FUDMA Journal of Sciences (FJS)
ISSN online: 2616-1370
ISSN print: 2645 - 2944
Vol. 10 No. 23, December, 2026, pp 103 – 112
DOI: https://doi.org/10.33003/fjs-2026-1023-6499

HausaHateData: A Manually Annotated Multi-Source Dataset for Hausa Hate Speech Detection with Benchmark Evaluation Using XLM-RoBERTa-Large, AfriBERTa and mBERT

Bashir Idris Sulaiman and Muhammad Aminu Ahmad

Department of Secure Computing, Faculty of Computing, Kaduna State University, Kaduna, Nigeria.

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

Received: 11 Aug. 2026
Accepted: 06 Nov. 2026
Published: 11 Nov. 2026

ABSTRACT

Automated hate speech detection has advanced considerably for high-resource languages, but the availability of manually annotated resources for Hausa remains limited. This study presents HausaHateData, a multi-source corpus of 6,509 unique Hausa-language text instances collected from Twitter, Facebook, the AfriHate corpus, and other publicly available sources. Each instance was assigned a binary label of hate speech or non-hate speech. Trigger-word annotations were added to hate-speech instances where a relevant word or phrase could be identified, and all instances were assigned a positive or negative sentiment label. Two annotators independently labelled the data, producing an observed agreement of 82.99% and a Cohen's kappa coefficient of approximately 0.65 before disagreements were resolved through discussion. The final corpus contains 2,565 hate-speech instances (39.41%) and 3,944 non-hate-speech instances (60.59%). To establish benchmark results, XLM-RoBERTa-Large, AfriBERTa and multilingual BERT (mBERT) were fine-tuned for binary classification. At the reported final epoch, mBERT achieved the highest accuracy (96.08%) and precision (98.73%), while AfriBERTa achieved the highest recall (94.15%), F1-score (94.89%) and ROC-AUC (0.977). XLM-RoBERTa-Large recorded an accuracy of 94.93% and an F1-score of 93.25%. These results provide initial benchmark evidence for Hausa hate-speech classification and demonstrate the potential value of HausaHateData for further research on low-resource African language processing. The experimental split and confusion-matrix values should be confirmed against the original prediction files before publication.

Keywords: Hausa, Hate speech detection, Low-resource languages, Dataset construction, African languages, XLM-RoBERTa-Large, AfriBERTa, mBERT

INTRODUCTION

Social media platforms have expanded opportunities for communication and public discussion, but they have also enabled the rapid circulation of harmful and hateful content. Hate speech can damage social cohesion and may contribute to discrimination and violence. The automatic identification of such content has therefore become an important task in natural language processing (NLP), with applications in content moderation, online safety and the study of harmful communication (Corazza et al., 2020).

Transformer-based language models have improved text classification across many languages. However, language coverage and the availability of annotated training data remain uneven. African languages are under-represented in many NLP resources, and this limits the development and evaluation of language technologies for their speakers. Hausa is widely used across West and Central Africa, including in education, broadcasting, commerce and online communication. Its digital text includes regional vocabulary, informal spelling, dialectal variation, code-switching and culturally specific expressions. These characteristics make hate-speech interpretation difficult, particularly when meaning depends on context rather than isolated words (Adam et al., 2023).

Existing Hausa-focused resources, including HausaHate and the Hausa-related material available through AfriHate, provide important foundations for research (Vargas et al., 2024; Muhammad et al., 2025). Nevertheless, further work is useful on multi-source curation, consistent binary labels and auxiliary annotations that can support research beyond a single classification label. In particular, trigger-word annotations may help researchers examine the lexical evidence associated with hateful interpretations, while sentiment labels make it possible to investigate the relationship between negative sentiment and hate speech. Negative sentiment alone should not be treated as evidence of hate speech.

This paper introduces HausaHateData, a manually annotated collection of 6,509 unique Hausa-language texts drawn from Twitter, Facebook, the AfriHate corpus and other public sources. Each record contains a hate-speech label, a sentiment label and, where applicable, a trigger-word annotation. The paper also reports benchmark experiments using XLM-RoBERTa-Large, AfriBERTa and mBERT. The contributions are: (1) the construction and description of a multi-source Hausa hate-speech dataset; (2) the inclusion of auxiliary sentiment and trigger-word annotations; and (3) benchmark results from three multilingual transformer models.

Related Work

Hate-speech research depends on datasets that define the task clearly and provide consistent annotations. For English and other high-resource languages, resources such as the dataset of Davidson et al. (2017), OffensEval (Zampieri et al., 2019) and HateXplain (Mathew et al., 2021) have supported the development and comparison of classification systems. Shared tasks, including HASOC, GermEval and TRAC, have extended research to additional languages and annotation settings (Arora et al., 2023). These resources also show that

hate speech, offensive language and abusive language are related but distinct categories, and that task definitions influence reported performance.

Annotation is particularly challenging because the interpretation of harmful language may depend on context, speaker intent, target identity and local norms. Multiple-annotator procedures and explicit disagreement-handling methods are therefore important in subjective classification tasks (Davani et al., 2022; Dehghan et al., 2025). Some datasets provide additional information, such as target-group labels or rationale spans, to support more detailed analysis and explainability (Davani et al., 2023; Mathew et al., 2021; Sawant et al., 2024).

For African languages, initiatives such as Masakhane have encouraged community participation in the development of language technologies (Nekoto et al., 2020). Pretrained models such as AfriBERTa and AfroXLMR have also expanded the availability of representations for African languages (Ogueji, 2022; Alabi et al., 2022). These efforts are important because models trained predominantly on high-resource languages may not adequately represent the vocabulary and linguistic patterns found in African-language text.

