FUDMA Journal of Sciences (FJS)
ISSN online: 2616-1370
ISSN print: 2645 - 2944
Vol. 10 No. 23, December, 2026, pp 180 – 183
DOI: https://doi.org/10.33003/fjs-2026-1023-6523

A Comparative Evaluation of Afribert Base and Afribert Large for Hausa Hate Speech Detection

***Bashir Idris Sulaiman and Muhammad Aminu Ahmad**

Department of Secure Computing, Faculty of Computing, Franco-British International University, Kaduna, Nigeria.

*Corresponding authors' email: bashir.idris@kasu.edu.ng

Received: 11 Aug. 2026
Accepted: 11 Nov. 2026
Published: 13 Nov. 2026

ABSTRACT
Hate speech detection is an important task in Natural Language Processing (NLP), particularly with the growth of online communication. Automatic detection remains challenging for low-resource African languages such as Hausa because annotated datasets and language-specific computational resources are limited. This study comparatively evaluated AfriBERT Base and AfriBERT Large (AfriBERTXL) for Hausa hate speech detection using 6,000 annotated Hausa text samples classified into hate speech and non-hate speech. Both configurations were trained for three epochs and evaluated using accuracy, precision, recall, F1-score and Receiver Operating Characteristic Area Under the Curve (ROC-AUC). On the final test set, AfriBERT Base achieved 79.32% accuracy, 77.98% precision, 66.23% recall, 71.63% F1-score and 87.03% ROC-AUC. AfriBERT Large achieved 80.45% accuracy, 79.94% precision, 67.27% recall, 73.06% F1-score and 88.41% ROC-AUC. Thus, AfriBERT Large recorded higher values across all five reported metrics, although the improvements were modest. The findings provide a direct comparison of the two AfriBERT configurations for Hausa hate speech classification. However, the results should be interpreted in light of the dataset size and the methodological details that require fuller reporting, including dataset sampling, annotation procedures and experimental configuration.

Keywords: Hausa, Hate Speech Detection, AfriBERT, AfriBERT Base, AfriBERT Large, Low-Resource Languages, Text Classification

INTRODUCTION

The rapid growth of social media and other digital communication platforms has increased the volume of user-generated text. Although these platforms support communication and information exchange, they can also facilitate the spread of hateful and abusive expressions. Hate speech detection has therefore become an important research area in Natural Language Processing (NLP), with the aim of identifying harmful language automatically and supporting safer online environments (Fortuna & Nunes, 2018).

Automatic hate speech detection is particularly challenging for low-resource languages. Such languages often have fewer annotated datasets, limited language-processing tools and less representation in large-scale pretrained language models. Hausa, one of the major languages spoken in West Africa, presents challenges associated with linguistic variation, spelling differences, informal expressions and limited task-specific annotated resources (Ogueji et al., 2021; Conneau et al., 2020).

Transformer-based language models have improved text representation and classification by learning contextual relationships between words (Vaswani et al., 2017; Devlin et al., 2019). Multilingual pretrained models can transfer knowledge across languages (Conneau et al., 2020), while models developed with greater attention to African languages may offer useful representations for African-language NLP tasks. AfriBERTa provides pretrained language-model configurations that can be adapted to downstream classification tasks (Ogueji et al., 2021).

Model size and configuration may influence classification performance because larger configurations generally have greater representational capacity. However, increased capacity does not necessarily guarantee better performance, particularly when the available task-specific dataset is limited; multilingual model research has documented trade-offs between model capacity and performance across languages (Conneau et al., 2020). It is therefore useful to compare model configurations under comparable experimental conditions rather than assume that a larger model will always perform better.

Although transformer-based approaches have been applied to text classification, direct experimental comparisons between AfriBERT Base and AfriBERT Large for Hausa hate speech detection remain useful for understanding their relative performance on this task. This study therefore compares the two configurations using the same Hausa dataset and evaluation measures.

This study aimed to comparatively evaluate AfriBERT Base and AfriBERT Large for Hausa hate speech detection using the same annotated dataset and experimental conditions to determine whether the larger model configuration provides improved classification performance.

MATERIALS AND METHODS

Dataset Description

The dataset used in this study consists of 6,509 Hausa text samples classified into two categories: hate speech and non-hate speech. The text data were annotated and preprocessed before being used for model training and evaluation. Both AfriBERT configurations were evaluated using the same dataset and binary classification task.

