SENTIMENT ANALYSIS OF NIGERIAN OPINIONS USING LOGISTIC REGRESSION AND RANDOM FOREST ALGORITHMS: A COMPARATIVE STUDY

Original Research Article

ABSTRACT

This study investigates the efficacy of Logistic Regression and Random Forest models in sentiment analysis using Nigerian-based datasets, namely "Gangs of Lagos" and "PeterObi Politics." Sentiment analysis, a vital component of Natural Language Processing (NLP), plays a crucial role in understanding public opinion and sentiment trends, particularly in the context of Nigerian socio-political discourse. Leveraging machine learning techniques, the study examines the performance of these models in predicting sentiment classes, including positive, negative, and neutral sentiments, within the datasets. The findings shed light on the strengths and limitations of Logistic Regression and Random Forest in discerning sentiment nuances prevalent in Nigerian language expressions with Logistic Regression outperforming Random Forest in both cases. This research contributes to the advancement of sentiment analysis methodologies tailored to Nigerian linguistic and cultural contexts, with implications for various applications, including social media monitoring, political analysis, and market research.

Keywords: Natural Language Processing, Machine Learning, Logistic Regression, Random Forest, Nigeria.

1.0 INTRODUCTION

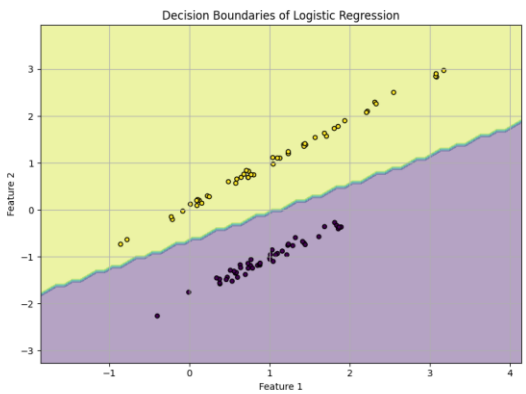
In today's era of information abundance, the proliferation of user-generated content across various digital platforms presents both a unique opportunity and a formidable challenge to address. Sentiments, encapsulating opinions on specific subjects, are prevalent within this vast pool of textual data, holding significant importance for businesses, researchers, and decision-makers across diverse domains [9]. Utilizing natural language processing and machine learning techniques, sentiment analysis, also known as opinion mining, aims to extract and decipher sentiments and attitudes expressed in textual, audio, and visual formats.

As social media platforms serve as dynamic hubs for user interactions, conversations, and reactions to current events, they emerge as invaluable repositories of unfiltered sentiments. The real-time nature of social media, coupled with its interactive features and multimedia elements, enriches the depth and authenticity of expressed sentiments, providing a nuanced understanding of public opinion dynamics [8]. With ongoing advancements in machine learning, a systematic exploration of sentiment analysis models becomes imperative for informed decision-making. Supervised learning algorithms, guided by labeled datasets, categorize input data into predefined classes, with examples including logistic regression, random forest and Naive bayes. These algorithms leverage patterns and relationships within the data to facilitate accurate classifications. Artificial intelligence (AI) and machine learning (ML) intersect within the realm of computer science, with AI encompassing broader cognitive functionalities and ML focusing on data-driven learning mechanisms. While AI seeks to emulate human thought processes, ML harnesses data analysis to enable computers to learn and adapt autonomously [13]. Within the machine learning landscape, algorithms serve as the foundational components that extract actionable insights from data, empowering machines to make informed predictions and decisions. Logistic Regression and Random Forest are two machine learning algorithms commonly employed in sentiment analysis tasks due to their unique strengths and capabilities.

Logistic Regression, a binary classification algorithm, is particularly suitable for sentiment analysis due to its assumption of a linear relationship between input features and the log-odds of the output [4]. This linear decision boundary makes it ideal for problems where sentiment classification can be approximated as linear or where there's a clear separation between positive and negative sentiments. Moreover, Logistic Regression provides a probabilistic interpretation of its predictions, offering insights into the likelihood of a given input belonging to a specific sentiment class. This probabilistic output is valuable in sentiment analysis applications where understanding the confidence of predictions is crucial.

