University of Ibadan

Journal of Science and Logics in ICT Research

Sentiment Analysis of Low-Resource Yorùbá Tweets Using Fine-Tuned Bert Models

[1]Odumuyiwa, V.T., and [2]Shoyemi, O.O., and Fagoroye A.E.

[1,2]Department of Computer Sciences, University of Lagos, Akoka, Nigeria
vodumuyiwa@unilag.edu.ng

Abstract

Sentiment analysis in low-resource languages poses a notable challenge because of the scarcity of labelled data and language-specific models. This study addresses this challenge of Yorùbá sentiment analysis using fine-tuned variants of Bidirectional Encoder Representations from Transformers (BERT) model. Yorùbá, being a low-resource language, lacks effective sentiment analysis tools for detecting the sentiment polarity of content written in the language. Solving this problem is important for understanding beliefs of the public, cultural sentiment, and enhancing communication analytics in Yorùbá-speaking communities. The paper employs transfer learning techniques to adjust pretrained models to the unique linguistic properties of Yorùbá. The chosen models include Bert Base (Uncased), African Bidirectional Encoder Representations from Transformers (AfriBERTa), Multilingual version of BERT (mBERT), and multilingual version of RoBERTa (XLM-RoBERTa). AfriBERTa model demonstrates a superior performance in capturing sentiment nuances specific to Yorùbá language tweets after comparative analysis was done on the performance of the four models on two different datasets.

Keywords: *Sentiment Analysis, Low-Resource Yorùbá Language, BERT, Natural Language Processing*.

1. Introduction

1.1 Background of the Study

Natural language processing (NLP) now heavily emphasises sentiment analysis (SA) since it makes it possible to glean insightful information from textual input. Sentiment analysis presents particular challenges for low-resource languages like Yorùbá as a result of a need for linguistic resources as well as data availability [1]. There is a noticeable vacuum in the literature about the efficient analysis of sentiments in low-resource languages like Yorùbá, despite the advances in sentiment analysis [2]. Sentiment analysis techniques that are currently in use frequently have trouble capturing the subtleties and cultural quirks unique to these languages. One major obstacle to accurately interpreting user attitudes in this linguistic setting is the dearth of studies and models specifically designed for sentiment analysis in Yorùbá. In order to close the gap, this work tests improved BERT model variations for sentiment analysis of Yorùbá tweets and reviews.

The rest of this paper is ordered as follows: a summary of pertinent literature on sentiment analysis, low-resource languages, and several BERT models is presented in Section 2; Section 3 discusses the methodology, covering data collection, model architecture, and experimental setup; Section 4 outlines and discusses the results gotten from the experiments; Section 5 offers a thorough conclusion, including limitations, future research directions, and the study's overall impact.

2. Literature Review

2.1 Sentiment Analysis

In NLP, sentiment analysis is the methodical identification, extraction, measurement, and

Odumuyiwa, V.T., Shoyemi, O.O., and Fagoroye A.E. (2024). Sentiment Analysis of Low-Resource Yorùbá Tweets Using Fine-Tuned Bert Models, *University of Ibadan Journal of Science and Logics in ICT Research (UIJSLICTR)*, Vol. 12 No. 1, pp. 110 – 122.

analysis of subjective data from textual sources [3]. The textual inputs for sentiment analysis encompass different formats including social media posts, online reviews and comments, survey responses, news articles, or any piece of writing that contains subjective opinions and attitudes [4].

The term sentiment implies viewpoints, emotions, speculations, judgments, or feelings expressed about specific entities, events and their attributes. These subjective expressions can indicate overall positive or negative sentiment towards an entity or event or topic of discussion. Sentiment polarities can also be classified into emotional states like "angry," "sad," or "happy" [5]. Beyond polarity and basic emotions, some sentiment analysis approaches extract detailed opinions and preferences of the author or speaker.

A major focus area within sentiment analysis is classifying the overall tone, tenor, feeling or temperament of a document into classes like "positive", "negative", or "neutral". This automated document-level sentiment categorization enables efficient analysis of sentiment trends for brands, products, issues or activities across thousands of textual records. Common use cases include public perception tracking, customer experience insights, campaign audience targeting, stock price prediction, and more [6].

Social media has become one of the most useful and practical application domains for sentiment analysis models and techniques. Platforms like Twitter, Facebook, and Reddit contain a vast, ever-growing stream of text data – hundreds of millions of microblogs, posts, comments, and online conversations daily – which enables near real-time monitoring of sentiment trends at immense scale [7].

Some application areas of sentiment analysis on social data include:

• **Brand monitoring** – companies track brand and product name mentions on social media to monitor consumer feedback and satisfaction, discover pain points, and understand how their marketing campaigns and product updates are being perceived. Automatically detecting sentiment towards brands enables rapid response.

• **Political polling prediction** – by following the volume and sentiment leaning of social chatter surrounding political races, prediction models have been built to show the outcomes of upcoming elections with competitive or better accuracy than traditional telephone polls [8].

• **Stock price movement modeling** – greater social media positivity and bullishness around public stocks have been statistically shown to positively lead financial trading volumes and stock returns. Sentiment analysis drives multiple profitable trading strategies [9].

The textual nature and emotional expression prevalent on social platforms make this rich data source ideal for real-world deployment of sentiment analysis techniques at scale to gain pulse-reading insights across massive populations and topical domains [10].

