**COMPARATIVE ANALYSIS OF MACHINE LEARNING ALGORITHMS FOR SENTIMENT ANALYSIS OF MULTILINGUAL NIGERIAN SOCIAL MEDIA COMMENTS**

**BY**

**NTUI SAMUEL AKWA**

**JULY 2024**

**DECLARATION**

I, Ntui Samuel Akwa hereby declare that the project work titled Comparative Analysis Of Machine Learning Algorithms For Sentiment Analysis Of Multilingual Nigerian Social Media Comments submitted towards partial fulfillment of requirements for the award of a Post Graduate Diploma in Computer Science is my original work and the dissertation has not formed the basis for award of any degree, fellowship or any similar title to the best of my knowledge. In instances where references of other works have been cited, full acknowledgement has been given.

**……………………………………………………………………………………….**

**Ntui Samuel Akwa**

**Date: ……………………………………………………………………………**

**CERTIFICATION**

This document confirms that the project titled Comparative Analysis Of Machine Learning Algorithms for Sentiment Analysis of Multilingual Nigerian Social Media Comments which was presented to the School of Postgraduate Studies at the University of Lagos to attain a Post Graduate Diploma in Computer Science, represents an authentic research endeavor conducted by Ntui Samuel Akwa within the Department of Computer Science.

**DR V.T. ODUMUYIWA**

PROJECT SUPERVISOR SIGNATURE AND DATE

**PROF A. P. ADEWOLE**

HEAD OF DEPARTMENT SIGNATURE AND DATE

**DEDICATION**

This project work is dedicated to the Almighty God, the giver of understanding, knowledge and wisdom.

**ACKNOWLEDGMENT**

My deepest appreciation goes to my mother, Late Mrs. Glory Ntui for her supportive words of encouragement, her fervent prayers and her relentless push towards my betterment before her passing, and to my father for his words of affirmation and prayers, both of which have greatly aided me throughout my programme.

I also want to express my warmest gratitude towards my loving siblings Mrs. Maria Oziabor, Mr Emmanuel (Kasid) Ntui, Mrs. Esther Otuwoye, especially Mrs. Hannah Obeten-Ekpo and Mrs. Peace Ukpaka for their significant contributions to the successful completion of this programme. Also, to my invaluable friends Mr. Christopher Azere, Mr. Ahmed Ehimare Audu and Ms. Ashioma Etukomeni for reminding me that the most valuable gift anyone can have, are true friends.

I extend my sincere appreciation and heartfelt gratitude to my exceptionally brilliant supervisor, Dr V.T Odumuyiwa, for his unwavering and substantial support. And also, my wonderful Head of Department, Prof. A.P. Adewole.

**ABSTRACT**

The study probes into sentiment analysis within the multilingual landscape of Nigerian social media comments by using machine learning algorithms as computational tools. Nigeria has a variety of languages which creates a complicated scenery for understanding sentiment dynamics in digital discourse. This research developed sentiment classification models across languages such as English, Hausa, Yoruba, Nigerian-Pidgin and Igbo. By utilizing machine learning algorithms such as Logistic Regression, Support Vector Machine (SVM), Random Forest Classification (RFC) and Long Short-Term Memory (LSTM), the project was aimed at providing insights into the varieties of expressions and sentiment in Nigerian social media. The research explores the domains of natural language processing and socio-linguistics in its study.

In summary, the research finds that SVM and Random Forest are the most effective 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. The choice of algorithm should consider the specific dataset and the desired balance between precision and recall.

