Harnessing Ensemble and Transformers for Sentiment Analysis and Emotion Detection in Hausa Text

Taiwo Kolajo, Kabir Garba

Department of Computer Science, Federal University Lokoja, P.M.B 1154 Lokoja, Kogi State, Nigeria.

DOI: https://doi.org/10.47772/IJRISS.2026.100601109

Received: 22 June 2026; Accepted: 27 June 2026; Published: 13 July 2026

ABSTRACT

Understanding emotional tone and sentiment in text has driven significant advancements in Natural Language Processing (NLP), particularly in sentiment analysis and emotion detection. This study addresses the challenge of developing effective NLP tools for low-resource languages, focusing on the Hausa language. By leveraging ensemble methods and pre-trained transformer models like BERT and XLM-R, along with traditional classifiers such as Logistic Regression, SVM, Naive Bayes, Random Forest, and XGBoost, we aim to improve sentiment analysis and emotion detection for Hausa text. Utilizing a balanced sentiment dataset (9,958 samples) and a complex multi-label emotion dataset (19,757 samples across 11 categories), we benchmarked individual classifiers, voting ensembles, and deep contextual models. For sentiment analysis, a Hard Voting Ensemble of TF-IDF-vectorized base learners achieved a highly competitive F1-score of 0.8748. However, Transformer models significantly outperformed traditional baselines, with Multilingual BERT (mBERT) achieving a peak F1-score of 0.8983. In the multi-label emotion detection task, individual traditional models struggled with label sparsity, yielding low Subset Accuracy scores (2.88% to 8.30%) and moderate Micro-F1 scores. Standard Hard Voting ensembles further underperformed due to discrete prediction conflicts. To resolve this, a Probability-based Majority Voting mechanism with calibrated thresholding (0.3) was introduced, boosting the Micro-F1 to 0.3825 and reducing the Hamming Loss to 0.1967. Ultimately, XLM-RoBERTa emerged as the superior architecture, achieving a Subset Accuracy of 0.1545, a Micro-F1 of 0.4275, and the lowest Hamming Loss of 0.1804. This research establishes a rigorous benchmark for Hausa NLP, highlighting the indispensable role of subword tokenization, contextual embeddings, and threshold calibration in handling the morphological richness and multi-label complexities of low-resource languages

Keywords: Sentiment analysis, Emotion detection, Transformers, Hausa text

INTRODUCTION

The problem of understanding the emotional tone and feelings conveyed in a text has spurred innovations in the field of Natural Language Processing (NLP) and led to the development of NLP tasks such as sentiment analysis and emotion detection (Al Maruf et al., 2024). Emotion detection and sentiment help computer systems understand the underlying sentiments and emotions expressed in text (Alslaity et al., 2024). Although datasets and linguistic resources for Hausa, a language spoken by millions in West Africa, have become more available, the development of effective NLP tools remains a significant challenge due to the limited depth and variety of annotated data (Shehu et al., 2024).

The field of sentiment analysis and emotion detection has evolved significantly over the past few years, with numerous approaches and techniques being developed to enhance the accuracy and efficiency of these tasks. Low-resource languages in NLP, are categorized as languages that have limited or insufficient annotated data for training and developing NLP applications (Magueresse et al., 2020). High-resource languages such as English typically have large text corpora; an extensive collection of text documents such as books, articles and web data. Sentiment-annotated datasets refer to text data where each piece is labelled with its sentiment (positive, negative). While emotion-annotated data are labelled with corresponding emotional tones such as anger, sadness,

disgust, fear, surprise, joy, trust, optimism, pessimism, anticipation and neutral. The availability of resources for the low-resourced Hausa language has improved, but the limited research in this area still presents challenges in developing NLP tools such as machine translation, speech recognition, sentiment analysis, and emotion detection for low-resource languages (Shehu et al., 2024).

This study tackles these challenges by leveraging ensemble methods and pre-trained transformers such as BERT (Bidirectional Encoder Representations from Transformers) and XLM-R (XLM-RoBERTa) to enhance sentiment analysis and emotion detection in Hausa text. Additionally, we explore the efficacy of traditional machine learning models, including Logistic Regression, Support Vector Machine, Naive Bayes, and Random Forest, to provide a comprehensive evaluation.

The structure of the rest of the paper is as follows. Section 2 discusses the background of the study and related work. The methodology is presented in Section 3. Section 4 presents the results and discussion while the conclusion and further work are presented in Section 5.

BACKGROUND AND RELATED WORK

An overview of the literature and techniques in the field is given in this section, with particular attention to sentiment analysis, emotion detection, transformers, ensemble approaches, and the particular challenges in processing Hausa language.

Hausa language

Spoken by about 40 million people, mostly in Nigeria and Niger, Hausa is a Chadic language that belongs to the Afro-Asiatic subgroup. Hausa is regarded as a low-resource language in the context of NLP despite being widely used because there aren't many linguistic tools and annotated corpora available (Inuwa-Dutse, 2021). The development of strong NLP applications, such as sentiment analysis and emotion recognition, is hampered by this scarcity (Yusuf et al., 2019).

Prior studies on Hausa text processing have concentrated on creating fundamental language resources and tools, like machine translation and part-of-speech taggers (Muhammad et al., 2020; Garba et al., 2024). Sentiment analysis and emotion recognition for Hausa have been explored to some extent, but further research is needed to fully develop these fields (Sani et al., 2022). Some of these challenges can be addressed by utilising multilingual transformer models, such as XRoBERTa and BERT, which are pre-trained on various languages and can transfer knowledge from high-resource languages to Hausa.

Sentiment Analysis and Emotion Detection

In sentiment analysis text is classified into positive and negative sentiments, while emotion detection aims to identify specific emotions such as joy, anger, sadness, etc. These tasks are essential for applications like public opinion mining, monitoring of social media space, and customer feedback analysis. Rule-based methods and lexicons were some of the earliest approaches relied upon, which were limited due to their inability to effectively handle ambiguity and context (Dejaeghere et al., 2024; Kolajo & Kolajo, 2018).

Supervised learning techniques utilizing algorithms such as Naive Bayes, Support Vector Machines (SVM), and Logistic Regression were introduced with the rise of machine learning and have been applied to sentiment analysis (Ali et al., 2023). Although these techniques performed better, training with them required huge, annotated datasets. Deep learning models have demonstrated great promise in capturing the intricate patterns found in text data more recently, especially those that are based on recurrent neural networks (RNNs) and convolutional neural networks (CNNs) (Dang et al., 2020; Shiri et al., 2023).

