English-Hausa Cyberbullying Detection on Facebook-Oriented Text Using Multilingual Transformers and Cross-Lingual Transfer

Dawud Muhammad Aisha[1], Umar Iliyasu[2], Rosemary Mrumun Dima[3]
[1,2,3]*Department of Computer Science, Faculty of Computing, Federal University Dutsin-Ma, Katsina State, Nigeria*

***Abstract*—Cyberbullying detection remains dominated by high-resource languages, while task-specific Hausa resources and controlled cross-lingual evidence remain limited. This study evaluates a Facebook-oriented English-Hausa benchmark for five message-level categories: none, insult, threat, identity-based and relational. The 10,000-message corpus contains 5,000 English texts (2,522 TRAC-1 Facebook records and 2,478 synthetic supplements) and an entirely synthetic 5,000-message Hausa component selected from a quality-controlled candidate pool. Labels are AI-assisted rather than independently established human gold labels. A TF-IDF with Logistic Regression baseline was compared with XLM-RoBERTa, DistilBERT, AfriBERTa and a language-aware ensemble on fixed 70/15/15 partitions. XLM-RoBERTa obtained 0.8010 macro-F1 on the bilingual test set, with 0.8198 for English and 0.7812 for Hausa. English-to-Hausa zero-shot macro-F1 was 0.3839, increasing to 0.5236 with 100 Hausa training examples and 0.7812 under full bilingual supervision. These single-run results support controlled benchmark analysis, not validated real-world deployment; authentic Hausa testing, independent human validation and statistical replication remain necessary.**

***Index Terms*—cyberbullying detection; English-Hausa; XLM-RoBERTa; low-resource NLP; cross-lingual transfer; synthetic data**

I. INTRODUCTION

Nigeria's sustained growth in digital connectivity has expanded communication, education, commerce and community participation, while also enlarging the space in which harmful interpersonal behaviour can occur. National Bureau of Statistics data recorded 163,838,439 active Internet subscriptions in the fourth quarter of 2023 (National Bureau of Statistics [NBS], 2024). Conventional cyberbullying definitions emphasise intentional aggression, repetition and power imbalance (Smith et al., 2008), although digital repetition may arise when a single message remains visible or is repeatedly shared (Slonje et al., 2013). In this study, however, the model operates on an individual message. The label cyberbullying is therefore an operational message-level proxy for language consistent with cyberbullying, not evidence that repetition, intent, power imbalance or the victim-target relationship has been independently established. The social significance of harmful online behaviour is well established. A UNICEF U-Report poll reported online-bullying experiences among one in three participating young people across 30 countries, while Nigerian evidence has documented substantial levels of cyberbullying perpetration (Owoade et al., 2023). Reviews associate cybervictimisation with depression, anxiety, reduced self-esteem, emotional distress and suicidal behaviour, while cautioning that much of the evidence is correlational rather than causal (Kowalski et al., 2014; Nixon, 2014). These harms create a need for scalable screening tools, but automated classifiers should support rather than replace contextual human judgement.

Automated harmful-content detection has progressed from hand-crafted lexical features and conventional classifiers to deep contextual representations. Transformer architectures are prominent because self-attention captures contextual relationships that simple bag-of-words features may miss. XLM-RoBERTa extends this capability across many languages through multilingual pretraining (Conneau et al., 2020). Nevertheless, multilingual representations do not establish that an English-trained model will reliably distinguish fine-grained harmful-content categories in

Hausa, where informal orthography, loanwords, code-switching and culture-specific expressions can influence meaning (Adam et al., 2024).

Hausa is used across a wide geographical and sociolinguistic continuum in northern Nigeria and neighbouring countries. Online Nigerian Hausa frequently contains non-standard spellings, omitted diacritics, Arabic-derived vocabulary, English lexical borrowing and code-switching; Nigerian Pidgin may also appear in mixed discourse. The present corpus retained naturally occurring orthographic variation in the authentic English source and intentionally varied synthetic Hausa style, but it did not attach dialect or regional-variety labels. Consequently, the study cannot claim equal coverage of Kano, Katsina, Sokoto or other regional varieties, nor can it quantify performance on English-Hausa-Pidgin mixtures as distinct subgroups.

Hausa-specific offensive-language resources and broader datasets for Hausa, Yoruba and Igbo are emerging (Adam et al., 2024; Aliyu et al., 2024), yet offensive-language detection is not equivalent to a five-class task that distinguishes insult, threat, identity-based attack and relational aggression from non-bullying text. Maikano (2024) provides rare Hausa cyberbullying evidence in a binary, Twitter-focused conventional machine-learning setting. Elsewhere, transformer-based cyberbullying studies have reported strong results for English Facebook and resource-constrained Bengali data (Khafajeh, 2024; Sihab-Us-Sakib et al., 2024), while cross-lingual harmful-content research demonstrates that transfer can be useful but remains language- and task-dependent (Liu et al., 2023).