2.2 Techniques in Sentiment Analysis

2.2.1 Rule-based Approaches

Rule-based approaches were some of the earliest methods applied for SA. They rely on a set of hand-crafted rules, semantic templates, lexical resources, and affect lexicons to detect emotion and sentiment in text [5]. However, hand-crafting accurate rule sets is challenging across domains.

A typical rule-based system works by having human experts define vocabularies and multiple types of rules encoding sentiments. These rules are coded into the system modules and then applied in a cascading fashion to piece together clause-level sentiment parsing, before aggregating an overall document sentiment based on frequencies.

Custom rules can integrate domain specifics for improved context-aware analysis. Rule-based methods have also been used in Aspect-based SA [10][11][12] towards extracting "aspect terms" referring to attributes of entities that opinions describe. Rule-based methods leverage dependencies between terms to recognize aspect expressions.

2.2.2 Machine Learning Approaches

Machine learning approaches for SA utilize statistical methods to learn sentiment patterns from labeled data. These methods have become popular to a large extent as a result of their ability to handle large amounts of data and their ability to adapt to new data without requiring extensive hand-crafting of rules.

Some machine learning algorithms for SA include Naïve Bayes, Support Vector Machines (SVM), Decision Trees etc.

Naive Bayes is a probabilistic classifier developed on Bayes' theorem [13]. It assumes that the features are not dependent on one another, which may not always be the case for sentiment analysis. Naive Bayes classification is a common technique for text categorization tasks, where documents must be assigned to predefined categories [14]. The approach implements Bayes' theorem, attributing categories based on the probabilistic occurrence of features [15]. Information retrieval systems widely incorporate Naive Bayes classifiers due to their relative simplicity and rapid training [16]. As a generative model, Naive Bayes learns distributions of features within each class rather than discriminating between classes directly. However, a limitation of Naive Bayes classifiers is reduced effectiveness for highly skewed, unbalanced class distributions.

Support Vector Machines are discriminative classifies that find a hyperplane which maximizes the margin between the classes. SVMs are well-suited for sentiment analysis tasks as they can handle high-dimensional data and are robust to noise [17]. The original SVM algorithm, developed by Vapnik and Chervonenkis [18], focused on binary classification problems. However, over the years, SVMs have been extensively adapted for multi-class classification. A variety of methods now exist to enable SVMs to handle problems with more than two target classes, representing an active area of machine learning research [19].

Decision Trees are tree-like structures that represent a series of decisions that lead to a final classification. Decision trees are easy to interpret and can handle both categorical and numerical data [20]. However, they can be prone to overfitting if not properly tuned. Decision tree models are now widely used and useful classification methods. Magerman [21] introduce it as a classification model but was later expanded as an inductive approach by Quinlan [22]

2.2.3 Deep Learning Approaches

In recent years, deep learning methods have made great progress in achieving state-of-the-art performance, revolutionising SA. Compared to typical machine learning techniques, deep learning algorithms, such Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs), are able to build more robust representations of sentiment and capture complex patterns in language.

Because RNNs can catch lengthy sequences in text data, they are especially well-suited for tasks involving sentiment analysis. RNNs analyse text in a sequential fashion, enabling them to take into consideration the context of each word whenever sentiment predictions are made [10]. Nevertheless, because gradients used to update network weights have a high propensity to either vanish or explode exponentially over lengthy sequences, typical RNNs have difficulty modelling long-term relationships [23].

More complex RNN variations, including gated recurrent units (GRUs) and long short-term memory (LSTM) units, have been created to solve this problem. While GRUs for machine translation jobs were just recently introduced [23], with a longer history, long-range dependencies are now frequently captured by LSTMs in sequence modelling and natural language processing issues. Gating mechanisms in the recurrent cells of both kinds are used to control gradient flow and improve contextual information retention.

CNNs are an additional potent deep learning architecture utilized in sentiment analysis. CNNs may recognize sentiment-bearing expressions by extracting local features from text input, such as word patterns and n-grams [25]. CNNs work especially well at interpreting short text messages, such tweets and posts on social media.

2.3 Sentiment Analysis in Low-Resource Languages

Yorùbá have its proper place in the Yoruboid subbranch of the Volta-Niger branch of the Niger-Congo language family [43]. It is spoken in the southwestern regions of Nigeria expanding into some regions of countries like Togo and Benin. The Yorùbá alphabet is founded on the Latin script consisting of eighteen consonants, seven oral vowels, five nasal vowels and syllabic nasal consonants with additional characters like e, o, s, gb [57]. Approximately 47 million people speaks this language [58], mostly in Nigeria, and Republic of Benin.

A vast majority of SA research has dwelled on high-resource languages with abundant datasets and linguistic resources, such as English, French, Spanish, Chinese, and certain European languages [26]. Significantly less attention has been devoted to low-resource languages, despite over 7,000 living languages globally.

SA for low-resource languages has recently attracted popularity [27][28][29] due to the increase in the large number of comments from tweets in such languages. Multiple studies have investigated using X (formerly Twitter) for SA—either by automatically creating a corpus or manually annotating one. Remarkable studies that automated the building of X corpora include Go et al. [30], Pak and Paroubek [31], and Wicaksono *et al.* [32]. Recently, Kwaik *et al.* [33] developed an Arabic Twitter SA corpus using distant supervision and self-training. In contrast, studies by Refaee and Rieser [34], Brum and Nunes [35], Mozetic *et al.* [36], Nakov *et al.* [37], and Moudjari *et al.* [38] utilized native speakers and expert annotators to add annotations to the corpus manually.