By offering contextual embeddings that are more adept at understanding linguistic subtleties than conventional techniques, transformer-based models like BERT have taken the discipline even further (Gupta, 2024). These models have achieved state-of-the-art results after being refined for a variety of NLP applications, such as

sentiment analysis and emotion identification. Abdullahi et al. (2024) underscore the significance of sentiment analysis in understanding user-generated texts, particularly in languages like Hausa. By adapting Multinomial Naïve Bayes (MNB) and Logistic Regression algorithms along with the count vectorizer, the research aims to develop an enhanced Hausa Sentiment Dataset. Results indicate a 4% improvement in accuracy compared to plain Hausa datasets, highlighting the importance of resolving ambiguity in abbreviation and acronym usage for more accurate sentiment analysis in Hausa language texts.

Muhammad et al. (2023) developed AfriSenti, a sentiment analysis benchmark comprising over 110,000 tweets across 14 African languages. AfriSenti addresses this gap by providing high-quality annotated datasets for languages including Hausa, Igbo, Nigerian Pidgin, and Yoruba spanning four language families. Salahudeen et al. (2023) worked on the SemEval-2023 Task 12 which focused on sentiment analysis for low-resource African languages leveraging data from X formally Twitter. The authors employed pre-trained models such as Afro-xlmr-large, AfriBERTa-Large, and BERT-base-arabic-camelbert-da-sentiment (Arabic-camelbert), among others, to analyze sentiment across 14 African languages. Their findings highlight the Afro-xlmr-large model's superior performance, especially in Nigerian languages like Hausa, Igbo, and Yoruba.

Ramanathan et al. (2023) carried out research which focused on monolingual sentiment classification in Hausa tweets. The authors proposed AfriSentiSemEvaaddresses, a framework for sentiment analysis for low-resource African languages using Twitter data. The framework approach demonstrates a tailored solution for sentiment analysis in Hausa, contributing to the broader effort of sentiment analysis in low-resource languages. Sani et al. (2022) highlighted the importance of understanding public opinion of Hausa language speakers. The authors identified the challenges of deciphering opinions and emotions in Hausa text sentiment analysis due to its informal nature and unstructured format. In addressing this, the authors employed machine learning and lexicon-based approaches by using Multinomial Naive Bayes (MNB) and Logistic Regression (LR) algorithms with Count Vectorizer and TF-IDF methods on a dataset sourced from BBC Hausa X handle formally Twitter. Results highlight LR's superior performance in categorizing Hausa language text sentiments.

Ensemble Approach

To combine the predictions of several models and enhance model performance, ensemble methods have been applied extensively in machine learning. These techniques, which include stacking, boosting, and bagging, aid in lowering bias and variance while enhancing the models' generalizability (Başarslan & Kayaalp, 2024).

Training several models on various subsets of the training data and averaging their predictions is known as bagging or bootstrap aggregating. Conversely, boosting concentrates on training models in a stepwise manner, with each model aiming to rectify the mistakes of the one before it. In stacking, a meta-learner is trained to integrate base model predictions in the best possible way (Tahir et al., 2020).

Recently, studies have shown that ensemble methods can enhance the performance of sentiment analysis systems significantly. An example is a study by Smith et al. (2019) demonstrating how an ensemble of traditional machine learning models and deep learning models outperformed individual models in sentiment classification tasks.

This study employed a multi-tiered ensemble framework beginning with five diverse base classifiers, namely Logistic Regression, Linear SVM, Naive Bayes, Random Forest, and XGBoost, each integrated into a One-vs-Rest decomposition strategy for the multi-label emotion task. For binary sentiment analysis, a standard Hard Voting Ensemble aggregated discrete class predictions to smooth individual model variances. For multi-label emotion detection, a Probability-based Majority Voting mechanism was engineered to aggregate continuous decision functions, with a logistic sigmoid transformation applied to the SVM outputs to enable seamless probabilistic averaging, followed by a calibrated classification threshold to aggressively capture sparse, underrepresented emotions.

Transformer-based models

Transformer was first introduced by Vaswani et al. (2017) in their paper, “Attention is all you need”.

Transformer-based models such as BERT and XLM-R (XLM-RoBERTa) are advanced neural network architectures designed for natural language understanding tasks. BERT uses a bidirectional approach to capture context from both the left and right of a token, enhancing its ability to understand the nuances of language (Wolf et al., 2020). XLM-R extends this by being multilingual, and capable of handling multiple languages with robust cross-lingual understanding. Both models excel in tasks like sentiment analysis and emotion detection due to their deep contextual embeddings and fine-tuning capabilities on specific datasets, providing state-of-the-art performance in various NLP applications.

METHODOLOGY

The ensemble learning methodologies employed encompass three distinct approaches: ensemble with majority voting, ensemble of individual classifiers, and transformer-based models for sentiment analysis and emotion detection of Hausa text. The development process of the model was in these phases: data collection and preprocessing, feature selection, training/fine-tuning, testing, validation, and evaluation. These steps are subsequently discussed. Fig. 1 showcases a development design highlighting the processes involved.

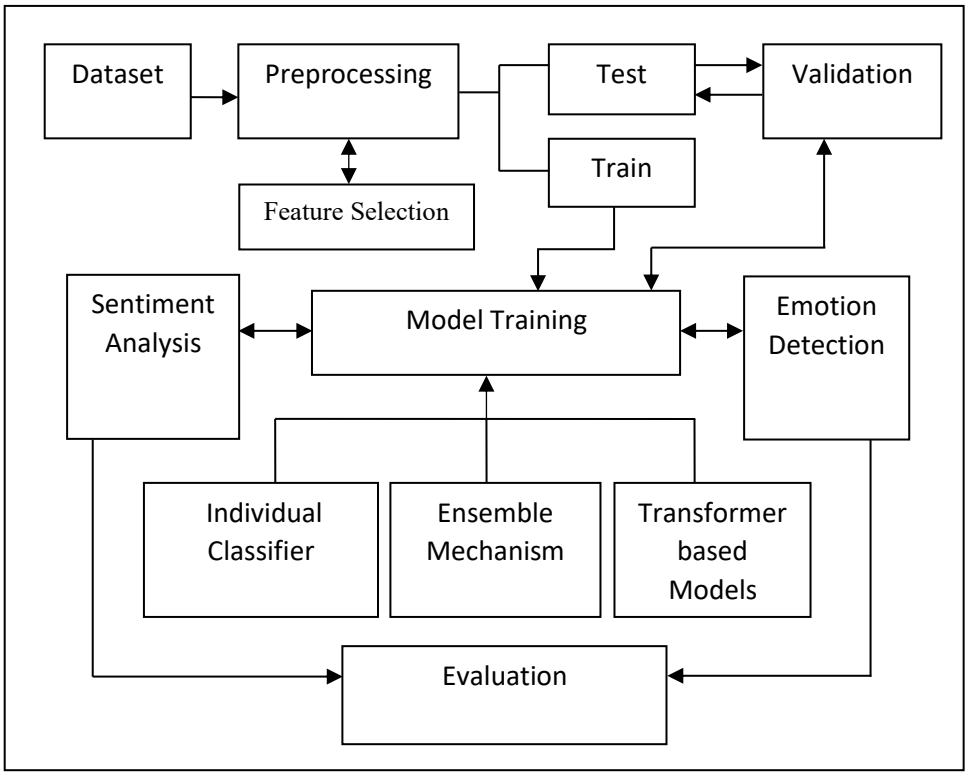

Fig. 1: Development Design

Data collection and preprocessing

The two separate annotated datasets used for this study were sourced from the NaijaSenti corpus developed by Muhammad et al. (2022) and the BRIGHTER dataset introduced by Sani et al. (2025). The dataset used for the sentiment analysis task had two columns labelled “Sentiment” and “Polarity”. Polarity consists of three sentiment labels, later encoded numerically: Positive and Negative. The dataset used for the emotion detection task also contained two columns labeled “Text” and “Labels”, the label columns contain five (5) emotion labels: Sadness, Joy, Love, Anger, Fear and Surprise. The labels were encoded numerically.