Data scarcity creates a second methodological challenge. Public English datasets do not necessarily provide balanced coverage of all cyberbullying phenomena, and a sufficiently large task-specific Hausa Facebook corpus was not available for the present taxonomy. Synthetic-data research shows that generated examples can support controlled model development, but synthetic resources can also introduce stylistic artefacts, source-label confounding and generator-specific shortcuts; they should not be equated with naturally occurring, independently human-labelled discourse (Cuzzocrea et al., 2025).

The novelty claim is therefore deliberately narrow. The study does not claim to be the first cyberbullying system for Hausa or the first multilingual harmful-content model. Its contribution is a controlled, fine-grained English-Hausa Facebook-oriented benchmark protocol with explicit authentic-versus-synthetic provenance, same-partition comparison of classical, multilingual and language-specialist models, and a direct comparison of zero-shot, fixed 100-example few-shot and fully supervised English-to-Hausa transfer. The empirical results are interpreted as internal benchmark performance rather than evidence of validated real-world Hausa moderation.

The General Aggression Model, Routine Activity Theory and Social Cognitive Theory provide interpretive foundations for the study (Anderson & Bushman, 2002; Bandura, 2001; Cohen & Felson, 1979). These theories are not measured causal variables in the machine-learning experiments. They motivate attention to contextual language, position automated screening as one possible element of digital guardianship and caution against reducing harmfulness to isolated toxic words.

II. RELATED WORK AND RESEARCH GAP

Recent large-scale reviews show that cyberbullying detection research remains fragmented across platforms, label definitions, languages and evaluation protocols. Philipo et al. (2026) synthesise traditional machine learning, deep learning and large-language-model approaches and identify continuing challenges around dataset quality, multilingual coverage, annotation consistency and comparability. The present benchmark is positioned within that evidence base as a controlled English-Hausa five-class experiment rather than a broad claim of field-level novelty.

African-language NLP has expanded through Hausa offensive-content work and multilingual Nigerian-language datasets (Adam et al., 2024; Aliyu et al., 2024), while multilingual pretraining provides shared representations that can support low-resource transfer (Conneau et al., 2020). Nevertheless, language coverage alone does not resolve domain-specific label semantics. Cyberbullying categories such as relational aggression require social-context interpretation that differs from generic toxicity or hate-speech labels.

Dataset construction is equally important. Human annotation provides a label set; clear decision rules and disagreement analysis are needed to make categorical judgements interpretable (Artstein & Poesio, 2008). Data statements and provenance

documentation reduce the risk that model performance reflects undocumented sampling or collection choices (Bender & Friedman, 2018). Fairness research further shows that harmful-content systems can reproduce linguistic and social biases, particularly where identity terms or dialectal features correlate spuriously with abusive labels (Blodgett et al., 2020; Sap et al., 2019). In synthetic corpora, these concerns extend to generator artefacts and contamination across train and test partitions. Exact-duplicate control is necessary but does not by itself establish distributional equivalence to authentic discourse.

Table 1. Structured comparison of representative related studies

Study	Language	Platform	Dataset	Classes	Model	Authentic/Synthetic	Main Result	Limitation
Kumar et al. (2018)	English/Hindi	Facebook, Twitter	TRAC-1 aggression corpus	NAG/CAG/OAG	Shared-task baselines	Authentic	Established a benchmark for aggression identification	Aggression labels differ from the present five-class cyberbullying proxy
Khafajeh (2024)	English	Facebook	Facebook comments	Cyberbullying classes	BERT/deep learning	Authentic	Shows strong value of deep contextual models	Single-language study; no Hausa transfer
fSihab-Us-Sakib et al. (2024)	Bengali	Social media	Resource-constrained Bengali corpus	Cyberbullying labels	XLM-R/mBERT/BanglaBERT	Authentic	Transformer models effective in a low-resource setting	Different language, platform mix and taxonomy
Adam et al. (2024)	Hausa	Online/social media	Hausa offensive-content data	Offensive/threatening content	Detection models	Authentic	Provides Hausa harmful-content evidence	Not a five-class cyberbullying benchmark
Aliyu et al. (2024)	Hausa/Yoruba/Igbo	Social media	Multilingual Nigerian dataset	Offensive/hate speech	Transformer evaluation	Authentic	Broadens Nigerian-language resources	Task and label schema differ from cyberbullying categories
Cuzzocrea et al. (2025)	Multilingual focus	Social media	Synthetic cyberbullying data	Cyberbullying	LSTM-autoencoder	Synthetic	Demonstrates utility of synthetic data for model development	External authenticity and generator artefacts remain concerns

Philipo et al. (2026)	Multiple	Multiple	Systematic review	Multiple	ML/DL/LLM	Mixed	Synthesises datasets, models and language challenges	Review only; does not provide English-Hausa empirical evidence
Present study	English/Hausa	Facebook-oriented	10,000-message benchmark	5 message-level classes	TF-IDF/LR, XLM-R, DistilBERT, AfriBERTa	Hybrid English; fully synthetic Hausa	Controlled supervision comparison and same-partition benchmarking	Synthetic/AI-assisted labels; no authentic Hausa external test

Table 1 makes the gap explicit: prior work provides important Hausa offensive-language resources, low-resource transformer evidence and synthetic-data precedents, but not the same combination of a five-class English-Hausa Facebook-oriented task, explicit source provenance and controlled zero-/few-/full-supervision comparison. At the same time, the table also shows why the present results should not be interpreted as directly superior to prior studies whose platforms, labels and data-generation regimes differ.