Despite some progress made in SA for indigenous Nigerian languages have attracted less sufficient attention. This is mostly because of the need for a freely accessible dataset in these languages. However, few studies exist on Nigerian code-mixed English [39], [40], [41], [42]. In this study, we are interested in SA of Yorùbá tweets and reviews.

Research on Yorùbá language SA is still in its early stage. Some work has been done on Yorùbá language sentiment analysis using machine learning [43], however, there's still a lot of space for development. The absence of labelled data is one of the primary obstacles in Yorùbá language sentiment analysis. For this reason, training machine learning models is challenging. The Yorùbá language's intricate morphology presents another difficulty. Given that Yorùbá is a morphologically complex language, words are created by joining smaller pieces known as morphemes. This complicates the task of teaching machine learning models word meanings.

Notwithstanding these difficulties, it is crucial to remember that Yorùbá language sentiment analysis ought to be regarded as a crucial area of study, given that more than 46 million people speaks the language globally. Sentiment analysis of Yorùbá language has several applications. For instance, it can monitor social media sentiment, examine client comments, and spot trends. Additionally, it can aid in the promotion of the language and culture, increasing linguistic accessibility for a worldwide audience.

2.4 Transfer Learning in Natural Language Processing

In NLP, transfer learning has emerged as a potent paradigm. Transfer learning is used by pretrained language models, such as BERT, to acquire broad language representations by pretraining on a huge unlabeled corpora and then fine-tuning on smaller downstream datasets [44]. This section examines BERT's architecture, tenets, and variations, highlighting their importance in sentiment analysis—particularly with regard to the Yorùbá language.

2.4.1 BERT Architecture

Devlin *et al.* [44] introduced BERT. It made use of a transformer design, which made two-way context understanding possible. While previous models only processed text in one direction, BERT has the advantage of training using both left and right context. BERT is very useful for a variety of NLP applications because of its ability to capture complex dependencies and relationships inside language thanks to its bidirectionality attribute.

2.4.2 Pretraining and Masked Language Model (MLM)

BERT is pretrained by unsupervised learning on big corpora. A masked language model (MLM) objective is used during training, in which the model's job is to predict words that are masked in respect to the context in which they appear. This pretraining phase equips BERT with a rich understanding of language nuances.

2.4.3 Existing BERT Models for Languages

2.4.3.1 AfriBERTa

AfriBERTa is a specialized BERT model designed for African languages. The model was trained on 11 languages, including Yorùbá, Amharic, Hausa, and Swahili. It builds on the original BERT architecture and is pretrained on diverse African languages to capture the linguistic nuances and specificities of the continent [48].

2.4.3.2 mBERT (Multilingual BERT)

mBERT, or Multilingual BERT, is a pretrained language model that encompasses a wide range of languages, including Yorùbá. It was trained on a multilingual corpus, ensuring that is captures cross-lingual relationships and representations. While mBERT is not specifically tailored for Yorùbá, its multilingual nature enables it to provide context-aware embeddings for sentiment analysis in various languages, including low-resource languages like Yorùbá [44].

2.4.3.3 XLM-ROBERTA

XLM-RoBERTa, a multilingual version of RoBERTa was pre-trained on 2.5TB of filtered CommonCrawl data containing 100 languages. RoBERTa is a transformer model that was pretrained on the raw texts (with no human labels) with an automatic process to generate inputs and labels from those texts.

2.5 Related Works

Some notable works on machine learning and deep learning algorithms for SA are presented in this section.

Saha *et. al.* [49] presented a study on sarcasm detection in Twitter, highlighting its significance in tweet analysis for understanding user sentiments towards products. The methodology involved data preprocessing using TextBlob for tasks like tokenization, part-of-speech tagging, and parsing, along with stop words removal. The study utilized Weka to evaluate tweet accuracy using Naïve Bayes and SVM classifiers, achieving accuracies of 65.2% and 60.1%, respectively.

The study by Qi *et. al.* [50] focused on extracting COVID-19 related data from Twitter users in major cities of England. It compared various machine learning models such as multinomial Naïve Bayes, Random Forest, and Support Vector Machine, with lexicon-based approaches such as Vader and Textblob. Two feature extraction methods, Word2Vec embedding and TF-IDF, were employed for analysis. The SVM model with TF-IDF exhibits superior accuracy (71%) compared to the other models.

Kumar *et. al.* [51] focused on the influence of gender and age on customer reviews. The study utilized various machine learning algorithms including Maximum Entropy (ME), SVM, and LSTM models. The Bag of Words (BOW) feature extraction technique was employed in the Naive Bayes, Maximum Entropy, and Support Vector Machine algorithms, whereas the LSTM model utilized word2vec for feature extraction. The highest accuracy (78%) in predicting female customer reviews was achieved within the age group over 50.

From a study by Memiş *et. al.* [52], financial tweets were categorized as "positive," "negative," or "neutral" and datasets were created for binary and multi-class classification. Different models including Neural Network, CNN, LSTM, GRU, and GRU-CNN were employed for analysis. Pre-trained word embeddings combined with the CNN model yielded the best results for both binary (83.02%) and multi-class (72.73%) datasets. However, when using word embeddings, the Neural Network model performed best for the multi-class dataset (63.85%), while the GRU-CNN model achieved the highest accuracy for the binary dataset (80.56%).

