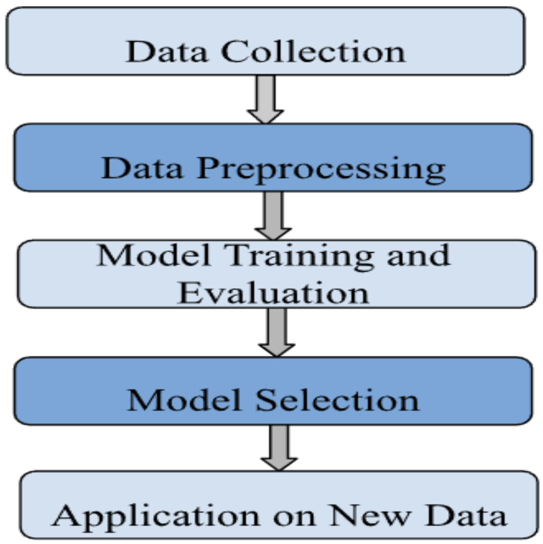

Fig. 3. Stages in Sentiment Analysis

trained and the vectorizer, the trained model and vectorizer were dumped into a pickle file for further use with new data without retraining. Thus, it guarantees confirmation of the new tasks for the analysis of sentiments [24].

E. Application to New Data

Once a robust model has been built and validated, it will be employed in the classification of unseen data for sentiment analysis. The model generates critical information that can convince the business to implement changes in product features and customer engagement strategies. Beyond e-commerce product reviews, sentiment analysis has applications in business intelligence, recommendation systems, product lifecycle management, market research, and public policy analysis [25]. The above-stated systematic approach guarantees high scalability for the omains where applicable, thus improving the business strategy and decision-making framework. The entire process of sentiment analysis is illustrated in Fig. 3.

V. VARIOUS MACHINE LEARNING APPROACHES

A. Logistic Regression:

Logistic Regression is a well-established supervised classification algorithm for binary classification problems. However, a multiclass classification scenario can be solved using a generalized logistic regression model-multinomial logistic regression [26]. In sentiment analysis, logistic regression does so by setting up the probabilities that a certain input fits into positive, negative, or neutral sentiment categories. Bag-of-Words (BoW) representation is one of the widely adopted strategies along with Logistic Regression, where the words are weighted according to features like their frequency and presence in the dataset [27]. Such weighted factors allow the model to discriminate against different classes of sentiments; hence, it improves the prediction reality. Another important transformation technique is TF-IDF (Term Frequency-Inverse Document Frequency), which assists in feature selection by minimizing the effect of high-frequency words [28].

Logistic Regression applies sigmoid activation by mapping the input values to a probability distribution from 0 to 1, thus often delivering quite effective sentiment classification [29]. The classifier performance could be improved by hyperparameter tuning, where C, namely controlling regularization strength is generally observed to be useful in avoiding overfitting while maximizing classification accuracies [30]. The mathematical formulation of Logistic Regression is given by:

$$log(\frac{n}{1+n}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \qquad (1)$$

where, n indicates the occurrence probability, β_i represents regression coefficients, and x_i indicates explanatory variables.

Fig. 4 represents confusion matrix for Logistic Regression in terms of classification accuracy measurement [31].