The textual data were preprocessed to remove noise, such as special characters and URLs, and tokenized using language-specific tools. The textual data which were in English were translated to Hausa with the help of Google Translate and their emotion and polarity labels were maintained. Table 1 presents the summary of the sentiment and emotion datasets while Figures 2 and 3 show a sample of the dataset, respectively.

Table 1: Summary of the sentiment and emotion datasets

Dataset	Description
Sentiment Analysis Dataset	A manually annotated Hausa sentiment dataset consisting of 6,287 social media texts labeled into 2 sentiment classes (Positive and Negative), split into 70% training (4,526 samples), 10% validation (503 samples), and 20% testing (1,258 samples) for binary sentiment classification.
Emotion Detection Dataset (HaEmoC_V1)	The HaEmoC_V1 dataset comprises 6,000 Hausa tweets annotated with 11 emotion categories (anger, sadness, disgust, fear, surprise, joy, trust, optimism, pessimism, anticipation, and neutral), using approximately 70% training (4,200 samples), 15% validation (900 samples), and 15% testing (900 samples) for multi-label emotion classification.

	ID	Hausa	Polarity
0	1	A ajizance	Positive
1	2	A amince	Positive
2	3	A arha	Positive
3	4	A banza aka buga	Negative
4	5	A banza aka buga buhun	Negative

Fig. 2: Dataset Sample for Sentiment Analysis Task

	ID	tweet	anger	sadness	disgust	fear	surprise	joy	trust	optimism	pessimism	antic
hausa_00001	ah su rahama za axha madara sa lemu fah mudai ...	0	0	0	0	0	1	1	1	0		
hausa_00002	bbc idan bakuda abin posting ku saka mana hoto...	1	1	1	1	1	0	0	0	1		

Fig. 3: Dataset Sample for Emotion Detection Task

Feature Selection and Feature Extraction

Feature selection is a dimension reduction technique used to select relevant features based on the target variable(s) for the machine learning task (Buyukkececi & Okur, 2023). Eliminating redundant and irrelevant features speeds up the process and increases the performance of machine learning algorithms. Unwanted features such as irrelevant columns were removed leaving only two columns each per dataset, one containing the Hausa texts and the other the polarity for sentiment analysis task and emotion tone for emotion detection task.

In this study, we utilized Term Frequency-Inverse Document Frequency (TF-IDF) vectorization for the traditional machine learning pipeline. To capture local contextual phrases and Hausa idiomatic expressions, we extracted both unigrams and bigrams. Strict feature selection parameters were applied to manage dimensionality and prevent overfitting: the vocabulary was constrained to a maximum of 3,099 features for the sentiment dataset and 5,000 features for the emotion dataset, with a minimum document frequency (min_df) of 2 and a maximum

document frequency (max_df) of 0.95 to eliminate both noise and ubiquitous stop-words. Sublinear TF scaling and L2 normalization were applied, resulting in a highly optimized, albeit extremely sparse, feature matrix (exhibiting 99.88% sparsity in the sentiment task). Crucially, a custom regular expression token pattern was engineered to explicitly preserve Hausa-specific hooked characters (ɓ, ɗ, ƙ) and apostrophes, ensuring that vital morphological distinctions were not obliterated by aggressive text cleaning.

Conversely, for the deep learning pipeline, we leveraged subword tokenization to inherently handle Hausa's agglutinative morphology and out-of-vocabulary (OOV) words. We utilized the WordPiece tokenizer for Multilingual BERT (mBERT), which features a vocabulary size of 119,547, and the SentencePiece Byte-Pair Encoding (BPE) tokenizer for XLM-RoBERTa, boasting a vastly larger vocabulary of 250,002. Exploratory data analysis revealed that Hausa language texts are highly concise, with a median token length of just 8. Consequently, the maximum sequence length was dynamically capped at 128 tokens for sentiment and 64 tokens for emotion detection. This strategic truncation preserved crucial contextual windows while optimizing GPU memory utilization during batch processing. Prior to feature extraction, the datasets underwent rigorous stratified splitting. The sentiment dataset was divided into training (70%), validation (15%), and testing (15%) sets using standard stratified sampling to maintain the 50/50 class balance. For the multi-label emotion dataset, standard stratification is insufficient due to overlapping label distributions; therefore, we employed the MultilabelStratifiedShuffleSplit algorithm. This ensured that the complex, co-occurring emotion distributions remained statistically consistent across all data partitions.

Handling Class Imbalance

This study deployed distinct task-specific methodological interventions due to the divergent natures of the two datasets to address class distribution, . The binary sentiment dataset, comprising 9,958 samples, exhibited a perfect 1:1 class balance, with exactly 4,979 instances (50.0%) for both the positive and negative polarities, yielding an imbalance ratio of 1.00. Because the target variable was inherently equitably distributed, algorithmic rebalancing techniques such as the Synthetic Minority Over-sampling Technique (SMOTE) were deemed unnecessary and intentionally omitted to preserve the authentic linguistic distribution.

Instead, the data was partitioned into training (6,970 samples), validation (1,494 samples), and testing (1,494 samples) using standard stratified sampling. This approach rigorously maintained the exact 747-to-747 class equilibrium across the validation and testing subsets, ensuring that the traditional machine learning and Transformer models were evaluated on a statistically unbiased foundation without introducing synthetic artifacts.

Conversely, the multi-label emotion dataset (19,757 samples) exhibited a mild but critical distribution variance across its 11 categories, ranging from a maximum of 25.07% (joy) to a minimum of 17.07% (neutral). Standard stratification algorithms are mathematically incapable of preserving the complex, overlapping co-occurrences of multi-label targets; therefore, we employed an iterative multi-label stratification algorithm to partition the data into 13,856 training, 2,955 validation, and 2,946 testing samples while strictly preserving the joint probability distribution of all emotion labels.

To mitigate label sparsity during inference without relying on noise-inducing synthetic text generation, we engineered a Probability-based Majority Voting mechanism. Because the Linear Support Vector Machine outputs raw decision boundaries rather than native probabilities, we applied a logistic sigmoid transformation to its decision function, mathematically normalizing its outputs to be seamlessly averaged alongside the native probability estimates of the other base learners.