Tan *et. al.* [53] introduced a new approach to SA by combining two powerful models: RoBERTa and GRU. Evaluation on three popular sentiment

analysis datasets (IMDb, Sentiment140, and Twitter US Airline Sentiment) demonstrated impressive accuracies: 94.63% on IMDb, 89.59% on Sentiment140, and 91.52% on Twitter US Airline Sentiment. These results highlight the efficiency of the hybrid model in SA tasks.

Odumuyiwa & Adedayo [24] proposed using Siamese gated recurrent neural networks, specifically LSTM and GRU architectures, to measure semantic similarity between texts. The goal is to capture the contextual meaning of words and sentences to determine if a sentence pair has similar or contrasting meaning. Ainapure et. al. [54] explored how social media platforms, particularly Twitter, serve as a medium for individuals to express their sentiments regarding the COVID-19 pandemic and vaccination efforts in India. The study utilized both deep learning and lexicon-based methods to analyze the sentiments conveyed in tweets.

Specifically, it employed VADER and NRCLex tools for lexicon-based sentiment analysis, while they utilized Bi-LSTM and GRU recurrent neural networks for deep learning sentiment classification. A recurrent neural network was trained employing Bi-LSTM and GRU methodologies, achieving accuracies of 92.70% and 91.24% on the COVID-19 dataset, respectively. For the classification of vaccination-related tweets, accuracy rates of 92.48% and 93.03% were attained using Bi-LSTM and GRU methods, respectively.

Alqarni & Rahman [55] focused on SA of Arabic tweets during the pandemic in Saudi Arabia collected data from Riyadh, Dammam, and Jeddah. Tweets were categorized as positive, negative, or neutral. CNN and Bi-directional Long Short-Term Memory (BiLSTM) algorithms were used for sentiment classification, achieving 92.80% accuracy for CNN and 91.99% for BiLSTM. The study found that there were a lot of negative sentiments during COVID-19 compared to pre-pandemic levels. Shode et. al. [56] performed experiments on YOSM dataset by using pre-trained language models -Afriberta and mBert. The f1-score gotten was 87.2% & 83.2% respectively.

3. Research Methodology

3.1 Data Set

In this study, two datasets -NaijaSenti and YOSM dataset were utilized. The NaijaSenti dataset consists 15127 Yorùbá tweets [43]. The dataset contains 6344 positive, 5487 neutral and 3296 negative tweets. These tweets were Yoruba-based comments on twitter. While YOSM dataset contained a balanced set of positive and negative movie reviews. The total reviews were 1,500 [56]. These reviews were gotten from IMDB, Rotten Tomatoes and, Letterboxd.

The data were preprocessed through the following steps:

i. changed all capitalization to lowercase,
ii. cleared punctuation and unnecessary marks
iii. cleaned extra spaces
iv. removed the html strips
v. removed the square brackets
vi. removed the noisy text
vii. cleaned out @user mentions, hashtags and URLs which contribute no useful to the affect knowledge.

3.2 Model Architecture

This research made use of four BERT pretrained models - Bert Base (Uncased), AfriBERTs, mBERT and XLM-roBERTa. The models were trained and fine-tuned with the NaijaSenti dataset and the YOSM dataset.

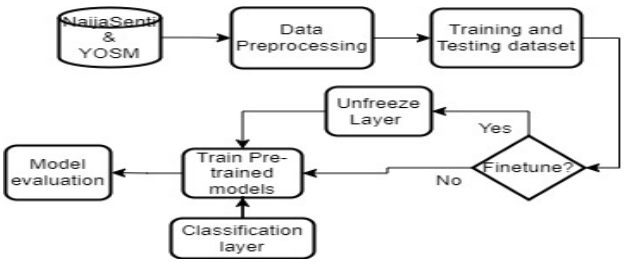

Figure 3.0 Diagram representation of the Methodology

The above architecture for Yoruba sentiment analysis makes use of pre-trained models and follows a structured workflow. The workflow consists of multiple stages, each crucial for the successful application of sentiment analysis on the Yoruba language datasets, NaijaSenti and YOSM.

3.2.1 Data Sources
The main sources of Yoruba text data are NaijaSenti and YOSM. While YOSM is a compilation of Yoruba language movie reviews, NaijaSenti is a multilingual dataset that contains text from twitter.

3.2.2 Data Preprocessing
To clean and get ready for model training, the raw text data from NaijaSenti and YOSM was preprocessed. Handling of special characters was performed on the dataset. The preprocessing makes sure that the text data is formatted appropriately for the next stages.

3.2.3 Data Preparation
The preprocessed data is split into training and testing sets

3.2.4 Model Training
The Yoruba datasets were used to train the pre-trained models. Unfreezing layers of the pre-trained model can lead to better performance by adjusting the weights based on the datasets.

3.2.5 Model Improvement
A classification layer is added on top of the pre-trained model consisting of a bidirectional LSTM layer, hybrid pooling layers, and a dense output layer.

3.2.6 Model Evaluation
The testing dataset was used to evaluate the performance of the trained model. Important metrics including recall, accuracy, precision, and F1-score were computed to assess how well the model is doing at classifying the sentiment in Yoruba texts.