On the other hand, Random Forest, an ensemble learning algorithm, offers distinct advantages for sentiment analysis tasks [13]. Unlike Logistic Regression, Random Forest can capture complex non-linear relationships between input features and sentiment labels. By aggregating predictions from multiple decision trees, each trained on a random subset of the data and features, Random Forest mitigates overfitting and improves generalization performance. This robustness to overfitting is particularly valuable in sentiment analysis tasks where the dataset may contain noise or outliers [4].

Furthermore, Random Forest provides insights into feature importance, indicating the contribution of each feature to the overall predictive performance. This information is crucial for understanding the most influential words or phrases in sentiment analysis and guiding feature selection efforts. Additionally, Random Forest's ability to handle imbalanced datasets effectively makes it suitable for sentiment analysis tasks where one sentiment class may be more prevalent than the other. Both Logistic Regression and Random Forest offer unique advantages for sentiment analysis tasks. The choice between the two algorithms depends on the specific characteristics of the sentiment analysis problem, including the linearity of the decision boundary, the complexity of the relationships between features and sentiments, and the interpretability of the model [7]. The diagrams below are the decision boundaries generated with synthetic datasets for Logistics Regression and Random Forest.



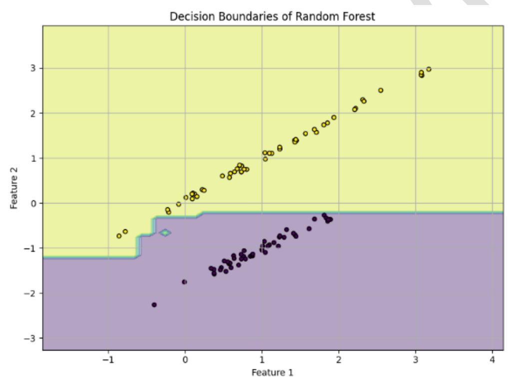


Figure 1.Logistics Regression

Figure 2. Random Forest

2.0 RELATED WORKS

Authors in [2] explored sentiment analysis, a technology within Natural Language Processing (NLP) that integrates Artificial Intelligence (AI) and Machine Learning (ML). The paper delves into various aspects of sentiment analysis, including its definition, algorithms, and procedural steps. The scope of coverage extends from the initial stages of sentiment analysis to the evaluation of sentiment classifiers' predictions. Authors in [6] investigate the application of sentiment analysis for predicting stock prices using machine learning techniques, particularly Random Forest and Multinomial Naive Bayes algorithms. The study utilizes the TFIDF technique for feature extraction and focuses on news headlines from Financial Times to classify stock price changes as positive or negative. The main objective is to evaluate the effectiveness of sentiment analysis in stock prediction and compare the performance of the two algorithms mentioned. Authors in [11] conducted a review with a focus on sentiment analysis of Twitter data, considering the unique nature of tweets as concise expressions. The study situates sentiment analysis within the domains of text data mining and natural language processing (NLP). Research on sentiment analysis of Twitter data is explored from various perspectives, covering different types and techniques. The survey offers a comparative analysis of various techniques and approaches specifically applied for sentiment analysis using Twitter data.

3.0 MATERIALS AND METHODS

In this study, the machine learning operations were conducted using Google Colab notebook, a cloud-based platform for Python programming and machine learning tasks. Leveraging the computational resources and collaborative features of Google Colab, the research seamlessly integrated data preprocessing, model training, and evaluation processes. The datasets stored in Google Drive were accessed directly from the Google Colab environment, ensuring efficient data handling and seamless integration with the machine learning workflow. This approach facilitated easy sharing and collaboration among researchers, eliminating the need for local data storage and management. Python programming language served as the primary tool for implementing machine learning algorithms and conducting data analysis. The extensive libraries available in Python, particularly Scikit-learn (Sklearn), Pandas, and Matplotlib, provided comprehensive support for data manipulation, model development, and visualization tasks. Sklearn, a powerful machine learning library in Python, offered a wide range of algorithms and tools for building and evaluating machine learning models. From preprocessing data to training classification algorithms such as Logistic Regression and Random Forest, Sklearn provided a user-friendly interface and efficient implementation for seamless experimentation. Pandas, another essential library in Python, facilitated data manipulation and preprocessing tasks. It enabled researchers to load, clean, and transform datasets efficiently, ensuring data readiness for model training and evaluation. Additionally, Pandas provided robust support for handling tabular data structures, making it well-suited for data analysis tasks. Matplotlib, a popular data visualization library in Python, enabled the creation of informative plots and visualizations to analyze and interpret the results effectively. From simple line plots to complex heatmaps and bar charts, Matplotlib offered versatile tools for conveying insights from the data and model evaluation metrics.