By applying a calibrated classification threshold of 0.3 deliberately lowered from the standard 0.5 to this aggregated probabilistic output, the ensemble was empowered to aggressively capture sparse, underrepresented emotions. Finally, for the Transformer architecture, sparsity and label overlap were handled natively by configuring the model with independent Sigmoid activations and Binary Cross-Entropy loss, allowing the network to learn the distinct frequency distributions of each emotion independently rather than forcing them into a mutually exclusive, imbalance-prone Softmax distribution.

Model Training

To ensure rigorous and unbiased evaluation, the datasets underwent systematic partitioning into training (70%), validation (15%), and testing (15%) sets. For the binary sentiment dataset, standard stratified sampling was employed to preserve the exact 50/50 class distribution across all partitions. Conversely, the multi-label emotion dataset required a more sophisticated approach; standard stratification is statistically insufficient when labels overlap. Therefore, we utilized iterative multi-label stratification, an algorithmic technique designed to ensure that the complex, co-occurring emotion distributions remained statistically consistent across the training, validation, and testing sets, thereby preventing label skew.

Building upon this data foundation, the experimental framework evaluated three distinct categories of predictive models, beginning with individual base classifiers. We trained five diverse algorithms Logistic Regression, Linear Support Vector Machines, Multinomial Naive Bayes, Random Forest, and XGBoost to capture varying linear and non-linear decision boundaries. To adapt these inherently binary classifiers for the multi-label emotion task, we integrated them into a One-vs-Rest decomposition strategy. This meta-estimator approach effectively trained an independent binary classifier for each of the eleven emotion categories, allowing the models to evaluate the presence or absence of each emotion in isolation.

The second category sought to aggregate the predictive strengths of these base learners through ensemble voting mechanisms. For the binary sentiment task, a standard Hard Voting mechanism was deployed, combining discrete class predictions via majority rule to smooth out individual model variances and provide a robust baseline. However, empirical testing revealed that standard hard voting proved detrimental in the highly sparse multi-label emotion space, as base models rarely agreed on the exact same subset of rare labels, causing performance degradation. To circumvent this, we engineered a Probability-based Majority Voting mechanism for the emotion task. This approach aggregated the continuous decision functions and probability estimates of the base models, averaging them across the ensemble. Aditionally, a calibrated classification threshold of 0.3 was applied—deliberately lowered from the standard 0.5 to aggressively capture sparse, underrepresented emotions that discrete voting would otherwise suppress.

The final category integrated deep contextual embeddings via pre-trained Transformer architectures, specifically Multilingual BERT (mBERT) and XLM-RoBERTa. These models were fine-tuned utilizing mixed-precision training to accelerate convergence, optimized with a learning rate of 2e-5 and weight decay, and governed by an early stopping protocol monitoring validation F1-scores to prevent overfitting. For the binary sentiment task, the models were fitted with a standard sequence classification head. For the multi-label emotion task, the architecture was fundamentally adapted to support simultaneous, non-mutually exclusive predictions. The standard softmax output layer was replaced with independent Sigmoid activations for each emotion category, and the model was optimized utilizing Binary Cross-Entropy with Logits Loss. This architectural adjustment ensured that the model prioritized the correct identification of individual label instances across the entire dataset, effectively navigating the complex, overlapping emotional landscape of the Hausa language.

RESULTS AND DISCUSSION

In this section, the results of the hyperparameter tuning for sentiment analysis (See Table 2) and emotion detection (see Table 3) are presented followed by the evaluation of the models' performance. The models were evaluated using the following steps:

Hold-Out Testing Protocol: Final performance was assessed exclusively on unseen test sets (1,494 samples for sentiment; 2,946 samples for emotion) that were completely isolated from the training and validation phases to ensure unbiased generalization metrics.

Binary Classification Metrics: The Sentiment Analysis models were evaluated using standard classification metrics: Accuracy, Weighted Precision, Recall, and F1-Score, providing a clear measure of binary polarity detection.

ISSN No. 2454-6186 | DOI: 10.47772/IJRISS | Volume X Issue VI June 2026

Multi-Label Evaluation Metrics: The Emotion Detection models were evaluated using a specialized suite of metrics Subset Accuracy (Exact Match Ratio), Micro-F1, Macro-F1, and Hamming Loss to rigorously account for overlapping labels and penalize partial misclassifications.

Inference Threshold Calibration: During the evaluation of the probabilistic ensemble, a customized decision threshold (0.3) was applied at inference time to optimize the capture of sparse, underrepresented emotion labels without altering the underlying model weights.

Table 2: Sentiment Analysis - Hyperparameter configuration and fine-tuning settings

Parameter	**BERT (bert-base-multilingual-cased)**	**XLM-R (xlm-roberta-base)**
Task	Binary sentiment classification	Binary sentiment classification
Number of Classes	2	2
Label Encoding	LabelEncoder	LabelEncoder
Random Seed	42	42
Tokenizer	bert-base-multilingual-cased tokenizer	xlm-roberta-base tokenizer
Maximum Sequence Length	128 tokens	128 tokens
Padding Strategy	Dynamic padding (DataCollatorWithPadding)	Dynamic padding
Learning Rate	2×10^{-5}	2×10^{-5}
Optimizer	AdamW (Transformers default)	AdamW (Transformers default)
Weight Decay	0.01	0.01
Training Epochs	5	5
Training Batch Size	32	16
Evaluation Batch Size	64	32
Mixed Precision	FP16	FP16
Evaluation Strategy	End of every epoch	End of every epoch
Checkpoint Saving	Every epoch	Every epoch
Best Model Selection	Highest weighted F1-score	Highest weighted F1-score
Early Stopping	Patience = 2 epochs	Patience = 2 epochs

Maximum Saved Checkpoints	2	2
Metrics	Accuracy, Precision, Recall, Weighted F1	Accuracy, Precision, Recall, Weighted F1

Table 3: Emotion Detection-Hyperparameter configuration and fine-tuning settings

Parameter	**Traditional Machine Learning Models**	**XLM-R Multi-label Fine-tuning**
Task	Multi-label emotion detection	Multi-label emotion detection
Number of Emotion Labels	11	11
Emotion Classes	Anger, Sadness, Disgust, Fear, Surprise, Joy, Trust, Optimism, Pessimism, Anticipation, Neutral	Same
Feature Extraction	TF-IDF	XLM-R tokenizer
TF-IDF Features	5,000 maximum features	Not applicable
N-gram Range	(1,2)	Not applicable
Models	Logistic Regression, Linear SVM, Multinomial Naïve Bayes, Random Forest, Majority Voting Ensemble	XLM-R Base
Random Forest Trees	150	Not applicable
Train/Test Split	80% / 20%	70% Train, 15% Validation, 15% Test
Random Seed	42	42
Maximum Sequence Length	Not applicable	64 tokens
Learning Rate	Default Scikit-learn	2×10^{-5}
Optimizer	Model defaults	AdamW (Transformers default)
Weight Decay	Not applicable	0.01
Training Epochs	Until model convergence	3
Training Batch Size	Not applicable	16

Evaluation Batch Size	Not applicable	32
Mixed Precision	No	FP16
Evaluation Strategy	Test set evaluation	End of every epoch
Checkpoint Saving	None	Every epoch
Best Model Selection	Not applicable	Highest Micro-F1
Problem Type	One-vs-Rest Multi-label Classification	multi_label_classification
Prediction Threshold	Majority Voting ≥ 0.5 (later 0.3 for probability voting)	Sigmoid ≥ 0.5

Cross-Architecture Benchmarking: The empirical results were systematically compared across three distinct architectural tiers (Individual Classifiers, Ensemble Mechanisms, and Transformers) to determine the optimal trade-off between computational efficiency and semantic depth in low-resource settings.