**TABLE OF CONTENTS**

Title Page i

Declaration ii

Certification iii

Dedication iv

Acknowledgement v

Abstract vi

Table of Content vii

List of Figure xi

**CHAPTER ONE: INTRODUCTION**

1.1 Introduction:

1.2 Statement of Problem

1.3 Aim and Objectives

1.4 Scope of Study

1.5 Methodology

1.6 Significance of The Project

1.7 Project Outline

**CHAPTER TWO: LITERATURE REVIEW**

2.1 Introduction to Sentiment Analysis

2.1.1 Significance of Sentiment Analysis

2.1.2 Overview of Sentiment Analysis Techniques

2.2 Sentiment Analysis in Social Media

2.2.1 Importance of Sentiment Analysis in Social Media

2.2.2 Previous Studies and Their Findings

2.2.3 Challenges Specific to Social Media Sentiment Analysis

2.3 Sentiment Analysis Techniques

2.3.1 Rule-Based Approaches

2.3.2 Machine Learning Approaches

2.3.3 Deep Learning Approaches

2.3.4 Comparison of Approaches

2.3.5 Applications of Sentiment Analysis Techniques

2.4 Sentiment Analysis on Multilingual Data

2.4.1 Challenges of Multilingual Sentiment Analysis

2.4.2 Techniques for Multilingual Sentiment Analysis

2.4.3 Applications in the Nigerian Context

2.4.4 Case Studies and Findings

2.5 Machine Learning Algorithms in Sentiment Analysis

2.5.1 Overview of Logistic Regression

2.5.2 Overview of Support Vector Machines (SVM)

2.5.3 Overview of Random Forests

2.5.4 Overview of Long Short-Term Memory (LSTMs)

## 2.6 Case Studies: Applying Machine Learning Algorithms to Sentiment Analysis

2.7 Sentiment Analysis in Multilingual Nigerian Context

2.7.1 Linguistic Diversity in Nigeria

2.7.2 Challenges in Multilingual Sentiment Analysis

2.7.3 Methodologies for Multilingual Sentiment Analysis

2.7.4 Applications and Impact

2.7.4 Applications and Impact

**CHAPTER THREE: METHODOLOGY**

3.1 Research Design

3.1.1 Research Design Objectives:

3.1.2 Overview of Research Design

3.1.3 Exploratory Aspects

3.1.4 Descriptive Aspects

3.1.5 Experimental Aspects

3.1.6 Alignment with Research Objectives

3.2 Data Collection

3.2.1 Data Acquisition Process

3.2.2 Sources and Nature of Data

3.2.3 Dataset Overview

3.3 Data Preprocessing Steps

3.3.1 Justification of Data Sources and Preprocessing Techniques

3.4 AI Models and Algorithms

3.4.1 Description of AI Models and Algorithms

3.4.2 Rationale Behind Model Selection

3.5 Experimental Design

3.5.1 Experiment Setup

3.5.2 Training and Testing Datasets

3.5.3 Partitioning Strategies

3.5.4 Cross-Validation Techniques

3.5.5 Performance Metrics

3.6 Implementation Details

3.6.1 Technical Implementation

3.5.2 Software Development Environment

3.7 Ethical Considerations

3.7.1 Data Collection and Usage

3.7.2 Addressing Bias and Fairness

3.7.3 Societal Impact

3.8 Limitations and Assumptions

3.8.1 Methodological Limitations

3.8.2 Data and Model Constraints

3.8.3 Implications on Research Findings

**CHAPTER FOUR: RESULTS**

4.1 For Individual Languages

4.1.1 Comparing Analysis for Nigerian-Pidgin Language Dataset

4.1.2 Comparing Analysis for Hausa Language Dataset

4.1.3 Comparing Analysis for Igbo Language Dataset

4.1.4 Comparing Analysis for Yoruba Language Dataset

4.2 Performance Metrics for the All Four (4) Languages Dataset Combined

4.3 Comparative Analysis

4.3.1 Individual Language Performance

4.3.2 Combined Dataset Performance

4.4 Discussion

**CHAPTER 5: CONCLUSION**

5.1 Summary of Findings

5.2 Summary of Work

5.3 Recommendations for Future Work

5.4 Conclusion

**REFERENCES**

**LIST OF FIGURES**

Fig 1: Nigerian-Pidgin Multilingual DataFrame before Preprocessing

Fig 2: Hausa Language Multilingual DataFrame before Preprocessing

Fig 3: Igbo Language Multilingual DataFrame after Preprocessing

Fig 4: Yoruba Language Multilingual DataFrame after Preprocessing

Fig 5: Total Multilingual Dataset while Preprocessing

**LIST OF TABLES**

Table 4.0 Performance Metrics on Pidgin Language Dataset

Table 4.1 Performance Metrics on Hausa Language Dataset

Table 4.2 Performance Metrics on Igbo Language Dataset

Table 4.3 Performance Metrics on Yoruba Language Dataset

Table 4.4 Performance Metrics for Combination of all Multilingual Language Datasets

**CHAPTER 1: INTRODUCTION**

**1.1 Introduction**

In an era defined by digital interconnectedness, social media platforms have become 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, a huge cultural diversity which promotes a world of social media dialogue using simultaneously several languages 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 et al., 2018; Huang et al., 2022).

As Nigeria's social media digital 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 an effort. This simply 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.

The aim of this study is to explore sentiment analysis in multilingual Nigerian social media, using computational analysis to shotlight the intricate relationship between language and culture in digital expressions. By utilizing logistic regression, support vector machine (SVM), random forest and long short term memory (LSTM), this research intends to classify sentiments effectively, providing simple insights into sentiment trends of Nigerian 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 are mostly effective as binary classification.

**1.2 Statement of Problem**

The increase in social media platforms has changed communication, information sharing and expression of opinion in the world at large. As social media platforms become huge revenue channels for individuals, making it engaging enough to revolutionize communication, information sharing, and a canvas to express opinions and sentiments. A country like Nigeria which is celebrated for its linguistic richness and cultural diversity, social media platforms have become an enabling component for multilingual commentary in languages like Nigerian-Pidgin, Hausa, Yoruba, and Igbo. Nevertheless, it has been difficult to discern the sentiments within this multilingual discourse due to language complexities like idioms, background meanings, culture and contextual variations (Olaleye et al., 2018; Bamman et al., 2014).

The challenge this research work seeks to address is a prevalent discourse of effective tools and methodologies for analyzing sentiments in multilingual Nigerian social media comments accurately. The existing sentiment analysis techniques are sometimes unable to give accurate meanings to some comments because of the cultural context and figure of speech within the Nigerian languages, which eventually leads to a reduced performance of analysis and inaccurate results. This study aims to investigate robust sentiment analysis models using machine learning algorithms like Logistic Regression, Support Vector Machine SVM, Random Forest and Long Short-Term Memory (LSTM) to classify sentiments expressed in Nigerian social media comments which are multilingual. Addressing this gap would enable this research to provide valuable insights into sentiment dynamics in Nigerian social media discourse (Mohammad et al., 2013).

**1.3 Aim and Objectives**

**Aim**

The aim of this study is to conduct sentiment analysis on multilingual Nigerian social media comments using machine learning algorithms, with the task of accurately classifying sentiments expressed in different linguistic domains.

**Objectives**

The objectives of the work are:

1. To curate and preprocess a diverse dataset of multilingual Nigerian social media comments from the NaijaSenti dataset and various platforms, including but not limited to Twitter, Facebook, and online forums.
2. To train classification models using four machine learning algorithms, segmenting sentiments expressed in the curated social media comments into positive, negative, or neutral categories.
3. To evaluate the performance of the sentiment classifiers developed in the second objective using metrics such as accuracy, precision, recall and F-score to compare which model is most effective.

**1.4 Scope of Study**

This research focuses on developing sentiment analysis systems, specifically put together to analyze multilingual Nigerian languages used in social media comments. The system aims to classify sentiments expressed in languages such as English, Hausa, Yoruba, Nigerian-Pidgin and Igbo on various social media platforms which include Twitter, Facebook and other online forums.

**1.5 Methodology**

Multilingual Data Collection: To curate social media comments written in a mixture of Nigerian-Pidgin, Hausa, Yoruba and Igbo from vital platforms.

Preprocessing: Text preprocessing techniques will be used to clean and balance the data collected after which tasks like tokenization, normalization and removal of noise from the dataset would be performed.

Feature Extraction: Specific Language and vital cultural features will be gotten from the preprocessed data to show the variations of sentiment expressions in different languages in Nigerian social media (Bamman et al., 2014).

Machine Learning Algorithms: The system will implement algorithms like logistic regression, Support Vector Machine (SVM) Random Forest Classifier (RFC), Long Short-Term Memory (LSTM) to classify sentiments into positive, negative or neutral.

Model Evaluation: The performance of the sentiment analysis models would be evaluated and compared using metrics like accuracy, precision, recall and F1-score (Devlin et al., 2019).

Interpretation and Comparison **:** Analysis and interpretation of the results from the four machine learning algorithms will be trained and tested to get insights of sentiment analysis, and models compared to shown which algorithm functions best.

**Limitations**

Language Coverage: The system will focus mainly on English, Hausa, Yoruba, Nigerian-Pidgin and Igbo languages as well as limiting its application to other Nigerian languages.

Data Availability: The effectiveness of sentiment analysis models may be reduced by the availability and the quality data and the various language social media commentary.

Cultural Deep Meanings: The cultural variety in sentiment expressions which would be obtained although the system may not be able to fully account for all cultural variations present in the Nigerian social media field.(Mohammad et al., 2013).

**1.5 Methodology**

This section outlines the methodology used in conducting the research on sentiment analysis of multilingual Nigerian social media comments using machine learning algorithms.

Data Collection: Curate a large dataset of social media comments in either English, Hausa, Yoruba, Igbo or Nigerian-Pidgin from NaijaSenti research work and other online forums and platforms such as Twitter, Facebook and Instagram.

Data Preprocessing: Clean and preprocess all data collected and perform other tasks like tokenization and handling emoticons and slang.

Feature Extraction: Extract language-specific and vital language features from the preprocessed data like TF-IDF and word embeddings (Huang et al., 2022).

Models Development: Implement logistic regression, support vector machine, random forest and long short-term memory for classification of sentiment analysis after which the dataset is splitted and distributed evenly to ensure the model is weighted properly.

Model Training: Each model is trained by its respective algorithm, accessing both individual language dataset and the combined language dataset using appropriate optimization techniques.

