



# Sentiment Analysis of Nigerian Opinions Using Logistic Regression and Random Forest Algorithms: A Comparative Study

Victor Mfon Abia <sup>a\*</sup>, E. Henry Johnson <sup>a</sup>  
and Akaninyene B. Obot <sup>b</sup>

<sup>a</sup> Department of Electrical and Electronics Engineering, Akwa Ibom State University, Nigeria.

<sup>b</sup> Akwa Ibom State University, Nigeria.

## Authors' contributions

This work was carried out in collaboration among all authors. All authors read and approved the final manuscript.

## Article Information

DOI: <https://doi.org/10.9734/jerr/2024/v26i101287>

## Open Peer Review History:

This journal follows the Advanced Open Peer Review policy. Identity of the Reviewers, Editor(s) and additional Reviewers, peer review comments, different versions of the manuscript, comments of the editors, etc are available here: <https://www.sdiarticle5.com/review-history/123246>

**Original Research Article**

**Received: 18/07/2024**

**Accepted: 21/09/2024**

**Published: 25/09/2024**

## 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

\*Corresponding author: Email: [vabia.mfon@gmail.com](mailto:vabia.mfon@gmail.com);

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. 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 [1]. 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 [2]. 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 [3]. 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.

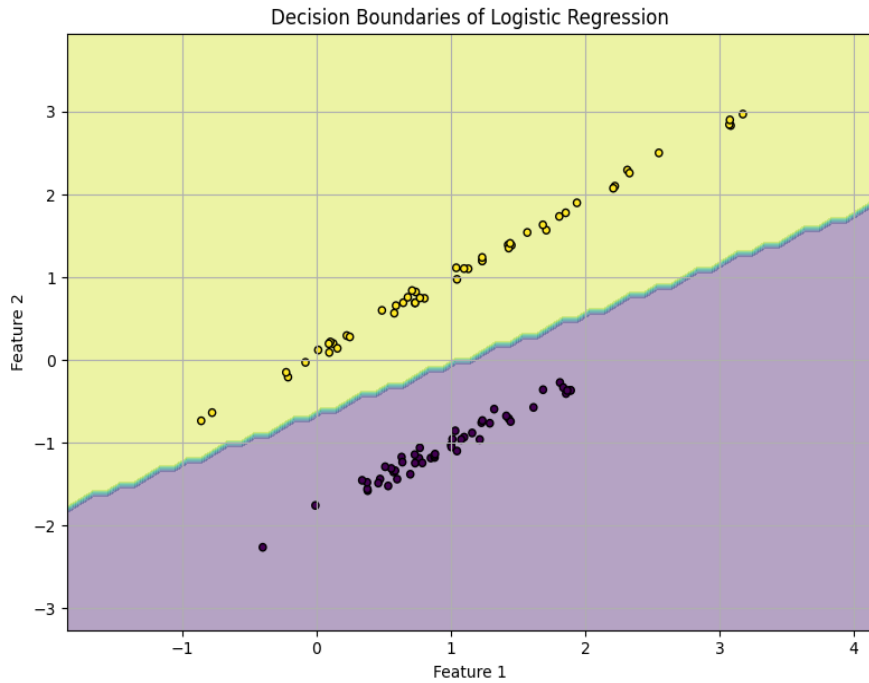
Logistic Regression, a binary classification algorithm, is particularly suitable for sentiment analysis due to its assumption of a linear relationship between input features and the log-odds of the output [4]. This linear decision boundary makes it ideal for problems where sentiment classification can 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 [3]. 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].

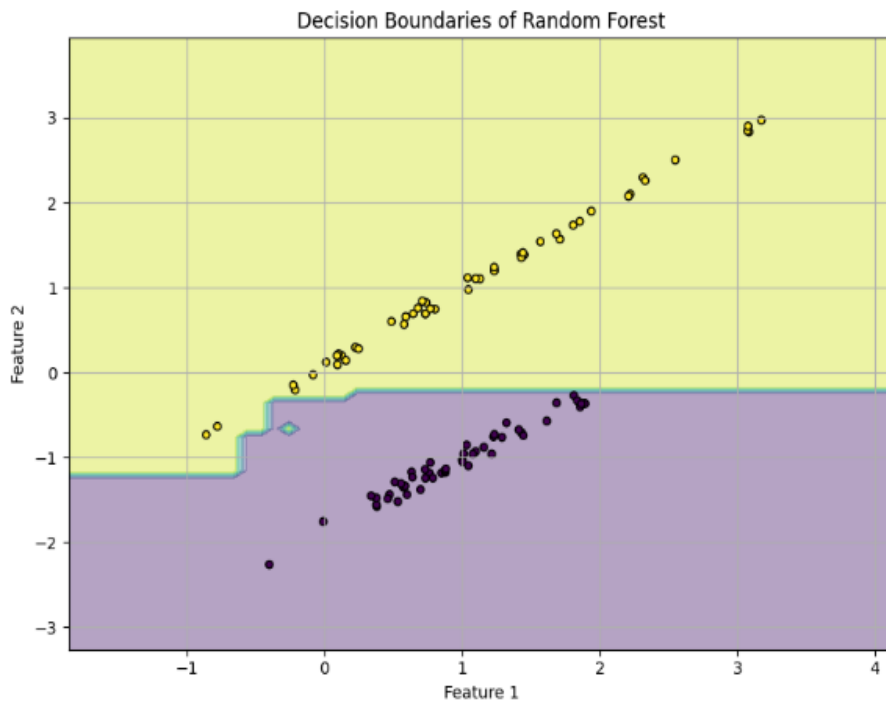
Furthermore, Random Forest provides insights into feature importance, indicating the contribution of each feature to the overall predictive performance. This information is crucial for understanding the most influential words or phrases in sentiment analysis and guiding feature selection efforts. Additionally, 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 [5]. The diagrams below are the decision boundaries generated with synthetic datasets for Logistics Regression and Random Forest.