Hausa-specific work includes the HausaHate corpus, which provides expert-annotated material for Hausa hate-speech detection (Vargas et al., 2024), and research on offensive online content in Hausa (Adam et al., 2023). The AfriHate collection further contributes multilingual African-language data, including Hausa material (Muhammad et al., 2025). HausaHateData builds on this research by combining several source categories and recording both sentiment and trigger-word information alongside the binary hate-speech label. Its purpose is to provide an additional curated resource and a set of baseline experiments, rather than to claim that earlier Hausa datasets are absent or unusable.

HausaHate Data: Dataset Construction

Dataset Design

HausaHateData was assembled with attention to source diversity, manual annotation, auxiliary labels, reproducibility and ethical curation. Multiple source categories were included to reduce dependence on a single platform. Each instance was independently reviewed by two annotators familiar with Hausa and its cultural context. In addition to the primary binary label, the dataset includes sentiment and trigger-word fields.

Data Sources

The final dataset contains 6,509 instances from four source categories. Table 1 reports the source distribution supplied for the corpus.

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%

Twitter contributed short informal posts, while Facebook contributed comments from public discussions. Hausa-language material from AfriHate was reviewed and harmonised with the other records. The remaining category comprises other publicly accessible material, described in the source manuscript as discussion forums, blogs and similar sources. The specific websites, collection dates and sampling procedure for this category should be documented in the dataset release or accompanying data statement.

Data Collection and Duplicate Removal

The initial collection contained 13,034 text instances. During curation, 2,610 exact duplicates and 3,915 near-identical instances were identified and removed. The resulting corpus contained 6,509 unique texts. These figures reconcile arithmetically: $13{,}034 - 2{,}610 - 3{,}915 = 6{,}509$. The manuscript treats the exact-duplicate and near-identical counts as separate removal categories; the deduplication procedure should therefore state how near-duplicates were identified and confirm that the two categories do not overlap.

The collection process involved identifying public sources, acquiring publicly available text, checking that Hausa was the primary language, removing duplicate and near-identical records, and consolidating the remaining instances for annotation. Multimedia, profile information and unrelated metadata were excluded. Instances with insufficient Hausa content, corrupted text or incomplete records were removed during validation. Duplicate removal was performed before annotation and model evaluation to reduce the risk that repeated text would inflate performance.

Annotation Methodology

Each text was assigned one of two mutually exclusive primary labels:

Hate speech: Text that attacks, demeans or dehumanises an individual or group, promotes hatred or discrimination against a group, encourages exclusion or violence, or uses derogatory stereotypes against a protected or socially identifiable community.

Non-hate speech: Neutral or factual statements, ordinary conversation, supportive language, and criticism or discussion that does not promote hatred or attack a protected group.

The binary scheme was selected to provide a manageable benchmark task. The distinction between hate speech and offensive language should be interpreted according to the annotation guidelines; not every rude, insulting or strongly negative expression is automatically hate speech.

Two auxiliary fields were also assigned. Trigger words record the word or phrase that contributed to a hateful interpretation when an annotator could identify one clearly. The trigger-word field may be empty where no single expression could be isolated. Sentiment records the overall polarity of the text as positive or negative. Sentiment and hate-speech labels are not interchangeable: a negative statement is not necessarily hateful, and a hateful statement may not always be strongly negative in tone.

Inter-Annotator Agreement

Two annotators independently labelled each instance before discussing disagreements. They worked separately during the initial annotation stage. Disagreements were then reviewed

jointly, with attention to context, implicit insults, sarcasm and ambiguous wording, until a consensus label was reached.
Table 2 presents the cross-tabulation of the two independent annotations. The annotators agreed on 5,402 of 6,509 instances, corresponding to an observed agreement of 82.99%. Using the marginal totals in Table 2, Cohen's kappa is approximately 0.654, which is reported as 0.65. This indicates substantial agreement under the interpretation scale of Landis and Koch (1977).

Table 2: Cross-Tabulation of Independent Annotations

	Annotator 1: Hate	**Annotator 1: Non-hate**	**Total**
Annotator 2: Hate	3,124	521	3,645
Annotator 2: Non-hate	586	2,278	2,864
Total	3,710	2,799	6,509

Observed agreement: (3,124 + 2,278) / 6,509 = 0.8299. Cohen's kappa calculated from the table is approximately 0.654. The final dataset labels were determined after consensus discussion.

Data Cleaning and Preprocessing
Text from multiple platforms varied in spelling, formatting and quality. Preprocessing was intended to reduce noise while preserving Hausa expressions and their meaning. The reported steps included removal of duplicate instances, hyperlinks and user mentions; conversion of semantically meaningful emojis and emoticons into text; conservative normalisation of whitespace, line breaks, punctuation and encoding; removal of isolated non-linguistic characters; and normalisation or removal of numerical and date/time expressions where they did not contribute semantic information. Language verification and a final completeness check were also performed.
The exact preprocessing implementation, including software versions and any regular expressions or normalisation rules, should be made available with the dataset or code to support reproducibility.

Dataset Schema and Class Distribution
Table 3 Summarises the Principal Fields Described for Each Record. Trigger-word Values may be Empty when no Distinct Trigger Expression was Identified

Table 3: Summary of Dataset Fields

Field	**Data type**	**Description**
ID	Integer	Unique record identifier
Text	String	Original Hausa text instance
Class label	Categorical	Hate speech or non-hate speech
Trigger word	String	Word or phrase associated with the hateful interpretation, where identifiable; otherwise, empty
Sentiment	Categorical	Positive or negative

Table 4: Distribution of Hate-Speech and Non-Hate-Speech Classes

Class	**Number of instances**	**Percentage**
Hate speech	2,565	39.41%
Non-hate speech	3,944	60.59%
Total	6,509	100.00%

The dataset is moderately imbalanced, with non-hate speech representing the majority class. A classifier that always predicts the majority class would therefore achieve an accuracy of 60.59%, not 94%. This majority-class baseline is the appropriate reference point when interpreting the reported model accuracies.