Model Evaluation: To investigate and decipher the performance of the trained models on the testing dataset using metrics such as accuracy, precision, recall and F1-score (Devlin et al., 2019).

Analysis and Comparison: Conduct analysis and interpretation of the results to reveal insights into sentiment dynamics in different across several languages and cultural contexts in Nigerian social media discourse deeply.

Ethical Considerations: Ensure ethical handling of social media data by respecting user privacy, gaining consent, acknowledging and addressing any biases or limitations in the data and methodology (Mohammad et al., 2013).

This research methodology provides a structured approach for conducting the study on sentiment analysis of multilingual Nigerian social media comments by ensuring diligence and reliability in the research process.

**1.6 Significance of The Project**

This research holds significant importance for stakeholders like academics, industry practitioners and policymakers and by developing a large sentiment analysis system for multilingual Nigerian social media comments, this project aims to:

Enhance Understanding: Provide insights into sentiment dynamics within Nigerian social media discourse and contribute to a deeper understanding of language, culture and digital expression (Olaleye et al., 2018).

Inform Decision-Making: Provide very rich information for market research, public opinion analysis and sociolinguistic studies which aids decision making processes in various domains.

Improve Technology**:** Improve the development of sentiment analysis techniques for diverse linguistic and cultural contexts which would pave way for more accurate, naturally and culturally-sensitive language processing solutions (Liu, 2012).

Foster Dialogue**:** Facilitate informed discussions on the role of social media in shaping public opinion and cultural identity in Nigeria which would improve dialogue and discourse on digital communication practices.

**1.7 Project Outline**

This streamlined project outline reflects the proposed chapter structure of the thesis by providing a clear roadmap for the reader.

1. Introduction

Provides an overview of the research topic, objectives and significance of the study.

1. Literature Review

Review existing literature on sentiment analysis, multilingual NLP and social media discourse analysis.

1. Research Design

Describes the research methodology which includes data collection, preprocessing and model development.

1. Implementation (Experiment, Results and Discussion)

Presents the experimental setup, results of sentiment analysis experiments and the analysis and interpretation of the findings.

1. Conclusion

Summarizes the key findings of the study by highlighting contributions to the field thereafter suggesting areas for future research.

**CHAPTER 2: LITERATURE REVIEW**

**2.1 Introduction to Sentiment Analysis**

Sentiment analysis is also known as opinion mining which is a field of study that is part of natural language processing (NLP), focusing on identifying 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 rise of social media and online reviews where a large amount of user-generated content is produced on a regular basis thereby showing valuable insights into public opinion(Liu, 2012).

Sentiment analysis aims to classify text as positive, negative or neutral and more advanced systems can as well detect the strength of the sentiment shown, if it is very positive or mildly negative and recognize emotions like joy, anger or sadness. The process usually involves several steps like text preprocessing, feature extraction, sentiment classification and evaluation (Liu, 2012).

**2.1.1 Significance of Sentiment Analysis**

The importance of sentiment analysis extends over many sectors such as businesses who 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 assess the reaction of the public to news, stories and events.

**2.1.2 Overview of Sentiment Analysis Techniques**

Sentiment analysis techniques can be classified broadly into rule-based, machine learning and deep learning approaches and each of these techniques has their own strengths and weaknesses.

Rule-based approaches are easy to implement but can be rigid and fail to generalize well. Machine learning models provide better performance but require labeled data and feature engineering. Deep learning models are highly accurate but demand high computational power and large amounts of data.

**2.2 Sentiment Analysis in Social Media**

The rapid increase of social media platforms has changed how people communicate, share ideas and influence public opinions greatly. Social networking platforms like Twitter, Facebook and Instagram are rich sources of user-generated content that provides 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 (Huang et al., 2022).

**2.2.1 Importance of Sentiment Analysis in Social Media**

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

* Business and Marketing: Companies use social media sentiment analysis to understand customer feelings towards their products and services which 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.
* Politics and Public Opinion: Politicians and policymakers analyze social media sentiment to understand the opinion of the public on policies, candidates and current events as the feedback obtained can guide campaign strategies and decision making on policies. During elections, social media sentiment analysis provides insights into voter preferences and what are their key issues of concern.
* Healthcare: In the healthcare system sentiment analysis helps understand the feedback of the services patients received in order 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.
* 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, showing also the feasibility of using tweets as a data source for extracting people's opinion(Mohammad et al., 2013). Others have investigated sentiment strength detection in social web content which gives insights into how effective various sentiment analysis techniques are (NaijaSenti 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 (Huang et al., 2022).

**2.2.3 Challenges Specific to Social Media Sentiment Analysis**

Despite its potential, social media sentiment analysis presents several challenges:

* Short and Informal Text: Social media posts are most times short and written in informal language like slang, abbreviations and emojis which can make it difficult for traditional NLP techniques to interpret the sentiment accurately.
* Sarcasm and Irony: Detecting sarcasm and irony is really 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.
* Multilingualism: Social media users frequently switch between languages or use multiple languages in a single post and it is essential that models that can handle multilingual data are developed for accurate sentiment analysis especially in regions like Nigeria where there are different languages (Huang et al., 2022).
* Noise and Spam: Social media platforms are filled with noise and spam which can reduce the quality of sentiment analysis results. Effective data cleaning and filtering techniques are necessary to ensure the quality of the analysis.

**2.3 Sentiment Analysis Techniques**

Sentiment analysis techniques have evolved increasingly over the years and has adapted to the complex human languages and the diverse ways in which sentiments are expressed. These techniques can be grouped broadly into three main approaches: rule-based methods, machine learning methods and deep learning methods.

**2.3.1 Rule-Based Approaches**

Rule-based approaches use manually coded rules to identify and classify sentiment through the use of sentiment lexicons, which are lists of words associated with positive or negative sentiments. When analyzing a text the system counts the number of positive and negative words to determine the overall sentiment and while it is simple and easy to implement, rule-based approaches many times have issues with context, sarcasm and the varieties of human language. They can be rigid and fail to generalize well in different domains and languages (Liu, 2012).

**2.3.2 Machine Learning Approaches**

Machine learning approaches have brought a huge improvement to sentiment analysis and it involves training models on annotated datasets where the sentiment of each text sample is labeled. Common algorithms used in sentiment analysis include Naive Bayes, Support Vector Machines (SVM) and Logistic Regression. These models learn patterns and features from the training data which enables them to classify sentiments in new and unseen texts. Machine learning approaches generally offer better performance than rule-based methods because they can capture more complex patterns and relationships in the data but require a very large amount of labeled data and careful feature engineering to achieve good results (Mohammad et al., 2013).

**2.3.3 Deep Learning Approaches**

Deep learning approaches represent the latest advancements in sentiment analysis which leverage on neural networks mainly in architectures like Long Short-Term Memory (LSTM) networks, Convolutional Neural Networks (CNN) and transformer-based models such as BERT and GPT. Deep learning models are capable of capturing important patterns and dependencies in the data most times doing better than traditional machine learning models. They can automatically learn things from raw text which reduces the need for manual feature engineering; however, deep learning models require significant computational resources and large datasets for training and need careful tuning and regularization to avoid overfitting (Devlin et al., 2019).