**Fig. 1. Logistics regression**



**Fig. 2. Random forest**

## 2. RELATED WORKS

Authors in [6] 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 [7] 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 [8] 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. 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 [9]. 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

	A	B	C	D	E	F	G	H	I
1	date	tweet	username	verification	displayname	description	device	location	No. o
2	2023-04-10	Watching ErlamLait		FALSE	'Laitan Fran	Freelance R		Lagos,Niger	
3	2023-04-10	@_uchayyyythefvnta		FALSE	Mr ican doi	I BELIEVE T		Lagos, Nige	
4	2023-04-10	AwwwwwwwPreshy_ous		FALSE	Nwa Nnu 裸	Dental Ther		http://twitter.co	
5	2023-04-10	A big thank Moyo_VIP		FALSE	Moy貌 of Le	We rise by li		Germany	
6	2023-04-10	@forlah_Gasixcenti1		FALSE	shaheed			http://twitter.co	
7	2023-04-10	OBALOLA!!! gidi_9ja		FALSE	GID19JA To	News		Lagos, Nige	
8	2023-04-10	Pls don't kawu90		FALSE	IB Textile	Smiling		Nigeria	
9	2023-04-10	Pls don't kawu90		FALSE	IB Textile	Smiling		Nigeria	
10	2023-04-10	I need Gang dragonfaya		FALSE	Mbamalu	there is not		United Kingd	
11	2023-04-10	@tobibakre MojeedOlad		FALSE	Oladipupo	MA Photojour		Lagos	
12	2023-04-10	That Zlatan tripl_a		FALSE	TRIP 鈇峴	Aspiring		http://twitter.co	
13	2023-04-10	Actually like0xwyz		FALSE	Glucose Guardian			htt...	
14	2023-04-10	@Zlatan_Ibi Big_Ridwan		FALSE	露瑟崇咏础崇	€纱		http://twitter.co	
15	2023-04-10	I got "Ify, theSlim_tee		FALSE	chocbaby 饒	Big baby饒		badan, Nige	
16	2023-04-10	Gangs of Lanadayar		FALSE	Nadayar	Co-Founder		Lagos, Nige	
17	2023-04-10	I knew jadeyosola		FALSE	Miracle no	cGod		Lagos, Nige	
18	2023-04-10	I got " clement_cyr		FALSE	Cyril *Dead	Wok N'Grill		Somewhere	
19	2023-04-10	Naturally, I cCCBenji		FALSE	Waffenmeis	鈇浹ch bin		Ass-gard	
20	2023-04-10	Yhemo leeuglyorubad		FALSE	Sneakers.Ni	Sell Sneake		http://twitter.co	
21	2023-04-10	#WKMUPLwoohleaven		FALSE	Omo Nna 饒	If the fear		http://twitter.co	
22	2023-04-10	@Zlatan_IbiEAtochi		FALSE	Haki	Arsenal    T		Rivers, Nige	
23	2023-04-10	鈇樸angs ofhowafrica		FALSE	How Africa	How Africa		Nigeria	
24	2023-04-10	Gangs of Lamaybatch30		FALSE	Deyemi Lag	Writer, singe		lagos	
25	2023-04-10	I got "Ify, thedatbackyard		FALSE	Nita	Don't		Lagos, Nige	
26	2023-04-10	@soles_fer TheArinola		FALSE	TheArinola	饒嚙饒嘉 饒		Frankfurt on	
27	2023-04-10	It irks me thBABBAFEM		FALSE	Babafemi	Audiophile.		badan   Lag	
28	2023-04-10	@adaugo_miam_harryy		FALSE	饒	Lsg. 饒脛.		http://twitter.co	
29	2023-04-10	@iiamDeji AMissPettyRe		FALSE	PettyBigBab	Don 鈇槲 ser		Rent free in	
30	2023-04-10	The castingBaro_of_Afr		FALSE	Rotimi	Business   A		http://twitter.co	
31	2023-04-10	I got "Teni" -pepple_ibim		FALSE	ibim_pepple	music 饒槩		http://twitter.co	
32	2023-04-10	If I hear anyL_Tomiwa_17		FALSE	饒峻	鈇 MUFC		Stade 17	
33	2023-04-10	My favouriteraychellerec		FALSE	Mocha饒	Nothing nea		Nigeria	
34	2023-04-10	I got "ObalolEyilayoo		FALSE	E 饒y 饒layo	Best in bada		Everything v	
35	2023-04-10	Gangs of Lalekan_Ddon		FALSE	Lanre OG饒	Omo, e be th		Jannah	
36	2023-04-10	Kinda enjoy_cStanley_O		FALSE	The Immort	Perfected th		New World.	
37	2023-04-10	That lord_popula		FALSE	Mr Popular!	Lasuite.		Lagos	
38	2023-04-10	I got "Ify, theAchidex		FALSE	afromandy	FBORN FOR C		Nigeria	
39	2023-04-10	because of OnyiiSequoi		FALSE	ObiaGeri	#sequoiase		http://twitter.co	
40	2023-04-10	I got "Ify, theMunibby		FALSE	Chandler Bir	Feminist		http://twitter.co	
41	2023-04-10	@vhic_tore wikkybee22		FALSE	Fasuyi Abim	I am a quee		http://twitter.co	
42	2023-04-10	Gangs of LaEtoBedlam		FALSE	.	蔚 s蔚 wara		http://twitter.co	
43	2023-04-10	Nigerian filnDoreenglm		FALSE	Doreen饒	饒writer. femir		Los Angeles	
44	2023-04-10	@gentle_pajrickydarey		FALSE	bLoNd_Rick	BIGWIZ		http://twitter.co	
45	2023-04-10	I got "Ify, thelamiscosco		FALSE	Lamide.	multifacete		Africa	
46	2023-04-10	now can webadboyzedd		FALSE	zedd	anatomy 饒		Georgia	
47	2023-04-10	Wow lord_popula		FALSE	Mr Popular!	Lasuite.		Lagos	
48	2023-04-10	Ageb where olu_petcom		FALSE	#EndSars	Follow me a		Nigeria.	
49	2023-04-10	ask me abohecallsmem		FALSE	pink oraimo	proLGBTQ+		http://twitter.co	
50	2023-04-10	@vhic_tore VictorChier		FALSE	Victor Chier	Chelsea FC		http://twitter.co	
51	2023-04-10	Gangs of LaEmperorTwi		FALSE	短悝短悝短悝	Live! 鈇膞膞		Everywhere	
52	2023-04-10	I got "Ify, theobaphemmy		FALSE	Ekundayo	BE KIND AL		Lagos, Nige	