3.1 Datasets

In this study, sentiment analysis was performed using two Nigerian-based datasets: Gangs of Lagos and PeterObi Politics. These datasets were chosen to evaluate the performance of Random Forest and Logistic Regression algorithms in sentiment prediction tasks. The Gangs of Lagos dataset comprises movie reviews extracted from Twitter, providing a large corpus of text data for sentiment analysis. On the other hand, the PeterObi Politics dataset consists of sentiments related to Peter Obi, a presidential candidate in the 2023 Nigerian Elections, also sourced from Twitter. This dataset, while smaller in size compared to Gangs of Lagos, offers insights into sentiments associated with political figures. The Gangs of Lagos dataset exhibits significant support for sentiment analysis, containing diverse opinions and expressions from Twitter users regarding movie reviews. In contrast, the PeterObi Politics dataset provides a narrower focus, with limited support for sentiment classes related to Peter Obi's political activities during the specified timeframe.

The sentiment classes in the PeterObi Politics dataset are categorized into three main categories, offering a nuanced perspective on public sentiments towards the political figure. These sentiment classes provide a structured framework for evaluating the performance of sentiment analysis algorithms [1]. Both Random Forest and Logistic Regression algorithms were chosen for their effectiveness in handling text classification tasks, particularly in sentiment analysis. Random Forest excels in handling high-dimensional data and capturing complex relationships between features, while Logistic Regression offers simplicity and interpretability, making it suitable for binary classification tasks. By leveraging these algorithms on the Gangs of Lagos and PeterObi Politics datasets, the study aimed to assess their performance in accurately predicting sentiments expressed in the text data [3]. This evaluation provided valuable insights into the efficacy of machine learning techniques for sentiment analysis in the Nigerian context, contributing to the advancement of sentiment analysis research in the region. Below are diagrams of the two datasets viewed with microsoft excel.