**2.3.4 Comparison of Approaches**

Each sentiment analysis technique has its strengths and weaknesses. Rule-based approaches are straightforward to implement and interpret but lack flexibility and cannot cover a vast area. Machine learning models strike a balance between simplicity and performance by providing improved accuracy and adaptability at the cost of requiring labeled data and feature engineering. Deep learning models are highly accurate and capable of handling complex language but demand very significant resources and expertise to develop and use effectively.

**2.3.5 Applications of Sentiment Analysis Techniques**

Sentiment analysis techniques have been applied across various fields with great success. In business they are used to monitor customer feedback, track brand sentiment and analyze market trends. It helps to understand public opinion on policies and candidates thereby providing valuable insights for campaign strategies in politics. In the healthcare system, they analyze patient feedback and social media discussions to improve healthcare services and identify public health concerns. Media organizations use it to measure the reactions of their audience to content and tailor what they offer accordingly.

**2.4 Sentiment Analysis on Multilingual Data**

As the internet connects a large number of people, it increases the need to analyze sentiments expressed in many languages which have changed overtime. Sentiment analysis on multilingual data creates unique challenges and opportunities in areas like Nigeria where multiple languages and dialects are common in social media interactions.

**2.4.1 Challenges of Multilingual Sentiment Analysis**

* Language Diversity: Handling multiple languages requires the development of models that decipher and interpret many language structures, idiomatic expressions and cultural variations. In Nigeria, language diversity spans across major dialects which include Nigerian-Pidgin, Hausa, Yoruba and Igbo as well as other minor languages.
* Code-Switching: It is the process by which social media users switch between languages in the same post or even sentence which can sometimes make it difficult to comprehend the context or determine which sentiments are expressed using traditional language models accurately.
* Lack of Resources: Many less spoken languages lack comprehensive language resources such as sentiment lexicons, annotated corpora and language models which reduces the development of large sentiment analysis systems for these languages (Huang et al., 2022).
* Translation Issues: Translating text to a single target language for analysis can introduce errors and lose context which affects sentiment accuracy as directly analyzing text in its original language is most times more accurate but requires specialized models for each language.

**2.4.2 Techniques for Multilingual Sentiment Analysis**

* Machine Translation: One common approach is to translate all text into a single language (usually English) and then perform sentiment analysis using models trained on that language. This can make the process simple but lead to loss of sentiment variations and introduce translation errors.
* Multilingual Embeddings: Techniques like multilingual word embeddings and language models (e.g., MUSE, mBERT) allow for the presentation of text in different languages within a common vector space which allows the development of models that can process multiple languages at the same time to preserve the semantic meaning and context.
* Transfer Learning: Trained models on large datasets in resource rich languages can be used with smaller specific language datasets so as to utilize the knowledge learned from the larger dataset to improve performance on less resourced languages (Devlin et al., 2019).
* Hybrid Approaches: Combining machine translation, multilingual embeddings and transfer learning can create a large system which can handle data from many languages. These hybrid models can adapt to the variations of different languages and provide more accurate sentiment analysis.

**2.4.3 Applications in the Nigerian Context**

In Nigeria, sentiment analysis on multilingual social media data can provide relevant insights into public opinion in different languages and cultural groups. It means understanding customer sentiments from a larger demographic and using more inclusive marketing strategies in business while for policymakers it provides a large understanding of public opinion which helps them to create policies that are in line with many communities. In the healthcare system, analyzing multilingual feedback can improve service delivery by addressing the concerns of different language groups.

**2.4.4 Case Studies and Findings**

Several studies have highlighted the effectiveness of multilingual sentiment analysis, for instance studies from multilingual social media data from Nigerian users has shown that incorporating local languages with English improves sentiment classification and greatly increased its accuracy (Huang et al., 2022). These studies show how important it is to consider various languages in sentiment analysis to get the true sentiment expressed by users.

**2.5 Machine Learning Algorithms in Sentiment Analysis**

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

**2.5.1 Overview of Logistic Regression**

Logistic regression is a type of regression analysis used to predict the outcome of a binary dependent variable based on one or more independent variables. It estimates the probability that a given input belongs to a particular class. The logistic function is also known as the sigmoid function which maps 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

* Simplicity and Interpretability: Logistic regression is easy to understand 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.
* Efficiency: Logistic regression is computationally efficient which is advantageous when dealing with large datasets. Its relatively low computational complexity allows it to scale well.
* Performance: Despite it being simple logistic regression often performs competitively with more complex models, especially when the relationship between features and the target is approximately linear.

Limitations of Logistic Regression

* Linearity Assumption: Logistic regression assumes a linear relationship between the independent variables and the log-odds of the dependent variable which can limit its performance when dealing with more complex and non-linear data.
* Feature Engineering: The effectiveness of logistic regression depends heavily on the quality of the input features and significant effort may be required to preprocess text data and engineer relevant features like n-grams or TF-IDF scores.
* Multicollinearity: High correlation among features can affect the model’s coefficients negatively which can lead to instability and reduced interpretability. Regularization techniques such as L1 (Lasso) or L2 (Ridge) can mitigate this issue.

Implementation in Sentiment Analysis

* Text Preprocessing: Before applying logistic regression the text data must be cleaned and transformed into a numerical way which can be done using steps such as tokenization, removal of stop words, stemming or lemmatization and converting text into feature vectors using methods like Bag of Words or TF-IDF.
* Feature Selection: Effective feature selection is very important for improving model performance and ability to interpret. Techniques such as chi-square testing or mutual information can help identify the most important features for sentiment classification.
* Model Training and Evaluation: Logistic regression models are trained on labeled datasets where the sentiment of each text sample is known. The training process involves optimizing the model’s coefficients to minimize the error in predicting the target sentiment. Model evaluation is typically performed using metrics such as accuracy, precision, recall and F1-score often with cross-validation to ensure versatility.

**2.5.2 Overview of Support Vector Machines (SVM)**

Support Vector Machines (SVMs) are a powerful machine learning technique that suits sentiment analysis tasks involving classification of text data into categories like positive, negative or neutral. SVMs aim to find a hyperplane in a high-dimensional feature space that best separates comments with various sentiment polarities.

Advantages of SVM for Sentiment Analysis

* High Accuracy**:** SVMs is good at finding maximum decision boundaries which leads to potentially high classification accuracy in sentiment analysis.
* 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.
* Works Well with High-Dimensional Data**:** Sentiment analysis involves analyzing text data which can be represented in a high-dimensional feature space and they are equipped to handle such data efficiently.

Limitations of SVM for Multilingual Social Media Sentiment Analysis

* 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.
* Limited Interpretability: SVMs can make it difficult to understand how they arrive at certain classifications.This can be a drawback for analyzing the variations of sentiment expressed in different languages.
* Scalability: Training SVMs on massive datasets of multilingual social media comments can be statistically expensive.