Fig. 3. Gangs of lagos

	A	B	C	D	E	F	G
1		id	text	clean_tweet	created_at	scores	sentiment
2	0	1.55E+18	The	[rejected, st	2022-07-29	0.2897342	positive
3	1	1.55E+18	The Light	[light, come,	2022-07-29	0.4277791	positive
4	2	1.55E+18	Lawrence O	[lawrence, o	2022-07-29	0.3583337	positive
5	3	1.55E+18	@BwalaDan	[afford, hanc	2022-07-29	0.7024115	negative
6	4	1.55E+18	Peter Obi FF	[peter, obi, fi	2022-07-29	0.3714602	positive
7	5	1.55E+18	@ChidiogoE	[obidient, or	2022-07-29	0.3318987	neutral
8	6	1.55E+18	When	[obidientlyyt	2022-07-29	0.5717822	positive
9	7	1.55E+18	Peter Obi is	[peter, obi, q	2022-07-29	0.3395852	negative
10	8	1.55E+18	@Babsolar	(haha, still, r	2022-07-29	0.4108631	positive
11	9	1.55E+18	Be OBidient	[obidient, st	2022-07-29	0.5915479	positive
12	10	1.55E+18	While @Offi	[saying, inhe	2022-07-29	0.3082714	neutral
13	11	1.55E+18	1. The	[1, governm	2022-07-29	0.684451	negative
14	12	1.55E+18	@jesfort_t	([egbo, u, nec	2022-07-29	0.3979866	negative
15	13	1.55E+18	They was a	[time, educa	2022-07-29	0.530725	negative
16	14	1.55E+18	See the face	[see, face, ir	2022-07-29	0.4616942	positive
17	15	1.55E+18	Am	[obidient, w	2022-07-29	0.6162133	positive
18	16	1.55E+18	"We Are The	[structure, o	2022-07-29	0.3619804	negative
19	17	1.55E+18	@renoomok	[main, juju, c	2022-07-29	0.3564359	negative
20	18	1.55E+18	Liars everyw	[liars, everyv	2022-07-29	0.4725714	negative
21	19	1.55E+18	@renoomok	[thought, bbi	2022-07-29	0.4067292	negative
22	20	1.55E+18	A sad comn	[sad, comm	2022-07-29	0.7522888	negative
23	21	1.55E+18	@HAHAYatu	[sir, please, i	2022-07-29	0.6089264	positive
24	22	1.55E+18	My people I	[think, una, e	2022-07-29	0.25392	negative
25	23	1.55E+18	Thank God	[thank, god, 2	2022-07-29	0.5797751	positive
26	24	1.55E+18	To all	[obidient, us	2022-07-29	0.5485727	positive
27	25	1.55E+18	Peter Obi	[peter, obi, fi	2022-07-29	0.3617279	positive
28	26	1.55E+18	I am	[automatica	2022-07-29	0.432456	positive
29	27	1.55E+18	Nigeria欵槌	[nigeria, vici	2022-07-29	0.6731955	negative
30	28	1.55E+18	@PeterObi	[sir, much, a	2022-07-29	0.6036842	positive
31	29	1.55E+18	@Bimbo_of	[abuja, thou	2022-07-29	0.5058554	negative
32	30	1.55E+18	YOU WON'T	[believe, voi	2022-07-29	0.9097149	negative
33	31	1.55E+18	You cannot	[cannot, wis	2022-07-29	0.4437698	positive
34	32	1.55E+18	Dear Youths	[dear, youth	2022-07-29	0.4652055	positive
35	33	1.55E+18	"If ye be will	[ye, willing, c	2022-07-29	0.2529216	negative
36	34	1.55E+18	Down here	[nigeria, say,	2022-07-29	0.3927607	positive
37	35	1.55E+18	@OfficialPD	[please, let,	2022-07-29	0.4466137	negative
38	36	1.55E+18	Aisha	[aisha, yesu	2022-07-29	0.4369968	neutral
39	37	1.55E+18	@Shehusky	[think, dollar	2022-07-29	0.8703759	negative
40	38	1.55E+18	@renoomo	[remember, i	2022-07-29	0.4541576	negative
41	39	1.55E+18	@steveose	[ng01, steen	2022-07-29	0.42921	positive
42	40	1.55E+18	@FS_Yusuf	[yusuf, girl, c	2022-07-29	0.6273055	positive
43	41	1.55E+18	Please let	[please, let,	2022-07-29	0.4466137	negative
44	42	1.55E+18	If someone	[someone, t	2022-07-29	0.2558435	negative
45	43	1.55E+18	@Cappadis	[call, obi, ob	2022-07-29	0.3963775	negative
46	44	1.55E+18	A nation tha	[nation, liste	2022-07-29	0.8213323	negative
47	45	1.55E+18	May the Lor	[may, lord, t	2022-07-29	0.4747134	negative
48	46	1.55E+18	If you're	[willing, obic	2022-07-29	0.2440648	positive
49	47	1.55E+18	@EyonkInfo	[moving, wir	2022-07-29	0.629258	negative

Fig. 4. PeterObi politics

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 [10]. 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.