3.2.7 Fine-tuning BERT for Yorùbá Sentiment Analysis
Fine-tuning BERT for Yorùbá sentiment analysis involves modifying BERT models to adapt to the sentiment classification in the Yorùbá language. In order to fine-tune the models, the weights of the pretrained models and weights of the head layer were optimized**.**

3.2.8 Comparison of Existing Models (AfriBERTa, mBERT, XLM-roBERTa)
These models were specifically chosen due to their relevance to Yorùbá sentiment analysis, as discussed in Section 2.4.2. Each model represents a different approach to adapting BERT for low-resource language. The models are compared against one another to identify variations in accuracy, precision, recall, and F1 score. This comparative analysis sheds light on the relative strengths and weaknesses of AfriBERTa, mBERT, and XLM-roBERTa in the context of Yorùbá sentiment analysis.

3.3 Experimental Setup
For the four models used, the following hyperparameters were kept consistent across the models:
- Maximum sequence length: 128
- Batch size: 32
- Number of epochs: 10

A custom data generator class was created to make it easier to handle and input NLP data into BERT-based models. This class takes in the Yoruba texts (tweets and reviews) and their corresponding labels, along with parameters such as batch size and shuffle option. During each epoch, the data generator shuffles the dataset and, for each batch, retrieves the encoded input features (input IDs, attention masks, and token type IDs) along with the corresponding labels. This approach ensures efficient data handling and model input preparation.

To leverage distributed training capabilities, we utilized TensorFlow's MirroredStrategy, enabling the use of multiple GPUs for training. The training process involved loading the BERT models with pre-trained weights and freezing these layers to preserve their learned features. To adapt the models for the sentiment classification task, a head layer consisting of a bidirectional LSTM layer, hybrid pooling layers, and a dense output layer was added. The models were compiled with the Adam optimizer and categorical cross-entropy loss function.

Experiments were performed on a balanced version of NaijaSenti as well as the original dataset (NaijaSenti and YOSM). Undersampling, a method

for addressing class imbalance by keeping all instances of the minority class and decreasing the number of the majority class, was used to build the balanced dataset for NaijaSenti. Following undersampling, the balanced dataset included 3296 tweets that were split among three classes, guaranteeing fair representation and reducing the effects of class imbalance.

Google Colab, a cloud-based Jupyter notebook environment that offers high-performance computational capabilities, was used for the tests. The deep learning models were implemented using TensorFlow, while the preparation and analysis of the data was done with NumPy and Pandas.

3.4 Evaluation Metrics

The evaluation metrics that were used in our experimentation include accuracy, precision, recall, and F1 score.

Accuracy: It measures the overall correctness of predictions. It is the ratio of correctly predicted instances to the total number of instances.
Accuracy = (True Positives + True Negatives) / (Total Samples)

Precision: It measures the accuracy of the positive predictions made by the model. It is the ratio of true positive predictions to the total number of predicted positive instances.
Precision = True Positives / (True Positives + False Positives)

Recall: It quantifies the model's ability to correctly identify positive instances. It is the ratio of true positive predictions to the total number of actual positive instances.
Recall = True Positives / (True Positives + False Negatives)

F1 Score: It provides a balanced measure of a model's overall performance, particularly when there is an imbalance between positive and negative instances.
F1-Score = 2 * Precision * Recall / (Precision + Recall)

4. Results and Discussion

4.1 Comparative Evaluation of the Models

We evaluated the performance of different sentiment analysis approaches for Yorùbá tweets, we conducted a comprehensive comparison between BERT base model and the other three pretrained models using both NaijaSenti and YOSM dataset

.Table 1: Comparison of models without fine-tuning with NaijaSenti dataset

Model	**Precision**	**Recall**	**F1 Score**	**Accuracy**
Bert Base (Uncased)	0.67	0.66	0.66	0.66
AfriBERTa	**0.80**	**0.80**	**0.80**	**0.80**
XLM-RoBERTa	0.64	0.65	0.64	0.65
mBERT	0.70	0.70	0.70	0.70

Table 2: Comparison of models after fine-tuning with NaijaSenti dataset

Model	**Precision**	**Recall**	**F1 Score**	**Accuracy**
Bert Base (Uncased)	0.73	0.73	0.73	0.73
AfriBERTa	**0.80**	**0.80**	**0.80**	**0.80**
XLM-RoBERTa	0.70	0.71	0.70	0.71
mBERT	0.75	0.74	0.74	0.74

Table 3: Comparison of models without fine-tuning on the balanced dataset with NaijaSenti dataset

Model	**Precision**	**Recall**	**F1 Score**	**Accuracy**
Bert Base (Uncased)	0.67	0.66	0.66	0.66
AfriBERTa	**0.90**	**0.90**	**0.90**	**0.90**
XLM-RoBERTa	0.64	0.63	0.63	0.63
mBERT	0.69	0.69	0.69	0.69

Table 4: Comparison of models after fine-tuning on the balanced dataset with NaijaSenti dataset

Model	**Precision**	**Recall**	**F1 Score**	**Accuracy**
Bert Base (Uncased)	0.71	0.70	0.71	0.71
AfriBERTa	**0.91**	**0.91**	**0.91**	**0.91**
XLM-RoBERTa	0.71	0.70	0.70	0.70
mBERT	0.73	0.72	0.72	0.72

Table 5: Comparison of models without fine-tuning with YOSM dataset

Model	**Precision**	**Recall**	**F1 Score**	**Accuracy**
Bert Base (Uncased)	0.80	0.80	0.80	0.80
AfriBERTa	**0.86**	**0.86**	**0.86**	**0.86**
XLM-RoBERTa	0.79	0.78	0.77	0.77
mBERT	0.82	0.81	0.81	0.81

Table 6: Comparison of models after fine-tuning with YOSM dataset

Model	**Precision**	**Recall**	**F1 Score**	**Accuracy**
Bert Base (Uncased)	0.84	0.84	0.84	0.84
AfriBERTa	**0.90**	**0.90**	**0.90**	**0.90**
XLM-RoBERTa	0.84	0.84	0.84	0.84
mBERT	0.87	0.86	0.86	0.86

This comparative evaluation highlights the superior performance of AfriBERTa model for Nigerian Twitter sentiment analysis, evident across all key metrics - accuracy, precision, recall and F1-score. As evident from the results, AfriBERTa model performs better than other models. The best score for each model and overall best scores is in **bold**.