**2.5.3 Overview of Random Forests**

Random Forests (RFs) have emerged as a prominent technique in sentiment analysis mainly for classifying the sentiment of social media comments. This type of learning method uses the collective strength of numerous decision trees of which each of them act as a classifier in order to achieve a large and accurate sentiment identification (Karthika et al., 2019).

Advantages of Random Forests

* Accuracy Prowess**:** Random Forests are well known for their ability to deliver high classification accuracy (Biba et al., 2017). This makes them highly reliable for discerning positive, negative or neutral sentiment within social media comments.
* Resilient to Noise: Social media platforms are known for their informality often composed mainly with slang, typos and casual language. Random Forests excel in such environments as they do not overfit to specific data points. Each decision tree casts a vote where the majority determine the final classification. This approach makes them less open to the noise associated with social media data (Sentiment Analysis Using Random Forest Algorithm-Online Social Media Based, 2019).
* Multilingual Potential: Random Forests demonstrate a degree of language independence. They can effectively work with various data types like text, without requiring extensive feature scaling (Wróblewski, L., & Czarnowski, P., 2020). This is a significant advantage for sentiment analysis tasks involving multilingual social media comments.

Limitations and Considerations

* Interpretability Challenges: Random Forests can be likened to a black box despite its accuracy. It is difficult to understand the reasoning behind their specific classifications. which causes a problem when trying to analyze the various sentiments expressed across different languages (Mohammad, S., & Khoo, F. S., 2017).
* Computational Demands: Training a Random Forest on a massive dataset of multilingual comments can be so much of statistical and computer work(Biba et al., 2017).
* **b**As with other machine learning approaches, crafting informative features from text data is vital for effective sentiment analysis. This is particularly true for multilingual comments where cultural and linguistic variations really impact sentiment expression (Wróblewski, L., & Czarnowski, P., 2020).

**2.5.4 Overview of Long Short-Term Memory (LSTMs)**

Long Short-Term Memory (LSTM) networks 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 (Jo & Song, 2017).

Advantages of LSTMs for Sentiment Analysis

* Capturing Context: LSTMs possess a unique architecture that allows them to effectively learn and retain long term dependencies within text 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 et al., 2016). For instance, sarcasm can be easily missed by simpler models that don't account for contextual cues.
* Multilingual Capabilities: LSTMs can be effectively applied to sentiment analysis tasks involving multiple languages that do not have the ability to learn from sequential data which makes them less reliant on language-specific features compared to traditional machine learning approaches (Zhang et al., 2018). This is advantageous for projects dealing with multilingual social media comments.
* Adaptability: LSTMs can be flexibly integrated with other deep learning architectures like convolutional neural networks (CNNs). This allows for the incorporation of additional features to greatly improve sentiment analysis performance (Wang et al., 2016).

Limitations and Considerations

* Computational Demands: Training LSTMs, especially on large datasets of multilingual social media comments can be expensive. This requires significant processing power and resources (Joulin et al., 2016).
* Data Dependence: The effectiveness of LSTMs depends highly on the quality and size of the training data. Insufficient or poorly labeled data can hinder their performance in sentiment analysis tasks (Yang et al., 2016).
* Hyperparameter Tuning: LSTMs involve various hyperparameters that greatly impact their performance. Optimizing these hyperparameters can be complex and consume time (Hochreiter & Schmidhuber, 1997).

## 2.6 Case Studies: Applying Machine Learning Algorithms to Sentiment Analysis

This section explores how various machine learning algorithms can be used for sentiment analysis and drawing on the example text you provided.

1. Logistic Regression:

* Case Study: A retail company wants to analyze customer reviews on their e-commerce platform to understand overall sentiment towards a new product line. Logistic regression can be used to classify reviews as positive, negative or neutral based on word features gotten from the text.
* Application: How simple and the interpretability of logistic regression make it ideal for establishing a baseline sentiment analysis model. By analyzing the coefficients assigned to different words the company can be able to know the specific factors increasing positive or negative sentiment (e.g., identifying features associated with praise or complaints). (Huang et al., 2022)

2. Support Vector Machine (SVM):

* Case Study**:** A social media monitoring company wants to track brand sentiment across various platforms in real time. An SVM can be trained on a labeled dataset of social media comments to classify sentiment effectively..
* Application: SVMs is able to handle high dimensional data like text features extracted from social media comments. Their ability to find optimal decision boundaries allows for accurate on site sentiment classification which allows the company to identify possible brand crises or positive trends fast. (Wang et al., 2009)

3. Random Forest:

* Case Study:A Random Forest can be trained on a large dataset of social media comments to analyze public sentiment on social media as regards a specific policy proposal.
* Application: Random Forests provide a large amount of noise and can handle large, messy datasets mostly seen in social media sentiment analysis. By using the wisdom of multiple decision trees, a Random Forest can provide a more accurate picture of public sentiment towards the policy proposal compared to simpler models. (Mohammad et al., 2017)

4. Long Short-Term Memory (LSTM):

* Case Study: An LSTM network can be trained on a dataset of labeled movie reviews to capture the variations of language context to analyze the sentiment of movie reviews that most times use sarcasm or humor to express opinions.
* Application: LSTMs is good at capturing long term dependencies within text data as it allows them to analyze the sentiment of movie reviews by considering the context of surrounding words which leads to more accurate understanding of if a positive review is actually sarcastic or negative (Tang et al., 2016).

**2.7 Sentiment Analysis in Multilingual Nigerian Context**

The multiple languages used in Nigeria makes it difficult to accurately do sentiment analysis as over 500 languages are spoken in the country where analyzing social media sentiments requires handling of 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.

**2.7.1 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.

**2.7.2 Challenges in Multilingual Sentiment Analysis**

* 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 use of code-switching, where users switch between languages in a single post or sentence.
* Resource Scarcity: Many Nigerian languages lack comprehensive language resources, such as annotated corpora, sentiment lexicons and pre-trained language models. This scarcity hinders the development of accurate sentiment analysis tools (Huang et al., 2022).
* Cultural Nuances: Sentiment expressions can really 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.
* Data Quality: Social media data is most times full of informal languages, slang, abbreviations and typographical errors. Cleaning and preprocessing this data to extract meaningful features for sentiment analysis is a significant challenge.

**2.7.3 Methodologies for Multilingual Sentiment Analysis**

* Language Detection and Segmentation: Advanced language detection algorithms are employed to identify and segment different languages within a single post. Tools like Google’s Compact Language Detector (CLD) or fastText can be used for this purpose.
* 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).
* 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 lose sentiment nuances.
* Hybrid Models: Combining multiple approaches like multilingual embeddings for initial representation and translation can improve accuracy of results. Hybrid models utilize the strengths of different techniques to handle complex multilingual data.

**2.7.4 Applications and Impact**

In the Nigerian context, sentiment analysis has significant applications across various sectors:

* Business: Companies can monitor customer feedback from different languages to have an understanding into consumer preferences and improve products and services to meet diverse linguistic groups.
* Politics: Politicians and policymakers can measure public opinion on policies and election candidates to get the sentiments expressed by different linguistic communities.
* Healthcare: Analyzing sentiments in health-related discussions can help identify public concerns, misinformation, and areas for improvement in healthcare services, accommodating the linguistic diversity of the population.
* Media and Entertainment: Media organizations can track audience reactions to content to ensure that their offerings tally with different language and cultural groups.