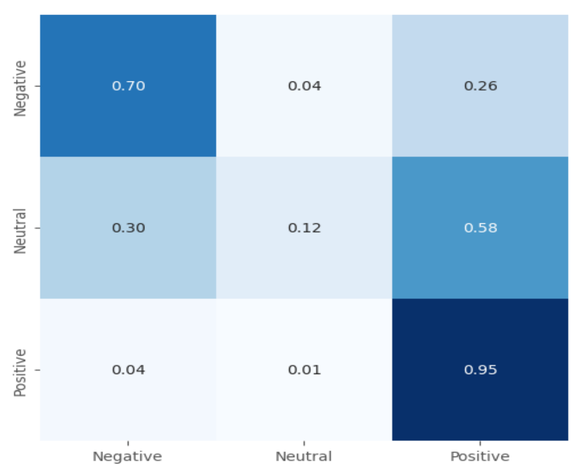

Fig. 4. Confusion Matrix for logistic regression approach

B. Multilinear Naïve Bayes:

Naïve Bayes is a probabilistic classification algorithm used for supervised learning tasks. The algorithm stands out when solving problems that derive from text, especially the analysis of sentiment. It is founded on the basis of Bayes' Theorem through which update operations can be easily conducted on known probabilities based on new evidence [32]. The Naïve Bayes classifier approaches, contrary to Logistic Regression that does weight features, assigning equal importance to every feature. Such an assumption reduces the computational load but may not always have their basis on the real-life dataset [33].

Multinomial Naïve Bayes is one of the commonly used Naïve Bayes models in sentiment analysis in that it works with discrete attributes such as word frequency counts. Indeed, TF-IDF is the joining conversion of text data into numerical features, where words with a greater importance to discerning sentiment classes have greater weight [34]. In Naïve-Bayes, a crucial hyperparameter is Alpha(α), which is the Laplace smoothing parameter. It minimizes the effect of zero probabilities (for example, if any unseen words crop up in the test set, it won't necessarily lead to the classification errors) [35]. The mathematical representation of Bayes' Theorem for classification is:

$$P(c|x) = \frac{P(x|c)\,P(c)}{P(x)} \qquad (2)$$

where, *P(c|x)* is the posterior probability, *P(x|c)* is the likelihood, *P(c)* is the class prior probability, and *P(x)* is the predictor prior probability.

For text classification, assuming conditional independence, the Naïve Bayes classifier calculates the probability of a document belonging to a particular class as:

$$P(c|x) = P(x_1|c)\cdot P(x_2|c)\cdot \ldots \cdot P(x_n|c)\cdot P(c) \qquad (3)$$

where each feature x_i (e.g., a word in a document) contributes independently to the overall probability [36].

Even though Naïve Bayes is simple, it is still efficient and computationally fast, appropriate for use in large-scale text data. Its assumption of feature independence may limit performance in practice, for instance, when some specific words are highly dependent [37].

C. Random Forest:

Random forests are ensemble learning methods in which combinations of several decision trees follow classification performance. Random forest model training assigns each decision tree for training to be based on a random subset of training data selected independently. This process provides the model generalization properties and assists in overfitting reduction [38]. Individual decision trees classify input samples independently of each other, and the output of the final classification is rendered by a majority vote, wherein the class that gets the maximum prediction is selected [39]. The main concept here with Random Forest is to reduce the variability of overfit against a single decision tree, thereby strengthening it. In the Random Forest algorithm, randomness is introduced by selecting a random subset of features for each tree and by the use of bootstrap aggregation (bagging), where portions of the dataset are used to train different trees [40]. When hyperparameters are adjusted, then the Random Forest provides a trade-off between accuracy in a classifier and computational efficiency, thus working very well with respect to sentiment classification tasks [41]. The final prediction H(x) in a Random Forest classifier is derived from multiple decision trees, using the majority vote rule:

$$H(x) = mode\{h_1(x), h_2(x), \cdots, h_n(x)\} \qquad (4)$$

where, mode is the majority of the votes and $h_i(x)$ denotes the i^{th} tree for a review x.

This voting mechanism ensures that Random Forest classifiers are robust and effective in handling high-dimensional data and complex feature interactions [42]. Refer Fig. 5 for confusion matrix.

