Comparative Analysis of Machine Learning Algorithms for Sentiment Analysis of Multilingual Nigerian Social Media Comments

*[1]Michael P. Asefon, [2] Ayomitope O. Isijola, [2] Ufuoma C. Ogude, and [2] Samuel A. Ntui
[1]Department of Computer Sciences, National Open University of Nigeria, Lagos, Nigeria
[2]Department of Computer Sciences, University of Lagos, Lagos, Nigeria
pelumiasefon@gmail.com | ayomitopeisijola@yahoo.com | uogude@unilag.edu.ng | aquaintuition@gmail.com

Received: 19-SEP-2024; Reviewed: 15-NOV-2024; Accepted: 27-NOV-2024
https://dx.doi.org/10.4314/fuoyejet.v9i4.8

ORIGINAL RESEARCH

Abstract— This study probes into sentiment analysis within the multilingual landscape of Nigerian social media comments by using machine learning algorithms as computational tools. Nigeria has various languages, creating a complicated scenario for understanding sentiment dynamics in digital discourse. This research developed sentiment classification models across languages such as English, Hausa, Yoruba, Nigerian-Pidgin, and Igbo. The choice of machine learning algorithms utilized in this paper was driven by algorithms suitability for the task, diversity of languages, and dataset characteristics. By utilizing machine learning algorithms such as Logistic Regression, Support Vector Machines (SVM), Random Forest Classification (RFC), and Long Short-Term Memory (LSTM), the study aims to provide insights into the varieties of expressions and sentiments in Nigeria's social media. The research explores the domains of natural language processing and socio-linguistics. In summary, the research finds that SVM and Random Forest are the most functional for the combined multilingual dataset, giving a superior precision of 76% and recall of 72% each. For individual language datasets, LSTM, due to its advanced sequence modeling capabilities, outperforms the other machine learning algorithms with model accuracy scores of 58% to 90% and 72% for the combined language datasets.

Keywords— Logistic regression, Multilingual, Random-forest, Sentiment, Social-Media.

1 INTRODUCTION

In an era defined by digital interconnectedness, social media sites have evolved into bustling cyberspaces where individuals engage in vibrant ideas, share opinions, and express heartfelt sentiments on reservations and so many topics. We focus on Nigeria, a nation filled with many different languages and a huge cultural diversity that promotes a world of social media dialogue using several languages simultaneously and indigenous dialects. Due to the constantly changing digital environment, the need to explain the underlying sentiments and opinions involved in multilingual Nigerian social media comments has increasingly become relevant (Olaleye, Adewole, & Odumuyiwa, 2018). As Nigeria's digital social media presence expands rapidly, it is difficult to have a firm understanding of the complexities of sentimental expression behind social media activities among different language regions. Starting from the busy streets of Lagos (metropolis) to the quiet rural communities in the Niger Delta, social media users make use of various languages such as English, Hausa, Yoruba, Nigerian-Pidgin (Naija), and Igbo, seamlessly navigating through them without effort. This illustrates how the culture in Nigeria is versatile, yet there exists a strong need to explain the sentiments present in the diverse, multilingual social media comments. This study aims to explore sentiment analysis in multilingual Nigerian social media, using computational analysis to spotlight the intricate relationship between language and culture in digital expressions.

*Corresponding Author: ayomitopeisijola@yahoo.com

Section B- ELECTRICAL/COMPUTER ENGINEERING & RELATED SCIENCES
Can be cited as:
Asefon M. P., Isijola A. O., Ogude U. C., and Ntui S. A. (2024).
FUOYE Journal of Engineering and Technology (FUOYE**JET), 9(4), 615-623.**
https://dx.doi.org/10.4314/fuoyejet.v9i4.8

By utilizing logistic regression, support vector machine (SVM), random forest, and long short-term memory (LSTM), this research intends to classify sentiments effectively, providing sentiments and simple insights into sentiment trends in Nigeria's social media space (Liu, 2012).

This research also aims to compare algorithms and their model performance leveraging on logistic regression, support vector machine (SVM), random forest, and long short-term memory (LSTM), which is mostly effective as binary classification.

2 RELATED WORKS

Many studies have highlighted the effectiveness of multilingual sentiment analysis, for instance, studies from multilingual social media data from Nigerian users have shown that incorporating local languages with English improves sentiment classification and greatly increases its accuracy (Ihsan, Ashraf, & Jhanjhi, 2023). These studies show how important it is to consider various languages in sentiment analysis to get the true sentiment expressed by users.

(Zhao et al., 2024) emphasized the necessity for a more in-depth view of cross-lingual sentiment analysis (CLSA) techniques to boost their analytical efficiency. CLSA faces challenges, including linguistic differences and limited materials, and it aims to assess sentiments throughout multiple languages.

(Qiao & Huang, 2024) recently used deep learning for cross-lingual sentiment analysis in natural language processing. The first section provides a definition and historical history of sentiment analysis and natural language processing (NLP). The use of deep learning in natural language processing and sentiment analysis is then examined, including information on text representation, feature extraction techniques, sentiment analysis case studies, and the fundamentals of deep learning algorithms.