III. MATERIALS AND METHODS

The study adopted a Design Science Research orientation because its central output was a computational artefact: a reproducible dataset-and-model pipeline evaluated through controlled experiments (Hevner et al., 2004). The unit of analysis was an individual text message rather than a human respondent. The workflow comprised candidate acquisition and generation, screening, annotation, deterministic sampling and partitioning, model development, validation-based checkpoint selection and final test evaluation.

Task Definition and Taxonomy

The five labels are mutually exclusive message-level content categories. They should not be read as complete behavioural diagnoses of cyberbullying because the model does not observe interaction history, repeated exposure, power imbalance or offline relationships. Severity is a descriptive annotation variable: 0 is reserved for none, while 1, 2 and 3 denote mild, moderate and severe harmful content respectively. Severity was not modelled as an ordinal or separate prediction target in the reported experiments; ordinal severity modelling is a future extension.

Table 2. Operational class definitions and representative examples

Class	Operational definition	Representative example	Severity
none	No direct interpersonal bullying in the message; may include disagreement, criticism or neutral discussion.	"I disagree with that argument."	0
insult	Direct degrading, ridiculing or humiliating attack on a person's character, appearance, intelligence or ability.	"You are completely useless at this."	1–3
threat	Intimidation that expresses or implies intended physical, psychological or social harm or retaliation.	"You will regret what you did."	1–3
identity-based	Hostility in which ethnicity, religion, gender, nationality, region or another identity characteristic is central to the attack.	"People from your group are all inferior."	1–3
relational	Exclusion, gossip, reputation damage or manipulation intended to harm social belonging or relationships.	"We will make sure nobody includes you."	1–3

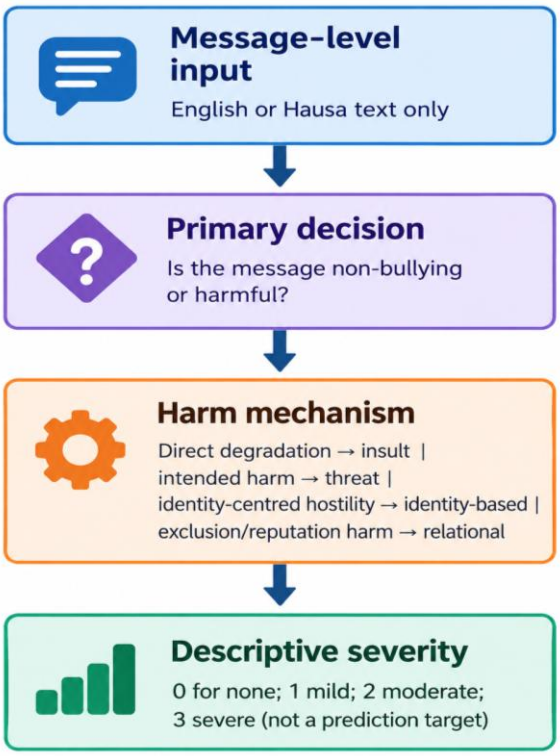

Figure 1. Operational taxonomy used for the five-class message-level task.

Dataset Construction and Annotation

The final corpus contained 10,000 messages, evenly divided between English and Hausa, with exactly 1,000 examples per class per language. The authentic English source was the TRAC-1 English Facebook Aggression corpus (Kumar et al., 2018). After source-specific screening, 14,457 candidate records were retained. Original TRAC-1 labels (NAG, CAG and OAG) were preserved as provenance metadata but were not treated as final labels because the present study used a different five-class taxonomy.

TRAC-1 supplied sufficient records for none and insult but insufficient examples for threat, identity-based and relational categories. A second Facebook-tagged candidate source was screened and excluded because diversity analysis revealed severe repetition around a small number of harmful textual cores. Instead, 4,200 synthetic English Facebook-style candidates were generated for underrepresented classes and re-annotated without exposing the intended generation family to the label decision. The final English component contained 2,522 TRAC-1-derived records and 2,478 synthetic records. Thus, 49.56% of the English benchmark is synthetic; importantly, all English relational examples and large proportions of the threat and identity-based classes are synthetic, creating a potential source-label confound that is discussed explicitly below.

Because a sufficiently large cyberbullying-specific Hausa Facebook corpus matching the study taxonomy was unavailable, the Hausa component was created through staged synthetic candidate generation. The initial pool contained 6,500 messages. After duplicate and near-duplicate review, 29 provisionally redundant items were removed. Targeted enrichment and top-up stages were used when post-annotation audits showed that threat, identity-based and relational categories remained below the required count. The final candidate pool contained 10,749 unique records, from which deterministic sampling with seed 42 selected exactly 1,000 records per class. The complete Hausa benchmark is therefore synthetic Facebook-style text, not authentic Facebook user content.