The Evaluation scores are showcased in Tables 2-6. In addition, two confusion matrices were showcased in Figures 6, 7, 8 and 9 to give an insight into the overall performance of the model and highlight areas for improvement in subsequent works. Table 4 presents the individual classifiers for sentiment analysis.

Table 4: Individual Classifiers-- Sentiment Analysis

Model Architecture	**Validation Accuracy**	**Precision**	**Recall**	**F1-Score**
Logistic Regression	0.8708	0.8716	0.8708	0.8707
Linear SVM	0.8701	0.8716	0.8701	0.8700
Naive Bayes	0.8541	0.8542	0.8541	0.8541
Random Forest	0.8708	0.8711	0.8708	0.8708
XGBoost	0.8273	0.8276	0.8273	0.8273

In the Sentiment Analysis task, the individual traditional classifiers demonstrated strong and relatively uniform performance. Logistic Regression, Random Forest, and Linear SVM achieved validation accuracies and F1-scores hovering around 0.8700. Random Forest achieved the highest training accuracy (0.9770) but exhibited moderate overfitting, as its validation accuracy plateaued at 0.8708. Conversely, XGBoost showed excellent generalization though it ultimately underperformed in absolute metrics (F1: 0.8273). This degradation is attributable to the extreme sparsity of the TF-IDF matrix (99.88%), which makes it exceedingly difficult for tree-based algorithms to find optimal, continuous split points without extensive hyperparameter tuning. Naive Bayes trailed slightly (F1: 0.8541), constrained by its rigid assumption of feature independence. The result for individual classifiers for emotion detection is presented in Table 5.

Table 5: Individual Classifiers-Emotion Detection

Model Architecture	**Subset Accuracy**	**Micro-F1**	**Macro-F1**
Logistic Regression	0.0418	0.2861	0.2454

Linear SVM	0.0665	0.3641	0.3395
Naive Bayes	0.0288	0.2564	0.2069
Random Forest	0.083	0.3137	0.2797

The Emotion Detection results present a much more complex analytical landscape. It is imperative to contextualize the Subset Accuracy metric in multi-label classification; it represents the Exact Match Ratio, requiring the model to predict the exact same set of active labels for a sample as the ground truth. Consequently, the Subset Accuracy for all traditional models is exceptionally low, ranging from 2.88% (Naive Bayes) to 8.30% (Random Forest). Therefore, Micro-F1 and Macro-F1 serve as the definitive metrics for evaluating holistic performance. Among the individual classifiers, Linear SVM emerged as the most effective traditional baseline, achieving a Micro-F1 of 0.3641. High-dimensional, sparse TF-IDF matrices align remarkably well with the linear hyperplanes constructed by SVMs, whereas Naive Bayes struggled significantly (Micro-F1: 0.2564). Table 6 presents ensemble mechanism for sentiment analysis.

Table 6: Ensemble Mechanism-Sentiment Analysis

Ensemble Strategy	Training Accuracy	Validation Accuracy	Precision	Recall	F1-Score	Training Time (s)
Hard Voting Ensemble	0.9403	0.8748	0.8752	0.8748	0.8748	24.639

For the binary Sentiment Analysis task, the Hard Voting Ensemble successfully mitigated individual model biases, achieving the highest traditional validation accuracy (0.8748) and the best weighted precision (0.8752) and recall (0.8748). By aggregating the discrete predictions of the base learners, the ensemble smoothed out individual variances, yielding a robust and computationally lightweight baseline that outperformed every individual traditional classifier. The ensemble mechanism is also presented in Table 7.

Table 7: Ensemble Mechanism - Emotion Detection

Ensemble Strategy	Micro-F1	Macro-F1	Hamming Loss
Hard Voting (Threshold 0.5)	0.3352	0.2927	N/A
Probabilistic Voting (Threshold 0.3)	0.3825	0.3419	0.1967

The Emotion Detection results expose a critical limitation of standard ensembles in multi-label spaces. The Hard Voting Ensemble actually performed worse than the best individual model (Micro-F1 dropped to 0.3352). In highly sparse multi-label datasets, base models rarely agree on the exact same subset of rare labels, and discrete voting mechanisms fail by amplifying conflicting negative predictions. To circumvent this, we engineered a Probability-based Majority Voting mechanism. By extracting continuous decision functions and applying a calibrated threshold of 0.3, the ensemble captured edge-case emotional markers, successfully boosting the Micro-F1 to 0.3825 and achieving a Hamming Loss of 0.1967 (see Table 8). This demonstrates that threshold calibration is a vital methodological intervention for combating label sparsity in low-resource multi-label datasets.

Table 8: Transformer Architectures

Transformer Model	Task	Validation Accuracy	Precision	Recall	F1-Score
mBERT	Sentiment Analysis	0.8984	0.8984	0.8984	0.8984

XLM-RoBERTa		0.8795	0.8795	0.8795	0.8795
XLM-RoBERTa	Emotion detection	0.1545	0.4275	0.3892	0.1804

Ultimately, the integration of deep contextual embeddings via Transformer-based architectures yields the most profound performance gains. In the Sentiment Analysis task, the Transformer models dominated the evaluation: mBERT achieved a peak validation accuracy of 0.8984, with near-perfect parity in weighted precision and recall, while XLM-RoBERTa scored 0.8795 across all metrics.

This superior performance can be attributed to mBERT's cased architecture and deep bidirectional contextual understanding, allowing it to differentiate between subtle sarcasm, negation particles, and idiomatic expressions that bag-of-words TF-IDF models inherently missed. Parallel to these findings, the Emotion Detection task demonstrated a similar reliance on deep context; here, XLM-RoBERTa vastly outperformed all traditional and ensemble methods, achieving a subset accuracy of 0.1545, a micro-F1 score of 0.4275, and the lowest Hamming loss of 0.1804. A critical analysis reveals a noticeable gap between its micro-F1 (0.4275) and macro-F1 (0.3892) metrics, indicating class-conditional performance variance.