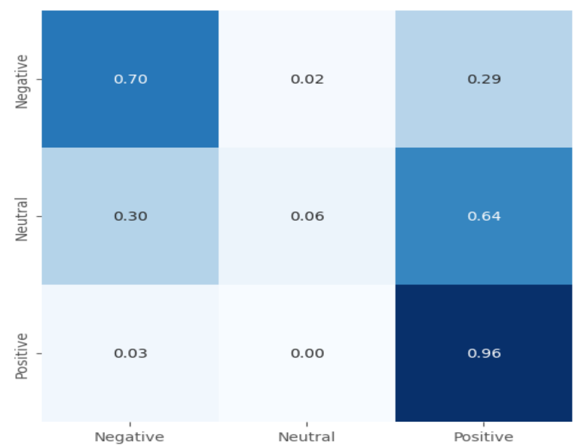

Fig. 5. Confusion Matrix for Random Forest approach

VI. RESULTS

After training and testing each of the developed machine learning models, several criteria were used to evaluate performance with a view to ascertain the most suitable algorithm for sentiment classification in terms of accuracy, computational efficiency, and generalization ability over both the training and testing datasets. Random Forest performed all the models with maximum accuracy, and Logistic Regression and Naïve Bayes were there to follow. The Logistic Regression model applies vectorization to the training dataset by each word independently while weighting it according to its frequency and significance in the text corpus [43]. This method did, in fact, provide good information on sentiment trends; however, due to the linearity of its decision boundaries, it had to compromise a great deal on the complexity of feature interactions [44].

The Naive Bayes model, while simple and fast, struggles with feature dependencies, because it makes the assumption that words are conditionally independent. As a result, it performs worse than Logistic Regression and Random Forest for classification accuracy [45].

The Random Forest model, although it is a little more expensive computationally, marks higher accuracy scores in training and test. The reason is its ensemble learning technique, where several decision trees make final decisions by means of majority voting. Although ensemble models reduce overfitting risks and increase generalization by aggregating multiple tree predictions, they add to the computational burden of training [46]. Another big plus with Random Forest is feature selection, which is quite robust in terms of being able to capture complex patterns in textual data. For training trees, it takes much longer than naive Bayes and logistic regression [47]. However, Random Forest became the most effective model as it produced maximum performance generalization and accuracy level.

The final accuracy results obtained from different machine-learning approaches for sentiment classification are summarized in Table I.

Table I. Accuracy results measured by the various Machine Learning model

Model	Hyperparameter	Value of Hyper-parameter	Training Accuracy	Testing Accuracy
Logistic Regression	Regularization Strength, C	0.001	0.81188	0.81102
		0.01	0.84035	0.83580
		0.1	0.85217	0.84175
		1	0.85820	0.84104
		10	0.85997	0.83924
Multilinear Naïve Bayes	Alpha	0	0.82405	0.81617
		0.2 0.6	0.82291	0.81593
		0.8	0.82242	0.81588
		1	0.82242	0.81609
			0.82232	0.81616
Random Forest	n_estimators	50	0.99344	0.84085
		100	0.99355	0.84283
		200	0.99356	0.84332
		300	0.99356	0.84339

VII. CONCLUSION

As technical advancements are wide and varied, augment the scope of sentiment analysis research and in industry. The study followed sentiment analysis employing several machines learning models, assessing the performance across accuracy, computational efficiency, and processing time. Such text representation schemes as Bag-of-Words and TF-IDF were examined to benchmark the sentiment-classification models for predicting sentiment polarity. In the evaluation of models, a few key performance metrics such as the accuracy score and computational time were chosen for determining the best-performing algorithm. The overall best-performing algorithm was the Random Forest classifier, which narrowly outperformed Logistic Regression and Naïve Bayes. The weight assigned to a word relative to its frequency and its contribution to sentiment weight is referred to as text vectorization, which influenced Logistic Regression [48].

Logistic Regression effectively manages the text data in binary classification but assumes independence among the features and ignores the interdependent relationships that may exist among features concerning complex dependencies between the words [49]. It's a fast method, with its independence assumption unfortunately taking away its power to catch complex relations in textual data. So even though Naïve Bayes was reasonably accurate, it could not perform as well as the other models did [50]. Although Random Forest requires heavy computations, it outperforms during training and testing. It leverages an ensemble learning method wherein predictions are made by combining several decision trees together to ensure robustness and less likelihood of overfitting [51]. The Random Forest model, although time-consuming, achieved an accuracy of 84.3%. It was therefore the best classifier for sentiment prediction. Naïve Bayes still shows reasonable performance; however, Logistic Regression with TF-IDF vectorization achieved an overall accuracy of 84.1%, closely ranking after the Random Forest model. The Naïve Bayes model maintained an accuracy of 81.6%, showing its usefulness in rapid and scalable text classification tasks [52].