Synthetic-Data Generation, Filtering and Reproducibility Audit

Synthetic generation was treated as candidate creation rather than ground-truth labelling. Generation scenarios were not accepted automatically as final classes; candidates were subsequently subjected to independent text-based annotation. Filtering included missing-text checks, unique identifier checks, normalised exact-duplicate detection, near-duplicate review and manual rejection of provisionally redundant items. Records containing direct personal identifiers were not intentionally generated, and the synthetic texts were used as research artefacts rather than representations of named Facebook users.

A revision-stage audit of the available project script record identified gemini-3.1-flash-lite as the model used in the documented Hausa enrichment. The enrichment prompt required distinct, primarily Nigerian Hausa Facebook-style texts; it allowed limited natural Hausa-English code-switching, prohibited real names, contact details, graphic violence and actionable harm instructions, and treated scenario families as generation situations rather than final labels. The exact prompt excerpts are reproduced in Appendix A. However, verifiable revision hashes, temperature, top-p, random seed and complete run metadata were not retained for every generation stage in the submitted research record. The initial 6,500 Hausa-generation and 4,200 English-generation

model/version records likewise cannot be reconstructed reliably from the submitted manuscript. These missing provenance fields are disclosed rather than inferred retrospectively.

Annotation Workflow, Human Oversight and Label Validity

English candidates received a single independent annotation pass. TRAC-1 source labels remained provenance metadata only, and the generation scenario of a synthetic message was not automatically accepted as its project label.

Hausa candidates received two independent annotation passes followed by adjudication of label or severity disagreements. Across the major Hausa generation stages, exact label agreement ranged from 89.74% to 94.17%, while Cohen's kappa ranged from 0.8337 to 0.8846. These values quantify consistency between two AI-assisted passes; they are not human inter-annotator reliability estimates and should not be interpreted as evidence of human gold-standard validity.

Independent human validation of a representative English-Hausa subset was requested during peer review but was not available in the submitted experimental record. No human Cohen's kappa or Krippendorff's alpha is therefore reported, and the paper does not attribute the frozen labels to human experts. This distinction is methodologically important because inter-coder agreement assesses the reliability of a labelling process, not the substantive validity of the label construct itself (Artstein & Poesio, 2008). The appropriate next validation step is blinded annotation by another two qualified English and native-Hausa annotators, followed by pre-specified disagreement coding and adjudication.

Preprocessing and Quality Control

Preprocessing was deliberately conservative. Original case, punctuation, emojis, spelling variation, informal orthography and code-switching were retained; no global lowercasing was applied to stored modelling text. Quality control included unique identifier checks and normalised exact-duplicate checks. The frozen corpus contained 10,000 unique identifiers and zero normalised exact duplicate texts. These controls reduce direct leakage, but they do not prove absence of semantically similar templates or generator-specific lexical signatures.

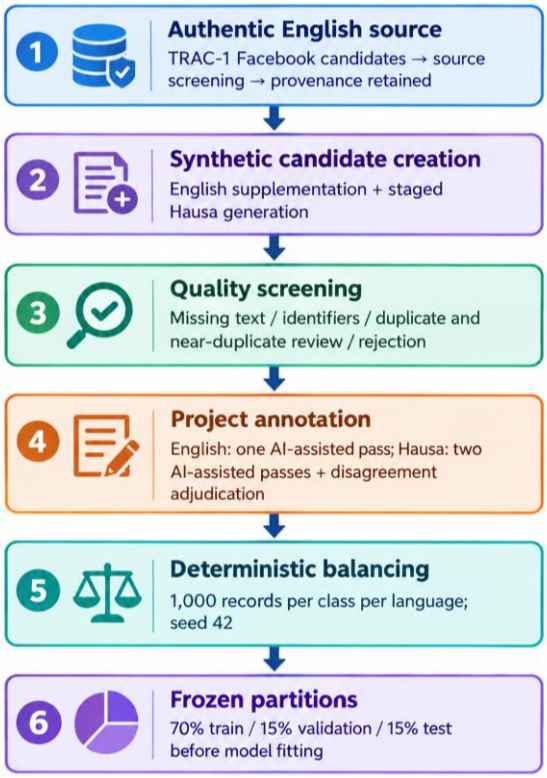

Figure 2. Data-flow from candidate acquisition and synthetic generation to the frozen benchmark partitions.

Table 3. Final bilingual dataset composition

Class	English: TRAC-1	English: synthetic	English total	Hausa total	Combined
None	1,000	0	1,000	1,000	2,000
Insult	1,000	0	1,000	1,000	2,000
Threat	198	802	1,000	1,000	2,000
Identity-based	324	676	1,000	1,000	2,000
Relational	0	1,000	1,000	1,000	2,000
Total	2,522	2,478	5,000	5,000	10,000

Data Partitioning, Models and Evaluation

The corpus was frozen before model training and divided deterministically with seed 42 into 70% training, 15% validation and 15% test partitions. Each language-class cell contributed 700 training, 150 validation and 150 test examples, giving 7,000 training

records and 1,500 records each for validation and test. No class weights were used because every class was exactly balanced within each language and split. Seed 42 governed corpus partitioning, Logistic Regression and the fixed few-shot sample. The archived run record does not establish that every stochastic transformer operation was globally seeded in every framework component; the reported transformer values are therefore treated as single-run point estimates.