**2.7.5 Case Studies and Research Findings**

Several studies have explored sentiment analysis in the Nigerian context. For example, research on analyzing Twitter data in multiple Nigerian languages has shown that incorporating local languages alongside English significantly enhances sentiment classification accuracy. These studies highlight the importance of adapting sentiment analysis methodologies to handle linguistic diversity effectively (Huang et al., 2022).

**CHAPTER THREE: RESEARCH DESIGN**

**3.1 Research Design**

The main objective of this research project is to conduct a comparative analysis 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).
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.
4. 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 on 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 design 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:

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

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

**3.1.4 Descriptive Aspects**

In the descriptive design, attempts to highlight and illustrate the emotions expressed in the multilingual comments by:

* Curating a dataset from social media comments leveraging platforms such as Twitter, Facebook and Instagram etc.
* Adding Labels to the dataset with sentiment meanings which are either positive, negative and neutral using both manual and automated methods.
* 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:

* 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 dataset individually.
* Concatenating all the mentioned datasets and develop models from each of the algorithms.
* Comparing the performance of logistic regression, support vector machine, random forest and long short-term memory to know the performance of each based on the metric gained.

**3.1.6 Alignment with Research Objectives**

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

* 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.
* The aim of the descriptive phase is to provide a detailed account of sentiment annotation and distribution, fulfilling the objective of portraying sentiments in a multilingual context.
* The experimental phase tests and validates the use of all four machine learning algorithms, meeting the objective of evaluating and comparing sentiment analysis techniques for Nigerian languages.

By combining exploratory, descriptive and experimental approaches, this research design 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 comment of widely spoken Nigerian language like Nigerian-Pidgin, Igbo, Hausa, Yoruba. The dataset was later annotated to include comments labeled as positive, negative, or neutral sentiment.

**3.2.1 Data Acquisition Process**

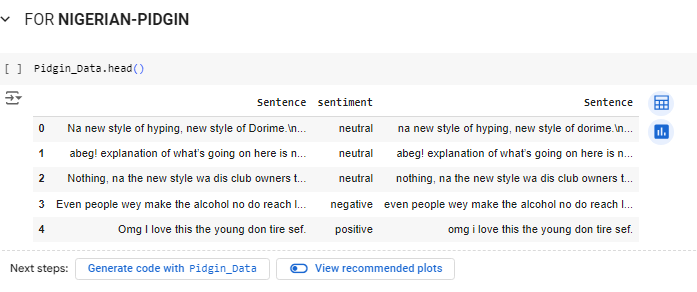
The process of data acquisition involved:

* Scraping social media platforms manually using Google Forms as a tool to collect and curate posts and comments.
* Filtering data to include only relevant posts written in the target languages and involving a mixture of more than Nigerian dialect.
* Preprocessing 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 platforms 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.

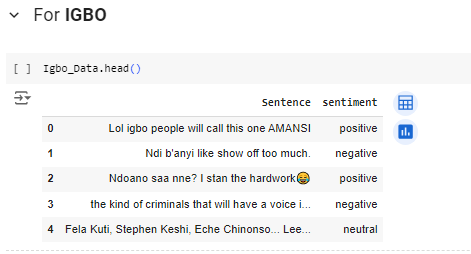
**3.2.3 Dataset Overview**



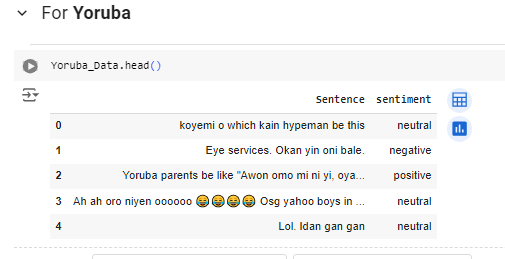
**Fig 1: Nigerian-Pidgin Multilingual DataFrame before Preprocessing**



**Fig 2: Hausa Language Multilingual DataFrame before Preprocessing**

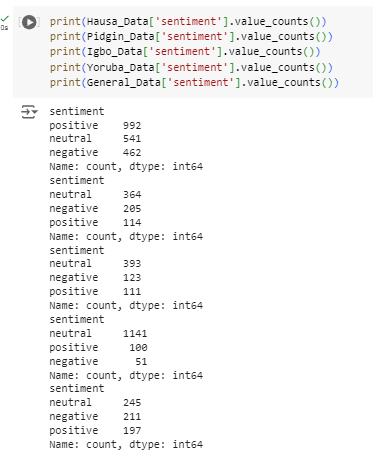


**Fig 3: Igbo Language Multilingual DataFrame after Preprocessing**



**Fig 4: Yoruba Language Multilingual DataFrame after Preprocessing**

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



**Fig 5: Total Multilingual Dataset while Preprocessing**

**3.3 Data Preprocessing 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 so as 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 which 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 preprocessing techniques are justified as relevant on the basis of:

* Large social networking platforms which 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.
* In preprocessing, crucial steps which have been listed earlier in this study were taken to ensure that the data is clean and standardized for accurate sentiment analysis.
* Feature extraction techniques capture language-specific and culturally relevant nuances, improving the model’s ability to classify sentiments accurately.

**3.4 AI Models and Algorithms**

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

**3.3.1 Description of AI Models and Algorithms**

* **Logistic regression** is a statistical method used for binary classification, predicting one of two possible outcomes. It uses the logistic (sigmoid) function to map a linear combination of input features to a probability between 0 and 1:

σ(z) = (1+e-Z)-1​

Over most other binary classification algorithms, Logistic regression is favored for its simplicity, interpretability and efficiency. It assumes a linear relationship between input features and the log-odds of the outcome. 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 tasks.

* **Support Vector Machine** (SVC) is a supervised machine learning algorithm that uses

binary classification tasks to solve complex classification, regression and outlier detection by finding the optimal hyperplane which maximally separates the data points of different classes in a high-dimensional space.

Key concepts of SVC include:

1. Hyperplane: This is a decision boundary that separates data points of two or more different classes. In an nnn-dimensional space, a hyperplane is an (n−1)(n-1)(n−1)-dimensional subspace.
2. Support Vectors: These are the data points closest 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 maximize this margin which is the distance between the hyperplane and the nearest data points from either class. Generally this leads to better summary on unseen data.

* **Random Forest Classification** is an ensemble supervised learning method used for both

classification and regression tasks. Random forest operates by constructing multiple decision trees during training with each node pointing to the next structurally and outputting the mode of the classes (for classification) or the mean prediction (for regression) of the individual trees.

