# A Text Filtering Based Approach for Multiple Language Sentiment Collection Considering Nigeria's Twitter Users

**Enesi Femi Aminu[1], Oluwaseun Adeniyi Ojerinde[1], Ayobami Ekundayo[1], Abdussamad Muhammad Jamiu[1], Opeyemi Aderike Abisoye[1], Hussaini Abubakar Zubairu[2]**

[1] Department of Computer Science, Federal University of Technology, Minna, Niger State.
[2] School of Information Technology, Carleton University, Canada.

## Abstract

In this digital age, there is a widespread trend and desire for people to have a social presence across a variety of social media platforms. Nigeria, for instance, is a multilingual country that aspires to have a social presence in the media, such as Twitter, for important languages including Hausa, Igbo, and Yoruba. This does not come without creating a research challenge for the sentiment analysis (SA) algorithms that are already in use owing to the complex nature of text data and filtration strategy adopted. Thus, this research aims to use text-filtering approach to improve the accuracy of the current model. This study made use of the African Language Bidirectional Encoder Representations from Transformers (AfriBERTa) language model, which was created especially for African languages by eliminating terms that are common to several sentiment classes. The algorithm's performance across the chosen languages is compared for both filtered and unfiltered datasets, and the results based on accuracy for Pigin for unfiltered is 0.69 and filtered is 0.75; accuracy for Hausa for unfiltered is 0.75 and filtered is 0.79. Similarly, accuracy for Yoruba for unfiltered is 0.75 and filtered is 0.80; while accuracy for Igbo for unfiltered is 0.77 and filtered is 0.76. These results show that the filtration strategy generally improves in terms of accuracy, precision, recall, and F1-scores. This implies that for efficient sentiment analysis in a variety of linguistic contexts, these customized data pretreatment approaches are essential because the proposed technique aids to improve sentiment classification. In addition to emphasizing the value of context-specific approaches in SA, this research lays the groundwork for future developments in multilingual sentiment analysis, which could find useful in a number of fields such as public opinion analysis and market research.

## Keywords:

Text Filtering Approach, Sentiment analysis, Twitter, Multiple languages, AfriBERTa

Submitted:

Accepted:

Published:

DOI:



## Highlights:

- To use text-filtering approach to improve the accuracy of the current model
- This study made use of the AfriBERTa language model
- The proposed technique helps to improve sentiment classification
- Nigeria, for instance, is a multilingual country that aspires to have a social presence in the media, such as Twitter, for important languages including Hausa, Igbo, and Yoruba.

## 1. Introduction

Sentiment analysis, often known as opinion mining, is the practice of using natural language processing to detect, extract, and organize sentiment from user-generated texts in blogs, social networks, and product reviews (Abdullahi et al., 2024). The technique known as sentiment analysis (SA) uses computers to recognize and categorize viewpoints, particularly in order to ascertain if the author is feeling positively, negatively, or neutrally about a certain subject (Raychawdhary & Das, 2023). There are three basic approaches used for sentiment analysis (SA) on linguistic data: the Lexicon based techniques, Machine learning technique and the Hybrid technique which requires the two techniques combined together to get better classification results (Sani et al., 2022). Opinion mining, also known as sentiment analysis, is another term for the subtle, complex ways in which emotions are represented in content. In addition to being expressed directly by the user in relation to a particular topic, users can express their emotions through a variety of other means (Chinyere et al., 2021).

Derogatory remarks, commonly known as hate speech, are characterised as "any expression that disparages an individual or a group based on certain attributes like race, color, ethnicity, gender, sexual orientation, country, religion, or other features." (Oladipo et al., 2020). The social media landscape has developed into incredibly intricate information-sharing platforms. As a result of these platforms' growing practices, interest in sentiment analysis as a paradigm for the mining and analysis of user opinions and sentiments based on their posts has grown (Oladipo et al., 2022). The identification and categorization of sentiment in texts is the focus of sentiment analysis (SA). Because SA has so many important applications, it has garnered a lot of attention in recent years. However, high-resource languages like English receive the majority of attention in SA research, whereas languages with little data are still underrepresented (Muhammad et al., 2021).

Consequently, it is imperative that hate speech be addressed and combated. This calls for an all-encompassing strategy that mobilizes society at large. It is the moral obligation of individuals and institutions, such as governments, businesses, media, Internet companies, religious leaders, schools, youth, and civil society, to vehemently denounce hate speech and to play a significant role in eradicating this evil (Onuora et al., 2024). There is an increase in internet and social media users. Social media is an online communication platform that lets users share and publish their activities with a network of others. In daily life and business, social media is extremely important, especially when it comes to online marketing and product promotion (Akuma et al., 2021).

In order to facilitate seamless collaboration, communication, microblogging, networking, socializing, and information sharing among geographically dispersed people, Twitter was developed. Lately, however, it has also been used to support businesses by providing a means of connecting with and keeping a large number of clients (Mutanga et al., 2022). Natural language processing finds it difficult to explore sentiments in social media due to the variety and complexity of dialect expression, noisy terms such as acronyms, emoticons, and abbreviations, as well as spelling errors and the constant stream of updated content (Chinedum & Ogochukwu, 2021).

Twitter has more than 300 million users worldwide as of October 2020; 91% of these users are over the age of 18. Politicians are drawn to the platform because it allows them to interact and use it as a tool for their campaigns. It is also a valuable resource for researchers looking to predict elections because it offers an API that allows users' public information and connections to be extracted (Khan et al., 2021).

The sentiment analysis of Twitter data from Nigeria presents a unique challenge due to the multilingual nature of the country. Existing sentiment analysis techniques often struggle to accurately analyze sentiments expressed in various Nigerian languages alongside English Language owing to adopted or adapted approach on how the data is filtered or not even filtered at all. However, this research aims to address the text filtering approach in capturing sentiments from multiple languages found on Nigerian Twitter, thereby contributing to the development of a more comprehensive sentiment analysis framework through the text filtering approach.

Finally, the results of the proposed sentiment collection based on text filtered approach for four Nigeria’s languages on Twitter social media were evaluated against text unfiltered approach. The remaining sections of this paper is organized as follows: section 2 gives account of the existing related literature to the subject matter, while the methodology and methods employed were accounted for by section 3, section 4 discussed the results obtained and finally the conclusion section.

## 2. Related Works

Arabic text sentiment analysis is a crucial task for many commercial applications, including Twitter. However, the study of Abo et al., (2021) presents a multi-criteria approach to evaluate and rank Arabic sentiment analysis classifiers empirically. Distinguished machine learning algorithms were utilized to construct Arabic sentiment analysis classifier classification models. In addition, a performance evaluation of the top five machine learning classifiers was examined in order to rank the classifier's performance. The evaluation measures of machine learning classifiers, including accuracy, recall, precision, F-measure, CPU time, classification error, and area under the curve (AUC), were merged with the top five ranking techniques. Five well-known classifiers were compared using Saudi Arabic product reviews to test the approach. Based on findings, it appears that the deep learning and support vector machine (SVM) classifiers outperform the others in the following areas: recall 88.41%, 83.89; F-measure 86.81, 83.87%; classification error 14.75, 17.70; and AUC 0.93, 0.90, respectively; accuracy 85.25%, 82.30%; precision 85.30, 83.87%. They perform better than Naïve Bayes classifiers, K-nearest neighbors (K-NN), and decision trees.