The classical benchmark represented text with TF-IDF word unigrams and bigrams using sublinear term frequency and a maximum vocabulary of 50,000 features. Logistic Regression used a maximum of 2,000 iterations and random seed 42.

FacebookAI/xlm-roberta-base was the principal bilingual transformer and used the checkpoint-associated tokenizer. The complete fine-tuned model contained approximately 278.05 million trainable parameters, including the pretrained encoder and classification head. Training used AdamW optimisation, a learning rate of 2×10^{-5}, per-device batch size 8, gradient accumulation of 2 (effective batch size 16), 10% warm-up, FP16 mixed precision, maximum sequence length 128 and up to four epochs. Validation macro-F1 determined checkpoint selection. The 128-token limit was selected as a memory-efficient operating point because tokenizer diagnostics showed that only 58 of 7,000 bilingual training records (0.83%) reached the truncation ceiling; 6,942 records did not. A longer-sequence sensitivity analysis was not performed.

Two language-specialist transformers were evaluated. distilbert-base-uncased, with its checkpoint-associated tokenizer and approximately 66.96 million trainable parameters, was trained on the fixed 3,500-record English training partition and selected on the 750-record English validation set. castorini/afriberta_base, with its checkpoint-associated tokenizer and approximately 111.46 million trainable parameters, was trained on the corresponding 3,500-record Hausa training partition and selected on the 750-record Hausa validation set. After both specialist models were frozen, a language-aware ensemble routed English test messages to DistilBERT and Hausa messages to AfriBERTa; no test-set weighting, probability blending or post-hoc tuning was used.

Approximate raw FP32 parameter storage is 1.11 GB for XLM-RoBERTa, 0.27 GB for DistilBERT and 0.45 GB for AfriBERTa. These values are derived from four bytes per parameter and are not measurements of peak training or inference memory, which also depends on activations, optimizer states, batch size, sequence length and framework overhead. Peak GPU memory, inference latency, wall-clock training duration, GPU model and exact Python/PyTorch/Transformers package versions were not preserved in the submitted run metadata and are therefore not retrospectively fabricated. This omission is part of the reproducibility limitation of the current study.

Table 4. Transformer configuration and reproducibility information

Model	Checkpoint	Tokenizer	Trainable params	Approx. FP32 weights	Reported training settings	Scope
XLM-RoBERTa	FacebookAI/xlm-roberta-base	Checkpoint tokenizer	278.05M	1.11 GB	2×10^{-5}; AdamW; effective batch 16; ≤4 epochs; max 128	Bilingual
DistilBERT	distilbert-base-uncased	Checkpoint tokenizer	66.96M	0.27 GB	Same fixed data protocol; specialist English training	English
AfriBERTa	castorini/afriberta_base	Checkpoint tokenizer	111.46M	0.45 GB	Same fixed data protocol; specialist Hausa training	Hausa

Macro-F1 was the principal model-selection and evaluation measure because it gives equal weight to all five classes. Accuracy, macro-precision, macro-recall, class-wise F1, language-specific performance, confusion patterns and descriptive post-hoc error counts were also used. Validation results are reported separately from final test results, and the final test partition was not used to select checkpoints or tune models after error inspection.



Figure 3. System architecture and evaluation pathway.

Prediction-level outputs needed to compute reliable bootstrap macro-F1 confidence intervals, full per-class precision/recall tables and separate English/Hausa confusion matrices were not available in the submitted manuscript artefacts; these statistics are therefore not reconstructed from aggregate values. Likewise, the main transformer systems were not re-run across multiple random seeds. All performance differences should consequently be interpreted descriptively rather than as statistically established superiority.

English-to-Hausa transfer was evaluated under three XLM-R conditions. Zero-shot training used the fixed 3,500 English training records and no Hausa training data, then evaluated on the unchanged 750-record Hausa test set. The few-shot condition added exactly 100 Hausa records sampled without replacement from the fixed Hausa training partition using seed 42, with 20 examples per class. Full bilingual training used all 3,500 English and 3,500 Hausa training records. Hausa validation and test partitions remained unchanged. Only one fixed 100-example subset was used; no 25/50/250/500/1,000-example learning curve, alternative 100-example sample or variance analysis was run. The selected 100 examples may therefore not represent the distribution of all possible few-shot samples.

Ethical and Data-Governance Considerations

No direct scraping of Facebook profiles or direct collection of personal user data was undertaken. Authentic English messages were obtained from an existing research corpus; the synthetic English supplements and the entire Hausa component were generated without direct personal identifiers. The system was evaluated as a research artefact rather than an autonomous moderation or disciplinary authority. Any future operational use would require authentic external validation, independent human label validation, human oversight, privacy-preserving deployment, bias assessment and a review or appeals mechanism for consequential errors.

IV. RESULTS AND DISCUSSION

Dataset Quality, Source Composition and Baseline Performance

The completed corpus achieved exact language and class balance: 5,000 English and 5,000 Hausa messages, with 1,000 examples of none, insult, threat, identity-based and relational content in each language. The English component contained 2,522 TRAC-1-derived records (50.44%) and 2,478 synthetic supplements (49.56%). TRAC-1 supplied all none and insult examples, 198 threat examples and 324 identity-based examples; synthetic supplementation provided the remaining 802 threat, 676 identity-based and all 1,000 relational examples. The Hausa component was fully synthetic and selected from the 10,749-record quality-controlled candidate pool.