Key characteristics of Random Forest include:

1. Ensemble of Trees: This starts off with a process called bootstrap sampling in which each decision tree is trained on a random subset of the data with replacement. Each tree in the forest is grown to the maximum extent without pruning leading to collection of decision trees.
2. Random Feature Selection: In order to introduce further randomness and help reduce correlation among trees, at each split in the decision tree, a random subset of features 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 replacement from the original dataset. This technique reduces variance and helps prevent overfitting.
4. Voting/Averaging: For classification tasks, the final prediction is made by majority vote of each individual tree. For regression tasks, the prediction is the average of the outputs of the individual trees.

* **Long Short-Term Memory (LSTM)** is a type of recurrent neural network (RNN)

architecture designed to model sequences and time-series data. It addresses the vanishing gradient problem commonly found in standard RNNs. LSTMs are particularly well-suited for tasks such as language modeling, machine translation and time-series prediction, enabling the network to capture long-range dependencies more effectively.

Key components of LSTM include:

1. Memory Cell: Maintains information over time, allowing the network to remember important information from earlier in the sequence.
2. Gates: Control the flow of information into and out of the memory cell. There are three types of gates:
   * Forget Gate: Decides what information to discard from the cell state.
   * Input Gate: Determines which new information to add to the cell state.
   * Output Gate: Controls what information to output based on the cell state.

**3.4.2 Rationale Behind Model Selection**

These all four mentioned machine learning algorithms were considered because they are supervised learning algorithms meaning that they require labeled data to learn during the model training, classification capability and efficiency in the area of natural language processing, making them suitable baseline models for sentiment analysis. Logistic regression, Support vector machine and Random forest are particularly effective for binary classification tasks. Other models are selected for their ability to capture complex patterns and provide comparative insights into the performance of all four algorithms.

**3.5 Experimental Design**

**3.5.1 Experiment Setup**

To conduct the sentiment analysis on multilingual comments using support vector machines, random forest, logistic regression and long short-term memory, a comprehensive experimental setup was established. This setup encompassed the data collection process, preprocessing steps and the implementation of the machine learning models. The primary tools and libraries used included Python, pandas, NumPy, scikit-learn, TensorFlow library for deep learning application and natural language processing (NLP) libraries such as NLTK.

**3.5.2 Training and Testing Datasets**

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

**3.5.3 Partitioning Strategies**

The dataset is partitioned into:

* 80% for training
* 20% for testing

**3.5.4 Cross-Validation Techniques**

K-fold cross-validation is employed to ensure robustness and prevent overfitting. This involves dividing the training set into k subsets and training the model k times, each time using a different subset as the validation set and the remaining k-1 subsets for training.

**3.5.5 Performance Metrics**

Performance metrics include accuracy, precision, recall and F1-score, providing a comprehensive evaluation of the model’s effectiveness in classifying sentiments.

**3.6 Implementation Details**

**3.6.1 Technical Implementation**

* Programming Languages: Python is used for its extensive libraries and frameworks for machine learning and natural language processing.
* Libraries and Frameworks: Scikit-learn for machine learning, NLTK and SpaCy for text preprocessing and TensorFlow/Keras for deep learning models.
* Hardware/Software Configurations: Experiments are conducted on a system with a high-performance GPU to accelerate model training and evaluation.

**3.5.2 Software Development Environment**

Google Colab Notebooks are used for iterative development and documentation. The development environment includes virtual environments to manage dependencies and ensure consistency across different setups.

**3.7 Ethical Considerations**

**3.7.1 Data Collection and Usage**

Ethical considerations related to data collection and usage include:

* Ensuring user privacy by anonymizing data and obtaining necessary permissions where required.
* Complying with relevant regulations and institutional policies on data usage.

**3.7.2 Addressing Bias and Fairness**

Efforts are made to address potential biases in the data and models:

* Ensuring a balanced representation of different languages and sentiments.
* Regularly evaluating and mitigating any biases detected in the model’s predictions.

**3.7.3 Societal Impact**

The potential societal impacts of this research are considered, including:

* The influence of sentiment analysis on public opinion and discourse.
* Ensuring the responsible use of sentiment analysis results in decision-making processes.

**3.8 Limitations and Assumptions**

**3.8.1 Methodological Limitations**

Acknowledged limitations include:

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

**3.8.2 Data and Model Constraints**

Constraints related to data and models include:

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

**3.8.3 Implications on Research Findings**

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

* Reduced accuracy and reliability in detecting nuanced sentiments.
* 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.

**CHAPTER FOUR: RESULTS**

The results of the sentiment analysis were measured using various metrics such model accuracy, precision, F1 score and recall.

**4.1 For Individual Languages**

**4.1.1 Comparing Analysis for Nigerian-Pidgin Language Dataset**

**Table 4.0** Performance Metrics on Pidgin Language Dataset

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Accuracy | Precision | F1 Score | Recall |
| Logistic Regression | 55% | 55% | 49% | 55% |
| SVM | 55% | 55% | 49% | 55% |
| Random Forest | 55% | 55% | 49% | 55% |
| LSTM | 58% | 58% | 57% | 58% |

Across the four metrics which are accuracy, precision, F1 score and recall, the LSTM model outperforms Logistic Regression, SVM and Random Forest. The improvement in F1 score from 49% for all three algorithms to 57% for LSTM indicates that it is better at handling the variations of the Nigerian-Pidgin language comments in the dataset, likely due to its ability to capture temporal dependencies and long-term patterns in sequential data. In contrast, the identical performance of Logistic Regression, SVM and Random Forest suggests that these traditional models may not be as well-suited for this specific dataset, possibly due to its built-in complexity or structure that these models cannot adequately capture.

Based on this comparative analysis, the LSTM model appears to be the most effective choice for this dataset, providing a more balanced and accurate representation of the Nigerian-Pidgin language.

**4.1.2 Comparing Analysis for Hausa Language Dataset**

**Table 4.1** Performance Metrics on Hausa Language Dataset

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Accuracy | Precision | F1 Score | Recall |
| Logistic Regression | 74% | 73% | 73% | 74% |
| SVM | 74% | 73% | 73% | 74% |
| Random Forest | 74% | 73% | 73% | 74% |
| LSTM | 74% | 73% | 74% | 74% |

The performance metrics for all four algorithms, Logistic Regression, SVM, Random Forest and LSTM were remarkably similar for the Hausa language dataset. Each model achieves an accuracy of 74%, precision of 73% and recall of 74%, with the F1 score being 73% for the traditional models and slightly higher at 74% for the LSTM. The slight improvement in the F1 score for the LSTM model simply suggests a marginally better balance between precision and recall, potentially indicating its superior ability to capture sequential patterns in the data.

However, given the minimal differences in performance metrics, it can be concluded that all four algorithms are comparably effective for this particular dataset.

**4.1.3 Comparing Analysis for Igbo Language Dataset**

**Table 4.2** Performance Metrics on Igbo Language Dataset

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Accuracy | Precision | F1 Score | Recall |
| Logistic Regression | 62% | 53% | 50% | 62% |
| SVM | 62% | 53% | 50% | 62% |
| Random Forest | 61% | 53% | 50% | 62% |
| LSTM | 64% | 63% | 63% | 64% |

In the Igbo language dataset, Logistic Regression, SVM and Random Forest performed worse than the LSTM model across all metrics (accuracy, precision, F1 score and recall). The improvement in F1 score from 50% to 63% indicates that LSTM is more effective at handling the intricacies of the Igbo language dataset, likely due to its ability to capture temporal dependencies and long-term patterns in sequential data.