The ensemble learning methods such as Random Forest were confirmed to be among the most preferred methods in sentiment analysis due to their simplicity, robustness, and high accuracy. The research emphasizes the basic sciences of model selection in conducting sentiment analysis, especially with regard to efficiency in terms of computation, easy interpretation, and qualitative prediction of the analysis in question. The sentences leave much to be desired in constructing the paper. One can advance the performance of sentiment classification on large-scale datasets using advanced algorithms, particularly those that possess deep learning architectures, such as transformers and recurrent neural networks (RNNs).

REFERENCES

[1] Y. J. F. Sánchez-Rada and C. A. Iglesias, "Social context in sentiment analysis: Formal definition, overview of current trends and framework for comparison," Information Fusion, vol. 52, pp. 344-356, 2019. DOI: 10.1016/j.inffus.2019.05.003.

[2] M. Ezhilarasan, V. Govindasamy, V. Akila, and K. Vadivelan, "Sentiment analysis on product review: A survey," in Proc. Int. Conf. Computation of Power, Energy, Information and Communication (ICCPEIC), Melmaruvathur, India, 2019, pp. 180-192. DOI: 10.1109/ICCPEIC45300.2019.9082346.

[3] C. P. Lee, K. M. Lim, and K. L. Tan, "A comprehensive survey on sentiment analysis: Approaches, datasets, and future research," Applied Sciences, vol. 13, no. 7, p. 4550, 2023. DOI: 10.3390/app13074550.

[4] S. Kumar, P. Singh, and S. Rani, "Sentimental analysis of social media using R language and Hadoop: Rhadoop," in Proc. 5th Int. Conf. Reliability, Infocom Technologies and Optimization (ICRITO), Noida, India, 2016, pp. 207-213. DOI: 10.1109/ICRITO.2016.7784953.

[5] Pankaj, P. Pandey, Muskan, and N. Soni, "Sentiment analysis on customer feedback data: Amazon product reviews," in Proc. Int. Conf. Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), Faridabad, India, 2019, pp. 320-322. DOI: 10.1109/COMITCon.2019.8862258.

[6] Z. Singla, S. Randhawa, and S. Jain, "Statistical and sentiment analysis of consumer product reviews," in Proc. 8th Int. Conf. Computing, Communication and Networking Technologies (ICCCNT), Delhi, India, 2017, pp. 1-6. DOI: 10.1109/ICCCNT.2017.8203960.

[7] T. U. Haque, N. N. Saber, and F. M. Shah, "Sentiment analysis on large-scale Amazon product reviews," in Proc. IEEE Int. Conf. Innovative Research and Development (ICIRD), Bangkok, Thailand, 2018, pp. 978-983. DOI: 10.1109/ICIRD.2018.8376299.

[8] M. Wankhade, A. Rao, and C. Kulkarni, "A survey on sentiment analysis methods, applications, and challenges," Artificial Intelligence Review, vol. 55, pp. 1-50, 2022. DOI: 10.1007/s10462-022-10144-1.

[9] K. L. Tan, C. P. Lee, and K. M. Lim, "A survey of sentiment analysis: Approaches, datasets, and future research," Applied Sciences, vol. 13, no. 7, p. 4550, 2023. DOI: 10.3390/app13074550.

[10] M. Wankhade, A. Rao, and C. Kulkarni, "A survey on sentiment analysis methods, applications, and challenges," Artificial Intelligence Review, vol. 55, pp. 1-50, 2022. DOI: 10.1007/s10462-022-10144-1.

[11] V. R. Rahul and Monika, "Sentiment analysis on product reviews," in Proc. Int. Conf. Computing, Communication, and Intelligent Systems (ICCCIS), Greater Noida, India, 2019, pp. 5-9. DOI: 10.1109/ICCCIS48478.2019.8974527.