Severity was retained descriptively rather than modelled as a separate target. The combined dataset contained 2,000 severity-0 records, 3,689 mild, 3,592 moderate and 719 severe records. The final corpus had zero normalised exact duplicate texts and was frozen before the deterministic 70/15/15 split.

The source composition itself is a validity warning. In English, relational content is entirely synthetic, while 80.2% of threat and 67.6% of identity-based examples are synthetic. A model could therefore learn source- or generator-associated cues that correlate with these classes. The submitted experiment did not include a synthetic-versus-authentic discrimination classifier or

a generator-leakage test, so distributional equivalence cannot be claimed. Exact-duplicate screening reduces direct leakage but does not rule out subtler template or lexical artefacts.

The TF-IDF with Logistic Regression baseline achieved 0.7567 accuracy and 0.7577 macro-F1 on the bilingual test set. English macro-F1 was 0.7577 and Hausa macro-F1 was 0.7552. Class-wise F1 was lowest for none (0.6777) and insult (0.6983), and stronger for threat (0.8206), identity-based (0.7762) and relational aggression (0.8156). This establishes a substantial lexical benchmark while exposing difficulty around the semantic boundary between ordinary negative discourse and direct personal abuse.

Bilingual and Specialist Transformer Performance

XLM-RoBERTa validation macro-F1 improved across all four epochs and reached 0.7893 at epoch 4, which was retained because macro-F1 was the pre-specified validation selection criterion. On the frozen bilingual test set, XLM-R obtained 0.8013 accuracy and 0.8010 macro-F1, an observed 0.0433 absolute increase over the classical baseline. Because the transformer experiment is a single run without confidence intervals or significance testing, this difference is reported as an observed benchmark gain rather than an inferential claim of superiority.

Language-specific test macro-F1 was 0.8198 for English and 0.7812 for Hausa, an observed 3.86-percentage-point gap. Class-wise bilingual F1 values were 0.7400 for none, 0.7174 for insult, 0.8479 for threat, 0.8514 for identity-based and 0.8481 for relational aggression. Hausa exceeded English on insult (0.7291 versus 0.7055) and identity-based attacks (0.8638 versus 0.8385), whereas English was much stronger on relational aggression (0.9272 versus 0.7677).

DistilBERT achieved 0.8161 macro-F1 on the English test set, 0.0037 below the XLM-R English score of 0.8198. AfriBERTa achieved 0.7814 macro-F1 on Hausa, only 0.0002 above XLM-R Hausa performance. These are descriptive near-ties; without paired prediction-level significance tests or multi-seed replication, neither difference supports a claim that one system is statistically better. The language-aware ensemble achieved 0.7986 macro-F1 on the bilingual test set, 0.0024 below XLM-R; again, this margin should be interpreted as nominal rather than conclusive.

Table 5. Comparative performance on the frozen test partitions (single-run point estimates)

System / test set	Accuracy	Macro-F1	Trainable parameters
TF-IDF + Logistic Regression / bilingual	0.7567	0.7577	—
XLM-RoBERTa / bilingual	0.8013	0.8010	278.05M
XLM-RoBERTa / English	0.8200	0.8198	278.05M
DistilBERT / English	0.8173	0.8161	66.96M
XLM-RoBERTa / Hausa	0.7827	0.7812	278.05M
AfriBERTa / Hausa	0.7827	0.7814	111.46M
Language-aware ensemble / bilingual	0.8000	0.7986	Two specialist models

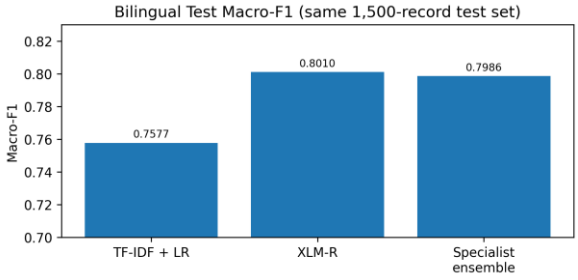

Figure 4. Macro-F1 comparison for systems evaluated on the same bilingual test set. Values are single-run point estimates.

Cross-Lingual Transfer

Zero-shot English-to-Hausa transfer was limited. XLM-R trained only on 3,500 English examples achieved 0.4267 accuracy and 0.3839 macro-F1 on the unchanged Hausa test set. Relational aggression transferred particularly poorly, with F1 of 0.0387, whereas threat was the strongest zero-shot class at 0.5822. Adding exactly 100 balanced Hausa training examples raised Hausa accuracy to 0.5293 and macro-F1 to 0.5236, an observed 13.97-percentage-point increase over zero-shot transfer. Relational F1 increased from 0.0387 to 0.5315; identity-based F1 increased from 0.4495 to 0.5804; and threat F1 increased from 0.5822 to 0.6579. Full bilingual XLM-R reached 0.7812 macro-F1. The pattern indicates that target-language supervision is useful, but a single fixed 100-example subset cannot establish the full annotation-efficiency curve or its variance.