Ethical Considerations
The dataset was constructed from publicly available textual content and publicly accessible corpora. The source manuscript states that private communications and restricted communities were not accessed. Personally identifiable information, user mentions, hyperlinks and unnecessary metadata were removed where applicable. Before public release, the authors should confirm the terms of service and licensing conditions of each source, describe any institutional ethics review or exemption that applies, and ensure that the released data do not expose personal information. Public availability of a post does not by itself remove privacy, platform-policy or copyright considerations.

Evaluation Method
Model Selection
Three pretrained transformer models were selected for the benchmark:
XLM-RoBERTa-Large: A large multilingual transformer pretrained on CommonCrawl-derived data across approximately 100 languages. It provides a strong multilingual baseline for transfer to languages with comparatively limited labelled data.
mBERT: A multilingual BERT model pretrained on Wikipedia text from 104 languages. It serves as a widely used multilingual baseline.
AfriBERTa: A pretrained language model developed for African languages. It provides a comparison with models trained on broader multilingual corpora and allows the study to examine the performance of an African-language-focused model on Hausa text.

Experimental Design
The manuscript reports an 80:20 training-test split. The dataset should be divided using stratification so that the class proportions are maintained across the partitions. The test partition should remain separate from model selection and

hyperparameter decisions. The original methods section also mentions stratified five-fold cross-validation within the training partition, whereas the original Table 5 describes a stratified train-test split without specifying cross-validation. This revised description therefore reports the 80:20 split but does not claim that five-fold cross-validation was completed. The authors must confirm the actual experiment logs and update this subsection if five-fold cross-validation was in fact performed.

Training Configuration

The three models were fine-tuned for binary sequence classification using the Hugging Face Transformers training framework. The narrative description in the source manuscript reports a maximum sequence length of 128 tokens, three training epochs, a learning rate of 2×10^{-5}, a per-device batch size of 2, gradient accumulation over 8 steps, weight decay of 0.01, 500 warm-up steps and a linear learning-rate scheduler. Class-weighted cross-entropy loss was used to account for the class distribution, and early stopping with a patience of two evaluation epochs was reported. F1-score was used as the model-selection metric.

Table 5 presents the configuration used in this revised description. The original manuscript's Table 5 contains conflicting values (batch size 4, gradient accumulation 4 and warm-up ratio 0.10). These values have not been combined with the narrative settings because they describe a different configuration. The authors must check the saved training arguments or scripts and retain only the values actually used for each model.

Table 5: 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

For full reproducibility, the final manuscript should also report the exact model identifiers/checkpoints, random seed, software and library versions, class-weight calculation, tokenizer settings, hardware, and whether each model used the same split and stopping rule.

Evaluation Metrics

The models were assessed using accuracy, precision, recall, F1-score and the area under the receiver operating characteristic curve (ROC-AUC). A confusion matrix was used to distinguish true-positive, true-negative, false-positive and false-negative predictions. Precision measures the proportion of predicted positive instances that are positive; recall measures the proportion of actual positive instances correctly identified; and the F1-score is the harmonic mean of precision and recall. ROC-AUC measures discrimination across classification thresholds. The positive class should be stated explicitly in the final results tables; this manuscript treats hate speech as the positive class.

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad (1)$$

$$Precision = \frac{TP}{(TP+FP)} \quad (2)$$

$$Recall = \frac{TP}{(TP+FN)} \quad (3)$$

$$F1 = 2 \times \frac{(Precision \cdot Recall)}{(Precision+Recall)} \quad (4)$$

RESULTS AND DISCUSSION

Annotation Agreement

The independent annotation cross-tabulation is presented in Table 2. The annotators agreed on 5,402 instances, giving an observed agreement of 82.99%. Cohen's kappa, calculated from the displayed marginal totals, is approximately 0.654. The result indicates substantial agreement, although it does not mean that every ambiguous instance was labelled without difficulty. The consensus discussion was therefore an important part of finalising the dataset.

Table 6: Annotation Agreement (Annotator 1 vs Annotator 2)

	Hate (Annotator 1)	Non-Hate (Annotator 1)	Total
Hate (Annotator 2)	3,124	521	3,645
Non-Hate (Annotator 2)	586	2,278	2864
Total	3,710	2,799	6,509

This corresponds to an observed agreement of

$$P_o = \frac{5{,}402}{6{,}509} = 0.83$$

The expected agreement by chance was calculated as

$$P_e = \left(\frac{3645}{6509} \times \frac{3710}{6509}\right) + \left(\frac{2864}{6509} \times \frac{2799}{6509}\right) = 0.51$$

Thus, the Cohen's Kappa coefficient is

$$K = \frac{P_o - P_e}{1 - P_e} = \frac{0.83 - 0.51}{1 - 0.51} = 0.65$$

According to the interpretation proposed by Landis and Koch (1977), a Kappa value of 0.65 represents substantial agreement (see Table 2), indicating that the annotation process produced reliable and consistent labels suitable for supervised learning.

Training and Evaluation Loss

The original manuscript includes training- and evaluation-loss plots for the three models. It reports that XLM-RoBERTa-Large evaluation loss decreased across the three epochs (0.699, 0.469 and 0.244). For AfriBERTa, the reported evaluation loss reached 0.162 at epoch 2 and increased to 0.198 at epoch 3. For mBERT, the reported evaluation loss increased between epochs 1 and 2 before reaching 0.253 at epoch 3. These values are retained as reported.