Table 1: Distribution of Data Sources

Source	Number of instances	Percentage
Twitter	1,302	20.0%
Facebook	651	10.0%
AfriHate	1,953	30.0%
Other publicly available sources	2,603	40.0%
Total	6,509	100.0%

Model Configurations

This study compared two configurations from the AfriBERT model family: AfriBERT Base and AfriBERT Large (referred to in the source study as AfriBERTXL). AfriBERT Base represents the smaller configuration, whereas AfriBERT Large represents the larger configuration. The principal difference investigated in this study is the model configuration. A larger transformer configuration generally provides greater capacity for learning contextual representations from text, which may help the model represent more complex linguistic patterns (Vaswani et al., 2017; Ogueji et al., 2021).

Both configurations were applied to the same binary classification task, with hate speech and non-hate speech as the target classes. The exact pretrained checkpoint identifiers and tokenizer configurations should be reported to make the experiment reproducible.

Table 2: Model Training Configuration

Parameter	Reported Configuration used in this Revision
Task	Binary sequence classification
Number of classes	2
Labels	Hate speech / Non-hate speech
Data split	80% training / 20% test
Maximum sequence length	128 tokens
Learning rate	2×10^{-5}
Per-device batch size	2
Gradient accumulation	8 steps
Effective batch size	16, subject to device and distributed-training setup
Training epochs	3
Weight decay	0.01
Warm-up	500 steps
Learning-rate scheduler	Linear
Loss	Class-weighted cross-entropy
Early stopping	Patience of 2 evaluation epochs
Model-selection metric	F1-score
Evaluation metrics	Accuracy, precision, recall, F1-score, ROC-AUC and confusion matrix

Experimental Procedure

Both AfriBERT configurations were trained and evaluated using the same Hausa hate speech dataset. The models were subjected to the same classification task and evaluated using the same performance measures to provide a consistent basis for comparison. Each model was trained for three epochs, and performance was evaluated after each epoch. Training and validation losses were recorded throughout training. The final performance of each model was subsequently assessed using the test set.

The evaluation measures were accuracy, precision, recall, F1-score and ROC-AUC. Confusion matrices were also generated to examine the numbers of correctly and incorrectly classified hate speech and non-hate speech samples.

Evaluation Metrics

Accuracy measures the proportion of all test instances classified correctly. Precision measures the proportion of instances predicted as hate speech that are correctly classified, while recall measures the proportion of actual hate speech instances identified by the model. The F1-score is the harmonic mean of precision and recall. ROC-AUC measures the model's ability to distinguish between the two classes across classification thresholds. These metrics were used together to provide a broader assessment than accuracy alone.

RESULTS AND DISCUSSION

AfriBERT Base Training and Validation Results

AfriBERT Base showed progressive improvement in classification performance across the three training epochs. The training and validation results are presented in Table 3.

Table 3: AfriBERT Base Training and Validation Results

Epoch	Training Loss	Validation Loss	Accuracy	Precision	Recall	F1-score	ROC-AUC
1	5.314198	0.624487	65.40%	61.58%	32.47%	42.52%	68.02%
2	4.648754	0.521216	74.31%	76.59%	50.13%	60.60%	81.25%
3	3.688423	0.452512	79.84%	76.70%	70.13%	73.27%	85.51%

The final test evaluation of AfriBERT Base produced an accuracy of 79.32%, precision of 77.98%, recall of 66.23%, F1-score of 71.63% and ROC-AUC of 87.03%. The confusion matrix is presented in Table 4.

Table 4: AfriBERT Base Confusion Matrix

Actual class	Predicted Non-Hate	Predicted Hate
Actual Non-Hate	520	72
Actual Hate	130	255

The confusion matrix shows that AfriBERT Base correctly classified 520 non-hate samples and 255 hate speech samples. It incorrectly classified 72 non-hate samples as hate speech and 130 hate speech samples as non-hate. The matrix contains 977 test samples.

AfriBERT Large Training and Validation Results

AfriBERT Large also demonstrated progressive improvement throughout the three training epochs. Its training and validation results are presented in Table 5.

Table 5: AfriBERT Large Training and Validation Results

Epoch	Training Loss	Validation Loss	Accuracy	Precision	Recall	F1-score	ROC-AUC
1	5.336479	0.617000	67.86%	66.36%	37.40%	47.84%	70.05%
2	4.579680	0.498773	77.18%	79.56%	56.62%	66.16%	82.65%
3	3.422859	0.430487	81.37%	80.48%	69.61%	74.65%	86.90%

The final test evaluation of AfriBERT Large produced an accuracy of 80.45%, precision of 79.94%, recall of 67.27%, F1-score of 73.06% and ROC-AUC of 88.41%. The corresponding confusion matrix is presented in Table 6.

Table 6: AfriBERT Large Confusion Matrix

Actual class	Predicted Non-Hate	Predicted Hate
Actual Non-Hate	527	65
Actual Hate	126	259

The confusion matrix shows that AfriBERT Large correctly classified 527 non-hate samples and 259 hate speech samples. It incorrectly classified 65 non-hate samples as hate speech and 126 hate speech samples as non-hate. The matrix contains 977 test samples.