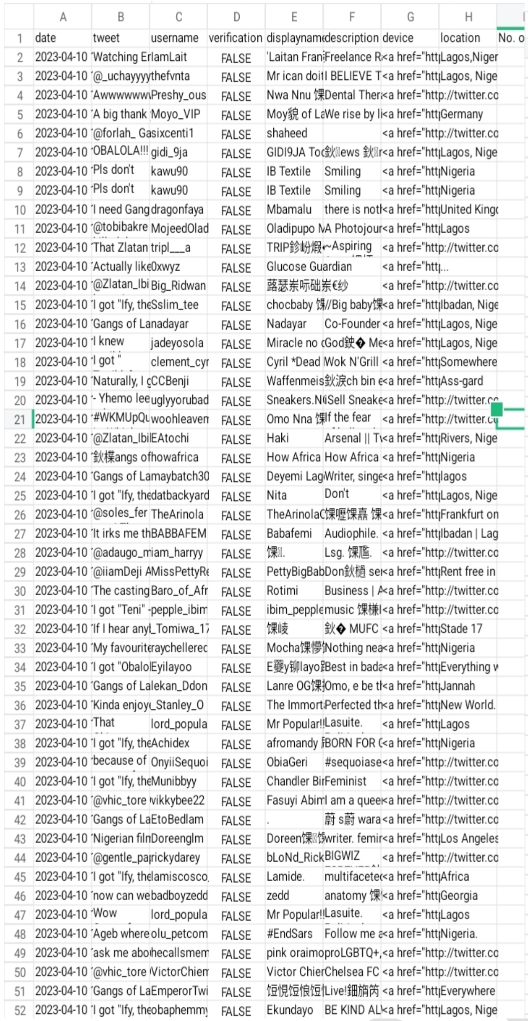


Figure 3. Gangs of Lagos

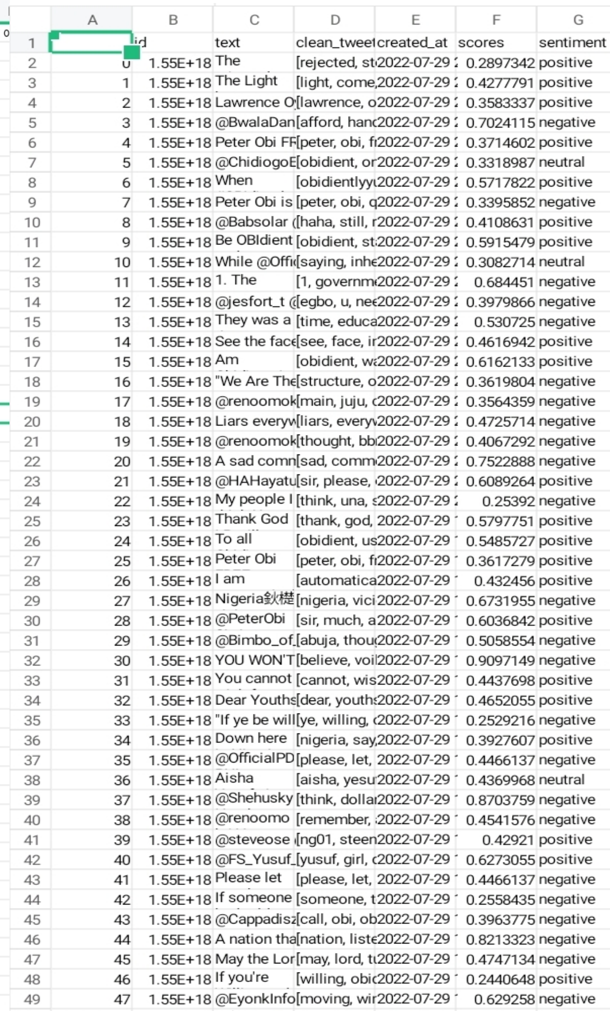


Figure 4. PeterObi politics

3.2 Procedure

In this study, sentiment analysis was conducted using Logistic Regression and Random Forest algorithms in Python, utilizing Google Colab for machine learning operations and accessing datasets stored in Google Drive. The datasets used were "Gangs of Lagos" for movie reviews from Twitter and "PeterObi Politics" for sentiments related to the Nigerian presidential candidate Peter Obi in the 2023 elections. The datasets were loaded into pandas dataframes after mounting Google Drive to Google Colab. Preprocessing steps, including cleaning, handling missing values, and feature engineering, were performed to prepare the data for analysis [10]. Text data was converted into numerical representations using techniques such as tokenization and TF-IDF.

The datasets were split into training and testing sets using the train_test_split function from scikit-learn. Logistic Regression and Random Forest classifiers were trained using the training data and evaluated using the testing data. Evaluation metrics such as accuracy, precision, recall, and F1-score were calculated to assess model performance [12]. Classification reports and confusion matrices were generated to analyze model predictions. Visualization techniques such as bar charts, line graphs, heatmaps, ROC curves, and precision-recall curves were employed to visualize the evaluation metrics and analyze classifier performance. The findings from the analysis were summarized, including insights gained from the evaluation metrics and visualization plots. The Flowchart is shown below:

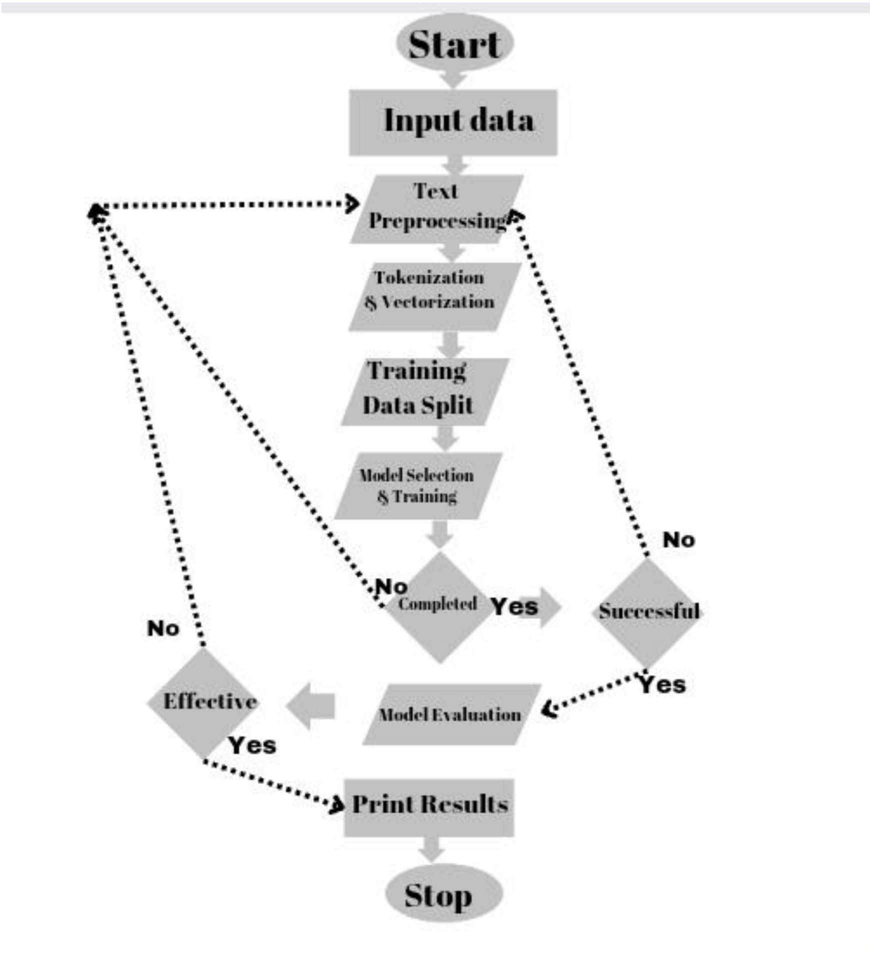


Figure 5. Flowchart for Sentiment Analysis Operation

4.0 RESULTS

It is essential to understand the performance of the sentiment analysis models using Logistic Regression and Random Forest algorithms. These evaluation metrics provide insights into the effectiveness of the classifiers in predicting sentiment labels for the given datasets. The classification report offers a comprehensive summary of the model's performance, including precision, recall, F1-score, and support for each sentiment class. Additionally, precision-recall curves and AUC curves visualize the trade-off between precision and recall and the model's ability to rank positive instances, respectively. Together, these metrics and visualizations provide a holistic view of the sentiment analysis models' performance, enabling informed decisions and further analysis. The classification report consists of several key components that provide insights into the performance of a classification model:

Precision: Precision measures the proportion of true positive predictions among all instances predicted as positive. It indicates the model's ability to correctly identify positive cases without misclassifying negative cases.

Recall: Recall, also known as sensitivity or true positive rate, measures the proportion of true positive predictions among all actual positive instances in the dataset. It indicates the model's ability to capture all positive instances [5].

F1-Score: The F1-score is the harmonic mean of precision and recall. It provides a balance between precision and recall and is particularly useful when the class distribution is uneven.

Support: Support represents the number of instances in each class in the dataset. It indicates the reliability of the evaluation metrics by providing context about the distribution of classes.

The results for the tests are shown below in classification report format:

For gangs of lagos movie:

Table 1: Classification Report - Logistic Regression for gangs of lagos movie

	precision	recall	f1-score	support
positive	0.90	0.44	0.59	399
neutral	0.85	0.94	0.89	1312
negative	0.85	0.89	0.87	1163
accuracy			0.85	2874
macro avg	0.87	0.76	0.78	2874
weighted avg	0.86	0.85	0.84	2874

The results above show that the logistic regression model exhibits robust performance, particularly in discerning neutral and negative sentiments, albeit with potential for refinement in identifying positive sentiments, as evidenced by the comparatively lower recall.

.