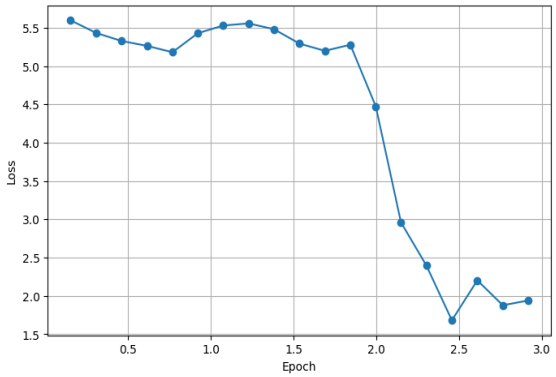


Figure 1: XLM-RoBERTa-Large Training Loss

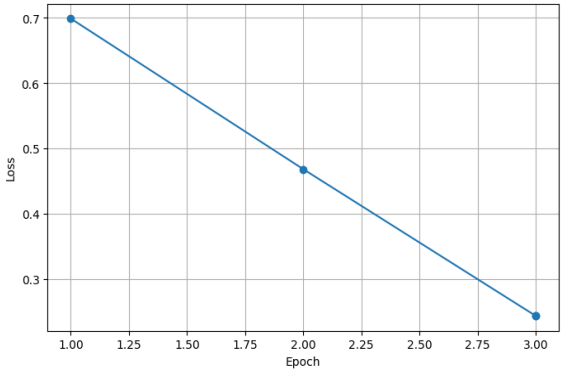


Figure 2: XLM-RoBERTa-Large Validation Loss

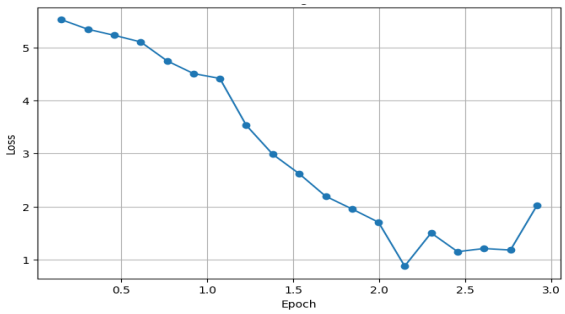


Figure 3: AfriBERTa Training Loss

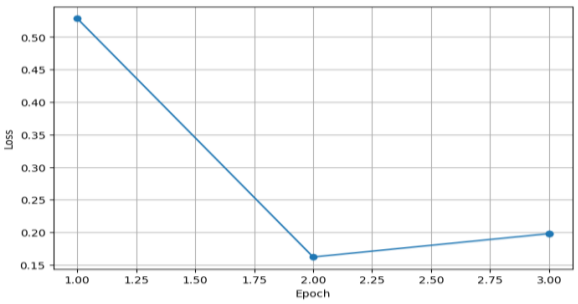


Figure 4: AfriBERTa Validation Loss

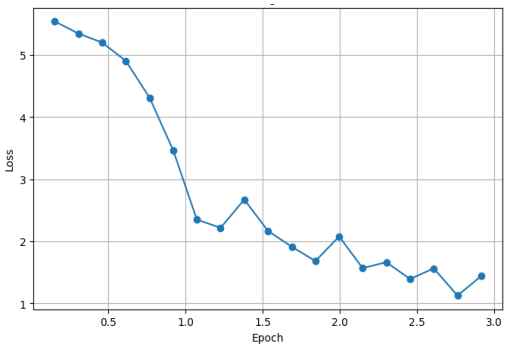


Figure 5: Mbert Training Loss

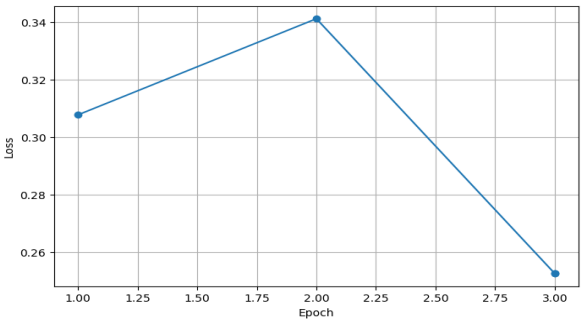


Figure 6: mBERT Validation Loss

The source manuscript sometimes refers to these values as test loss and elsewhere describes a separate validation procedure. To avoid implying that the held-out test set was repeatedly used during training, the plots and values should be labelled as validation loss if they were calculated on the validation partition. If they were calculated on the test set after every epoch, the authors should rerun the evaluation or clearly disclose that limitation. The final test set should ideally be used once after model selection.

Classification Performance

Table 6 summarises the reported class-wise precision, recall and F1-score. All three models performed better on the non-hate class in terms of recall than on the hate-speech class. AfriBERTa showed the most balanced hate-speech precision and recall among the three models, while XLM-RoBERTa-Large and mBERT had higher precision than recall for the hate-speech class.

Table 7: Class-wise Classification Performance of the Models

Model	Hate precision	Hate recall	Hate F1	Non-hate precision	Non-hate recall	Non-hate F1
XLM-RoBERTa-Large	0.98	0.89	0.93	0.93	0.99	0.96
AfriBERTa	0.96	0.94	0.95	0.96	0.97	0.97
mBERT	0.99	0.91	0.95	0.95	0.99	0.97

The reported epoch-wise performance is shown in Tables 8–10. These values are reproduced from the original manuscript and should be checked against the saved evaluation outputs before submission.

Table 8: XLM-RoBERTa-Large Performance by Epoch

Epoch	Accuracy	Precision	Recall	F1-score
1	0.568356	0.321168	0.085770	0.135385
2	0.813364	0.837500	0.653021	0.733844
3	0.949309	0.980645	0.888889	0.932515

Table 9: XLM-RoBERTa-Large Performance by Epoch

Epoch	Accuracy	Precision	Recall	F1-score
1	0.722734	0.611437	0.812865	0.697908
2	0.943932	0.948980	0.906433	0.927218
3	0.960061	0.956436	0.941520	0.948919

Table 10: AfriBERTa Performance by Epoch

Epoch	Accuracy	Precision	Recall	F1-score
1	0.893241	0.934884	0.783626	0.852598
2	0.940860	0.957983	0.888889	0.922144
3	0.960829	0.987342	0.912281	0.948328

At the reported final epoch, XLM-RoBERTa-Large achieved an accuracy of 94.93%, precision of 98.06%, recall of 88.89% and F1-score of 93.25%. AfriBERTa achieved an accuracy of 96.01%, precision of 95.64%, recall of 94.15% and F1-score of 94.89%. mBERT achieved an accuracy of 96.08%, precision of 98.73%, recall of 91.23% and F1-score of 94.83%. Thus, mBERT recorded the highest reported accuracy and precision, while AfriBERTa recorded the highest reported recall and F1-score.