While these works highlight the potential of machine learning algorithms for sentiment analysis, they often focus on monolingual datasets, overlook the complexities of multilingual contexts, and lack emphasis on culturally nuanced expressions. There is a gap regarding the comparative performance of algorithms on multilingual Nigerian social media comments, which this research aims to fill.

2.1 INTRODUCTION TO SENTIMENT ANALYSIS

Sentiment analysis can as well be referred to as opinion prospecting, which is a domain of research that is part of natural language processing (NLP), focusing on recognizing and grouping opinionated text to determine the writer's attitude towards a particular subject. The field of opinion mining has become significantly useful with the increase in social media reviews, where a large amount of user-generated content is produced regularly, thereby showing valuable insights into public opinion (Liu, 2012).

2.1.1 SIGNIFICANCE OF SENTIMENT ANALYSIS

The importance of sentiment analysis extends over many sectors, such as businesses that evaluate satisfied customers, monitor the reputation of their brands, and get insights from consumers. In politics, it is used to track the opinion of the public on policies and political candidates, while healthcare professionals use it to analyze patient feedback to improve their services. And also, media organizations can evaluate the reaction of the public to news, stories, and events.

2.1.2 OVERVIEW OF SENTIMENT ANALYSIS TECHNIQUES

Methods for sentiment analysis can be largely classified into rule-based and deep-learning methods; each has its strengths and weaknesses.

Rule-based approaches are easy to implement but can be rigid and fail to generalize well.

Machine learning approaches provide better performance but require labeled data and feature engineering.

Deep learning approaches are highly accurate but demand high computational power and large amounts of data.

2.2 SENTIMENT ANALYSIS IN SOCIAL MEDIA

The swift increase of social media sites has changed how people communicate, share ideas, and influence public opinions greatly. Social networking sites like Twitter, Facebook, and Instagram are abundant sources of content created by users that provide unique opportunities for sentiment analysis, and by examining the sentiments expressed in social media posts, researchers can gain insights into public opinion, current trends, and consumer behavior (Ihsan, Ashraf, & Jhanjhi, 2023).

2.2.1 IMPORTANCE OF SENTIMENT ANALYSIS IN SOCIAL MEDIA

Sentiment analysis in social media holds high importance across various fields:

1. Commercial and promotional: Companies employ social media sentiment analysis to understand client feelings toward their goods and services. This helps them track brand reputation, evaluate customer satisfaction, and develop new marketing strategies. Positive sentiments can be used for promotional campaigns, while negative sentiments can promote customer service improvements and product renovations and changes.
2. Politics and Public Opinion: Politicians and policymakers analyze social media sentiment to understand the public's opinion on policies, candidates, and current events. The feedback obtained can guide campaign strategies and policy decision-making. During elections, social media sentiment analysis provides insights into voter preferences and key issues of concern.
3. Healthcare: In the healthcare system, sentiment analysis helps understand the feedback on the services patients receive to identify public concerns about health issues and monitor the spread of health-related misinformation, which can help healthcare providers improve patient care and communication strategies.
4. Media and Entertainment: Media organizations use sentiment analysis to measure the public reaction to news stories, television shows, movies, and celebrities, which provide guides for content creation, decision-making, and audience engagement strategies.

2.2.2 PREVIOUS STUDIES AND THEIR FINDINGS

Many studies have done sentiment analysis in social media contexts, and some have developed ways to analyze Twitter messages for sentiment classification, also showing the feasibility of using tweets as a data source for extracting people's opinions (Mohammad, Kiritchenko, & Zhu, 2013). Others have investigated sentiment strength detection in social web content, which gives insights into how effective various sentiment analysis techniques are (Muhammad et al., 2022) (Devlin et al., 2019).

Specific applications have also been explored for example, some studies have focused on political tweets by showing patterns in public opinion during election periods, and in business domains, it was used to assess customer satisfaction on services like airlines to show how sentiment analysis can inform service improvements (Ihsan, Ashraf, & Jhanjhi, 2023).

2.2.3 CHALLENGES SPECIFIC TO SOCIAL MEDIA SENTIMENT ANALYSIS

In spite of its potential, social media sentiment analysis presents multiple difficulties:

1. Short and Informal Text: Social media posts are most times short and written in an informal language like slang, abbreviations, and emojis, which may hinder the accuracy of sentiment interpretation by conventional NLP algorithms.
2. Sarcasm: Recognizing sarcasm is quite challenging, as these expressions can invert the real sentiment of a message, and more advanced techniques and contextual understanding are needed to address this issue.
3. Multilingualism: Social media users routinely flip between languages or post in more than one language, and models that can handle multilingual data must be developed for accurate sentiment analysis, especially in regions like Nigeria where there are different languages (Ihsan, Ashraf, & Jhanjhi, 2023).
4. Noise and Spam: Social media sites are crowded with noise and spam, which can reduce the quality

http://.doi.org/10.46792/fuoyejet.v9i4.8 engineering.fuoye.edu.ng/journal

of sentiment analysis results. Effective data cleaning and filtering techniques are necessary to ensure the quality of the analysis.