Opinionated content about a wide range of goods, services, occasions, and political parties is actively created by users in multiple languages as a result of social media's ongoing and rapid growth. The study of Abubakar et al., (2021) presents an enhanced feature acquisition method (EFAM) for multilingual sentiment analysis of English and Hausa tweets. In order to measure classification performance and create a more accurate sentiment classification process, the method integrates two newly created Hausa features (Hausa Lexical Feature and Hausa Sentiment Intensifiers) with an English feature using a machine learning methodology. The effectiveness of the approach in improving feature integration for multilingual sentiment analysis has been assessed through a series of experiments involving various classifiers in both monolingual and multilingual datasets. Similarly, by utilizing features derived from multiple languages, we can create machine learning classifiers with an average precision of over 65%.

Due to its usefulness in extracting users' thoughts, attitudes, and emotions from big textual data sets, sentiment analysis has attracted a lot of study attention lately. However, this study of Adewole et al., (2021) proposes a hybrid feature selection framework based on the fusion of filter- and wrapper-based feature selection techniques. To determine which feature subsets are most discriminative for sentiment analysis, Boruta and Recursive Feature Elimination (RFE) are hybridized with Correlation Feature Selection (CFS). The performance of the suggested hybrid feature selection approach is assessed using four publicly accessible sentiment analysis datasets: Amazon, Yelp, IMDB, and Kaggle. To determine whether the suggested method is better, this study compares the performance of three classification algorithms: Random Forest, Naïve Bayes, and Support Vector Machine (SVM). The study's datasets, which illustrate various scenarios, demonstrate that the combination of CFS and Boruta yielded encouraging outcomes, particularly when the chosen features were fed into the Random Forest classifier. Predicting users' opinions and emotions effectively while considering predictive accuracy is made possible by the hybrid framework that has been proposed. Because of the suggested hybrid feature selection framework, the final model has a shorter computing time.

By gathering tweets from Nigerian university students on Twitter regarding their opinions of the country's existing educational system, the research aims to investigate data mining tools. The tweet data gathered in the study of Alade & Nwankpa, (2022) was pre-processed using the Twitter application before being converted from text to vector form using a feature extraction method like Bag-of-Words. The Naïve Bayes classifier (NBC) approach, a straightforward yet efficient classifier to ascertain the polarity of the education dataset, was used in the paper's proposed sentiment analysis architectural design to compute the class probabilities. Moreover, the Naïve Bayes classifier classified the tweets' polarity as positive or negative based on their phrasing. The assessment metrics used in this study are: precision, recall, F1-score, and balanced accuracy, for the models' evaluation because they were trained on unbalanced data. The classifier's prediction accuracy, misclassification error rate, recall, precision, and f1-score were 63 %, 37%, 63%, 62%, and 62% respectively.

The gathering and examination of unstructured textual data enables decision-makers to examine the increasing number of posts and comments on our social media networks. Hence, to get past the noise and unreliability of these unstructured datasets from digital media platforms, autonomous big data analysis is required. The machine learning algorithms in use today, however, are performance-driven and concentrate on the accuracy of categorization and prediction utilizing properties that have been learned from training samples. In the study of Asogwa et al., (2021) two supervised machine learning algorithms

are combined with text mining techniques to produce a hybrid model which consists of Naïve Bayes and support vector machines (SVM). Additionally, the system offers an open forum where people with similar interests may exchange messages and comments, with the comments being automatically categorized as either lawful or illegal. As a result, user conversations are of higher quality. Using Java programming language and WEKA tools, the hybrid model was created. According to the results, the hybrid model produced an accuracy of 96.76%, whereas the Naïve Bayes and SVM models produced accuracy of 61.45% and 69.21%, respectively.

The absence of a generic architecture, imprecision, threshold settings, and fragmentation problems are the main obstacles to automated hate speech classification on Twitter. Therefore, in order to address issues with hate speech classification, a probabilistic clustering model for Twitter was proposed by Ayo et al., (2021) . Using a metadata extractor, tweets containing hate speech keywords were gathered, and crowdsourcing experts then labeled the collected hate tweets into two groups: opposing and complementary discourse. Using the Term Frequency-Inverse Document Frequency (TF-IDF) model, features were represented and then further augmented with themes that were inferred by a Bayes classifier. Real-time tweets were automatically classified into the appropriate topic categories using a rule-based clustering technique. The classification of hate speech was then accomplished by fuzzy logic employing a score computation module and semantic fuzzy rules. The evaluation results showed that, when utilizing a 5-fold cross validation, the constructed model outperformed the others in terms of hate speech identification, with an F1-score of 0.9256. Comparing the generated model to comparable models, the hate speech categorization model scored better, with an F1-score of 91.5. Comparing the constructed model to other comparable approaches, it also shows an even more flawless test with an AUC of 0.9645. The effectiveness of the created model for the classification of hate speech was confirmed using the Paired Sample t-Test.

Despite the extensive usage of computer tools in Nigerian educational institutions, online collaboration tools are rarely or never used for academic planning in postsecondary educational establishments. Therefore, the purpose of the study of Efuwape et al., (2022) is to use text mining to extract emotions from the viewpoints of stakeholders in the academic research business regarding the practical applications of collaborative tools for academic planning. A sentiment analysis method based on VADER is modeled in the natural language processing use case's opinion mining study. The uni-gram and bi-gram tokenized dictionary of known words were assigned negative, positive, neutral, and compound values. Thus, the experimental result reveals a mean sentiment negative score of -0.10, which represents a 17.27% cluster of respondents who are not favorably disposed to the idea, while a 22.7% cluster of highly convinced respondents express positive sentiments about the use of collaborative tools with a mean sentiment score of 0.49. With a mean score of 0.39, a 60.01% cluster of average respondents who indicated neutral thoughts really leans toward a positive mood.

These days, the task of labeling textual data from a set of theme labels has gained significant importance. The purpose of Ibrahim et al., (2021) is to thoroughly compare, for text categorization, the performance of several ensemble learning strategies with foundational supervised machine learning techniques. We employed two encoding paradigms for feature engineering in the experiments: the Term Frequency Inverse Document Frequency, or TF-IDF, and the Bag-of-Words approach. The supervised learning algorithms (ensemble learning techniques, among others) received the effective feature vectors that were obtained as input. Therefore, using 5-fold and 10-fold cross-validation, each algorithm was trained on the YouTube Spam Collection Dataset. The findings demonstrate that Adaboost and LightGBM perform better on both assessments than alternative methods. The findings suggest that some Ensemble learning strategies frequently produce superior text categorization results as compared to the foundational techniques analyzed.