The reported ROC-AUC values were 0.966 for XLM-RoBERTa-Large, 0.977 for AfriBERTa and 0.969 for mBERT. AfriBERTa therefore recorded the highest reported ROC-AUC. These results suggest that all three models learned useful patterns for the binary classification task, while their relative strengths differed across metrics. In particular, the higher recall reported for AfriBERTa indicates that it identified a larger proportion of hate-speech examples, whereas mBERT's higher precision indicates that a larger proportion of its hate-speech predictions were correct.

Confusion Matrices and Error Analysis

The confusion matrices for the three models are shown in Figures 7 to 9. The results providing summarises the correctly and incorrectly classified hate speech and non-hate speech instances. XLM-Roberta-Large correctly classified 1236 instances (780 non-hate and 456 hate) while misclassifying 67 instances (9 non-hate speech as hate speech and 57 hate speech as non-hate). AfriBERT correctly classified 1250 instances (768 non-hate and 483 hate) while misclassifying 52 instances (22 non-hate speech as hate speech and 30 hate speech as non-hate). mBERT correctly classified 1255 instances (783 non-hate and 468 hate) while misclassifying 51 instances (6 non-hate as hate speech and 45 hate speech and non-hate).

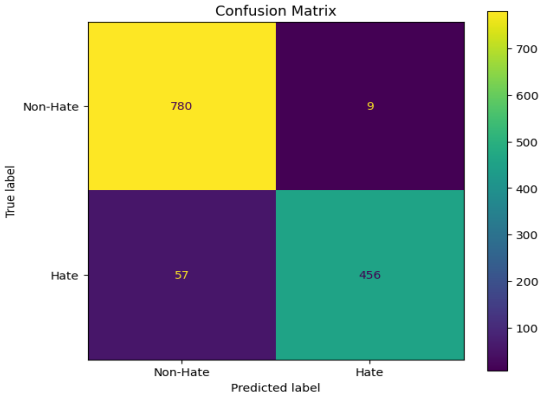


Figure 7: XLM-RoBERTa-Large Confusion Matrix

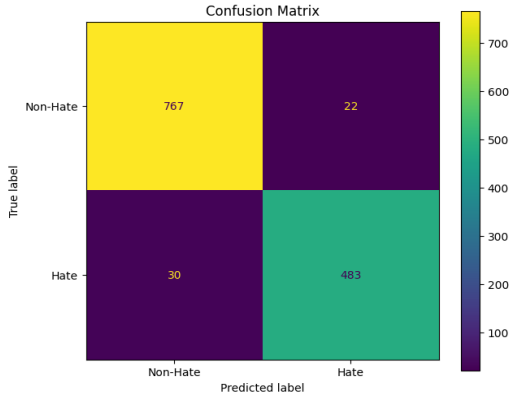


Figure 8: AfriBERTa Confusion Matrix

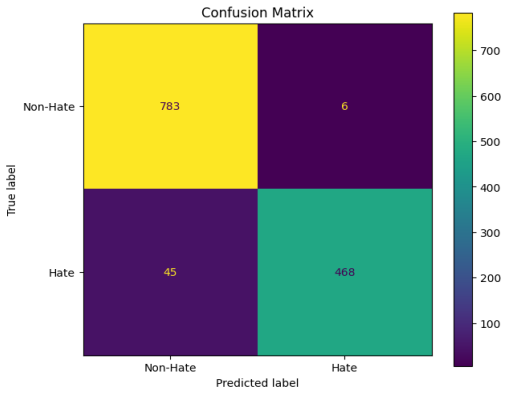


Figure 9: mBERT Confusion Matrix

The misclassified instances revealed that many remaining errors involved implicit hate speech, sarcasm, irony, culturally specific expressions, code-switched Hausa-English sentences, limited conversational context, and ambiguous lexical usage. These observations highlight the inherent difficulty of hate speech detection in natural language, where contextual interpretation often extends beyond the literal meaning of individual words.

The Receiver Operating Characteristic (ROC) curves for the three models are presented in Figures 10 to 12. AfriBERTa's curve sits closest to the upper-left corner of the ROC space, consistent with its highest Area Under the Curve (AUC) value of 0.977, followed by mBERT (0.969) and XLM-RoBERTa-Large (0.966). All three values indicate strong discrimination between hate speech and non-hate speech across a range of classification thresholds, with AfriBERTa achieving the highest ROC-AUC. AfriBERTa is the only model that was exclusively pretrained on Hausa and related African languages, utilising well under 1 GB of text, in contrast to XLM-RoBERTa-Large (pretrained on about 100 languages) and mBERT (pretrained on 104 languages). Thus, AfriBERTa (the smallest of the three language-targeted models) achieved its best generalisation, along with the lowest test loss and highest ROC-AUC with a fraction of the pretraining data. This suggests that language-specific pretraining offers a significant advantage on this dataset. This is significant, particularly in light of the fact that mBERT produced the highest raw classification scores due to its bigger general-purpose pretraining corpus.