The model excels on high-frequency emotions (e.g., joy, anticipation, and trust) but struggles with low-frequency or highly contextual ones like neutral states or pessimism. The sheer superiority of XLM-RoBERTa over the TF-IDF pipelines underscores a fundamental linguistic reality: emotion in Hausa is heavily dependent on syntax, verb conjugations, and contextual framing. Because TF-IDF completely ignores word order and semantic relationships, it is fundamentally incapable of capturing syntactic emotional triggers. XLM-RoBERTa's self-attention mechanism successfully captures these deep contextual dependencies, proving that cross-lingual Transformers are indispensable for mapping the complex emotional landscape of morphologically rich languages like Hausa.

In summary, the empirical results demonstrate that while traditional machine learning models and ensemble strategies offer reliable baselines, deep contextual Transformer architectures are indispensable for capturing the linguistic nuances of Hausa. For binary sentiment analysis, traditional classifiers like Logistic Regression and Linear SVM achieved strong validation accuracies hovering around 0.8700. They were subsequently optimized to 0.8748 by a Hard Voting Ensemble.

However, these bag-of-words models struggled significantly with the sparse, multi-label complexity of emotion detection, where a custom Probabilistic Voting Ensemble (calibrated to a 0.3 threshold) was required to lift the micro-F1 score to 0.3825. Ultimately, Transformer architectures dominated both tasks; mBERT leveraged its bidirectional contextual awareness to achieve a peak sentiment accuracy of 0.8984, while XLM-RoBERTa vastly outperformed all traditional pipelines in emotion detection with a micro-F1 score of 0.4275 and a low Hamming loss of 0.1804.

These findings underscore that analyzing sentiment and emotion in a morphologically rich language like Hausa fundamentally relies on the self-attention mechanisms of cross-lingual Transformers to decode syntax and context-dependent emotional triggers. Figure 4 presents model accuracy for sentiment analysis tasks.

Fig. 4 represents the model accuracy comparison for the sentiment analysis task. The accuracy results for the sentiment analysis pipeline demonstrate a clear progression from traditional feature-based methods to modern contextual language models. Conventional classifiers such as Logistic Regression, Linear SVM, and Random Forest achieved strong baseline performance, showing that TF-IDF features remain effective for binary sentiment classification.

In contrast, XGBoost and Naive Bayes produced lower accuracies, reflecting the challenges posed by sparse feature representations and the simplifying assumptions of these algorithms. Combining the traditional classifiers through a Hard Voting Ensemble further improved performance by reducing the weaknesses of individual models, making it the strongest conventional approach.

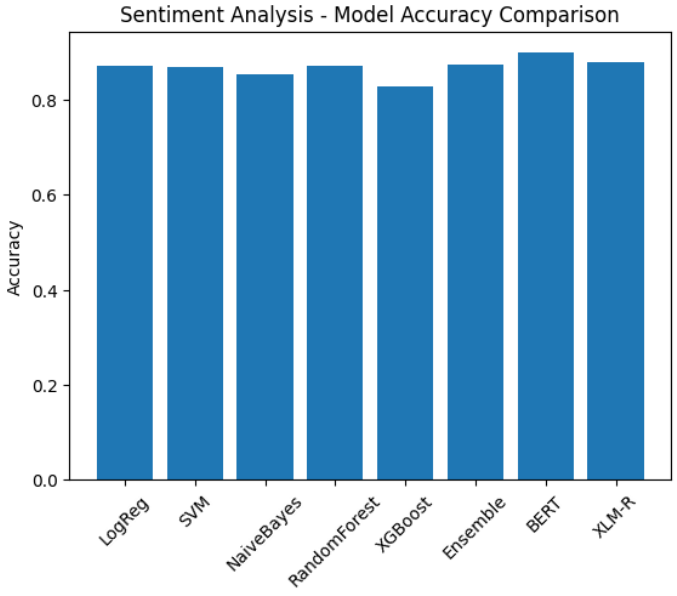

Fig. 4: Model Accuracy Comparison Bar Chart (Sentiment Analysis)

However, the Transformer models consistently outperformed all traditional methods. By processing text within its full context rather than relying on isolated word frequencies, Multilingual BERT (mBERT) was able to capture the complex grammatical structure, local expressions, and subtle contextual cues present in Hausa text. This enabled it to achieve the highest accuracy in the study, highlighting the value of contextual deep learning for sentiment analysis in low-resource languages. The results of the F1 comparison is presented in Figure 5.

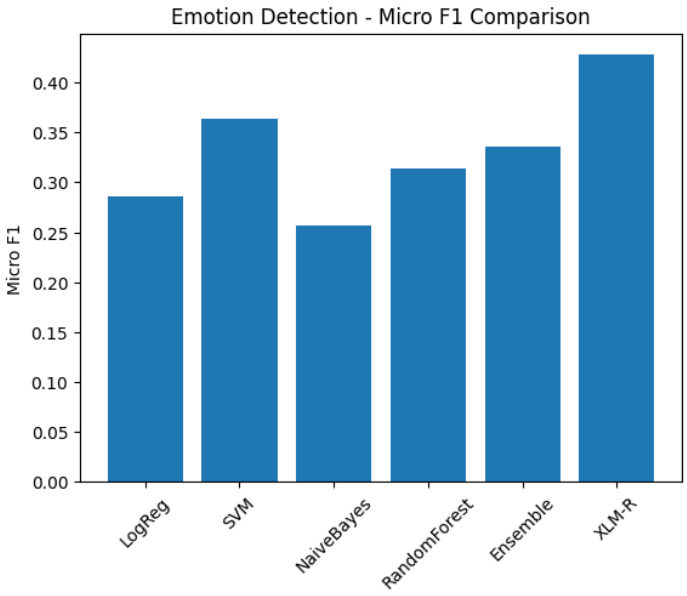

Fig. 5: Micro F1 Comparison Bar Chart (Emotion Detection)

In contrast, as shown in Fig. 5, the transition from binary sentiment analysis to multi-label emotion detection presents a much greater challenge, revealing the limitations of traditional machine learning methods. Unlike binary classification, emotion detection requires models to identify multiple emotions that may occur simultaneously, making the task considerably more complex. As a result, frequency-based approaches performed less consistently, with Naive Bayes showing the weakest results due to its assumption that features are independent. Among the traditional models, Linear SVM delivered the strongest performance because it is better suited to handling sparse, high-dimensional feature spaces than tree-based methods such as Random Forest. Although the Hard Voting Ensemble improved robustness in sentiment analysis, it performed less effectively

for emotion detection because conflicting votes often led to incorrect predictions for infrequent emotion labels. To address this limitation, a Probability-based Ensemble with a calibrated decision threshold was developed, allowing the model to better detect underrepresented emotional categories. Despite this improvement, XLM-RoBERTa achieved the best overall performance. Its self-attention mechanism enabled it to capture rich contextual relationships beyond the capabilities of TF-IDF-based models, resulting in the highest Micro F1-score (see Figure 6) and demonstrating the importance of deep contextual representations for multi-label emotion detection in Hausa