4.2 Discussion

From Table 1, AfriBERTa achieved an accuracy of 80% showing that it performed better than other

models because the model was pretrained on multilingual dataset including Yoruba language. However, the Bert Base (Uncased) without fine-tuning, gave a poor result because it was primarily trained on English text. Also, XLM-roBERTa performed poorly because though it was pretrained on a multilingual dataset, it was not pretrained on Nigeria Languages. In Table 2, fine-tuned AfriBERTa achieves the highest accuracy of 80% reflecting its ability to correctly classify positive, negative and neutral sentiment tweets in this low-resource Yoruba. The 80% F1-score balances both precision and recall strengths, underscoring its reliable detection of nuanced sentiment signals.

Comparing Table 2 to Table 1 shows that all the models when fine-tuned, performed better than the non-fine-tuned versions in terms of precision, recall and F1 capabilities. This demonstrates the significance of extensive adaptation on target task data rather than just lexical overlap.

From Table 3 & 4, the result looks similar to Table 1 & 2 in that AfriBERTa performed best both on finetuning and without finetuning with the balanced dataset.

From table 5 and 6, AfriBERTa maintained dominance in its perfomance. Performance on YOSM dataset, though way smaller compared to NaijaSenti, is better because it had only 2 classes while the latter had 3 classes.

Comparing Shode et. al. [56] results with ours, both Afriberta and mBert in this study performed better than their result. Reason being that we added a classification layer on top the pretrained models.

In summary, extensive fine-tuning enables pretrained AfriBERTa model to capture subtleties in informal, low-resource languages even outperforming counterparts with existing pretrained models. For Nigerian Yorùbá Twitter sentiment analysis, a robust F1-score along with high accuracy highlights fine-tuned AfriBERTa's reliability in distinguishing positive, negative and neutral tweets.

5. Conclusion

The findings of this study carry significant implications for sentiment analysis in low-resource languages, particularly Yorùbá language. The adaptability of BERT variants, as demonstrated through our fine-tuning approach, showcases the potential in enhancing sentiment analysis accuracy in linguistic contexts with limited resources. The comparative analysis contributes insights into the suitability of fine-tuning BERT models for Yorùbá sentiment analysis tasks.

While this study provides valuable contributions, it is crucial to acknowledge its limitations. The primary constraints include the availability of labeled data for Yorùbá sentiment analysis, potential biases in the pretrained BERT models, and the challenge of achieving optimal fine-tuning with limited resources. Our study mainly focused on the NaijaSenti Yorùbá annotated dataset and the YOSM dataset. These limitations may impact the generalizability of the findings, emphasizing the need for caution in applying the results to broader contexts.

Future works should explore expanding labeled datasets for Yorùbá sentiment analysis and the development of more tailored models. In addition, attempts should be made at investigating cross-domain sentiment analysis.

In conclusion, this research, in addition to advancing the understanding of sentiment analysis in low-resource Yorùbá language contexts, sets the stage for continued exploration and improvement in sentiment analysis methodologies for under-represented languages.

References

[1] Nasim, Z. and Ghani, S. (2020). *Sentiment analysis on urdu tweets using markov chains*. SN Computer Science, 1(5):1–13.

[2] Adelani, D. I., Abbott, J., Neubig, G., D'souza, D., Kreutzer, J., Lignos, C., Palen-Michel, C., Buzaaba, H., Rijhwani, S., Ruder, S., Mayhew, S., Azime, I. A., Muhammad, S. H., Emezue, C. C., Nakatumba-Nabende, J., Ogayo, P., Anuoluwapo, A., Gitau, C., Mbaye, D., Alabi, J., Yimam, S. M., Gwadabe, T. R., Ezeani, I., Niyongabo, R. A., Mukiibi, J., Otiende, V., Orife, I., David, D., Ngom, S., Adewumi, T.,

Rayson, P., Adeyemi, M., Muriuki, G., Anebi, E., Chukwuneke, C., Odu, N., Wairagala, E. P., Oyerinde, S., Siro, C., Bateesa, T. S., Oloyede, T., Wambui, Y., Akinode, V., Nabagereka, D., Katusiime, M., Awokoya, A., MBOUP, M., Gebreyohannes, D., Tilaye, H., Nwaike, K., Wolde, D., Faye, A., Sibanda, B., Ahia, O., Dossou, B. F. P., Ogueji, K., DIOP, T. I., Diallo, A., Akinfaderin, A., Marengereke, T., and Osei, S. (2021). *MasakhaNER: Named Entity Recognition for African Languages*. Transactions of the Association for Computational Linguistics, 9:1116–1131, 10.

[3] Pang, B. and Lee, L. (2007). *Opinion mining and sentiment analysis*. Found. Trends Inf. Retr., 2:1–135.

[4] Liu, B. (2022). *Sentiment analysis and opinion mining*. Springer Nature.