Comparative Test Performance

The final test performance of AfriBERT Base and AfriBERT Large is presented in Table 7.

Table 7: Final Test Performance Comparison

Evaluation Metric	AfriBERT Base	AfriBERT Large	Difference (Percentage Points)
Accuracy	79.32%	80.45%	+1.13
Precision	77.98%	79.94%	+1.96
Recall	66.23%	67.27%	+1.04
F1-score	71.63%	73.06%	+1.43
ROC-AUC	87.03%	88.41%	+1.38

The results show that AfriBERT Large achieved higher performance than AfriBERT Base across all five reported evaluation measures. Accuracy increased from 79.32% to 80.45%, a difference of 1.13 percentage points. Precision increased by 1.96 percentage points, while recall increased by 1.04 percentage points. The F1-score increased from 71.63% to 73.06%, representing a difference of 1.43 percentage points. ROC-AUC increased from 87.03% to 88.41%, a difference of 1.38 percentage points.

The confusion matrices provide further detail about the classification outcomes. AfriBERT Base correctly classified 520 non-hate samples and 255 hate speech samples, while AfriBERT Large correctly classified 527 non-hate samples and 259 hate speech samples. AfriBERT Large therefore produced fewer incorrect classifications in both categories on the reported test set.

Discussion of Model Differences

The primary difference examined in this study is the model configuration. AfriBERT Base represents the smaller configuration, whereas AfriBERT Large (AfriBERTXL) represents the larger configuration. A larger transformer configuration generally provides greater capacity for learning contextual representations from text. This increased capacity may be beneficial when dealing with complex linguistic patterns, although a larger model does not automatically guarantee better performance.

In this experiment, AfriBERT Large achieved higher accuracy, precision, recall, F1-score and ROC-AUC than AfriBERT Base. The improvements were consistent but relatively modest: accuracy increased by 1.13 percentage points and F1-score by 1.43 percentage points. These results suggest that the larger configuration provided a measurable performance advantage under the experimental conditions used in this study.

The findings should nevertheless be interpreted cautiously. The study used one dataset containing 6,509 samples, and the results do not establish that AfriBERT Large will outperform AfriBERT Base across all Hausa text sources or hate speech datasets. Dataset composition, annotation decisions, class distribution and the experimental configuration may influence model performance. Furthermore, no statistical significance test or confidence interval is reported in the source results; therefore, the observed differences should not be described as statistically significant.

CONCLUSION

This study comparatively evaluated AfriBERT Base and AfriBERT Large (AfriBERTXL) for Hausa hate speech detection using a dataset comprising 6,509 annotated Hausa text samples. The experimental results showed that AfriBERT

Large achieved better performance than AfriBERT Base across all five evaluation metrics, recording an accuracy of 80.45%, precision of 79.94%, recall of 67.27%, F1-score of 73.06% and ROC-AUC of 88.41%. Although the improvements were relatively modest, the larger configuration consistently recorded higher values under the experimental conditions adopted. The study is limited by the size and scope of the dataset and the absence of reported external validation and statistical significance testing. Future research should investigate larger and more diverse Hausa datasets, document annotation procedures in greater detail, evaluate additional transformer configurations and test model generalisability on independent data. Overall, this study provides a direct comparison of two AfriBERT configurations and contributes to research on transformer-based NLP for low-resource African languages.

Ethics Statement

The dataset used in this study was prepared for research purposes and was used to investigate automatic detection and classification of hateful and non-hateful Hausa text. The study focuses on text classification at the dataset level and does not seek to identify, profile or target individual users. Where hateful or offensive expressions occur in the dataset, they are treated as linguistic data for the purpose of evaluating hate speech detection models.

REFERENCES

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 8440–8451). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.747

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (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, Volume 1 (Long and Short Papers) (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Fortuna, P., & Nunes, S. (2018). A survey on automatic detection of hate speech in text. ACM Computing Surveys, 51(4), Article 85, 1–30. https://doi.org/10.1145/3232676

Mozafari, M., Farahbakhsh, R., & Crespi, N. (2020). Hate speech detection and racial bias mitigation in social media based on BERT model. PLOS ONE, 15(8), Article e0237861. https://doi.org/10.1371/journal.pone.0237861

Ogueji, K., Zhu, Y., & Lin, J. (2021). 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). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.mrl-1.11

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems (Vol. 30, pp. 5998–6008).

©2026 This is an Open Access article distributed under the terms of the Creative Commons Attribution 4.0 International license viewed via https://creativecommons.org/licenses/by/4.0/ which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is cited appropriately.