The pervasive utilization of social media platforms, including Facebook, Instagram, LinkedIn, and Twitter, has significantly influenced everyday human interactions and decision-making. In order to reduce context loss, the study of Ijairi et al., (2023) proposes a text pre-processing method that looks at negation words and emoji characteristics in text data by translating these attributes into single contextual words in tweets. Using four deep learning algorithms Long Short-Term Memory (LSTM), Recurrent Neural Network (RNN), Artificial Neural Network (ANN), and others. The suggested preprocessing was assessed on benchmark Twitter datasets. LSTM outperformed previously stated techniques in the literature, as evidenced by its accuracy of 96.36%, 88.41%, and 95.39%.

Complexity in language, unstructured data, and ambiguity make semantic processing of social media data difficult. The study of Kolajo et al., (2021) proposed the Social Media Analysis Framework for

Event Detection (SMAFED). Improved representation of social media stream content, enhanced summarization of event clusters in social media streams, and enhanced semantic analysis of noisy phrases in social media streams are the three main goals of SMAFED. We used important ideas like semantic representation of social media streams, integrated knowledge base, ambiguity resolution, and Semantic Histogram-based Incremental Clustering based on semantic relatedness to achieve this. Two assessment trials were carried out to verify the methodology. Initially, the effect of SMAFED's data enrichment layer. On the first dataset, SMAFED fared better than other pre-processing frameworks, with a lower loss function of 0.05, while on the second, it was0.15. Second, SMAFED's efficacy ascertained in identifying events from social media feeds. This second experiment's outcome demonstrated that SMAFED performed better than other event detection techniques, with improved Precision (0.922), Recall (0.793), and F-Measure (0.853) metric scores.

Because Pidgin and English are mixed together with regional cultural allusions, identifying irony in Pidgin tweets presents particular difficulties. In order to overcome these obstacles, the study of Ladoja & Afape,( 2024) intends to develop a model for precise sarcasm identification in Pidgin tweets. Accuracy, precision, recall, and F1-score metrics were used to evaluate the effectiveness of Vanilla Artificial Neural Network (ANN), XGBoost, Random Forest, and logistic regression classifiers on sarcasm data that was gathered by curating and pre-processing a dataset of Nigerian Pidgin tweets. With an accuracy of 85.78%, precision of 88.57%, recall of 94.44%, and F1-score of 91.41%, the XGBoost model showed impressive performance. These results demonstrated how well the model could distinguish between sardonic and non-sarcastic utterances.

Numerous established and emerging types of racism have been observed on social media due to its dominant position in the socio-political sphere. The study of Lee et al., (2022) uses sentiment analysis of tweets to identify those that include racist language. Due to deep learning's improved performance, gated recurrent units (GRU), convolutional neural networks (CNN), and recurrent neural networks (RNN) are combined to create a stacked ensemble deep learning model known as Gated Convolutional Recurrent-Neural Networks (GCR-NN). In the GCR-NN model, GRU is the best at extracting relevant and salient characteristics from unprocessed text, while CNN gathers data that is crucial for RNN to provide precise predictions. Naturally, a number of experiments are carried out to look into and assess the suggested GCR-NN's performance in relation to machine learning and deep learning models, showing that GCR-NN performs better with an increased accuracy of 0.98. 97% of tweets with racist remarks can be identified using the suggested GCR-NN model.

Particularly during elections, social media sites like Facebook, LinkedIn, and Twitter have been utilized as a means of organizing demonstrations, conducting surveys, developing campaign agendas, creating agitation, and serving as a forum for the expression of opinions. The research of Olabanjo et al., (2023) uses the Twitter dataset to create a Natural Language Processing framework that will analyze popular opinion in the Nigerian presidential election of 2023.

Sentiment analysis was carried out on the preprocessed dataset using three machine learning models: Long Short-Term Memory (LSTM) Recurrent Neural Network, Bidirectional Encoder Representations from Transformers (BERT), and Linear Support Vector Classifier (LSVC) models. Two million tweets with eighteen features were collected from Twitter containing public and personal tweets of the three top contestants Atiku Abubakar, Peter Obi, and Bola Tinubu. The sentiment models reported results based on the evaluated metrics and the models used.

As social media usage increases, some unstructured data has been available. If data is cleaned, organized, and analyzed, it can be useful. The research of Popoola et al., (2024) suggests a brand-new machine learning-based method for sentiment analysis of social media data analysis. There are three steps in the method that is being given. Preprocessing is the first step, during which the tweets are filtered and polished. Using the Inverse Document Frequency (TF-IDF) and Term Frequency, features were extracted in the second stage. In the third stage, machine learning techniques are used to predict utilizing the features that have been extracted.

The study included three machine learning models: k-nearest neighbor (KNN), naive bayes (NB), and random forest classifier (RF). The evaluation findings demonstrate that in terms of accuracy, precision, recall, and F1-score measures, both NB and RF outperform KNN. These findings also demonstrate a resoundingly favorable perception of financial news.

By adopting the novel text filtering approach, this research seeks to bridge the gap between existing sentiment analysis techniques and the diverse linguistic landscape of Nigeria. The reviewed

literature present different techniques for sentiment analysis on one or two languages. However, this study presents an adoption of a new text filtering strategy for four Nigeria's languages by comparing the results with an unfaltering approach.

## 3. The Methodology: Text Filtering Based Approach

An iterative hybrid design approach is considered in this study that is, top down and bottom up approaches. The bottom up approach first consider the collection of the textual twitter data for the four languages. The next segment of the approach was the preprocessing activity of the data, which include removal of stop words and emoji. The last segment of this approach was to pass the supposedly clean twitter text for model training using the AfriBERTa. Conversely, the stages of the design or modification could be considered in reverse order from model training down to reexamination of the datasets obtained hence, the top down approach. The suggested approach is centered on sentiment analysis using text-filtering method. The Jupyter Notebook IDE is used to train the suggested model. The programming language Python 3.11 was utilized in this study. It is essential to install different Python modules as dependencies in order to guarantee effective implementation. These modules include Matplotlib for graphical data visualization, which includes histograms, bar charts, scatter plots, and pie charts; NumPy for data manipulation in arrays; SkLearn and Tensorflow for machine learning models; and Pandas for tabular data visualization. Figure 1 shows the conceptual framework of the proposed model.