2.3 MACHINE LEARNING ALGORITHMS IN SENTIMENT ANALYSIS

Algorithms like Logistic regression and Random-forest classifier are fundamental statistical techniques used in binary classification tasks, which make both algorithms highly relevant for sentiment analysis in which the objective is to categorize text as either good or bad sentiment. Also, this section explores the application of Machine Learning Algorithms like Support Vector Machines and Long Short-Term Memory in sentiment analysis, showing its advantages, limitations, and implementation.

2.3.1 OVERVIEW OF LOGISTIC REGRESSION

Logistic regression is a kind of regression analysis utilized to predict, from one or more independent variables, the outcome of a binary dependent variable. It estimates the probability that a given input pertains to a specific class. The logistic function is also known as the sigmoid function, which depicts predicted values to probabilities and constrains them between 0 and 1. This probabilistic output makes logistic regression particularly suitable for classification tasks.

Advantages of Logistic Regression

1. Clarity and Comprehensibility: Logistic regression is simple to comprehend and implement. The model's coefficients indicate the relationship between each feature and the probability of the target outcome, which makes it easy to interpret and transparent.
2. Efficiency: Logistic regression is computationally efficient, which is advantageous when dealing with large datasets. Its relatively low computational complexity allows it to scale well.
3. Performance: Despite it being simple, logistic regression often performs competitively with more complex models, especially when there is roughly a linear relationship between the target and the features.

Limitations of Logistic Regression

1. Linearity Assumption: When working with more complicated and non-linear data, logistic regression may not perform as well because it is assumed that the log chances of the dependent and independent variables have a linear relationship.
2. Feature Engineering: The effectiveness of logistic regression is contingent on the nature of the input characteristics, and significant effort may be required to pre-process text data and engineer relevant features like n-grams or TF-IDF scores.
3. Multicollinearity: Excessive feature correlation may have a detrimental impact on the model's coefficients, resulting in instability and diminished interpretability.

Implementation in Sentiment Analysis

1. Text Pre-processing: Before applying logistic regression, the text data must be cleaned and transformed into a numerical way, which can be done using steps including tokenization, elimination of stop words, stemming or lemmatization, and converting text into feature vectors using methods like Bag of Words or TF-IDF.
2. Feature Selection: A key component of enhancing model performance and interpretability is feature selection. The most crucial features of sentiment classification can be found in methods like mutual information or chi-square testing.
3. Model Training and Evaluation: Logistic regression models are disciplined on classified datasets where the sentiment of each example of text is known. During the training process, the model's coefficients are optimized to reduce the error in predicting the target sentiment. Evaluation of the model is usually done utilizing metrics such as accuracy, precision, recall, and F1-score, often introducing cross-validation to ensure its ambidexterity.

2.3.2 OVERVIEW OF SUPPORT VECTOR MACHINES (SVM)

Support Vector Machines (SVMs) are among the powerful machine learning techniques that work well for sentiment analysis works that require classifying text data as positive, negative, or neutral. In a high-dimensional feature space, SVMs seek to identify the hyperplane that best divides comments with different sentiment polarities.

Advantages of SVM for Sentiment Analysis

1. High Accuracy: SVMs are good at finding maximum decision boundaries, which leads to potentially high classification accuracy in sentiment analysis.
2. Robust to Noise: SVMs focus on identifying the most critical data points (support vectors) for classification, making them less prone to irrelevant information and noise, which are most times present in social media comments like typos or slang.
3. Works Well with High-Dimensional Data: Sentiment analysis involves analyzing text data that can be represented in a high-dimensional feature space, and they are capable of effectively managing such data.

Limitations of SVM for Multilingual Social Media Sentiment Analysis

1. Feature Engineering: Effective sentiment analysis requires crafting informative features from the text data. This can be difficult for multilingual comments as sentiment can be expressed differently in different languages and cultures.
2. Limited Interpretability: SVMs can make it hard to comprehend how they reach specific classifications. This could be a drawback for analyzing the variations of sentiment expressed in different languages.
3. Scalability: Training SVMs on massive datasets of multilingual social media comments can be statistically expensive.

2.3.3 OVERVIEW OF RANDOM FORESTS

Random Forests (RFs) have emerged as a prominent technique in sentiment analysis, mainly for classifying the

http://.doi.org/10.46792/fuoyejet.v9i4.8 engineering.fuoye.edu.ng/journal

sentiment of social media comments. This type of learning method uses the collective strength of numerous decision trees, each of which acts as a classifier to achieve a large and accurate sentiment identification (Karthika et al., 2019).

2.3.4 OVERVIEW OF LONG SHORT-TERM MEMORY (LSTMS)

Long Short-Term Memory (LSTM) systems represent a powerful approach in deep learning for sentiment analysis tasks. Unlike traditional neural networks that struggle with capturing wide dependencies in sequential data, LSTMs excel at this very aspect, which makes them more suitable for analyzing social media comments where sentiment can often be influenced by the context of preceding or following words.