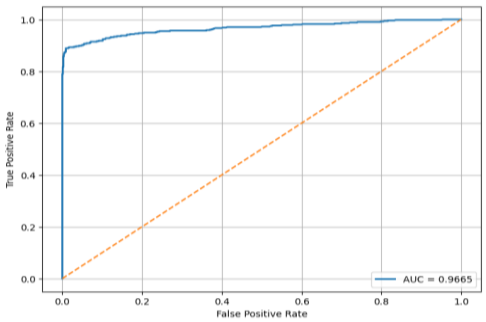


Figure 10: XLM-RoBERTa-Large ROC Curve

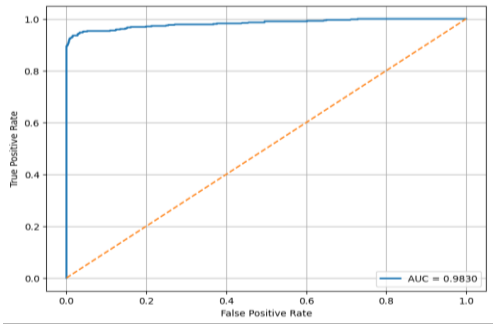


Figure 11: AfriBERTa ROC Curve

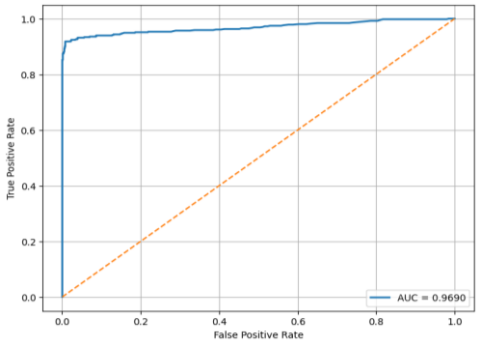


Figure 12: mBERT ROC Curve

The experiments demonstrate that HausaHateData provides sufficient annotation quality and linguistic diversity for training multilingual transformer models, with all three architectures reaching accuracy above 94% and F1-scores above 92% on the test set. mBERT has the highest accuracy and precision values, AfriBERTa and XLM-RoBERTa-Large has intermediary and lowest values, respectively, on those same metrics. AfriBERTa has the lowest test loss and highest recall, F1-score, ROC-AUC values, demonstrating its generalization of unseen data. This spread of results confirms that the dataset is discriminating enough to separate model performance rather than saturating uniformly across architectures.

Class Imbalance and Baseline

Hate-speech instances account for 39.41% of the corpus, while non-hate instances account for 60.59%. Consequently, a majority-class-only classifier would achieve 60.59% accuracy. The reported final-epoch accuracies of 94.93% to 96.08% are higher than this baseline. Accuracy alone is not sufficient for an imbalanced classification task, so precision, recall, F1-score and ROC-AUC are also reported. The recall values for the hate-speech class are particularly important because false negatives represent hateful instances that the classifier fails to identify.

Discussion

Contribution of HausaHateData to Existing Hausa Hate-Speech Research

The development of HausaHateData contributes to ongoing efforts to address the limited availability of annotated datasets for hate-speech detection in African languages. The present study assembled 6,509 unique Hausa-language text instances from Twitter, Facebook, AfriHate and other publicly available sources. Of these, 2,565 instances (39.41%) were classified as hate speech, while 3,944 (60.59%) were classified as non-hate speech. The integration of multiple data sources and the inclusion of sentiment and trigger-word annotations provide a foundation for examining harmful language beyond conventional binary classification.

These findings complement the work of Vargas et al. (2024), who developed HausaHate, an expert-annotated corpus for Hausa hate-speech detection. While HausaHate established an important resource for studying harmful expressions in Hausa, the present study extends this research direction by combining multiple source categories and incorporating auxiliary annotations. Similarly, Muhammad et al. (2025) introduced AfriHate, a multilingual collection addressing hate speech and abusive language across African languages. HausaHateData builds upon this broader effort by concentrating on Hausa-language material and evaluating three transformer architectures using a common binary classification task.

The present study also relates to Adam et al. (2023), who investigated offensive online content in Hausa. Both studies recognise the importance of language-specific resources for identifying harmful communication. However, the distinction between offensive language and hate speech remains important. As Davidson et al. (2017) demonstrated, offensive expressions do not necessarily constitute hate speech. Consequently, the binary annotation criteria used in HausaHateData should be interpreted carefully, particularly where insults, political criticism or strongly negative statements do not target protected or socially identifiable groups.

Annotation Reliability and Comparison with Previous Studies

The two annotators achieved an observed agreement of 82.99% and a Cohen's kappa coefficient of approximately 0.65. According to the conventional interpretation proposed by Landis and Koch (1977), this represents substantial agreement. The result suggests that the annotation guidelines supported a reasonably consistent distinction between hate speech and non-hate speech, although disagreement remained for 1,107 instances before consensus resolution.

This finding is consistent with the concerns raised by Davani et al. (2022), who emphasised that disagreement in subjective annotation tasks may reflect differences in interpretation rather than simple annotation errors. Hate speech is particularly sensitive to linguistic context, cultural knowledge, speaker intent and the identity of the targeted group. Therefore, the disagreement observed in the present study may reflect the difficulties of interpreting sarcasm, implicit hostility and culturally specific Hausa expressions.

Mathew et al. (2021), through the HateXplain dataset, also demonstrated the value of supplementing classification labels with explanations of the textual evidence underlying hateful interpretations. The trigger-word annotations included in HausaHateData represent a related approach, although they are not equivalent to fully annotated explanatory rationales. Their usefulness will depend on annotation completeness and consistency, which should be evaluated separately.

These comparisons suggest that HausaHateData provides a useful annotation framework, while also highlighting the importance of documenting difficult cases and preserving information about annotator disagreement.

Comparative Performance of Transformer Models

The benchmark evaluation demonstrated that all three transformer models achieved high classification performance. At the reported final epoch, mBERT recorded the highest accuracy (96.08%) and precision (98.73%), while AfriBERTa achieved the highest recall (94.15%), F1-score (94.89%) and ROC-AUC (0.977). XLM-RoBERTa-Large achieved an accuracy of 94.93% and an F1-score of 93.25%.

These results are consistent with the broader literature demonstrating the effectiveness of pretrained transformer models for multilingual text classification. Corazza et al. (2020) highlighted the importance of multilingual evaluation in hate-speech detection, particularly where language-specific data resources are limited. The present findings provide additional evidence that multilingual representations can support Hausa hate-speech classification after task-specific fine-tuning.