### 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 [11]. 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:

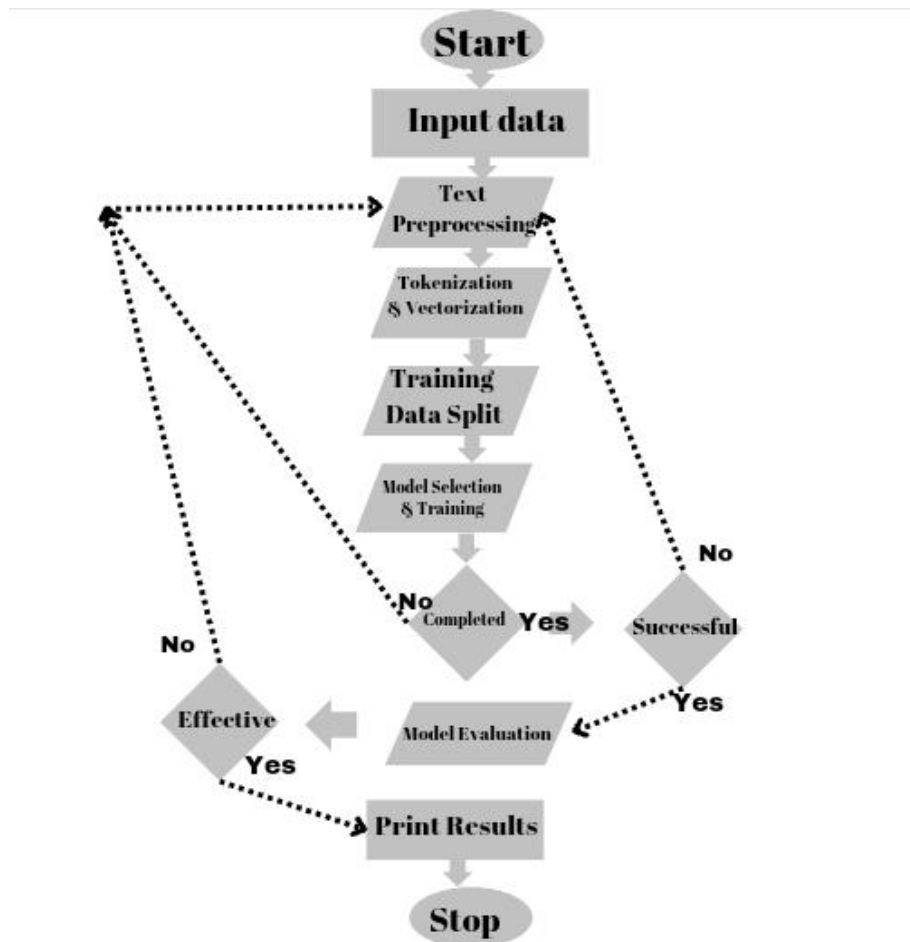


Fig. 5. Flowchart for sentiment analysis operation

#### 4. 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 [13].

**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.

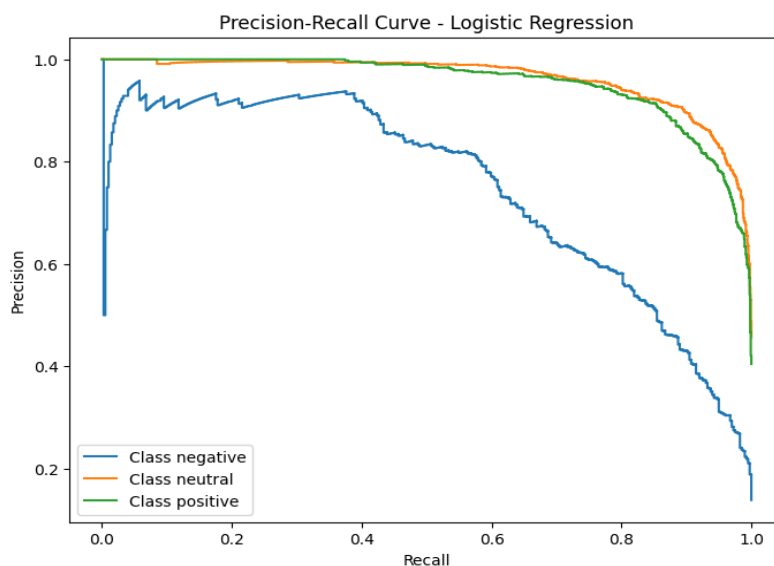
**Support:** Support represents the number of instances in each class in the dataset. It indicates the reliability of the evaluation metrics by providing context about the distribution of classes.

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

For gangs of lagos movie:

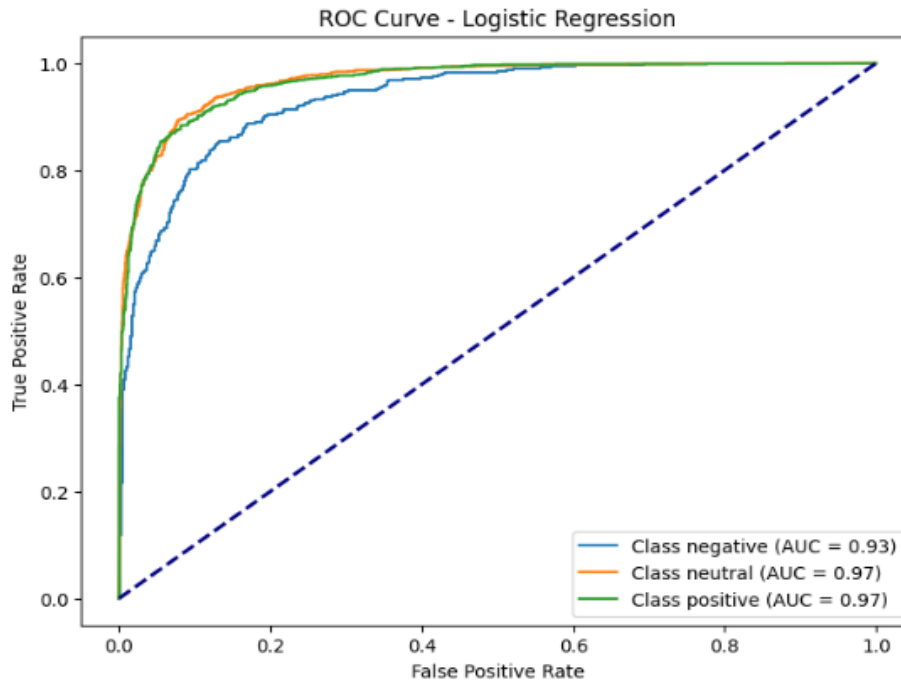
**Table 1. Classification report - logistic regression for gangs of lagos movie**

	<b>Precision</b>	<b>Recall</b>	<b>F1-score</b>	<b>Support</b>
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



**Fig. 6. Precision - recall curves logistics regression**





**Fig. 7. ROC curve logistic regression**

The results above Table 1 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.