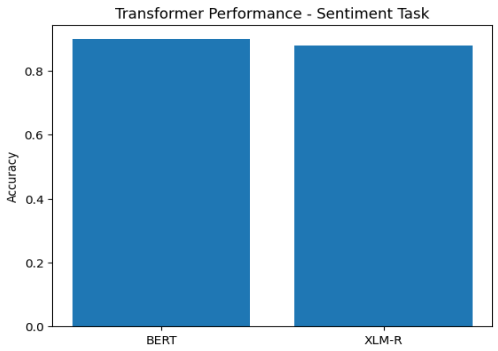

Fig. 6: Transformer Performance Bar chart (Sentiment Analysis)

A comparison of the top-performing sentiment analysis models shows a close competition between the two multilingual Transformer architectures, with mBERT achieving a slight advantage over XLM-RoBERTa. Although both models significantly outperformed the traditional machine learning approaches, mBERT's cased architecture and WordPiece tokenization were better able to capture the contextual and morphological characteristics of Hausa. This enabled it to more effectively interpret features such as negation, local expressions, and subtle sentiment cues, resulting in the highest accuracy achieved in this study. The marginally lower performance of XLM-RoBERTa does not indicate a weakness in the model itself but rather reflects the challenges of adapting a broadly multilingual model to the unique linguistic characteristics of a low-resource language such as Hausa.

In Fig. 7, XLM-RoBERTa's superior performance in the emotion detection task highlights the importance of contextual language understanding for modelling complex emotions in Hausa. By achieving the highest Micro F1-score among all models, it demonstrated the effectiveness of self-attention in capturing how the meaning of words changes with context rather than relying on word frequencies alone. This enabled the model to identify nuanced emotional categories such as joy, anticipation, and trust more accurately than traditional and ensemble methods. Although these results establish a strong benchmark for emotion detection in low-resource languages, they also reveal opportunities for further improvement, particularly in recognising rare and highly context-dependent emotions such as pessimism and neutral expressions. The confusion matrix for individual classifiers in Figure 8 while that of Transformers and Majority Voting are presented in Figure 9.

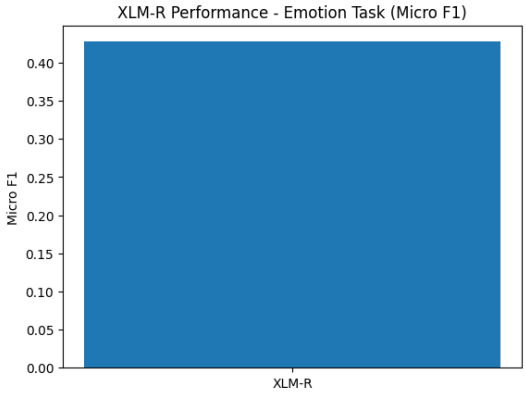

Fig. 7: XLM-R Micro F1 Performance (Emotion Analysis)

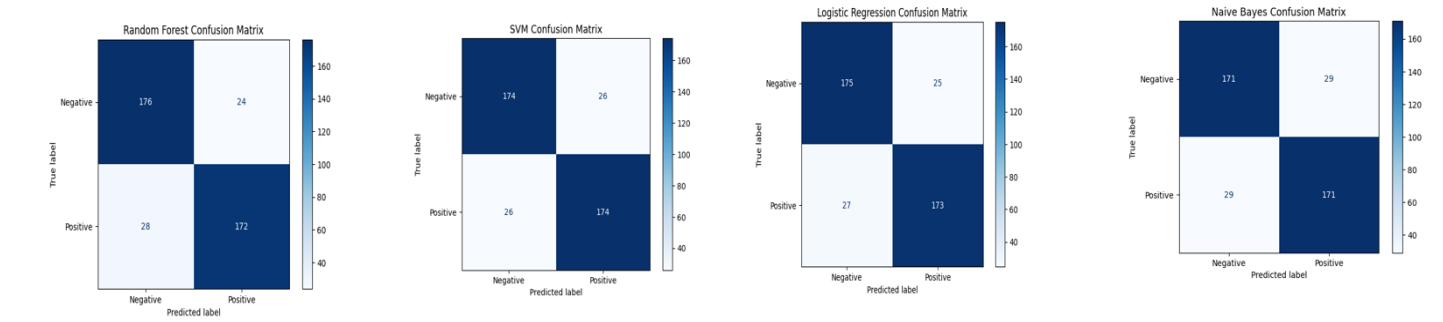

Fig 8: Confusion Matrix for Individual classifiers

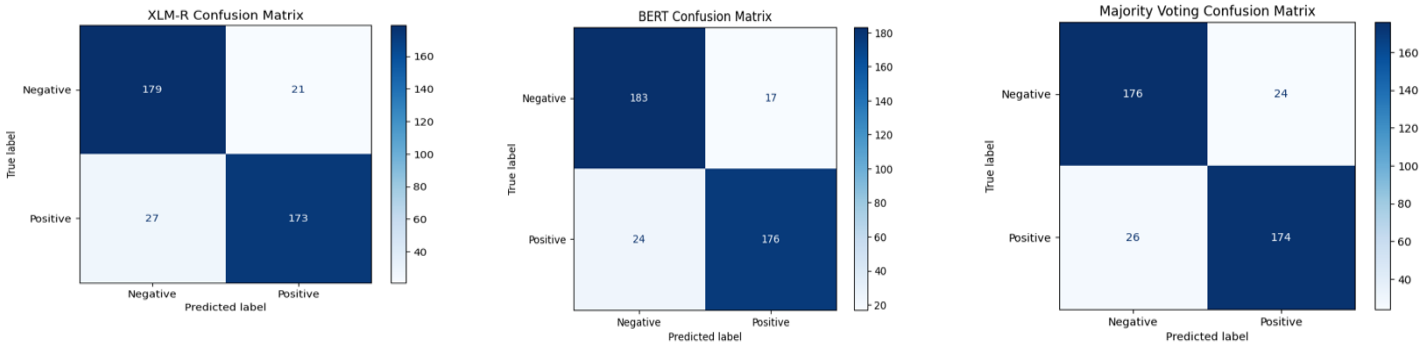

Fig 9: Confusion Matrix of the Transformers and Majority Voting

CONCLUSION AND FURTHER WORK

The findings of this study establish a strong benchmark for sentiment analysis and emotion detection in low-resource African languages, particularly Hausa. The results show that Transformer-based models are better equipped than traditional machine learning methods to capture the complex morphological and syntactic characteristics of the Hausa language.