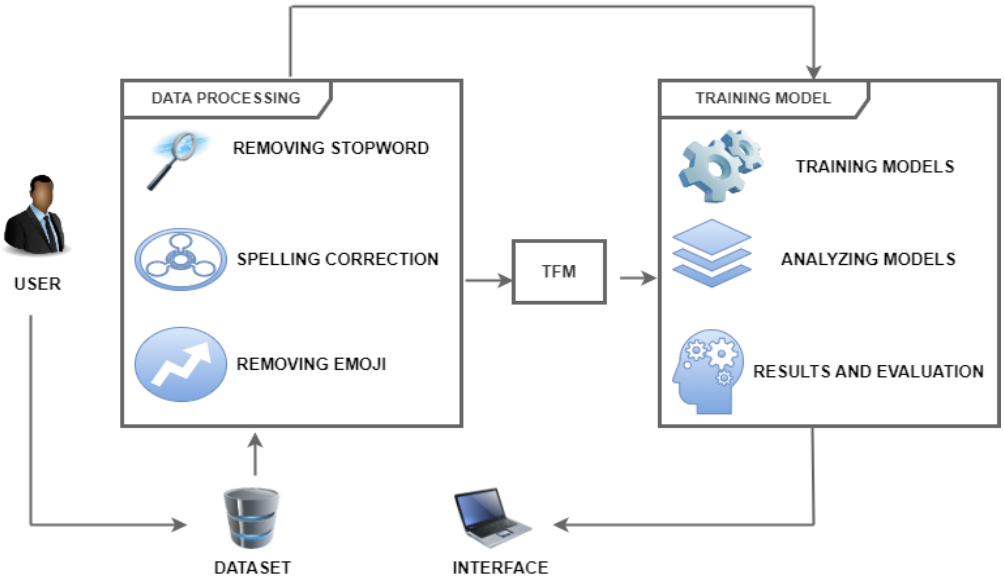


**Figure 1.** Conceptual Framework of the Proposed Model

The structure of the framework is represented by Figure1, which shows how the user obtained the dataset from the GitHub repository. Subsequently, the dataset undergoes preprocessing steps like as data cleaning, stop word removal, spelling correction verification, and lemmatization. Additionally, the dataset is subjected to the enhanced text filtering method before language models are used and the dataset is trained. Test data is used to evaluate the trained models, and accuracy values are obtained. Each structure of the framework is explained in details as follows.

### *3.1 Dataset Used*

The datasets used in this study is the same dataset used in the study of Muhammad et al., (2022). The dataset was obtained from GitHub and the dataset is named "NaijaSenti2". GitHub is a web-based platform that provides hosting for software development and version control using Git. It allows developers to collaborate on projects, track changes to code, and manage software development workflows efficiently. GitHub offers features like code repositories, issue tracking, pull requests, and project management tools

to facilitate collaboration among developers and teams. The NaijaSenti2 is an open-source Twitter sentiment dataset for the four most spoken languages in Nigeria Hausa, Igbo, Pidgin, and Yorùbá. This is the largest labelled sentiment dataset in these languages to date. As the Twitter API does not support these languages, a method to enable the collection was proposed, filtering, and annotation of such low-resource language data. Overall, 30,000 tweets in Hausa, Igbo, Yorùbá and Nigerian Pidgin (also known as Naija) were annotated.

### *3.2 Data Preprocessing*

The text data is preprocessed before the Language Model is used to train the datasets. During the preparation stage, each dataset will have its stop words removed, punctuation removed, text converted to lowercase, and spelling problems corrected. The remaining words in the texts are then downsampled, and the resulting text document is utilized to train the language model on each dataset.

Typical problems like missing words and overuse of letters in words are typical in the world of social media texts. Therefore, before supplying the text document into any model, it is imperative to make sure that the text spelling correction process is completed.

### *3.3 Text Filtering Method*

The text filtering method (TFM) removes words that elicit opposite emotions in order to improve the accuracy of positive, negative, and neutrally labeled data in datasets. First, preprocessed labels for positive, neutral, and negative texts are divided into various groups within the training set. Every word's frequency in texts with labels both positive and negative is tallied separately. Positively labeled texts' 'k' most frequent terms are chosen to be removed from negatively labeled texts, and vice versa. The goal of this procedure is to eliminate words from both positive and negative sentences that elicit opposing feelings. It is important to remember that some of the top "k" words may be neutral terms that don't express any emotion. Moreover, these top 'k' words appear in both texts with positive and negative labels. Removing these terms contributes to a smaller dataset overall.

### *3.4 Training Models*

The datasets sentiment is analyzed using the AfriBERTa language model. These language models include customized models created for the primary language of Nigeria. After the model has been trained, it is examined to assess its accuracy and evaluate the outcomes. The processes involved in the overall model is presented by the algorithmic design of Table 1.

**Table 1:** Algorithmic Design for the Proposed Text Filtering Approach

***Input:*** *t*
***Output:*** *results: o_d, f_d*
***Parameters:*** *datasets(t), google drive(gd), text(t), sentiment label data(sld), data cleaning(dc), stopwords(sw), validation(vn), test sets(ts), AfriBERT Tokenizer(a_t), AfriBERTa model(a_m), optimizer(o), loss function(l_f), metrics(m), confusion matrix(c_m), classification report(c_r), top words(t_w), sentiment class(s_c), filteration process(f_p), tokenizatio (t_n), model training(m_t), filtered dataset(f_d), original dataset(o_d), language dataset(l_d).*

**Procedure:**

```
1.  Initialize and import libraries
2.  Input t
3.  For t != n
4.      Read t and perform(sld)
5.      Preprocess(sld)
6.          If sw !=Ø
7.              Preprocess(sw)
8.          EndIf
9.      Initialize a_t = 0
10.     totenizeTest(sw)Define a function 't_t'
11.     Load a_m with o, l_f,
12.     Train m on each l_d
13.         Evaluate a_ m using c_m and c_r
14.         Extract t_w from s_c in t
15.         Apply f_p to remove t_w from t
16.       Repeat t_n and m_t for f_d
17.      Evaluate performance of m on f_d
18.      Returns a_m on o_d and f_d
19. EndLoop
```

From lines 1 to 8 of Table 1 the python libraries are imported and initialized. The datasets obtained from google drive are inputted. The inputted dataset is tested to begin the loop and perform sentiment label data from multiple TSV files for different languages. The sentiment label data is read and function is performed to preprocess the data if it contain stop words. The results that would be passed to the model (AfriBERT) is initialized as indicated by line 9. From lines 10 to 18, the clean data is loaded and trained. Results are evaluated based on the original and filtered datasets.

In addition, Figure 2 depict further the proposed model based on sequence diagram. Within the framework of UML (Unified Modeling Language), a sequence diagram is similar to a visual narrative that illustrates how various system components interact with one another in a predetermined order. Understanding who is interacting in a system, with whom, and when is made possible with this excellent tool.