The competitive performance of AfriBERTa is particularly noteworthy when considered alongside Ogueji (2022), who investigated the development of multilingual language models for low-resource African languages. AfriBERTa's relatively high recall and F1-score in the present study are compatible with the proposition that African-language-focused pretraining can provide useful linguistic representations for downstream classification tasks.

Similarly, Alabi et al. (2022) demonstrated the relevance of adapting pretrained language models to African languages. The present findings support the broader importance of African-language representation and adaptation. However, the results do not establish that African-language-focused pretraining is universally superior to large multilingual pretraining, since mBERT achieved slightly higher accuracy and precision.

The difference between mBERT and AfriBERTa also illustrates the importance of selecting evaluation metrics according to the intended application. In automated content moderation, higher precision may reduce the likelihood of incorrectly flagging legitimate communication, whereas higher recall may reduce the proportion of hateful messages that remain undetected. Consequently, the selection of a preferred model should consider the relative consequences of false-positive and false-negative predictions rather than accuracy alone.

Classification Errors and Linguistic Challenges

The confusion matrices indicate differences in the types of classification errors produced by the three models. In particular, AfriBERTa recorded fewer false negatives for hate speech than mBERT, whereas mBERT produced fewer false

positives. This pattern is consistent with their respective strengths in recall and precision.

The manuscript identifies implicit hate speech, sarcasm, culturally specific expressions, Hausa–English code-switching and ambiguous vocabulary as possible sources of misclassification. These difficulties closely resemble the findings of Adam et al. (2023), whose investigation of offensive Hausa content highlighted the importance of idiomatic expressions, local linguistic conventions and contextual interpretation. Their work demonstrates why recognising harmful content in Hausa cannot depend exclusively on identifying offensive words.

The findings also relate to Mathew et al. (2021), who developed HateXplain to support explainable hate-speech detection. The trigger-word annotations in HausaHateData could provide a starting point for examining model decisions, although further evaluation would be required to establish whether the identified words adequately explain each classification.

Future research should therefore consider contextual annotation, error analysis by linguistic category and explainability methods to determine whether models rely on meaningful contextual information or superficial lexical associations.

Implications of Class Imbalance and Model Evaluation

The class distribution indicates a moderate imbalance, with 60.59% non-hate speech and 39.41% hate speech. A majority-class classifier would consequently achieve an accuracy of 60.59%. The reported accuracies of 94.93–96.08% substantially exceed this simple baseline, indicating that the trained models captured information beyond the majority-class distribution.

Nevertheless, the evaluation should not rely exclusively on accuracy. This is consistent with the broader concerns raised by Davidson et al. (2017) regarding the difficulty of distinguishing hate speech from related forms of offensive communication. The relatively high F1-scores suggest promising classification performance, but their interpretation depends on the reliability of the annotations and the independence of the evaluation data.

An additional concern is that 30% of HausaHateData originated from AfriHate. Because the study combines previously available corpus material with newly collected texts, careful attention should be given to potential overlap between training and evaluation instances, including paraphrased or near-identical expressions. The present results should also be tested against an independent Hausa dataset, such as HausaHate, after harmonising the label definitions.

Taken together, the findings demonstrate that HausaHateData has potential as an additional resource for developing and evaluating Hausa hate-speech classifiers. The study complements the contributions of Vargas et al. (2024) and Muhammad et al. (2025) by introducing a multi-source corpus with auxiliary sentiment and trigger-word annotations. Its evaluation of XLM-RoBERTa-Large, AfriBERTa and mBERT also provides an opportunity to compare African-language-focused and broader multilingual transformer architectures.

The strong performance of AfriBERTa is compatible with earlier findings on the value of African-language pretraining, particularly those reported by Ogueji et al. (2021) and Alabi et al. (2022). However, the slightly higher accuracy and precision of mBERT demonstrate that the benefits of language-focused pretraining depend on the evaluation metric and task conditions.

CONCLUSION

This study presented HausaHateData, a manually annotated multi-source dataset containing 6,509 unique Hausa-language texts collected from Twitter, Facebook, the AfriHate corpus and other publicly available sources. Each record was assigned a binary hate-speech label, with sentiment labels and trigger-word annotations included as auxiliary information. The two annotators achieved 82.99% observed agreement and a Cohen's kappa of approximately 0.65 before consensus resolution. The final dataset contains 2,565 hate-speech instances and 3,944 non-hate-speech instances. Three multilingual transformer models were evaluated. At the reported final epoch, mBERT achieved the highest accuracy (96.08%) and precision (98.73%), while AfriBERTa achieved the highest recall (94.15%), F1-score (94.89%) and ROC-AUC (0.977). XLM-RoBERTa-Large achieved an accuracy of 94.93% and an F1-score of 93.25%. The results provide initial benchmark values for Hausa hate-speech classification and indicate that an African-language-focused model can perform competitively with broader multilingual models on this task. The study has limitations. The dataset is moderate in size, the class labels are binary, and hate-speech annotation is inherently context-sensitive. The collection draws on selected public sources and may contain source-specific or temporal biases. The present evaluation does not establish performance on independent external datasets or unseen platforms. Future work should expand the corpus, document the collection and annotation protocols in greater detail, introduce fine-grained categories such as ethnic or religious hate, threats and harassment, evaluate the auxiliary annotations, and test models on external datasets. The authors should also release the dataset and evaluation code where source licences and privacy requirements permit.

ACKNOWLEDGEMENTS

The authors thank the annotators for their work in labelling HausaHateData and for the linguistic and cultural judgement required by the task. The authors also acknowledge the Hugging Face Transformers ecosystem and the creators of the AfriHate corpus, whose tools and data contributed to this research.

REFERENCES

Adam, F. M., Zandam, A. Y., & Inuwa-Dutse, I. (2023). Detection and analysis of offensive online content in Hausa language. arXiv preprint arXiv:2311.10541.

Alabi, J., Adelani, D. I., Mosbach, M., & Klakow, D. (2022). Adapting pre-trained language models to African languages via multilingual adaptive fine-tuning. In Proceedings of the 29th International Conference on Computational Linguistics (pp. 4336–4349).