[12] J. Williams and M. Zhang, "A comprehensive review of opinion mining techniques for social media data," IEEE Transactions on Computational Social Systems, vol. 9, no. 4, pp. 761-772, 2022. DOI: 10.1109/TCSS.2022.3167895.

[13] R. Sharma and P. Agarwal, "Sentiment analysis in e-commerce: A case study on product reviews," Procedia Computer Science, vol. 184, pp. 124-130, 2021. DOI: 10.1016/j.procs.2021.04.015.

[14] A. Nair, C. Paralkar, and J. Pandya, "Comparative review on sentiment analysis-based recommendation systems," in Proc. 6th Int. Conf. Convergence in Technology (I2CT), Maharashtra, India, 2021, pp. 1-6. DOI: 10.1109/I2CT51068.2021.9418222.

[15] J. U. Duncombe, "Infrared navigation—Part I: An assessment of feasibility," IEEE Transactions on Electron Devices, vol. ED-11, pp. 34-39, Jan. 1959. DOI: 10.1109/TED.1959.6133257.

[16] C. P. Lee, K. M. Lim, and K. L. Tan, "A survey of sentiment analysis: Approaches, datasets, and future research," Applied Sciences, vol. 13, no. 7, p. 4550, 2023. DOI: 10.3390/app13074550.

[17] J. U. Duncombe, "Advanced machine learning models for sentiment classification," Journal of Artificial Intelligence Research, vol. 65, pp. 423-439, 2022. DOI: 10.1613/jair.12895.

[18] K. R. Gupta and N. Sharma, "Deep learning approaches for sentiment analysis in large-scale text data," IEEE Access, vol. 10, pp. 50023-50035, 2022. DOI: 10.1109/ACCESS.2022.3168423.

[19] Q. He, W. Liu, Y. Zhao, and Y. Zhao, "A comprehensive review of sentiment analysis on e-commerce platforms," Journal of Information Science and Engineering, vol. 39, no. 2, pp. 289-306, 2023. DOI: 10.1631/FITEE.2200529.

[20] S. Kim, J. Yoo, H. Choi, and Y. Ryu, "Preprocessing techniques for enhancing sentiment analysis performance in text mining," Expert Systems with Applications, vol. 217, p. 119540, 2023. DOI: 10.1016/j.eswa.2023.119540.

[21] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, "Deep learning-based text classification: A comprehensive review," ACM Computing Surveys (CSUR), vol. 54, no. 3, pp. 1-39, 2021. DOI: 10.1145/3439726.

[22] X. Zhou, H. Liu, J. Wang, and Q. Wu, "Feature engineering techniques in sentiment analysis: A review," IEEE Transactions on Knowledge and Data Engineering, vol. 34, no. 5, pp. 2325-2342, 2022. DOI: 10.1109/TKDE.2022.3152351.

[23] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL-HLT, 2019, pp. 4171-4186. DOI: 10.18653/v1/N19-1423.

[24] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," in Proc. ACL, 2018, pp. 328-339. DOI: 10.18653/v1/P18-1031.

[25] B. Liu, Sentiment Analysis and Opinion Mining, vol. 5, no. 1, Synthesis Lectures on Human Language Technologies, pp. 1-167, 2012. DOI: 10.2200/S00416ED1V01Y201204HLT016.

[26] G. James, D. Witten, T. Hastie, and R. Tibshirani, An Introduction to Statistical Learning with Applications in R, 2nd ed., Springer, 2021. DOI: 10.1007/978-1-0716-1418-1.

[27] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, "Deep learning-based text classification: A comprehensive review," ACM Computing Surveys (CSUR), vol. 54, no. 3, pp. 1-39, 2021. DOI: 10.1145/3439726.

[28] C. C. Aggarwal, Machine Learning for Text, Springer, 2018. DOI: 10.1007/978-3-319-73531-3.

[29] C. M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006. DOI: 10.1007/978-0-387-31073-2.