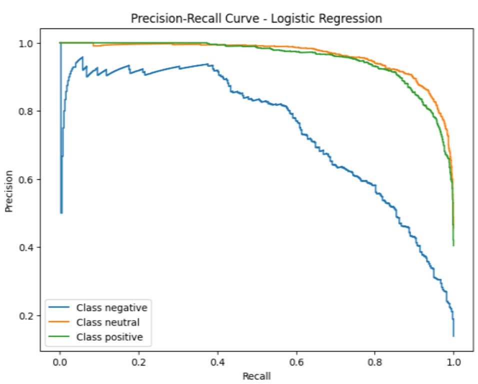


Figure 6. Precision - Recall Curves Logistics Regression

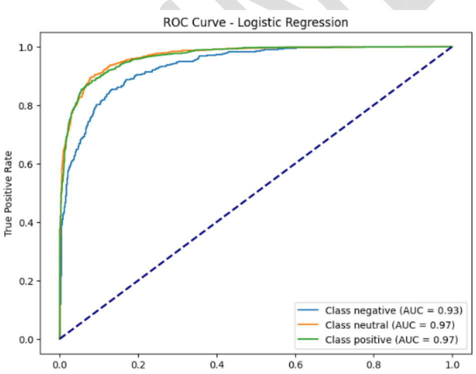


Figure 7. ROC curve Logistic Regression

The imbalanced distribution of sentiment classes poses a notable challenge, particularly impacting recall metrics, particularly evident in the case of positive sentiments. Exceptionally high Area Under the Curve (AUC) values for both neutral and positive sentiments underscore the model's discriminative prowess. Regarding the first dataset, the sentiment analysis model leveraging logistic regression demonstrates overall efficacy, showcasing proficiency in discriminating between neutral and negative sentiments.

Table 2: Classification Report - Random Forest for gangs of lagos movie

	precision	**recall**	**f1-score**	**support**
positive	**0.94**	**0.20**	**0.33**	**399**
neutral	**0.82**	**0.89**	**0.85**	**1312**
negative	**0.74**	**0.87**	**0.80**	**1163**
accuracy			**0.79**	**2874**
macro avg	**0.83**	**0.65**	**0.66**	**2874**
weighted avg	**0.80**	**0.79**	**0.76**	**2874**

The results above show that the random forest model showcases moderate performance, particularly excelling in classifying neutral sentiments.

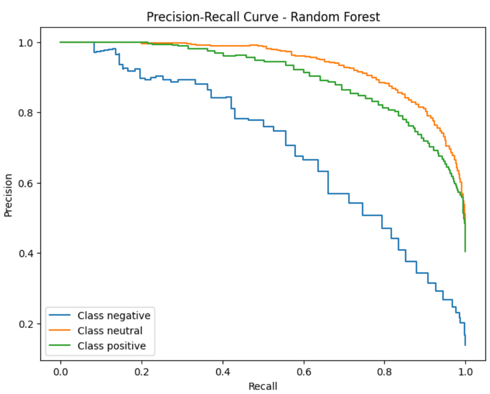


Figure 8. Precision - Recall Curves Random Forest

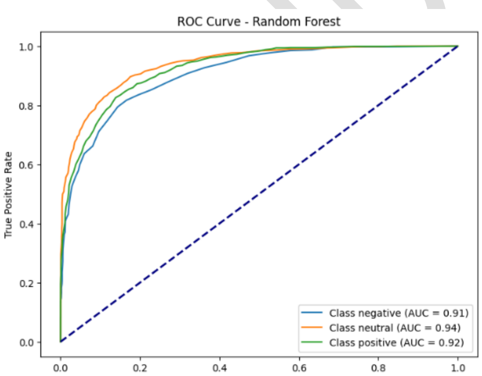


Figure 9. ROC curve Random Forest

However, there exists considerable room for improvement in detecting positive sentiments, with notable implications for recall metrics, owing partly to the imbalanced class distribution. The commendable AUC scores for neutral and positive sentiments further affirm the Random Forest model's adeptness in discriminating between sentiment classes.

While both logistic regression and random forest models yield satisfactory results, Logistic Regression emerges as the superior performer across various metrics for this dataset.

For Peter Obi political sentiment:

Table 3: Classification Report - Logistic Regression For Peter Obi political sentiment

	precision	**recall**	**f1-score**	**support**
positive	**0.76**	**0.84**	**0.80**	**546**
negative	**0.50**	**0.03**	**0.06**	**61**
neutral	**0.72**	**0.72**	**0.72**	**393**
accuracy			**0.74**	**1000**
macro avg	**0.66**	**0.53**	**0.52**	**1000**
weighted avg	**0.73**	**0.74**	**0.72**	**1000**

The results above show that the logistic regression model exhibits proficiency in predicting positive sentiments but faces challenges in accurately identifying negative sentiments, leading to lower recall values.