Advantages of LSTMs for Sentiment Analysis

1. Capturing Context: LSTMs possess a unique architecture that allows them to effectively learn and preserve long-term relationships in textual data. This is necessary for sentiment analysis, as the sentiment of a social media comment can be highly influenced by the context of surrounding words (Tang, Qin, & Liu, 2015). For instance, sarcasm can be easily missed by simpler models that don't account for contextual cues.
2. Multilingual Capabilities: LSTMs can be effectively applied to sentiment analysis tasks involving multiple languages that cannot learn from sequential data, which makes them less reliant on language-specific features compared to traditional machine-learning approaches. This is beneficial for researchers handling multilingual social media comments.
3. Adaptability: LSTMs can be combined with convolutional neural networks (CNNs) and other deep learning architectures. This makes it possible to add more features to significantly enhance sentiment analysis performance (Wang & Manning, 2012).

2.4 SENTIMENT ANALYSIS IN THE MULTILINGUAL NIGERIAN CONTEXT

The multiple languages used in Nigeria make it difficult to accurately do sentiment analysis as over 500 languages are spoken in the country where analyzing social media sentiments requires handling many language inputs, dialects, and code-switching practices. This section explores the specific context of sentiment analysis in Nigeria by focusing on the challenges and methods adapted to Nigeria.

2.4.1 LINGUISTIC DIVERSITY IN NIGERIA

Nigeria is a country with many languages, with three major languages (Hausa, Yoruba, and Igbo) where English is the official language. Aside from these, there are many other languages and dialects spoken by various ethnic groups, and this diversity is shown in social media interactions, where users most times express their sentiments in multiple languages and frequently switch between them.

2.4.2 CHALLENGES IN MULTILINGUAL SENTIMENT ANALYSIS

1) Language Identification: Identifying the language of each social media post in an accurate manner is a vital first step, which is complicated by the frequent usage of "code-switching," in which users navigate between languages within a single statement or post.
2) Resource Scarcity: Numerous Nigerian languages lack broad language resources, including annotated corpora, sentiment lexicons, and pre-trained language models. This scarcity hinders the development of accurate sentiment analysis tools (Ihsan, Ashraf, & Jhanjhi, 2023).
3) Cultural Nuances: Sentiment expressions can vary in various cultures and languages. Capturing these variations is important for accurate sentiment analysis but requires a deep understanding of each language's specific context and idiomatic expressions.
4) Data Quality: Social media data is most times full of informal language, slang, abbreviations, and typographical errors. Cleaning and pre-processing this data to extract meaningful features for sentiment analysis is a significant challenge.

2.4.3 METHODOLOGIES FOR MULTILINGUAL SENTIMENT ANALYSIS

1) Language Detection and Segmentation: Advanced language detection algorithms are used to recognize and distinguish between different languages in a single post. Tools like Google's Compact Language Detector (CLD) or fastText can be used for this purpose.
2) Multilingual Embeddings: Techniques such as multilingual word embeddings and models like mBERT (multilingual BERT) enable the representation of text from different languages in a shared vector space. These embeddings facilitate the training of models that can handle multiple languages simultaneously (Devlin et al., 2019).
3) Translation-Based Approaches: Translating all text into a single target language (e.g., English) allows the use of previous sentiment analysis tools, which makes the process simple, but it may introduce translation errors and loss of sentiment nuances.
4) Hybrid Models: Combining multiple approaches like multilingual embeddings for initial representation and translation can improve the accuracy of results. Hybrid models utilize the strengths of different techniques to handle complex multilingual facts.

3 METHODOLOGY

3.1 RESEARCH DESIGN

The main target of this research is to carry out a comparative evaluation of the four mentioned machine learning algorithms for sentiment analysis of social media comments in four Nigerian languages such as Nigerian-Pidgin, Igbo, Hausa, and Yoruba.

3.1.1 RESEARCH DESIGN OBJECTIVES

1) To compare the performance of Logistic Regression, Support Vector Machine (SVM), Random Forest, and Long Short-Term Memory (LSTM) algorithms for sentiment analysis on individual datasets of each Nigerian language (Nigerian-Pidgin, Igbo, Hausa, and Yoruba).

http://.doi.org/10.46792/fuoyejet.v9i4.8 engineering.fuoye.edu.ng/journal

2) To make comparisons and see the effectiveness of these algorithms across the four languages based on metrics like precision, recall, and F1-score.
3) To analyze and observe the performance of the algorithms on a combined dataset containing all four languages.

Identify and recommend the most suitable machine learning algorithm for sentiment analysis of multilingual Nigerian social media comments.

3.1.2 OVERVIEW OF RESEARCH DESIGN

A combination of exploratory, descriptive, and experimental research designs are employed in the research work to analyze sentiment in multilingual Nigerian social media comments using algorithms like random forest classifier (RFC), support vector machine (SVM), logistic regression, and long short-term memory (LSTM). Structured to address the intricacies of linguistic diversity and code-switching, this research study aims to develop and refine sentiment analysis techniques suited to the Nigerian context despite the limited resources for certain Nigerian languages.