While Logistic Regression, SVM and Random Forest demonstrate similar performance, with Random Forest showing a slight drop in accuracy, they were all outperformed by LSTM. The significant improvement in LSTM's precision and F1 score highlights its superior capability in balancing precision and recall, making it the most effective choice for this dataset.

**4.1.4 Comparing Analysis for Yoruba Language Dataset**

**Table 4.3** Performance Metrics on Yoruba Language Dataset

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Accuracy | Precision | F1 Score | Recall |
| Logistic Regression | 89% | 87% | 85% | 89% |
| SVM | 89% | 87% | 85% | 89% |
| Random Forest | 89% | 87% | 85% | 89% |
| LSTM | 90% | 90% | 90% | 90% |

Once again the LSTM model outperforms Logistic Regression, SVM and Random Forest across all metrics which were accuracy, precision, F1 score and recall, with improvement in F1 score from 85% to 90% indicates that LSTM is more effective at handling the Yoruba language dataset, likely due to its ability to capture temporal dependencies and long-term patterns in sequential data.

While Logistic Regression, SVM and Random Forest demonstrate strong and consistent performance, they are all slightly outperformed by LSTM. The improvement in LSTM's precision and F1 score highlights its superior capability in balancing precision and recall, making it the most effective choice for this dataset.

Ultimately, for theYoruba language dataset, LSTM model is recommended for providing the best overall model performance and offering a more balanced and accurate representation compared to the other models.

**4.2 Performance Metrics for the All Four (4) Languages Dataset Combined**

**Table 4.4** Performance metrics for combination of all multilingual language datasets

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 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% |

In the combined multilingual language dataset, 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 76%, with a strong F1 score 72% and recall 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 F1 score, indicating some imbalance between precision and recall. LSTM, on the other hand, demonstrates consistent performance with equal metrics across accuracy, precision, F1 score and recall (72%), but does not outperform SVM and Random Forest.

**4.3 Comparative Analysis**

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

#### 4.3.1 Individual Language Performance

1. Nigerian-Pidgin Language Dataset: Accuracy and precision for all models except LSTM achieved 55% accuracy and precision. Recall and F1 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% F1 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% F1 score and 89% recall. However LSTM performed best with 90% across all metrics, making it the most ideal model in this research work.

#### 4.3.2 Combined Dataset Performance

Logistic Regression: 72% accuracy, 72% precision, 71% F1 score, 72% recall.

SVM and Random Forest: Superior performance with 74% accuracy, 76% precision, 72% F1 score, 74% recall.

LSTM: Consistent but not leading, with 72% across all metrics.

### 4.4 Discussion

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

**CHAPTER FIVE: CONCLUSION**

**5.1 Summary of Findings**

The primary objective of this thesis was to compare machine learning algorithms for sentiment analysis Of multilingual Nigerian social media comments. The study successfully demonstrated the viability of the machine learning algorithms such as SVM, Random forest, logistic regression and LSTM in accomplishing this task of analyzing the sentiments behind Nigerian social media comments and achieving a significant accuracy rate. The comparative analysis with other models unveils the advantages each model presents and their unique strengths and weaknesses. Key findings include:

* Logistic regression achieved an accuracy of 72% on the general combined dataset.
* SVM and Random Forest both achieved an accuracy of 73.81% on the combined language dataset.
* LSTM outperformed other algorithms, with the accuracy of its model scoring highest in all individual languages. However, an accuracy of 72.% was achieved on the combined language dataset after 10 epochs.

These results indicate that machine learning algorithms which can be used for natural language processing are reliable to train models for sentiment analysis in a multilingual context, though there is room for improvement with more complex models like LSTM.

**5.2 Summary of Work:**

In summary, the project research has probed into the intersection between sentiment analysis in a linguistically diverse context such as the Nigerian social media space and the comparison of computational techniques for understanding natural language processing. Key highlights include:

* Demonstrating the effectiveness of machine learning algorithms for sentiment analysis in a multilingual setting.
* Exploring a comparative analysis between other models from algorithms like logistic regression, SVM, Random Forest and LSTM.
* Highlighting the challenges of sentiment analysis in a multilingual environment, including code-switching and dialectal variations.
* Offering a dataset of multilingual social media comments from Nigeria, which can serve as a valuable resource for future research.

**5.3 Recommendations for Future Work:**

In multilingual context, future research could focus on the following areas to enhance sentiment analysis:

* Advanced Models: Explore more advanced deep learning models such as transformers (e.g., BERT, GPT) may capture better linguistic nuances than traditional models.
* Larger Datasets: Curating a larger and more diverse datasets with more features will certainly improve model training and testing, evaluation and performance.
* Real-time Analysis**:** Implementation of real-time sentiment analysis systems with synchronous frameworks to provide instant sentiment insights from social media data.
* Multilingual NLP Techniques: Development, integration and deployment of advanced natural language processing techniques, tailored for multilingual data which have been built to addressing challenges like code-switching and dialectal variations.
* Cross-Cultural Studies: Carry out cross-cultural studies to understand language nuances, figures of speech and sentiment expressed in various areas across different cultural and linguistic groups within Nigeria.

**5.4 Conclusion:**

To conclude, this research work has investigated the four previously mentioned machine learning algorithms for the field of sentiment analysis in multilingual context on Nigerian social media comments. In spite of the challenges faced by a limited availability of data, linguistic diversity, the models achieved promising results, highlighting its 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, hence this research work supports valuable insights to classification. 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.

**REFERENCES**

1. Bamman, D., Eisenstein, J., & Schnoebelen, T. (2014). "Gender identity and lexical variation in social media." Journal of Sociolinguistics, 18(2), pp.135-160.
2. 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.
3. Huang, M., Zhu, X., & Gao, J. (2022). "Multilingual sentiment analysis with deep learning: A comprehensive survey." ACM Transactions on Information Systems (TOIS), 40(2), pp.1-30.
4. Liu, B., 2012. Sentiment Analysis and Opinion Mining. Synthesis Lectures on Human Language Technologies, 5(1), pp.1-167.
5. Mohammad, S.M., Kiritchenko, S., & Zhu, X., 2013. NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets. Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), pp.321-327.
6. NaijaSenti et al. (2022). "A Comprehensive Dataset for Sentiment Analysis of Multilingual Nigerian Social Media Comments."
7. Olaleye, O., Adewole, A., & Odumuyiwa, V. (2018). "Sentiment analysis of social media content using machine learning techniques." Journal of Computational Science, 28, pp.171-181.
8. Tang, D., Qin, B., & Liu, T. (2016). "Document modeling with gated recurrent neural network for sentiment classification." Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp.1422-1432.
9. Wang, S., Manning, C. D. (2012). "Baselines and Bigrams: Simple, Good Sentiment and Topic Classification."Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pp.90-94.