For binary sentiment analysis, traditional classifiers such as Logistic Regression and Linear Support Vector Machine (SVM) produced competitive baseline results, with performance further improved through a Hard Voting Ensemble that reduced the limitations of individual models. However, these approaches struggled with the more complex multi-label emotion detection task due to sparse feature representations and overlapping emotion categories. To overcome this challenge, a Probability-based Majority Voting strategy with a calibrated decision threshold of 0.3 was introduced, improving the detection of subtle and infrequent emotional expressions.

Overall, Transformer models achieved the best performance across both tasks. Their subword tokenization and bidirectional self-attention mechanisms enabled them to capture contextual information, including negation, sarcasm, and context-dependent emotions, that traditional bag-of-words methods often missed. Multilingual BERT (mBERT) recorded the highest performance for sentiment analysis, while XLM-RoBERTa achieved the best results for emotion detection with the lowest Hamming Loss. However, the gap between the micro- and macro-averaged metrics indicates that identifying minority emotion categories, such as neutral states and pessimism, remains challenging.

Future research should incorporate Hausa-specific linguistic resources, including dedicated lemmatizers, part-of-speech taggers, and a native Hausa tokenizer, to better capture the language's unique grammatical features. Further improvements may also be achieved through advanced techniques for handling class imbalance, such as context-aware data augmentation, synthetic sample generation using Large Language Models (LLMs), and cost-sensitive learning methods like Focal Loss. Finally, parameter-efficient fine-tuning approaches, including Low-

Rank Adaptation (LoRA) and language-specific adapter modules, should be explored to develop accurate yet computationally efficient models suitable for real-world deployment in resource-constrained settings.

Authors' Declarations

Competing Interests: We, the authors, declare that we have no competing interests, financial or non-financial, that are directly or indirectly related to the work submitted for publication.

Funding: We would like to disclose that this research received no external funding. We, the authors, independently conducted this study without financial support from any organization or entity.

We affirm that these statements accurately reflect our circumstances and commitments regarding competing interests and funding for the research presented in this paper.

REFERENCES

1. Ali, A., & Mashwani, W. K. (2023). A supervised machine learning algorithms: applications, challenges, and recommendations. Proceedings of the Pakistan Academy of Sciences: A Physical and Computational Sciences, 60(4), 1-12.
2. Al Maruf, A., Khanam, F., Haque, M. M., Jiyad, Z. M., Mridha, F., & Aung, Z. (2024). Challenges and opportunities of text-based emotion detection: a survey. IEEE Access, 12, 18416-18450.
3. Alslaity, A., & Orji, R. (2024). Machine learning techniques for emotion detection and sentiment analysis: current state, challenges, and future directions. Behaviour & Information Technology, 43(1), 139-164.
4. Başarslan, M. S., & Kayaalp, F. (2024). Sentiment analysis using a deep ensemble learning model. Multimedia Tools and Applications, 83(14), 42207-42231.
5. Buyukkececi, M., & Okur, M. C. (2023). A comprehensive review of feature selection and feature selection stability in machine learning. Journal of Science, 36(4), 1506-1520.
6. Dejaeghere, T., Singh, P., Lefever, E., & Birkholz, J. (2024, May). Exploring aspect-based sentiment analysis methodologies for literary-historical research purposes. In Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA)@ LREC-COLING-2024 (pp. 129-143).
7. Dang, N. C., Moreno-García, M. N., & De la Prieta, F. (2020). Sentiment analysis based on deep learning: a comparative study. Electronics, 9(3), 483.
8. Garba, K., Kolajo, T., & Agbogun, J. B. (2024). A transformer-based approach to Nigerian Pidgin text generation. International Journal of Speech Technology, 27(4), 1027-1037.
9. Gupta, R. (2024). Bidirectional encoders to state-of-the-art: a review of BERT and its transformative impact on natural language processing. Информатика. Экономика. Управление/Informatics. Economics Management, 3(1), 0311-0320.
10. Inuwa-Dutse, I. (2021). The first large scale collection of diverse Hausa language datasets. arXiv preprint arXiv:2102.06991.
11. Kolajo, T., & Kolajo, J.O. (2018). Sentiment analysis on Twitter health news. FUDMA Journal of Sciences, 2(2), 14-20.
12. Shehu, H. A., Majikumna, K. U., Suleiman, A. B., Luka, S., Sharif, H., Ramadan, R. A. (2024). Unveiling Sentiments: A deep dive into sentiment analysis for low-resource languages–a case study on Hausa texts. IEEE Access, 12, 98900-98916.
13. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv preprint arXiv:1907.11692.
14. Muhammad, A., Abdulmumin, I., & Bello, H. (2020). Developing language resources for Hausa language. International Journal of Advanced Computer Science and Applications, 11(7), 30-36.
15. Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2(1-2), 1-135.

16. Ramanathan, N., Sivanaiah, R., & Thanagathai, M. T. N. (2023, July). TechSSN at SemEval-2023 Task 12: Monolingual Sentiment Classification in Hausa Tweets. In Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023) (pp. 1190-1194).
17. Salahudeen, S. A., Lawan, F. I., Wali, A. M., Imam, A. A., Shuaibu, A. R., Aliyu, Y., ... & Jamoh, A. Y. (2023, April). HausaNLP at SemEval-2023 task 12: leveraging African low resource TweetData for sentiment analysis. In 4th Workshop on African Natural Language Processing (pp. 50-57). Association for Computational Linguistic: Toronto, Canada
18. Sani, M., Ahmad, A., & Abdulazeez, H. S. (2022). Sentiment analysis of hausa Language Tweet using machine learning approach. Journal of Research in Applied Mathematics, 8(9), 07-16.
19. Shehu, H. A., Majikumna, K. U., Suleiman, A. B., Luka, S., Sharif, M. H., Ramadan, R. A., & Kusetogullari, H. (2024). Unveiling sentiments: a deep dive into sentiment analysis for low-resource languages–a case study on Hausa texts. IEEE Access, 12, 98900-98916.
20. Smith, A., Jones, B., & Wang, X. (2019). Ensemble methods for sentiment analysis: an empirical study. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (pp. 1234-1244).
21. Shiri, F. M., Perumal, T., Mustapha, N., & Mohamed, R. (2023). A comprehensive overview and comparative analysis on deep learning models: CNN, RNN, LSTM, GRU. arXiv preprint arXiv:2305.17473.
22. Tahir, M. F., Haoyong, C., Mehmood, K., Larik, N. A., Khan, A., & Javed, M. S. (2020). Short term load forecasting using bootstrap aggregating based ensemble artificial neural network. Recent Advances in Electrical & Electronic Engineering (Formerly Recent Patents on Electrical & Electronic Engineering), 13(7), 980-992.
23. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. 31st Conference on Neural Information Processing Systems, (pp. 1-11). Long Beach, CA, USA.
24. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... & Rush, A. M. (2020, October). Transformers: State-of-the-art natural language processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (pp. 38-45).
25. Yusuf, M., Ahmad, K., & Bala, A. (2019). Sentiment analysis for Hausa language: challenges and future directions. Journal of Information Technology & Software Engineering, 9(3), 237.