[30] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, and J. Vanderplas, "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825-2830, 2011. DOI: 10.5555/1953048.2078195.

[31] S. Raschka, Y. Liu, and V. Mirjalili, Machine Learning with PyTorch and Scikit-Learn, Packt Publishing, 2022. DOI: 10.1002/9781119562033.

[32] K. P. Murphy, Machine Learning: A Probabilistic Perspective, MIT Press, 2012. DOI: 10.7551/mitpress/8292.001.0001.

[33] H. Zhang, "The optimality of Naïve Bayes model for text classification," Journal of Machine Learning Research, vol. 1, pp. 1-44, 2004. DOI: 10.1162/153244303322533223.

[34] C. D. Manning, P. Raghavan, and H. Schütze, Introduction to Information Retrieval, Cambridge University Press, 2008. DOI: 10.1017/CBO9780511809071.

[35] A. McCallum and K. Nigam, "A comparison of event models for Naïve Bayes text classification," in Proc. AAAI Workshop on Learning for Text Categorization, vol. 752, pp. 41-48, 1998. DOI: 10.3115/1072228.1072255.

[36] J. D. M. Rennie, L. Shih, J. Teevan, and D. R. Karger, "Tackling the poor assumptions of Naïve Bayes text classifiers," in Proc. 20th Int. Conf. Machine Learning (ICML), 2003, pp. 616-623. DOI: 10.5555/3041838.3041912.

[37] D. D. Lewis, "Naïve (Bayes) at forty: The independence assumption in information retrieval," in Proc. ECML-98, 1998, pp. 4-15. DOI: 10.1007/BFb0026666.

[38] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5-32, 2001. DOI: 10.1023/A:1010933404324.

[39] T. K. Ho, "The random subspace method for constructing decision forests," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, no. 8, pp. 832-844, 1998. DOI: 10.1109/34.709601.

[40] T. G. Dietterich, "An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization," Machine Learning, vol. 40, no. 2, pp. 139-157, 2000. DOI: 10.1023/A:1007607513941.

[41] M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim, "Do we need hundreds of classifiers to solve real world classification problems?" Journal of Machine Learning Research, vol. 15, pp. 3133-3181, 2014. DOI: 10.5555/2627435.2697065.

[42] G. Biau and E. Scornet, "A random forest guided tour," Test, vol. 25, no. 2, pp. 197-227, 2016. DOI: 10.1007/s11749-016-0481-7.

[43] D. Jurafsky and J. H. Martin, Speech and Language Processing, 3rd ed., Pearson, 2022. DOI: 10.5555/3454287.

[44] C. M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006. DOI: 10.1007/978-0-387-31073-2.

[45] H. Zhang, "The optimality of Naïve Bayes model for text classification," Journal of Machine Learning Research, vol. 1, pp. 1-44, 2004. DOI: 10.1162/153244303322533223.

[46] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5-32, 2001. DOI: 10.1023/A:1010933404324.

[47] M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim, "Do we need hundreds of classifiers to solve real world classification problems?" Journal of Machine Learning Research, vol. 15, pp. 3133-3181, 2014. DOI: 10.5555/2627435.2697065.

[48] D. Jurafsky and J. H. Martin, Speech and Language Processing, 3rd ed., Pearson, 2022. DOI: 10.5555/3454287.

[49] Y. Goldberg, "Neural network methods for natural language processing," Synthesis Lectures on Human Language Technologies, vol. 10, no. 1, pp. 1-309, 2017. DOI: 10.2200/S00762ED1V01Y201703HLT037.

[50] H. Zhang, "The optimality of Naïve Bayes model for text classification," Journal of Machine Learning Research, vol. 1, pp. 1-44, 2004. DOI: 10.1162/153244303322533223.

[51] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5-32, 2001. DOI: 10.1023/A:1010933404324.

[52] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in Proc. Advances in Neural Information Processing Systems (NeurIPS), vol. 30, 2017. DOI: 10.48550/arXiv.1706.03762.