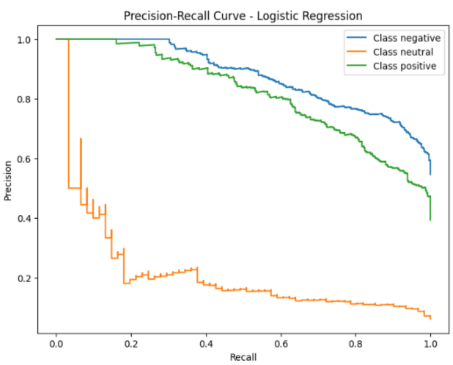


Figure 10. Precision - Recall Curves Logistic Regression

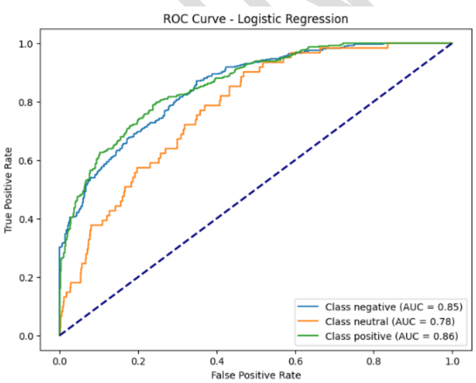


Figure 11. ROC curve Logistics Regression

The imbalanced distribution of sentiment classes, particularly for negative sentiments, contributes to this disparity in recall metrics. Robust discrimination is observed for positive sentiments, as evidenced by the high Area Under the Curve (AUC) values. Overall, the sentiment analysis model employing Logistic Regression demonstrates satisfactory performance, particularly excelling in predicting positive and neutral sentiments.

Table 4: Classification Report - Random Forest For Peter Obi political sentiment

	precision	**recall**	**f1-score**	**support**
positive	**0.75**	**0.82**	**0.79**	**546**
negative	**0.67**	**0.07**	**0.12**	**61**
neutral	**0.69**	**0.69**	**0.69**	**393**
accuracy			**0.73**	**1000**
macro avg	**0.70**	**0.53**	**0.53**	**1000**
weighted avg	**0.72**	**0.73**	**0.71**	**1000**

Similarly, the random forest model displays strengths in predicting positive sentiments but struggles with negative sentiments, resulting in comparatively lower recall rates.

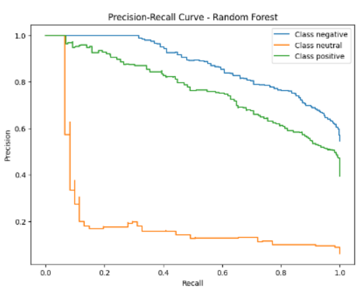


Figure 12. Precision - Recall Curves Random Forest

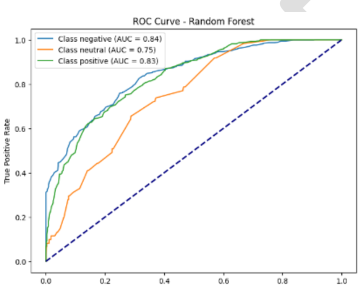


Figure 13. ROC curve Random Forest

The imbalanced nature of sentiment classes further exacerbates this issue, particularly affecting the recall for negative sentiments. Despite these challenges, the random forest model showcases robust discrimination for positive sentiments, as indicated by the high AUC values. The sentiment analysis model utilizing Random Forest delivers acceptable performance, particularly demonstrating proficiency in predicting positive and neutral sentiments.

In comparison, the performance of the logistic regression model is superior for the second dataset, particularly in terms of predicting sentiment classes.

5.0 CONCLUSION

This study explored the effectiveness of Logistic Regression and Random Forest models in sentiment analysis tasks using two Nigerian-based datasets, "Gangs of Lagos" and "PeterObi Politics." The findings revealed that both models exhibited strengths and weaknesses in predicting sentiment classes, with Logistic Regression demonstrating proficiency in identifying positive sentiments and Random Forest excelling in discriminating neutral sentiments. However, challenges were encountered in accurately identifying negative sentiments, attributed to the imbalanced distribution of sentiment classes within the datasets. Despite these limitations, both models achieved acceptable performance overall, with logistic regression performance being better in both cases, underscoring their potential utility in sentiment analysis tasks. Moving forward, further research is warranted to address the challenges posed by imbalanced datasets and enhance the performance of sentiment analysis models in real-world applications.