3.1.3 EXPLORATORY ASPECTS

The exploratory design is structured to probe the current field of multilingual sentiment analysis in Nigeria. It includes:

1. Pinpointing the main languages and dialects used in Nigerian social media.
2. Investigating common patterns of code-switching and mixed-language usage.
3. To comprehend several methods and tools for sentiment analysis in Nigeria's multilingual comments.

While the exploratory approach is necessary to gather preliminary insights and uncover underlying patterns that inform the growth and refinement of the model of sentiment analysis, other designs would have to be considered to further this research. This phase is crucial given the limited existing research on sentiment analysis of multiple languages in the Nigerian context.

3.1.4 DESCRIPTIVE ASPECTS

The descriptive design aims to accentuate the emotions expressed in the multilingual comments by:

1. Curating a dataset from social media comments leveraging sites including Twitter, Facebook, Instagram, etc.
2. Adding Labels to the dataset with sentiment meanings that are either positive, negative, or neutral using both manual and automated methods.
3. Evaluating sentiments for several languages and social networking sites to recognize if there are similarities in opinion patterns.

The descriptive approach provides the basis for understanding the broader context of public opinion and sentiment in Nigeria by enabling a detailed and systematic account of sentiment trends and patterns.

3.1.5 EXPERIMENTAL ASPECTS

The experimental approach compares algorithms such as logistic regression, random forest, support vector machine, and long short-term memory and aims to accurately classify sentiments and signify which of the algorithms are effective in multilingual Nigerian social media comments. It involves:

1. Training and testing machine learning models of algorithms such as logistic regression, support vector machine, random forest, and long short-term memory on Nigerian-Pidgin, Hausa, Igbo, and Yoruba language datasets individually.
2. Concatenating all the mentioned datasets and developing models from each of the algorithms.
3. Comparing the performance of logistic regression, support vector machine, random forest, and long short-term memory to determine each person's performance using the obtained metric.

3.1.6 ALIGNMENT WITH RESEARCH OBJECTIVES

The chosen research design aligns with the research objectives in several ways:

1. The exploratory phase points out the opportunities of multilingual sentiment analysis and its key challenges, addressing the objective of understanding the area of Nigerian social media.
2. The descriptive phase aims to provide a detailed account of sentiment annotation and distribution, fulfilling the objective of portraying sentiments in a multilingual context.
3. The experimental phase tests and validates the use of all four algorithms of machine learning, meeting the objective of evaluating and comparing sentiment analysis strategies for Nigerian languages.

By combining exploratory, descriptive, and experimental approaches, this research study ensures a comprehensive investigation of multilingual sentiment analysis in Nigeria, offering valuable insights and practical investigation of the field.

3.2 DATASET COLLECTION

For this research, a balanced dataset was curated comprising multilingual social media comments of widely spoken Nigerian languages like Nigerian-Pidgin, Hausa, Igbo, and Yoruba. The dataset was later annotated to include comments labeled as positive, negative, or neutral sentiment.

3.2.1 DATA ACQUISITION PROCESS

The process of data acquisition involved:

1. Scraping social media platforms manually using Google Forms as a tool to collect and curate posts and comments.
2. Filtering data to include only relevant posts written in the target languages and involving a mixture of more than Nigerian dialect.
3. Pre-processing the data by cleaning, normalizing the text to make it easier for training, and handling noise (e.g., removing special characters, etc).

3.2.2 SOURCES AND NATURE OF DATA

For this research, the data is curated from social media sites such as Twitter, Facebook, and Instagram, focusing on comments and posts written in English, Hausa, Yoruba, Nigerian-Pidgin, and Igbo. These platforms were chosen due to their widespread use in Nigeria and the diverse linguistic representation they offer.

http://.doi.org/10.46792/fuoyejet.v9i4.8 engineering.fuoye.edu.ng/journal

3.2.3 DATASET OVERVIEW

Figure 1: Nigerian-Pidgin Multilingual DataFrame before Preprocessing

Figure 2: Hausa Language Multilingual DataFrame before Preprocessing

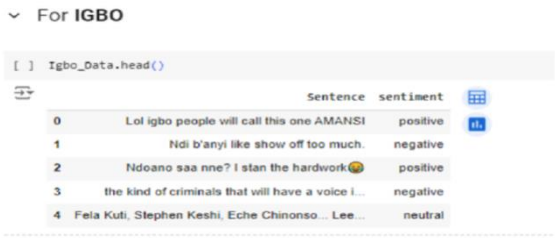

Figure 3: Igbo Language Multilingual DataFrame after Preprocessing

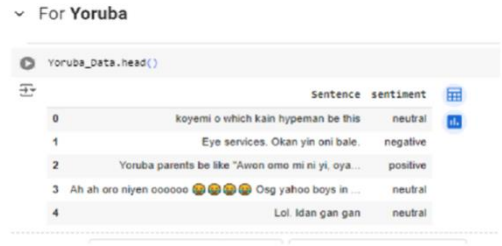

Figure 4: Yoruba Language Multilingual DataFrame after Preprocessing

1) Volume: Above 5,200 comments were collected carrying sentiment labels and language categories, ensuring a substantial dataset for training and testing each of the machine learning models.
2) Languages: The dataset included comments in English, Hausa, Yoruba, Nigerian- Pidgin and Igbo, reflecting the linguistic diversity of Nigeria's social media.