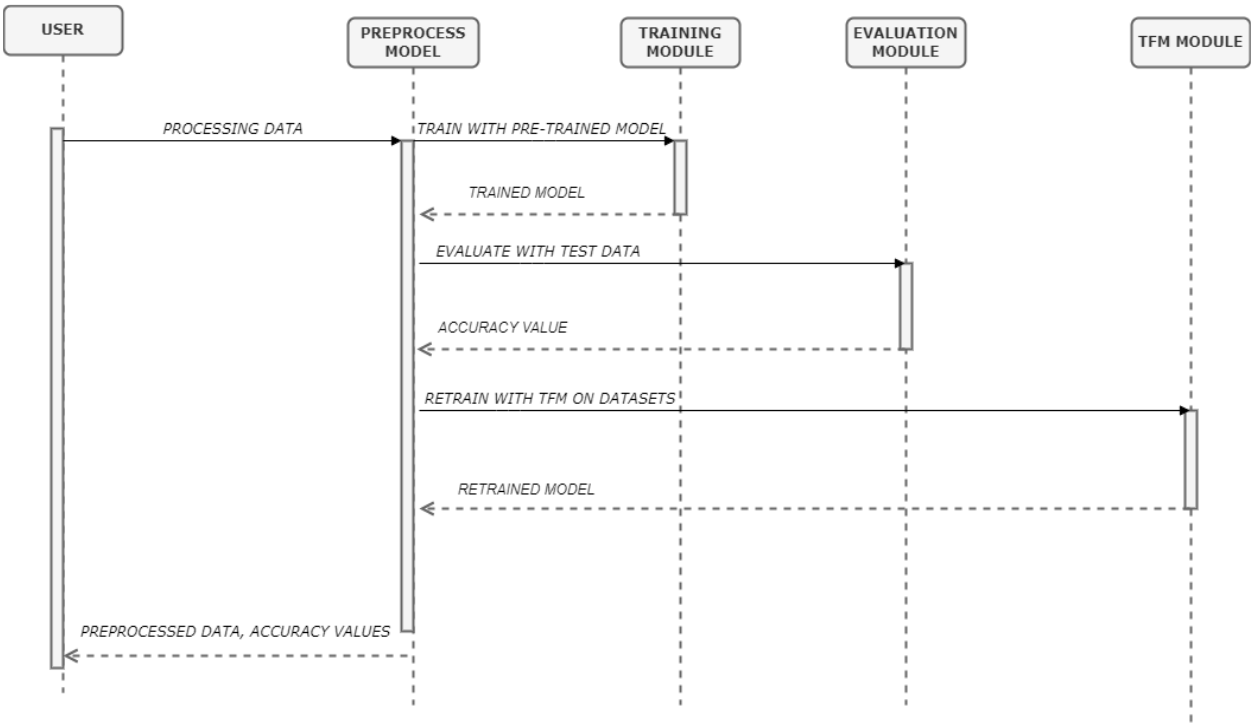


**Figure 2.** The Sequence Diagram of the proposed model

The processes from how the user processes the dataset to training the text filtering method on the dataset, training the models, and assessing the models' output are depicted in Figure 2.

## 4. Results and Discussion

Setting up a supportive atmosphere that meets the study's specific needs is essential throughout the first stages of implementation. This entails selecting the appropriate platforms, libraries, and tools that support the project's objectives and guarantee its successful completion.

The selection of libraries is crucial for any data-driven project since it determines the functionality and ease of implementation. Libraries like pandas, numpy, nltk, and seaborn were loaded for this study. All of these libraries include strong data processing, natural language processing, and visualization features that are necessary for sentiment analysis.

The platform of choice for data-intensive projects is Google Colab because of its smooth connection with Google Drive. Easy access to big datasets without the inconvenience of manual uploads is ensured by the decision to connect Google Drive. The integration of Google Drive with Colab facilitates a more efficient

data intake process, which in turn leads to more effective processing in the future.

After mounting Google Drive, it was simple to find the dataset's directory. From the google colab, the content is mounted from the local drive, in this case, the local hard disk. The researchers could read the dataset with the pandas library by giving the path to it. The dataset for this study was saved in CSV format, which is a standard and widely used format for tabular data storage. The next action initiated is to preprocess the dataset. Data preparation is an essential stage in every Natural Language Processing (NLP) study. This stage converts unprocessed data into a format that makes analysis simple and efficient.

The study's final result can be greatly impacted by the accuracy and caliber of the data preprocessing. This study's preparation procedures included label encoding, data downsampling, stopword removal, and data cleaning. Noise in raw text input, particularly from Twitter and other sites, can make NLP algorithms perform worse. Emojis, punctuation, URLs, and other superfluous characters are examples of this kind of noise. In order to create a cleaner, more consistent dataset, these components were to be eliminated during the data cleaning process. This guarantees that the model concentrates on the key passages that convey semantic significance in the text. Words that are often used in a language have little or no semantic significance when it comes to text analysis is known as stopwords. The dimensionality of the data can be decreased by eliminating them, which will free up the model to focus on more informative terms. Because the dataset is multilingual, specific stopword lists were created for each language to take into account its distinct linguistic characteristics. The custom stop words list for pidgin languages for instance is preprocessed by a user-defined function named set, which is invoked with an inner function named stopwords words, which finally returns list of stop words.

Predictions made by the model may be skewed by imbalanced datasets, in which one class predominates over another. The data was down sampled to ensure that all sentiment classes were equally represented. In order to keep the model from underfitting to the minority classes and overfitting to the majority class, this step was essential. Text labels and other categorical data need to be transformed into a numerical format in order for machine learning models to process them. This study utilized a simple mapping approach, turning the labels "positive," "negative," and "neutral" into integers. In order for the model to interpret the labels during training and evaluation, this transformation is necessary. Any sentiment analysis's effectiveness depends on the model's capacity to recognize and accurately categorize the input text's underlying sentiment. This part explores the core of the project, which is the model selection, training, and performance evaluation in all four Nigerian languages.

There is also a case of Imbalanced datasets for example Igbo text, where one class dominates over others, which could potentially gives room for outlies in the model's predictions. To counter this, the data was down sampled to ensure equal representation of all sentiment classes. This step was crucial in preventing the model from overfitting to the majority class and under fitting to the minority classes. Figures 3 and 4 present the original imbalanced Igbo dataset and the balanced version respectively.