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**

	<b>Precision</b>	<b>Recall</b>	<b>F1-score</b>	<b>Support</b>
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.

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.

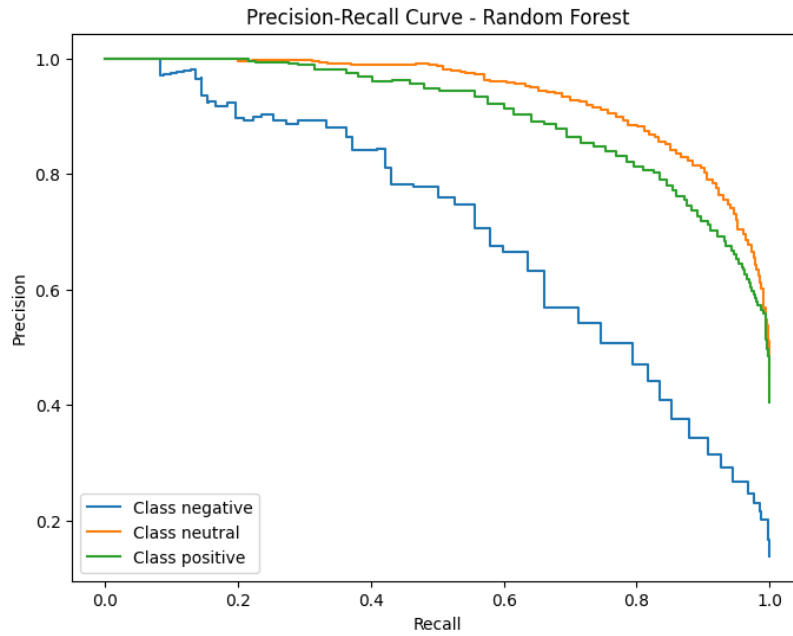


Fig. 8. Precision - recall curves random forest

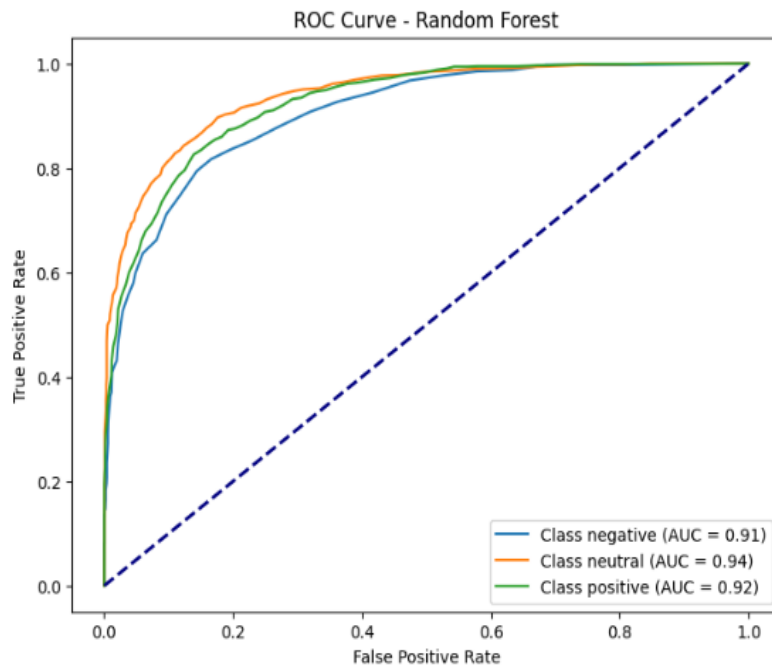


Fig. 9. ROC curve random forest

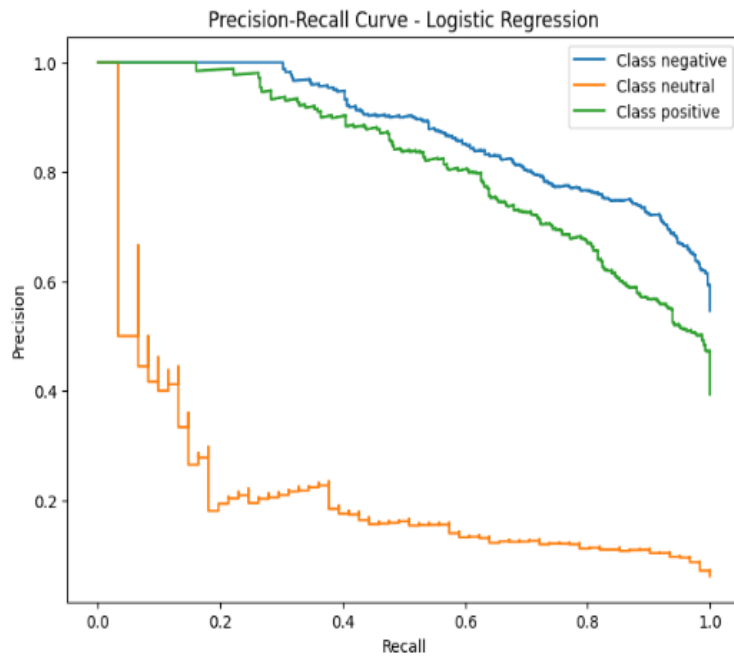
For Peter Obi political sentiment:

Table 3. Classification report - logistic regression for peter obi political sentiment

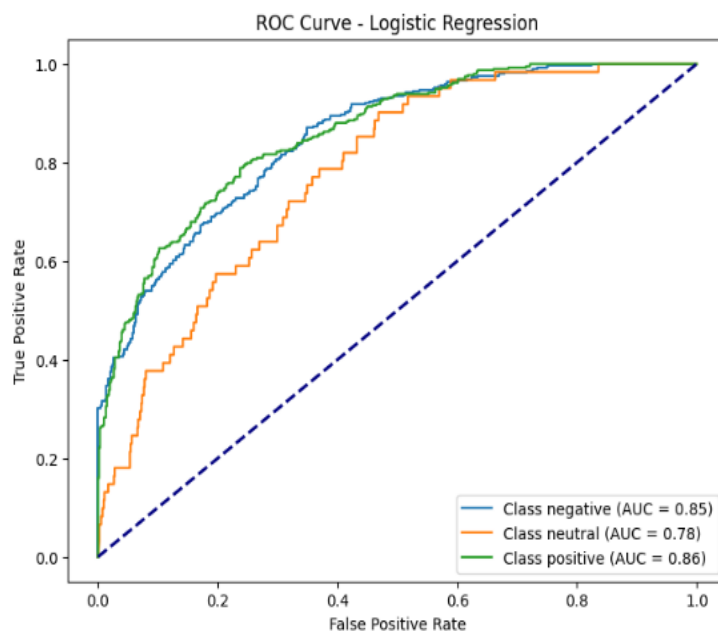
	Precision	Recall	F1-score	Support
Positive	0.76	0.84	0.80	546
Neutral	0.50	0.03	0.06	61
Negative	0.72	0.72	0.72	393

	Precision	Recall	F1-score	Support
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.