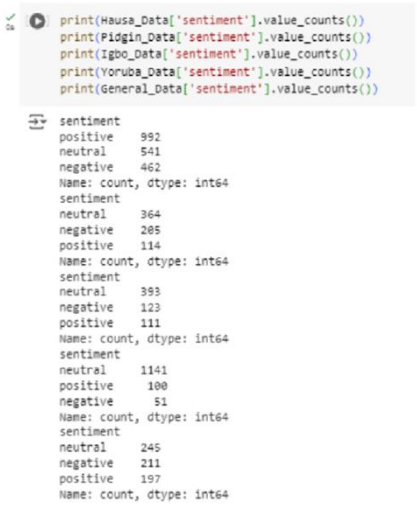

Figure 5: Total Multilingual Dataset while Preprocessing

3.3 DATA PRE-PROCESSING STEPS

1) Data Cleaning: The data curated was cleaned to remove duplicates, spam, and irrelevant content. Non-textual elements such as URLs were also removed to focus solely on textual data.
2) Language Detection: Routine language detections were carried out to classify and segment posts and comments by their languages. This step was crucial for handling code-switching instances where multiple languages were used within a single comment.
3) Normalization: Text normalization was performed to convert all text to lowercase to correct common misspellings of column titles and abbreviations. This step helped in standardizing the data for further analysis.
4) Tokenization: The cleaned and normalized text was tokenized into individual words or tokens, which are the basic units of analysis for sentiment classification.
5) Handling Code-Switching: Concerning comments that had multiple languages, they were tagged using language labelings like maybe English, Pidgin, and Hausa or English, Pidgin, and Yoruba, which made them stand out and were well represented.

3.3.1 JUSTIFICATION OF DATA SOURCES AND PREPROCESSING TECHNIQUES

The importance of the data sources and pre-processing techniques are justified as relevant based on the following:

1. Large social networking platforms that gave access to the general populace were considered because they provided a rich dataset with varieties of content reflecting real-world sentiments across different languages in Nigeria.
2. In pre-processing, crucial steps that have been listed earlier in this study were taken to guarantee that the data is standardized for precise sentiment analysis.
3. Feature extraction techniques capture language-specific and culturally relevant nuances, strengthening the model's capacity to classify sentiments accurately.

http://.doi.org/10.46792/fuoyejet.v9i4.8 engineering.fuoye.edu.ng/journal

3.4 AI MODELS AND ALGORITHMS

The primary Al models employed in this research are Logistic regression, Support vector machines (SVM), random forest classifier, and Long short-term memory (LSTM) networks, which are being used for comparative analysis.

3.4.1 DESCRIPTION OF AI MODELS AND ALGORITHMS

1) A statistical technique for binary classification, logistic regression predicts one of two probable outcomes. The logistic (sigmoid) function is used to translate a linear combination of input data into a probability ranging from 0 to 1:

$$\sigma(z) = (1+e^{-z})^{-1}$$

Over most other binary classification algorithms, Logistic regression is favored for its clarity, comprehensibility, and efficiency. The log chances of the result and the input attributes are assumed to have a linear relationship. Regularization techniques like L1 and L2 are gradient methods that can be applied to prevent overfitting. Despite its assumptions, logistic regression is a foundational algorithm for binary classification exercises.

2) Support Vector Machine (SVC) is a supervised machine learning algorithm that uses binary classification exercises to solve complex classification, regression, and outlier detection by locating the best hyperplane in a high-dimensional space that maximally divides the data points of various classes.

Key concepts of SVC include:

1. Hyperplane: This is a decision-making limit that differentiates data points of two or more different classes. In n-dimensional space, a hyperplane is an (n-1) (n-1) (n-1) dimensional subspace.
2. Support Vectors: These are the data points nearest to the broken lines on the hyperplane and influence its position and orientation. Note that these points are critical in defining the optimal hyperplane.
3. Margin: SVC aims to extend this margin, which represents the separation between the nearest data points from each class and the hyperplane. Typically, this leads to an improved summary of unseen data.
4. Random Forest Classification is a group-supervised learning method utilized for assignments involving both regression and classification. During training, Random Forest builds many decision trees, with each node giving the next structural indication and displaying the classes' mode.

Key characteristics of Random Forest include:

1. Ensemble of Trees: This starts with a process called bootstrap sampling wherein a random sample of the data is used to train each decision tree using replacement. Each tree in the forest is grown to the maximum level without being shortened, resulting in the aggregation of decision trees.
2. Random Feature Selection: To introduce further randomness and help reduce correlation among trees, at each break in the decision tree, a random subset of characteristics is considered for splitting rather than all features.
3. Bagging (Bootstrap Aggregating): Each tree is trained on a bootstrap sample, which happens to be a random sample drawn with a substitute from the initial dataset. This method reduces variance and assists in preventing excessive fitting.
4. Voting/Averaging: For classification tasks, the final prediction is made by a majority vote of each tree. The average of each tree's output is the forecast for regression tasks.
5. An architecture of recurrent neural networks (RNNs) called Long Short-Term Memory (LSTM) was created to model sequences and time-series data. It tackles the vanishing gradient problem mostly found in standard RNNs. LSTMs are particularly well-suited for exercises involving language modeling, machine translation, and time-series prediction, making it possible for the network to detect long-range dependencies more efficiently.