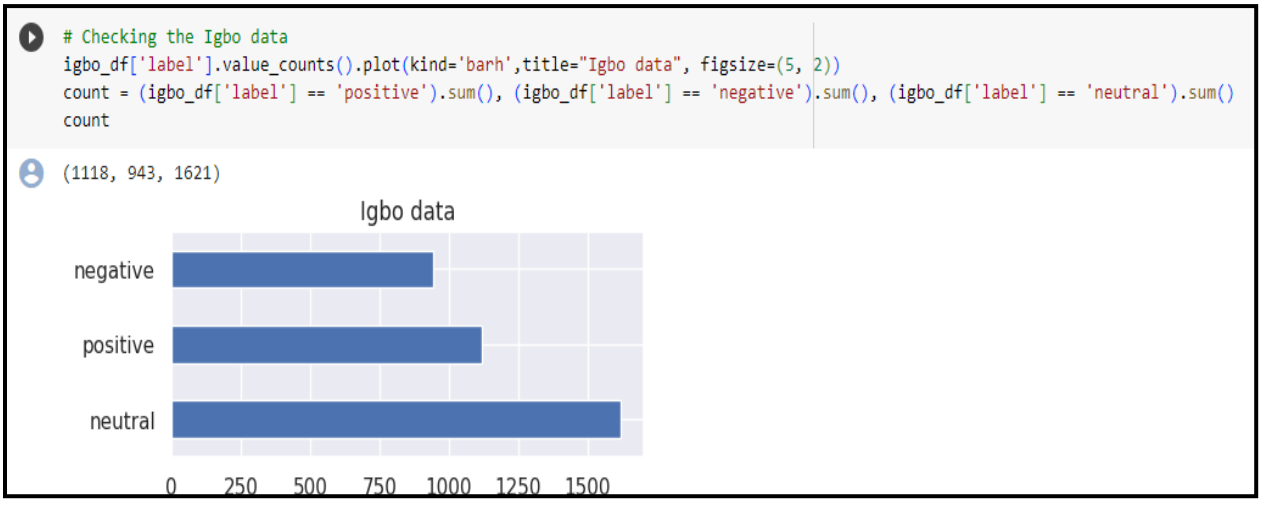


**Figure 3.** Original Imbalanced Igbo Dataset

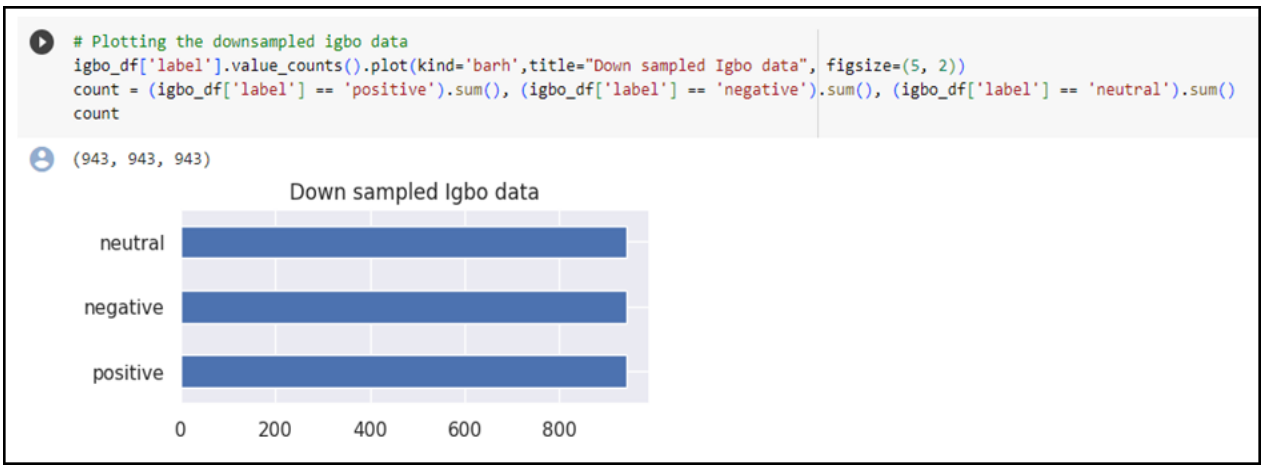


**Figure 4.** Balanced Igbo Dataset

As obviously shown by Figure 3, the dataset for the given language is imbalanced with three distinct values, which can potentially affect the accuracy of the training model. Figure 4 presents a balanced dataset.

The balanced dataset as shown by Figure 4 was obtained after employing the technique of down sampled to ensure equal representation of all sentiment classes.

An important development in the field of African language processing is the AfriBERTa model. AfriBERTa was created as a variation of the well-known BERT model and pretrained using a substantial amount of African language data. This makes it an excellent option for projects like ours that seek to study sentiments in numerous Nigerian languages since it enables it to capture the subtleties and peculiarities present in several African languages. We can guarantee better accuracy and relevance in our sentiment forecasts by utilizing AfriBERTa's power, closing the gap in NLP systems specifically designed for African languages. Furthermore, one of the fundamental stages of natural language processing is tokenization. Tokenization is the process of dividing text into smaller units known as chunks. These tokens range in length from a single character to a word. Tokenization for deep learning models involves more than just splitting, particularly for transformers like AfriBERTa. It entails mapping each token to a large vocabulary and transforming it into a distinct integer. This conversion makes it easier for the model to understand the text, which enables it to process and produce predictions. Considering how important this step is, you should use a tokenizer that is specific to the model you are using. The AfriBERTa tokenizer was utilized for our project in order to guarantee efficacy and compatibility.

Teaching humans is similar to training a machine learning model, particularly a deep learning model. Because of learning from data, the model modifies its internal parameters to improve outcome prediction. The Adam optimizer, a well-liked deep learning optimization technique, was used in the training of our model. Adam is renowned for its effectiveness and minimal memory needs. For multiclass classification jobs such as ours, the SparseCategoricalCrossentropy loss function was selected. To make sure our model learned from the training data and generalized well, we trained it over several iterations, or epochs. In order to keep track of the model's development and avoid overfitting, its performance was also routinely assessed using a different validation dataset.

Text filtration is the pivotal and underline emphasis in this research work. The text filtration technique is employed as a way to improve the accuracy of sentiment categorization by lowering noise levels in the dataset. This procedure is predicated on the idea that some words, which crop up frequently in a variety of sentiment classes, may obscure the sentiment as understood by the model. It is possible that the model will gain a more sophisticated knowledge of each feeling class by eliminating these frequently used terms. Three unique sentiment classifications were carefully distinguished from each language dataset: positive, negative, and neutral. Understanding the distinctive qualities of each sentiment class and determining which words are most frequently used within them required this division. In doing so, the groundwork for the filtration that followed was established. Figure 5 depicts a code snippet illustrating the process of separating the datasets into their respective sentiment classes.

```
[ ]  # Separating the positive and Negative and Neutral dataframes
     igbo_positive_df = igbo_df[igbo_df['label']=='positive']
     igbo_negative_df = igbo_df[igbo_df['label']=='negative']
     igbo_neutral_df = igbo_df[igbo_df['label']=='neutral']

     hausa_positive_df = hausa_df[hausa_df['label']=='positive']
     hausa_negative_df = hausa_df[hausa_df['label']=='negative']
     hausa_neutral_df = hausa_df[hausa_df['label']=='neutral']

     yoruba_positive_df = yoruba_df[yoruba_df['label']=='positive']
     yoruba_negative_df = yoruba_df[yoruba_df['label']=='negative']
     yoruba_neutral_df = yoruba_df[yoruba_df['label']=='neutral']

     pidgin_positive_df = pidgin_df[pidgin_df['label']=='positive']
     pidgin_negative_df = pidgin_df[pidgin_df['label']=='negative']
     pidgin_neutral_df = pidgin_df[pidgin_df['label']=='neutral']
```

**Figure 5.** Separation of datasets into sentiment classes.

Figure 5 presents the results sample of the process of separating the datasets into their respective sentiment classes for the four Nigeria's languages on twitter social media consider on this research.

The next step was to find the top 50 recurrent terms from each sentiment class using the segregated datasets. These terms offer a brief overview of the most prevalent topics in each feeling category. The filtration procedure benefited greatly from the extraction of these top words since it made it possible to identify terms that might be noise in various sentiment classifications. The text filtration process immediately followed this action. The screening procedure started when the top words from each sentiment class were determined. Making each sentiment class as different as feasible was the aim on this work. Common terms from the neutral and negative classes, for instance, were taken out of the positive class. This kept the distinctiveness of each feeling class and might improve the model's classification performance. Thus, the filtered data undergone training using the model.

Training the AfriBERTa model on the filtered data came after the dataset had been refined by the filtration procedure. To be clear, the model's architecture, training procedure, hyperparameters, and other relevant configurations were all kept the same as they were for the unfiltered data. In order to provide a uniform and equitable comparison of the performances on the filtered and unfiltered datasets, it was imperative to preserve this consistency. Without adding any more factors that would distort the results, the goal was to determine how the filtration system affected the model's accuracy and effectiveness. By contrasting the classification reports and confusion matrices of the models trained on both the unfiltered and filtered datasets, the effectiveness of the filtration system in sentiment classification is shown. Verifying the premise that removing common terms from sentiment classes would improve classification performance requires this comparison. Table 2 shows a contrasts findings from the unfiltered and filtered datasets and offers a thorough comparison of critical performance measures for each of the four languages.