[5] Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A. Y., & Potts, C. (2013, October). *Recursive deep models for semantic compositionality over a sentiment treebank*. In Proceedings of the 2013 conference on empirical methods in natural language processing (pp. 1631-1642).

[6] Hutto, C., & Gilbert, E. (2014, May). Vader: *A parsimonious rule-based model for sentiment analysis of social media text*. In Proceedings of the international AAAI conference on web and social media (Vol. 8, No. 1, pp. 216-225).

[7] Soha, M. (2012). *Networked, collaborative, and activist news communities online: a case study of Reddit and Daily Kos*.

[8] Franch, F. (2013). *(Wisdom of the Crowds) 2: 2010 UK election prediction with social media*. Journal of Information Technology & Politics, 10(1), 57-71.

[9] Zheludev, I., Smith, R., & Aste, T. (2014). *When can social media lead financial markets?* Scientific reports, 4(1), 4213.

[10] Cui, Y., Chen, Z., Wei, S., Wang, S., Liu, T., & Hu, G. (2016). *Attention-over-attention neural networks for reading comprehension*. arXiv preprint arXiv:1607.04423.

[11] Hassan, S. H. and Mihalcea, R. (2011). *Semantic relatedness using salient semantic analysis*. In Twenty-Fifth AAAI Conference on Artificial Intelligence.

[12] Odumuyiwa, V., & Olatunbosun, A., (2020) *An enhanced rule based approach for target extraction in aspect based sentiment analysis*. Journal of Scientific Research and Development (2020) Vol. 19 (2) 270-280

[13] Pang, B., & Lee, L. (2008). *Opinion mining and sentiment analysis*. Foundations and trends in information retrieval, 2(1-2), 1-135.

[14] Kaufmann, S., (2010). *CUBA: Artificial conviviality and user-behaviour analysis in webfeeds* (Doctoral dissertation, University of Luxembourg, Luxembourg, Luxembourg).

[15] Pearson, E. S. (1925). *Bayes' Theorem, Examined in the Light of Experimental Sampling*, Biometrika, 17(3/4), 388, 1925.

[16] Qu, Z. Song, X., Zheng, S., Wang, X., Song, X., & Li, Z. (2018). *Improved Bayes Method Based on TF-IDF Feature and Grade Factor Feature for Chinese Information Classification*, In Proceedings of the 2018 IEEE International Conference on Big Data and Smart Computing (BigComp), Shanghai, China, 15–17 January 2018; pp. 677–680.

[17] Pang, B., & Lee, L. (2008). *Opinion mining and sentiment analysis*. Foundations and trends in information retrieval, 2(1-2), 1-135.

[18] Vapnik, V. & Chervonenkis, A. Y. (1964). A *class of algorithms for pattern recognition learning*, Avtomat. i Telemekh, 25(6), 937–945.

[19] Hsu, C. W., & Lin, C. J. (2002). *A comparison of methods for multiclass support vector machines*. IEEE transactions on Neural Networks, 13(2), 415-425.

[20] Buche, A., Chandak, D., & Zadgaonkar, A. (2013). *Opinion mining and analysis: a survey*. arXiv preprint arXiv:1307.3336.

[21] Magerman, D. M. (1995). *Statistical DecisionTree Models for Parsing* in 33rd Annual Meeting of the Association for Computational Linguistics, pp. 276–283.

[22] Quinlan, J. R. (1986). *Induction of decision trees*, Mach. Learn., 1(1), 81–106.

[23] Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H. and Bengio, Y. (2014). *Learning phrase representations using RNN encoder-decoder for statistical machine translation*. arXiv preprint arXiv:1406.1078.

[24] Odumuyiwa, V., & Adedayo, A. (2020). *Measuring the semantic similarity between texts using siamese gated recurrent architectures*

[25] Kim, Y. (2014). *Convolutional neural networks for sentence classification*. arXiv preprint arXiv:1408.5882.

[26] Joshi, M., Chen, D., Liu, Y., Weld, D. S., Zettlemoyer, L., & Levy, O. (2020). *Spanbert: Improving pre-training by representing and predicting spans*. Transactions of the association for computational linguistics, 8, 64-77.

[27] Yimam, S. M., Alemayehu, H. M., Ayele, A., and Biemann, C. (2020). *Exploring amharic sentiment analysis from social media texts: Building annotation tools and classification models*. In

Proceedings of the 28th International Conference on Computational Linguistics, pages 1048–1060.

[28] Xia, M., Zheng, G., Mukherjee, S., Shokouhi, M., Neubig, G., & Awadallah, A. H. (2021). *MetaXL: Meta representation transformation for low-resource cross-lingual learning*. arXiv preprint arXiv:2104.07908.

[29] Jovanoski, D., Pachovski, V., and Nakov, P. (2021). *Sentiment analysis in twitter for macedonian*. arXiv preprint arXiv:2109.13725.

[30] Go, A., Bhayani, R., and Huang, L. (2009). *Twitter sentiment classification using distant supervision*. CS224N project report, Stanford, 1(12):2009.

[31] Pak, A. and Paroubek, P. (2010). *Twitter as a corpus for sentiment analysis and opinion mining*. In LREC, volume 10, pages 1320–1326.

[32] Wicaksono, A. F., Vania, C., Distiawan, B., and Adriani, M. (2014). *Automatically building a corpus for sentiment analysis on indonesian tweets*. In Proceedings of the 28th Pacific Asia Conference on Language, Information and Computing, pages 185–194.