Key components of LSTM include:

1. Memory Cell: Stores information over time, enabling the network to recall significant details from earlier in the sequence.
2. Gates: Manage the information that enters and exits the memory cell. There are three types of gates:
   - Forget Gate: Decides what information to remove from the cell state.
   - Input Gate: Decides which new information to include in the cell state.
   - Output Gate: Manages what information to output based on the cell state.

3.4.2 RATIONALE BEHIND MODEL SELECTION

The aforementioned machine learning algorithms were taken into consideration due to their classification ability, suitability as baseline models for sentiment analysis, and supervised learning nature, which necessitates labeled data for learning during model training. Logistic regression, Support vector machine, and random forest are mostly effective for binary classification exercises. Additional models are selected for their capability to detect complex patterns and provide comparative insights into the performance of all four algorithms.

3.5 TRAINING AND TESTING DATASETS

The dataset is categorized into two sections for training and testing. A stratified sampling method ensures that each set represents the distribution of sentiments and languages in the overall dataset.

3.5.1 PARTITIONING STRATEGIES

The dataset is partitioned into:

- 80% for training
- 20% for testing

3.5.2 PERFORMANCE METRICS

Performance metrics encompass accuracy, precision, recall, and Fl-score, providing an extensive evaluation of the model's effectiveness in classifying sentiments.

3.6 TECHNICAL IMPLEMENTATION

1) Programming Languages: Python is utilized for its comprehensive libraries and frameworks for machine learning and natural language processing.
2) Libraries and Frameworks: Scikit-learn for machine learning, NLTK and SpaCy for pre-processing of

http://.doi.org/10.46792/fuoyejet.v9i4.8 engineering.fuoye.edu.ng/journal

text, and TensorFlow/Keras for deep learning models.

3) Hardware/Software Configurations: To improve the training and evaluation of models, tests are carried out on a system equipped with a powerful GPU.

3.7 LIMITATIONS AND ASSUMPTIONS

3.7.1 METHODOLOGICAL LIMITATIONS

Acknowledged limitations include:

1. The focus on four major languages (Nigerian-Pidgin, Hausa, Yoruba, Igbo) may limit the generalizability to other Nigerian languages.
2. The reliance on social media data may not represent the entire population's sentiments.

3.7.2 DATA AND MODEL CONSTRAINTS

Constraints related to data and models include:

1. Limited availability of annotated datasets for certain languages.
2. The inherent limitations of logistic regression in capturing complex patterns compared to deep learning models.

3.7.3 IMPLICATIONS OF LIMITATIONS ON RESEARCH FINDINGS

The potential implications of these limitations on the research findings include:

1. Reduced accuracy and reliability in detecting nuanced sentiments.
2. Limited applicability of the findings to broader contexts beyond the specific languages and platforms studied.

By acknowledging these limitations and assumptions, the research maintains transparency and sets the stage for future improvements and extensions.

4 RESULTS AND DISCUSSION

4.1 PERFORMANCE METRICS FOR THE ALL FOUR (4) LANGUAGES DATASET COMBINED

	Accuracy	Precision	F1 Score	Recall
Logistic Regression	72%	72%	71%	72%
SVM	74%	76%	72%	74%
Random Forest	74%	76%	72%	74%
LSTM	72%	72%	72%	72%

Table 1: Performance metrics for the synthesis of all datasets in multilingual languages

In the synthesis of all datasets in multilingual languages, SVM and Random Forest models outperform Logistic Regression and LSTM across most metrics. Both SVM and Random Forest achieve the highest accuracy of 74% and precision of 76%, with a strong F1 score of 72% and recall of 74%. This indicates that SVM and random forest classifier models are better at handling the combined dataset of four languages, possibly due to their ability to manage complex patterns and feature interactions more effectively. To sum up, Logistic Regression, while showing good balance, has a slightly lower Fl score, indicating some imbalance between precision and recall. LSTM, on the other hand, demonstrates consistent performance with equal metrics across accuracy, precision, Fl score, and recall (72%) but does not outperform SVM and Random Forest.

4.2 COMPARATIVE ANALYSIS

This area of this study evaluates and emphasizes the performance of machine learning algorithms utilized in this research across four language datasets (Nigerian-Pidgin, Hausa, Igbo, and Yoruba) and a combined dataset.