**Fig. 10. Precision - recall curves logistic regression**



**Fig. 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.

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

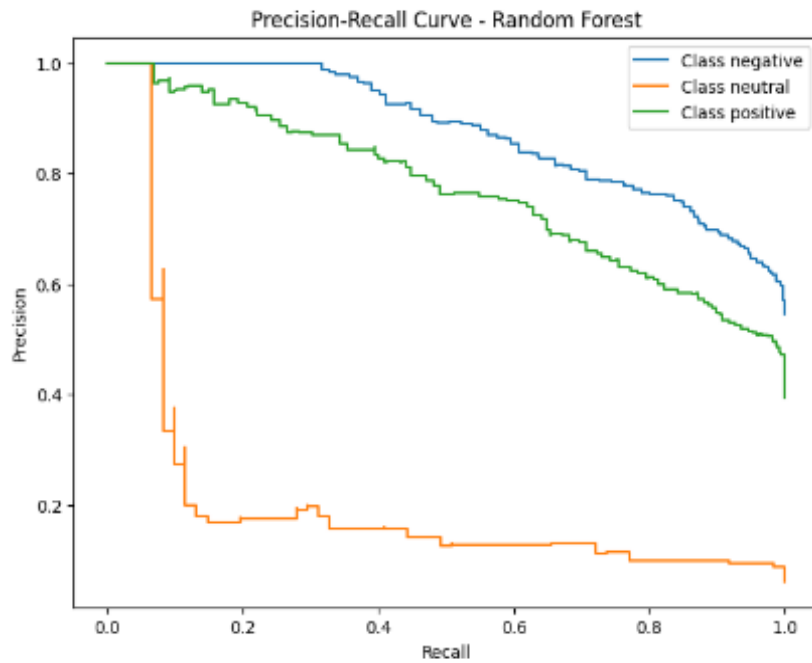


Fig. 12. Precision - recall curves random forest

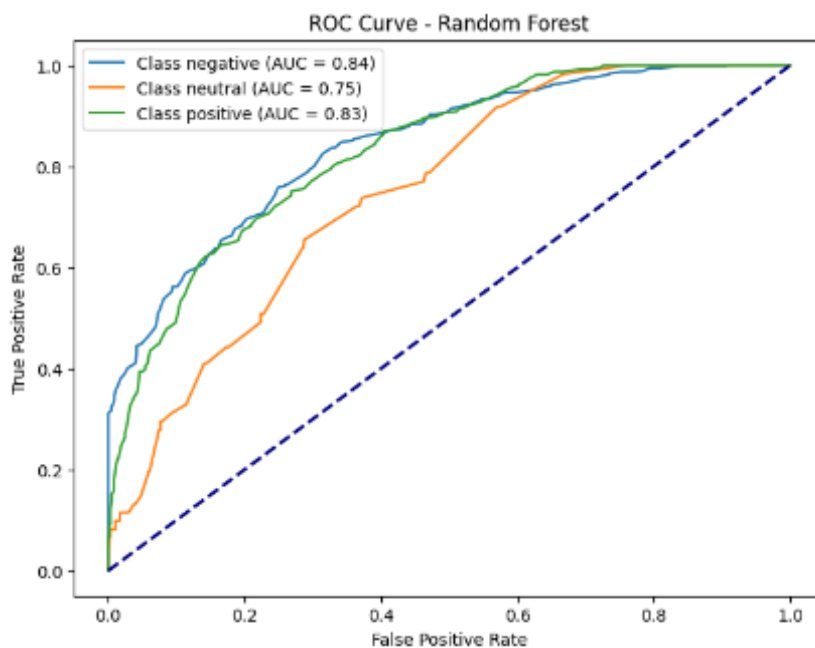


Fig. 13. ROC curve random forest

**Table 4. Classification report - random forest for peter obi political sentiment**

	<b>Precision</b>	<b>Recall</b>	<b>F1-score</b>	<b>Support</b>
Positive	0.75	0.82	0.79	546
Neutral	0.67	0.07	0.12	61
Negative	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

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. 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 this manuscript.

## COMPETING INTERESTS

Authors have declared that no competing interests exist.

## REFERENCES

1. Sindhu S, Kumar S, Noliya A. A review on sentiment analysis using machine learning. Proceedings of the International Conference on Innovative Data Communication Technologies and Application (ICIDCA), Uttarakhand, India. 2023;138-142. DOI:10.1109/ICIDCA56705.2023.1009966
2. Rawat A, Maheshwari H, Khanduja M, Kumar R, Memoria M, Kumar S. 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. 2022; 225-230. DOI:10.1109/COM-IT-CON54601.2022.9850629
3. Zahoor K, Bawany NZ, Hamid S. Sentiment analysis and classification of restaurant reviews using machine learning. Proceedings of the 21st International Arab Conference on Information Technology (ACIT), Giza, Egypt. 2020;1-6. DOI: 10.1109/ACIT50332.2020.9300098
4. Khanvilkar G, Vora D. Smart recommendation system based on product reviews using random forest. Proceedings of the International Conference on Nascent Technologies in Engineering (ICNTE), Navi Mumbai, India. 2019;1-9. DOI: 10.1109/ICNTE44896.2019.8945855.
5. Phulare P, Deshmukh SN. 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. 2021;141-145.  
DOI: 10.1109/CENTCON52345.2021.9688197