Arora, A., Nakov, P., Hardalov, M., Sarwar, S. M., Nayak, V., Dinkov, Y., ... & Augenstein, I. (2023). Detecting harmful content on online platforms: What platforms need vs. where research efforts go. ACM Computing Surveys, 56(3), 1–17. https://doi.org/10.1145/360339

Bäumler, J., Blöcher, L., Frey, L. J., Chen, X., Bayer, M., & Reuter, C. (2026). A survey of machine learning models and datasets for the multi-label classification of textual hate speech in English. ACM Transactions on Knowledge Discovery from Data, 20(6), 1–39. https://doi.org/10.1145/382077

Corazza, M., Menini, S., Cabrio, E., Tonelli, S., & Villata, S. (2020). A multilingual evaluation for online hate speech detection. ACM Transactions on Internet Technology, 20(2), 1–22.

Davani, A. M., Díaz, M., & Prabhakaran, V. (2022). Dealing with disagreements: Looking beyond the majority vote in subjective annotations. Transactions of the Association for Computational Linguistics, 10, 92–110. https://doi.org/10.1162/tacl_a_00449

Davani, A. M., Atari, M., Kennedy, B., & Dehghani, M. (2023). Hate speech classifiers learn normative social stereotypes. Transactions of the Association for Computational Linguistics, 11, 300–319. https://doi.org/10.1162/tacl_a_00550

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. Proceedings of the International AAAI Conference on Web and Social Media, 11(1), 512–515.

Dehghan, S., Umut Sen, M., & Yanikoglu, B. (2025). Dealing with annotator disagreement in hate speech classification. arXiv preprint arXiv:2502.08266. https://doi.org/10.48550/arXiv.2502.08266

Hershcovich, D., Frank, S., Lent, H., De Lhoneux, M., Abdou, M., Brandl, S., ... & Søgaard, A. (2022). Challenges and strategies in cross-cultural NLP. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 6997–7013). https://doi.org/10.18653/v1/2022.acl-long.482

Katyshev, A., Anikin, A., & Zubankov, A. (2023). Bidirectional transformers as a means of efficient building of knowledge bases: A case study with XLM-RoBERTa. In Novel & Intelligent Digital Systems Conferences (pp. 292–297). Springer Nature Switzerland.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. Biometrics, 33(1), 159–174.

Li, Y., Zhang, H., & Chen, X. (2026). Large language models for abusive language detection: Advances and challenges. IEEE Access, 14, 56789–56805.

Maity, K., Ghosh, N., Jain, R., Saha, S., & Bhattacharyya, P. (2025). StereoHate: Toward identifying stereotypical bias and target group in hate speech detection. Natural Language Processing, 31(2), 415–434. https://doi.org/10.1017/nlp.2024.29

Mathew, B., Saha, P., Yimam, S. M., Biemann, C., Goyal, P., & Mukherjee, A. (2021). HateXplain: A benchmark dataset for explainable hate speech detection. In Proceedings of the AAAI Conference on Artificial Intelligence, 35(17), 14867–14875.

Muhammad, S. H., Abdulmumin, I., Ayele, A. A., Adelani, D. I., Ahmad, I. S., Aliyu, S. M., ... & Ousidhoum, N. (2025). AfriHate: A multilingual collection of hate speech and abusive language datasets for African languages. In Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers) (pp. 1854–1871). https://doi.org/10.48550/arXiv.2501.08284

Muhammad, F. A., Hassan, S. M., & Inuwa-Dutse, I. (2025). Dataset creation and baseline models for sexism detection in Hausa. arXiv preprint arXiv:2510.27038. https://doi.org/10.48550/arXiv.2510.27038

Nekoto, W., Marivate, V., Matsila, T., Fasubaa, T., Fagbohungbe, T., Akinola, S. O., ... & Bashir, A. (2020). Participatory research for low-resourced machine translation: A case study in African languages. In Findings of the Association for Computational Linguistics: EMNLP 2020 (pp. 2144–2160).

Ogueji, K. (2022). AfriBERTa: Towards viable multilingual language models for low-resource languages. http://hdl.handle.net/10012/18662

Saleh, H., Alhothali, A., & Moria, K. (2023). Detection of hate speech using BERT and hate speech word embedding with deep model. Applied Artificial Intelligence, 37(1). https://doi.org/10.1080/08839514.2023.2166719

Sawant, M., Younus, A., Caton, S., & Qureshi, M. A. (2024). Using explainable AI (XAI) for identification of subjectivity in hate speech annotations for low-resource languages. In Proceedings of the 4th International Workshop on Open Challenges in Online Social Networks (pp. 10–17). https://doi.org/10.1145/3677117.3685006

Vargas, F., Guimarães, S., Muhammad, S. H., Alves, D., Ahmad, I. S. I., Abdulmumin, I., ... & Benevenuto, F. (2024). HausaHate: An expert annotated corpus for Hausa hate speech detection. In Proceedings of the 8th Workshop on Online Abuse and Harms (pp. 52–58).

Yu, Z., Sen, I., Assenmacher, D., Samory, M., Fröhling, L., Dahn, C., ... & Wagner, C. (2025). The unseen targets of hate: A systematic review of hateful communication datasets. Social Science Computer Review, 43(5), 1114–1144.

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., & Kumar, R. (2019). SemEval-2019 Task 6: Identifying and categorizing offensive language in social media (OffensEval). In Proceedings of the 13th International Workshop on Semantic Evaluation (pp. 75–86).

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

Vargas, F., Guimarães, S., Muhammad, S. H., Alves, D., Ahmad, I. S., Abdulmumin, I., Mohamed, D., Pardo, T., & Benevenuto, F. (2024). HausaHate: An expert annotated corpus for Hausa hate speech detection. In Proceedings of the 8th Workshop on Online Abuse and Harms (pp. 52–58). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.woah-1.5

©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.