4.2.1 INDIVIDUAL LANGUAGE PERFORMANCE

1) Nigerian-Pidgin Language Dataset: Accuracy and precision for all models except LSTM achieved 55% accuracy and precision. Recall and Fl score had similar patterns with 49% F1 score and 55% recall, while LSTM slightly performed better with 58% across all metrics.
2) Hausa Language Dataset: When evaluating the Hausa dataset, Logistic regression, SVM, and random forest witnessed a uniform performance highlighting identical metrics with 74% accuracy, 73% precision, 73% F1 score, and 74% recall, while LSTM was slightly better with 74% accuracy, 73% precision, 74% F1 score, and 74% recall.
3) Igbo Language Dataset: In the Igbo dataset logistic regression, SVM and random forest had uniform performances of 62% accuracy, 53% precision, 50% Fl score, and 62% recall, while LSTM outperformed others with 64% accuracy, 63% precision, 63% F1 score, and 64% recall.
4) Yoruba Language Dataset: The Yoruba dataset witnessed the highest overall performance from all models, that is, logistic regression, SVM, and random forest all had similar scores, which were 89% accuracy, 87% precision, 85% Fl score, and 89% recall. However, LSTM performed best with 90% across all metrics, making it the ideal model for this research work.

4.2.2 COMBINED DATASET PERFORMANCE

1) Logistic Regression: 72% accuracy, 72% precision, 71% Fl score, 72% recall.
2) SVM and Random Forest: Superior performance with 74% accuracy, 76% precision, 72% Fl score, and 74% recall.
3) LSTM: Consistent but not leading, with 72% across all metrics.

4.3 DISCUSSION

1) Logistic Regression: Demonstrates consistent performance across individual datasets but slightly lower in combined datasets with balanced metrics.
2) SVM and Random Forest: Show strong and consistent performance, particularly excelling in the combined dataset, making them suitable for complex, multilingual data.
3) LSTM: Exhibits the highest performance in individual datasets, particularly in the Igbo and Yoruba datasets, due to its capability to capture temporal dependencies but is slightly less effective in the combined dataset.

http://.doi.org/10.46792/fuoyejet.v9i4.8 engineering.fuoye.edu.ng/journal

5 CONCLUSION

This research investigates sentiment analysis in Nigeria's multilingual social media space, highlighting the effectiveness of models of machine learning like Logistic Regression, SVM, Random Forest, and LSTM. It addresses challenges such as code-switching and dialectal variations while contributing a curated multilingual dataset for upcoming research in languages with limited resources. Subsequent studies in sentiment analysis across multiple languages can concentrate on advanced deep learning models like transformers (e.g., BERT, GPT) for better linguistic nuance capture, curating larger and more diverse datasets, and implementing real-time sentiment analysis systems. It should also prioritize developing multilingual NLP techniques tailored to handle challenges like code-switching and dialectal variations and conducting cross-cultural studies to understand linguistic and cultural sentiment nuances within Nigeria. To conclude, this research work has investigated the four previously mentioned machine learning algorithms for the field of sentiment analysis in a multilingual context on Nigerian social media comments. Despite the challenges faced by a limited availability of data and linguistic diversity, the models achieved promising results, highlighting their potential for real-world applications. The comparative analysis with other models has shown the need for continued research and innovation in this field. A foundation for future advancement has been laid by addressing the challenges posed by multilingual sentiment analysis. The study draws to attention the valuable insights that can be gained from leveraging computational techniques to decipher the vast world of sentiments communicated in Nigeria's vibrant activities on social media platforms.

ACKNOWLEDGMENT

We would like to sincerely thank everyone who contributed to the completion of this research. We thank Ayomitope Isijola for his invaluable contribution to this research. Finally, we acknowledge the efforts and dedication of all authors involved in this work. Their collaboration, expertise, and hard work have been instrumental in every stage of the research process.

REFERENCES

Devlin, J., Chang, M.W., Lee, K., & Toutanova, K. (2019). BERT: Pre-Training of deep bidirectional transformers for language understanding. *NAACL-HLT,* pp. 4171-4186. doi: 10.4236/jss.2018.612029

Ihsan, U., Ashraf, H., & Jhanjhi, N.Z. (2023). Multilingual Sentiment Analysis Using Deep Learning: A Comprehensive Survey. doi: 10.20944/preprints202312.1990.v1

Karthika, P., Murugeswari, R., & Manoranjithem, R. (2019). Sentiment Analysis of Social Media Networks Using Random Forest Algorithm, pp. 1-5. doi: 10.1109/INCOS45849.2019.8951367

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., Jeorge, A., & Brazdil, P. (2022). NaijaSenti: A Nigerian Twitter Sentiment Corpus for Multilingual Sentiment Analysis. doi: 10.48550/arXiv.2201.08277

Qiao, R. & Huang, X. (2024). Application of Deep Learning in Cross-lingual Sentiment Analysis for Natural Language Processing. *Journal of Artificial Intelligence Practice,* 7(1), pp. 1-6. doi: 10.23977/jaip.2024.070101

Zhao, C., Wu, M., Yang, X., Zhang, W., Zhang, S., Wang, S., & Li, D. (2024). A Systematic Review of Cross-Lingual Sentiment Analysis: Tasks, Strategies, and Prospects. *ACM Computing Surveys,* 56(7), pp. 1-37. doi: 10.1145/3645106