**Table 2:** Results Comparative Analysis

| language | Evaluation Metrics | Unfiltered Data | Filtered Data |
|---|---|---|---|
| **Pidgin** | Accuracy | 0.69 | 0.75 |
| | Precision (avg) | 0.70 | 0.70 |
| | Recall (avg) | 0.70 | 0.76 |
| | F1-score (avg) | 0.70 | 0.73 |
| **Hausa** | Accuracy | 0.75 | 0.79 |
| | Precision (avg) | 0.76 | 0.79 |
| | Recall (avg) | 0.75 | 0.79 |
| | F1-score (avg) | 0.75 | 0.79 |
| **Yoruba** | Accuracy | 0.75 | 0.80 |
| | Precision (avg) | 0.75 | 0.80 |
| | Recall (avg) | 0.75 | 0.80 |
| | F1-score (avg) | 0.75 | 0.80 |
| **Igbo** | Accuracy | 0.77 | 0.76 |
| | Precision (avg) | 0.78 | 0.77 |
| | Recall (avg) | 0.77 | 0.76 |
| | F1-score (avg) | 0.77 | 0.76 |

The models trained on filtered data appear to be generally improving in terms of accuracy, precision, recall, and F1-score, particularly in languages like Pidgin, Hausa, and Yoruba. This shows that the filtering procedure was successful in removing noise or superfluous data from the dataset, allowing the model to concentrate on properties that are more important. However, it is crucial to examine the subtleties of these advancements in more detail. For example, the filtered data for Pidgin exhibits a discernible improvement in accuracy, but not so much for Igbo. The filtration process's specifics, the language used, or the original datasets' inherent quality could explain this discrepancy. Perhaps Pidgin had more common terms that overlapped across sentiment classes than Igbo; thus, the filtering worked better on that.

Furthermore, some metrics like precision, recall, and F1-score offer a more detailed picture of performance than accuracy, which only gives a broad perspective. The general gains in these metrics indicate that the model produced fewer false positives and negatives in addition to making more accurate predictions overall. Overall, the text-filtering approach significantly improve the accuracy of the current model in capturing sentiments from multiple languages found on Nigerian Twitter, thereby contributing to the development of a more comprehensive sentiment analysis framework through the text filtering approach. Finally, the results of the proposed sentiment collection based on text filtered approach for four Nigeria's languages on Twitter social media were evaluated against text unfiltered approach.

## 5. Conclusion

The study sought to analyze sentiment data collected from Nigerian Twitter users in several languages using a text filtering approach. Analyzing attitudes is crucial since internet platforms continue to influence public opinion, particularly in a linguistically varied nation like Nigeria. This study set out to comprehend the existing approaches, pinpoint their shortcomings, and develop an improved methodology to raise the accuracy of sentiment analysis. The underlying assumption of this study was that social media networks, particularly Twitter, can provide noisy raw data.

This noise has the potential to misrepresent the text's genuine sentiment, particularly when working with several languages, each with unique nuances. Pidgin, Hausa, Yoruba, and Igbo are four important Nigerian languages that were the subject of this study in order to guarantee a wide representation of opinions throughout the linguistic landscape of the country. The foundation of the study's methodology was an understanding of the raw datasets structure and sentiment distribution. Subsequently, a filtration process was put in place to remove noise and improve the sentiments in the tweets' clarity. To ascertain the effect of the filtration method, the filtered and unfiltered data were subsequently fed into the AfriBERTa model, a cutting-edge model for African language processing.

The performance indicators were found to be generally improved across all languages by the filtered datasets. This emphasizes how crucial it is to filter and preprocess data before beginning any sentiment analysis assignment, particularly when working with a variety of languages and the informal nature of tweets. Due of Twitter's dynamic nature, typical sentiment analysis models may miss certain nuances, slang terms, and mixed languages. This research has opened the door for more precise sentiment analyses that may be used in a variety of contexts, such as political analysis and market research, by developing a filtration procedure specifically designed to account for the linguistic quirks of Nigerian languages. It's crucial to realize, nevertheless, that there is no ideal model or process. Even though the filtration process has improved, there is a chance that it will be overly thorough and remove some context. It's a delicate balance to maintain meaning while filtering out noise, but this research has definitely made progress in that direction.

However, the development of iterative filtration systems that can be improved with time should be the focus of future study. The filtration procedure should be flexible enough to react to changes in languages and the emergence of new slang terms. Also, there can be a feedback loop set up where people can offer explanations for attitudes that have been incorrectly identified. The underlying model and the filtration procedure can both be enhanced with the help of this input. Although over 500 languages are spoken in Nigeria, this study concentrated on the four most common languages spoken there. To ensure a more thorough sentiment analysis, future research should think about enlarging the linguistic scope to include new languages.

Based on the findings and conclusions drawn from this research, future research should consider developing iterative filtration mechanisms that can be refined over time. As languages evolve and new slangs

or terminologies emerge, the filtration process should be adaptive enough to accommodate these changes. A feedback loop can be established where users can provide insights on wrongly classified sentiments. This feedback can be used to improve both the filtration process and the underlying model. While this research focused on four major Nigerian languages, there are over 500 languages spoken in Nigeria. Future studies should consider expanding the linguistic scope to include more languages, ensuring a more comprehensive sentiment analysis. Collaboration with linguists and cultural experts can provide deeper insights into the linguistic intricacies of each language, further refining the filtration process.

## References


[1]. Abdullahi, H. I., Ahmad, M. A., & Haruna, K. (2024). Twitter sentiment analysis for Hausa abbreviations and acronyms. Science World Journal, 19(1), 101–104. https://doi.org/10.4314/swj.v19i1.13

[2]. Abo, M. E. M., Idris, N., Mahmud, R., Qazi, A., Hashem, I. A. T., Maitama, J. Z., Naseem, U., Khan, S. K., & Yang, S. (2021). A multi-criteria approach for arabic dialect sentiment analysis for online reviews: Exploiting optimal machine learning algorithm selection. Sustainability (Switzerland), 13(18), 1–20. https://doi.org/10.3390/su131810018

[3]. Abubakar, A. I., Roko, A., Bui, A. M., & Saidu, I. (2021). An Enhanced Feature Acquisition for Sentiment Analysis of English and Hausa Tweets. International Journal of Advanced Computer Science and Applications, 12(9), 102–110. https://doi.org/10.14569/IJACSA.2021.0120913

[4]. Adewole, K. S., Balogun, A. O., Raheem, M. O., Muhammed, K., Jimoh, R. G., Mabayoje, M. A., Usman-hamza, F. E., Akintola, A. G., & Asaju-gbolagade, A. W. (2021). Hybrid Feature Selection Framework for Sentiment Analysis on Large Corpora. Jordanian Journal of Computers and Information Technology (JJCIT), 07(02), 130–151.