[33] Kwaik, K. A., Chatzikyriakidis, S., Dobnik, S., Saad, M., and Johansson, R. (2020). *An arabic tweets sentiment analysis dataset (atsad) using distant supervision and self training*. In Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection, pages 1–8.

[34] Refaee, E. and Rieser, V. (2014a). *An arabic twitter corpus for subjectivity and sentiment analysis*. In LREC, pages 2268–2273.

[35] Brum, H. B. and Nunes, M. d. G. V. (2017). *Building a sentiment corpus of tweets in brazilian portuguese*. arXiv preprint arXiv:1712.08917.

[36] Mozetic, I., Grecar, M., and Smailovic, J. (2016). *Multilingual twitter sentiment classification: The role of human annotators*. PloS one, 11(5): e0155036.

[37] Nakov, P., Kozareva, Z., Ritter, A., Rosenthal, S., Stoyanov, V., and Wilson, T. (2019). *Semeval-2013 task: Sentiment analysis in twitter*.

[38] Moudjari, L., Akli-Astouati, K., and Benamara, F. (2020). *An algerian corpus and an annotation platform for opinion and emotion analysis*. In Proceedings of the 12th Language Resources and Evaluation Conference, pages 1202–1210.

[39] Nwofe, E. S. (2017*). Pro-biafran activists and the call for a referendum: A sentiment analysis of 'biafraexit'on twitter after UK's vote to leave the European Union*. Journal of Ethnic and Cultural Studies, 4(1):65.

[40] Olaleye, S. A., Sanusi, I. T., and Salo, J. (2018). *Sentiment analysis of social commerce: a harbinger of online reputation management*. International Journal of Electronic Business, 14(2):85–102.

[41] Oyebode, O. and Orji, R. (2019). *Social media and sentiment analysis: The nigeria presidential election 2019*. In 2019 IEEE 10th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), pages 0140–0146. IEEE.

[42] Kolajo, T., Daramola, O., and Adebiyi, A. (2019). *Sentiment analysis on naija-tweets*. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, pages 338–343.

[43] Muhammad, S. H., Adelani, D. I., Ruder, S., Ahmad, I. S., Abdulmumin, I., Bello, B. S., ... & Brazdil, P. (2022). *Naijasenti: A nigerian twitter sentiment corpus for multilingual sentiment analysis*. arXiv preprint arXiv:2201.08277.

[44] Devlin, J., Chang, M. W., Lee, K. et al. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 4171–4186).

[45] Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). *DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter*. arXiv preprint arXiv:1910.01108.

[46] Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019). *Albert: A lite bert for self-supervised learning of language representations*. arXiv preprint arXiv:1909.11942.

[47] Liu, X., He, P., Chen, W., & Gao, J. (2019). *Improving multi-task deep neural networks via knowledge distillation for natural language understanding*. arXiv preprint arXiv:1904.09482.

[48] Ogueji, K., Zhu, Y., & Lin, J. (2021, November). *Small data? no problem! exploring the viability of pretrained multilingual language models for low-resourced languages*. In Proceedings of the 1st Workshop on Multilingual Representation Learning (pp. 116-126).

[49] Saha, S., Yadav, J., & Ranjan, P. (2017). *Proposed approach for sarcasm detection in twitter*. Indian Journal of Science and Technology.

[50] Qi, Y., & Shabrina, Z. (2023). *Sentiment analysis using Twitter data: a comparative application of lexicon-and machine-learning-based approach*. Social Network Analysis and Mining, 13(1), 31.

[51] Kumar, S., Gahalawat, M., Roy, P. P., Dogra, D. P., & Kim, B. G. (2020). *Exploring impact of age and gender on sentiment analysis using machine learning*. Electronics, 9(2), 374.

[52] Memiş, E., Akarkamçı, H., Yeniad, M., Rahebi, J., & Lopez-Guede, J. M. (2024). *Comparative Study for Sentiment Analysis of Financial Tweets with Deep Learning Methods*. Applied Sciences, 14(2), 588.

[53] Tan, K. L., Lee, C. P., & Lim, K. M. (2023). *Roberta-GRU: a hybrid deep learning model for enhanced sentiment analysis*. Applied Sciences, 13(6), 3915.

[54] Ainapure, B. S., Pise, R. N., Reddy, P., Appasani, B., Srinivasulu, A., Khan, M. S., & Bizon, N. (2023). *Sentiment analysis of COVID-19 tweets using deep learning and lexicon-based approaches*. Sustainability, 15(3), 2573.

[55] Alqarni, A., & Rahman, A. (2023). *Arabic Tweets-based Sentiment Analysis to investigate the impact of COVID-19 in KSA*: A deep learning approach. Big Data and Cognitive Computing, 7(1), 16.

[56] Shode, I., Adelani, D. I., & Feldman, A. (2022). *yosm: A new yoruba sentiment corpus for movie reviews*. arXiv preprint arXiv:2204.09711.

[57] Eludiora, S. I., & Akinbande, O. A. (2017). *Implementation of Yorùbá Language Multimedia Learning System. Transactions on Machine Learning and Artificial Intelligence*, 4(6), 01.

[58] Ahia, O., Aremu, A., Abagyan, D., Gonen, H., Adelani, D. I., Abolade, D., ... & Tsvetkov, Y. (2024). *Voices Unheard: NLP Resources and Models for Yor\ub\'a Regional Dialects*. arXiv preprint arXiv:2406.19564.