Table 6. Cross-lingual and Hausa-supervised performance

Experiment	English train	Hausa train	Hausa test accuracy	Hausa macro-F1
Zero-shot XLM-R	3,500	0	0.4267	0.3839
Few-shot XLM-R	3,500	100	0.5293	0.5236
Full bilingual XLM-R	3,500	3,500	0.7827	0.7812
AfriBERTa Hausa specialist	0	3,500	0.7827	0.7814

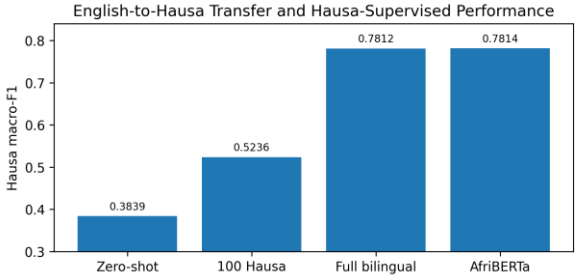

Figure 5. Hausa test macro-F1 under zero-shot, fixed 100-example few-shot and full-supervision conditions, with AfriBERTa shown for reference.

Error Analysis and Interpretation

Post-hoc analysis of XLM-R found 1,202 correct and 298 incorrect predictions among 1,500 test records, corresponding to an overall error rate of 19.87%. The English error rate was 18.00% and the Hausa error rate 21.73%. Insult had the highest overall class error rate (29.33%), followed by none (26.00%). The two largest confusion directions were insult predicted as none (42 cases) and none predicted as insult (35 cases).

The language-by-class analysis showed that English insult had the highest English error rate (31.33%), while English relational aggression had the lowest (6.67%). For Hausa, none was most difficult (32.67%), followed by insult (27.33%) and relational aggression (24.00%).

Hausa threat and identity-based attacks were comparatively strong, with error rates of 11.33% and 13.33%, respectively. The Hausa performance gap was therefore concentrated in particular discourse types rather than uniformly distributed across all classes.

Table 7. XLM-RoBERTa error rates by true class

True class	Total	Correct	Errors	Error rate
None	300	222	78	26.00%
Insult	300	212	88	29.33%
Threat	300	262	38	12.67%
Identity-based	300	252	48	16.00%
Relational	300	254	46	15.33%

The reviewer requested verbatim misclassified examples. The aggregate manuscript artefacts available for this revision do not contain prediction-level text linked to errors, so no test example is invented or retrospectively labelled as a model error. The operational examples in Table 2 instead illustrate the principal decision boundaries. In particular, the none-versus-insult confusion is consistent with the difference between criticism of an idea ("That argument is poor") and direct degradation of a person ("You are useless"). Future evaluation should preserve prediction-level records specifically for auditable qualitative error analysis.

Robustness, Code-Switching, Fairness and Sociolinguistic Interpretation

The retained text includes spelling variation, emojis, informal orthography and some code-switching, but the study did not construct dedicated perturbation test sets for misspellings, deliberate character substitutions, slang, emojis, English loanwords or English-Hausa code-switching. It therefore cannot quantify robustness to those phenomena. A stronger follow-up protocol should stratify naturally occurring and controlled perturbation sets and report performance changes relative to the unperturbed test set.

Fairness likewise cannot be inferred from the identity-based class alone. The corpus does not contain validated demographic attributes of authors or targets, regional-variety labels or sufficient subgroup metadata for demographic parity, equalised error-rate or dialect-fairness analysis. The identity-based F1 values describe one content category, not fairness across protected groups. This distinction is important because harmful-content systems may over-associate identity terms or dialectal markers with abusive labels (Blodgett et al., 2020; Sap et al., 2019).

The sociolinguistic limitation is particularly relevant for Hausa. Nigerian online Hausa can vary by region, education, orthographic convention, borrowing

patterns and interactional context. The synthetic design attempted stylistic variety and retained code-switching where it occurred, but no dialect-balanced sampling frame or region-specific external test set was available. Claims should therefore be restricted to the constructed benchmark rather than generalised to all Hausa-speaking Facebook communities.

V. LIMITATIONS

The study acknowledges that the use of synthetic data, particularly for the Hausa benchmark, presents an opportunity for further validation using authentic social-media data. The English dataset also combines synthetic and authentic examples, reflecting a practical approach to developing and evaluating multilingual cyberbullying detection models. The reported transformer results represent performance under the evaluated experimental settings, while the transfer-learning experiments provide an initial assessment of how training-data availability influences cross-lingual performance.

Further research could extend the framework through evaluations involving code-switching, dialectal variation, robustness and fairness, as well as broader interaction-level indicators of cyberbullying, including repetition, intent and power imbalance. Additional experiments involving repeated training runs, expanded data-scaling settings and detailed computational profiling could also strengthen statistical confidence and reproducibility. These extensions would enhance the framework's generalisability and practical applicability across diverse Nigerian digital environments.