[5]. Akuma, S., Obilikwu, P., & Ahar, E. (2021). Sentiment Analysis of Social Media Content for Music Recommendation. Nigerian Annals of Pure and Applied Sciences, 4(1), 110–120. https://doi.org/10.46912/napas.225

[6]. Alade, M. S., & Nwankpa, J. M. (2022). Sentiment Analysis of Nigerian Students' Tweets on Education: A Data Mining Approach. International Journal of Computer (IJC), 45(1), 1–27. https://www.researchgate.net/profile/Nwankpa-Joshua/publication/364110706_Sentiment_Analysis_of_Nigerian_Students'_Tweets_on_Education_A_Data_Mining_Approach/links/633a6db99cb4fe44f3f91b12/Sentiment-Analysis-of-Nigerian-Students-Tweets-on-Education-A-Data

[7]. Asogwa, D. ., Anigbogu, S. ., Onyenwe, I. ., & Sani, F. . (2021). Text Classification Using Hybrid Machine Learning Algorithms on Big Data. International Journal of Trend in Research and Development, 6(5), 128–134. www.ijtrd.com

[8]. Ayo, F. E., Folorunso, O., Ibharalu, F. T., Osinuga, I. A., & Abayomi-Alli, A. (2021). A probabilistic clustering model for hate speech classification in twitter. Expert Systems with Applications, 173, 1–21. https://doi.org/10.1016/j.eswa.2021.114762

[9]. Chinedum, A., & Ogochukwu, O. C. (2021). Application of Naïve Bayes and SVM Techniques on Sentiment Analysis on Nigerian Tweets. International Research Journal of Modernization in Engineering Technology and Science, 03(03), 931–935. www.irjmets.com

[10]. Chinyere, E., Christopher, B., & Eleonu, O. F. (2021). Automated Hybrid Model for Classifying Human Emotions Using Sentiment Analysis and Text Classification. International Journal of Advances in Engineering and Management (IJAEM), 3(7), 2956–2971. https://doi.org/10.35629/5252-030729562971

[11]. Efuwape, T. . O., Abioye, T. . E., & K-K. Abdullah, A. (2022). Text Analytics of Opinion-Poll on Adoption of Digital Collaborative Tools for Academic Planning Using Vader-Based Lexicon Sentiment Analysis. FUDMA Journal of Sciences (FJS), 6(1), 152–159. https://doi.org/10.33003/fjs-2022-0601-874

[12]. Ibrahim, Y., Okafor, E., Yahaya, B., Yusuf, S. M., Abubakar, Z. M., & Bagaye, U. Y. (2021). Comparative Study of Ensemble Learning Techniques for Text Classification. 2021 1st International Conference on Multidisciplinary Engineering and Applied Science, ICMEAS 2021, 1–6. https://doi.org/10.1109/ICMEAS52683.2021.9692306

[13]. Ijairi, M. U., Abdullahi, M., & Hassan, I. H. (2023). Sentiment Classification of Tweets with Explicit Word Negations and Emoji Using Deep Learning. International Journal of Software Engineering & Computer Systems (IJSECS), 9(2), 93–104. https://doi.org/https://doi.org/10.15282/ijsecs.9.2.2023.3.0114

[14]. Khan, A., Zhang, H., Boudjellal, N., Ahmad, A., Shang, J., Dai, L., & Hayat, B. (2021). Election Prediction on Twitter : A Systematic Mapping Study. Hindawi Complexity, 2021, 1–27. https://doi.org/https://doi.org/10.1155/2021/5565434

[15]. Kolajo, T., Daramola, O., & Adebiyi, A. A. (2021). SMAFED : Real-Time Event Detection in Social Media Streams. Research Square, 1–28.

[16]. Ladoja, K. T., & Afape, R. T. (2024). Sarcasm Detection in Pidgin Tweets Using Machine Learning Techniques. Asian Journal of Research in Computer Science, 17(5), 212–221. https://doi.org/10.9734/AJRCOS/2024/v17i5450

[17]. Lee, E., Rustam, F., Washington, P. B., Barakaz, F. E. L., Aljedaani, W., & Ashraf, I. (2022). Racism Detection by Analyzing Differential Opinions Through Sentiment Analysis of Tweets Using Stacked Ensemble GCR-NN Model. IEEE Access, 10, 9717–9728. https://doi.org/10.1109/ACCESS.2022.3144266

[18]. Muhammad, S. H., Adelani, D. I., Ruder, S., Ahmad, I. S., Abdulmumin, I., Bello, B. S., Choudhury, M., Emezue, C. C., Abdullahi, S. S., Aremu, A., Jorge, A., & Brazdil, P. (2021). NaijaSenti: A Nigerian Twitter Sentiment Corpus for Multilingual Sentiment Analysis. 1–13.

[19]. Mutanga, R. T., Naicker, N., & Olugbara, O. O. (2022). Detecting Hate Speech on Twitter Network Using Ensemble Machine Learning. (IJACSA) International Journal of Advanced Computer Science and Applications, 13(3), 331–340.

[20]. Olabanjo, O., Wusu, A., Afisi, O., Asokere, M., Padonu, R., Olabanjo, O., Ojo, O., Folorunso, O., Aribisala, B., & Mazzara, M. (2023). From Twitter to Aso-Rock : A sentiment analysis framework for understanding Nigeria 2023 presidential election. Heliyon, 9(5), 1–14. https://doi.org/10.1016/j.heliyon.2023.e16085

[21]. Oladipo, F., Akarah, P., & Ohieku, A. (2022). SENTIMENT ANALYSIS MODEL FOR TWITTER ON COVID-19

VACCINE. Journal of Information Systems & Operations Management, 16(1), 209–230.

[22]. Oladipo, F. O., Afolabi, S. P., & Ariwa, E. (2020). A dataset of abusive comments on the Nigerian web. International Journal of Computing and Artificial Intelligence, 2(1), 1–5.

[23]. Onuora, A. C., Ana, P. O., Otiko, A. O., & Maidoh, E. (2024). Machine Learning Architecture for Combating Hate Speech in Igbo Language. Academic Staff Union of Polytechnic (ASUP) Akanu Ibiam Federal Polytechnic Unwana Chapter 3rd International Conference 2024, 1–27.

[24]. Popoola, G., Abdullah, K.-K., Fuhnwi, G. S., & Janet, A. (2024). Sentiment Analysis of Financial News Data using TF-IDF and Machine Learning Algorithms. 3rd IEEE International Conference on AI in Cybersecurity (ICAIC), February, 1–7. https://doi.org/10.1109/ICAIC60265.2024.10433843

[25]. Raychawdhary, N., & Das, A. (2023). Seals _ Lab at SemEval-2023 Task 12 : Sentiment Analysis for Low-resource African Languages , Hausa and Igbo. Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023), 12, 1508–1517.

[26]. Sani, M., Ahmad, A., & Abdulazeez, H. S. (2022). Sentiment Analysis of Hausa Language Tweet Using Machine Learning Approach. Journal of Research in Applied Mathematics, 8(9), 7–16.