Disclaimer (Artificial intelligence)

Author(s) hereby declare that NO generative AI technologies such as Large Language Models (ChatGPT, COPILOT, etc) and text-to-image generators have been used during writing or editing of manuscripts.

5.0 REFERENCES

[1] Christian, P. H., & Desanti, R. I. (2022). The comparison of sentiment analysis algorithm for fake review detection of the leading online stores in Indonesia. Proceedings of the Seventh International Conference on Informatics and Computing (ICIC), Denpasar, Bali, Indonesia, pp. 01-04. doi: 10.1109/ICIC56845.2022.10006984.

[2] Dobhal, D. C., Kumar, B., & Das, P. (2023). Involvement of functional programming in language processing and machine learning. Proceedings of the 2nd International Conference for Innovation in Technology (INOCON), Bangalore, India, pp. 1-4. doi: 10.1109/INOCON57975.2023.10101253.

[3] Harjadinata, A., & Sibaroni, Y. (2022). Multi-aspect sentiment analysis on TikTok using random forest classifier and Word2Vec. Proceedings of the 1st International Conference on Software Engineering and Information Technology (ICoSEIT), Bandung, Indonesia, pp. 261-266. doi: 10.1109/ICoSEIT55604.2022.10029958.

[4] Khanvilkar, G., & Vora, D. (2019). Smart recommendation system based on product reviews using random forest. Proceedings of the International Conference on Nascent Technologies in Engineering (ICNTE), Navi Mumbai, India, pp. 1-9. doi: 10.1109/ICNTE44896.2019.8945855.

[5] Nazir, A., Rao, Y., Wu, L., & Sun, L. (2022). Issues and challenges of aspect-based sentiment analysis: A comprehensive survey. IEEE Transactions on Affective Computing, 13(2), 845-863. doi: 10.1109/TAFFC.2020.2970399.

[6] Parashar, D., DSilva, M., & Kulshreshtha, S. (2023). A machine learning framework for stock prediction using sentiment analysis. Proceedings of the 4th IEEE Global Conference for Advancement in Technology (GCAT), Bangalore, India, pp. 1-5. doi: 10.1109/GCAT59970.2023.10353541.

[7] Phulare, P., & Deshmukh, S. N. (2021). Cricket Twitter data sentiment analysis and prediction exerted machine learning. Proceedings of the International Conference on Disruptive Technologies for Multi-Disciplinary Research and Applications (CENTCON), Bengaluru, India, pp. 141-145. doi: 10.1109/CENTCON52345.2021.9688197.

[8] Rawat, A., Maheshwari, H., Khanduja, M., Kumar, R., Memoria, M., & Kumar, S. (2022). Sentiment analysis of Covid-19 vaccines tweets using NLP and machine learning classifiers. Proceedings of the International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON), Faridabad, India, pp. 225-230. doi: 10.1109/COM-IT-CON54601.2022.9850629.

[9] Sindhu, S., Kumar, S., & Noliya, A. (2023). A review on sentiment analysis using machine learning. Proceedings of the International Conference on Innovative Data Communication Technologies and Application (ICIDCA), Uttarakhand, India, pp. 138-142. doi: 10.1109/ICIDCA56705.2023.10099665.

[10] Subramanian, R. R., Akshith, N., Murthy, G. N., Vikas, M., Amara, S., & Balaji, K. (2021). A survey on sentiment analysis. Proceedings of the International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, pp. 70-75. doi: 10.1109/Confluence51648.2021.9377136.

[11] Wagh, R., & Punde, P. (2018). Survey on sentiment analysis using Twitter dataset. Proceedings of the Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, pp. 208-211. doi: 10.1109/ICECA.2018.8474783.

[12] Yadav, P., Kashyap, I., & Bhati, B. S. (2023). A systematic survey on deep neural networks for sentiment analysis. Proceedings of the 10th International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India, pp. 1378-1383.

[13] Zahoor, K., Bawany, N. Z., & Hamid, S. (2020). Sentiment analysis and classification of restaurant reviews using machine learning. Proceedings of the 21st International Arab Conference on Information Technology (ACIT), Giza, Egypt, pp. 1-6. doi: 10.1109/ACIT50332.2020.9300098.