6. Dobhal DC, Kumar B, Das P. Involvement of functional programming in language processing and machine learning. Proceedings of the 2nd International Conference for Innovation in Technology (INOCON), Bangalore, India. 2023;1-4.  
DOI: 10.1109/INOCON57975.2023.10101253
  7. Parashar D, DSilva M, Kulshreshtha S. A machine learning framework for stock prediction using sentiment analysis. Proceedings of the 4th IEEE Global Conference for Advancement in Technology (GCAT), Bangalore, India. 2023;1-5.  
DOI: 10.1109/GCAT59970.2023.10353541
  8. Wagh R, Punde P. Survey on sentiment analysis using Twitter dataset. Proceedings of the Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India. 2018;208-211.  
DOI: 10.1109/ICECA.2018.8474783.
  9. Christian PH, Desanti RI. 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. 2022;01-04.  
DOI: 10.1109/ICIC56845.2022.10006984
  10. Harjadinata A, Sibaroni Y. 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. 2022;261-266.  
DOI: 10.1109/ICoSEIT55604.2022.10029958
  11. Subramanian RR, Akshith N, Murthy GN, Vikas M, Amara S, Balaji K. A survey on sentiment analysis. Proceedings of the International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India. 2021;70-75.  
DOI: 10.1109/Confluence51648.2021.9377136.
  12. Yadav P, Kashyap I, Bhati BS. 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. 2023;1378-1383.
  13. Nazir A, Rao Y, Wu L, Sun L. Issues and challenges of aspect-based sentiment analysis: A comprehensive survey. IEEE Transactions on Affective Computing. 2022;13(2):845-863.  
DOI: 10.1109/TAFFC.2020.2970399

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of the publisher and/or the editor(s). This publisher and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

© Copyright (2024): Author(s). The licensee is the journal publisher. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Peer-review history:

The peer review history for this paper can be accessed here:  
<https://www.sdiarticle5.com/review-history/123246>