VI. CONCLUSION

Within the fixed constructed benchmark, XLM-RoBERTa produced the highest observed single-model bilingual macro-F1 (0.8010), compared with 0.7577 for TF-IDF with Logistic Regression and 0.7986 for the language-aware specialist ensemble. DistilBERT was close to XLM-R on English, and AfriBERTa was effectively tied with XLM-R on Hausa. Because these are single-run estimates without inferential uncertainty, the small differences between transformer systems should not be interpreted as statistically established superiority. The clearest empirical pattern is the dependence of Hausa performance on target-language supervision. English-only XLM-R obtained 0.3839 macro-F1 on the Hausa test set; a fixed 100-example Hausa sample increased this to 0.5236; and full bilingual supervision reached 0.7812. This supports the limited claim that multilingual pretraining can facilitate adaptation while direct task-specific Hausa supervision remains important in this five-class benchmark. The principal contribution is therefore a controlled benchmark-construction and evaluation protocol for fine-grained English-Hausa Facebook-oriented text, not a production-ready moderation system. The synthetic Hausa corpus, hybrid English source, AI-assisted labels, absence of an authentic Hausa external test set and lack of multi-seed statistical evaluation substantially constrain external validity. Independent native-Hausa human validation, ethically obtained authentic Hausa testing, source-distribution analysis, richer few-shot curves, robustness and code-switching evaluation, fairness analysis and reproducible compute logging are required before operational or population-level claims can be justified.

ACKNOWLEDGEMENT

The author acknowledges Dr Umar Iliyasu for supervision, guidance and constructive feedback during the dissertation research.

CONFLICT OF INTEREST STATEMENT

The authors declare no competing interests.

FUNDING ACKNOWLEDGEMENT

No specific external research funding was received.

REFERENCES

[1] F. M. Adam, A. Y. Zandam, and I. Inuwa-Dutse, "Detection and analysis of offensive online content in Hausa language," in *Proc. 2nd IJCAI AI for Good Symp. Africa*, 2024, pp. 2–14, doi: 10.24963/ijcai.aai4g.2024/1.

[2] S. M. Aliyu, G. M. Wajiga, and M. Murtala, "A multilingual dataset for offensive language and hate speech detection for Hausa, Yoruba and Igbo languages," *Nigerian Journal of Computing, Engineering and Technology*, vol. 3, no. 1, pp. 217–227, 2024. [Online]. Available: https://arxiv.org/abs/2406.02169

[3] C. A. Anderson and B. J. Bushman, "Human aggression," *Annual Review of Psychology*, vol. 53, pp. 27–51, 2002, doi: 10.1146/annurev.psych.53.100901.135231.

[4] R. Artstein and M. Poesio, "Inter-coder agreement for computational linguistics," *Computational Linguistics*, vol. 34, no. 4, pp. 555–596, 2008, doi: 10.1162/coli.07-034-R2.

[5] A. Bandura, "Social cognitive theory: An agentic perspective," *Annual Review of Psychology*, vol. 52, pp. 1–26, 2001, doi: 10.1146/annurev.psych.52.1.1.

[6] E. M. Bender and B. Friedman, "Data statements for natural language processing: Toward mitigating system bias and enabling better science," Transactions of the Association for Computational Linguistics, vol. 6, pp. 587–604, 2018, doi: 10.1162/tacl_a_00041.

[7] S. L. Blodgett, S. Barocas, H. Daumé III, and H. Wallach, "Language (technology) is power: A critical survey of 'bias' in NLP," in Proc. 58th Annu. Meeting Assoc. Comput. Linguistics, 2020, pp. 5454–5476, doi: 10.18653/v1/2020.acl-main.485.

[8] L. E. Cohen and M. Felson, "Social change and crime rate trends: A routine activity approach," American Sociological Review, vol. 44, no. 4, pp. 588–608, 1979, doi: 10.2307/2094589.

[9] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," in Proc. 58th Annu. Meeting Assoc. Comput. Linguistics, 2020, pp. 8440–8451, doi: 10.18653/v1/2020.acl-main.747.

[10] A. Cuzzocrea, S. Akter, H. Shahriar, and P. García Bringas, "Cyberbullying detection, prevention, and analysis on social media via trustable LSTM-autoencoder networks over synthetic data: The TLA-NET approach," Future Internet, vol. 17, no. 2, Art. no. 84, 2025, doi: 10.3390/fi17020084.

[11] S. Ghorbanpour, D. Dementieva, and A. Fraser, "Data-efficient hate speech detection via cross-lingual nearest neighbor retrieval with limited labeled data," arXiv preprint arXiv:2505.14272, 2025. [Online]. Available: https://arxiv.org/abs/2505.14272

[12] A. R. Hevner, S. T. March, J. Park, and S. Ram, "Design science in information systems research," MIS Quarterly, vol. 28, no. 1, pp. 75–105, 2004, doi: 10.2307/25148625.

[13] H. Khafajeh, "Cyberbullying detection in social networks using deep learning," International Arab Journal of Information Technology, vol. 21, no. 6, pp. 1023–1034, 2024.

[14] R. M. Kowalski, G. W. Giumetti, A. N. Schroeder, and M. R. Lattanner, "Bullying in the digital age: A critical review and meta-analysis of cyberbullying research among youth," Psychological Bulletin, vol. 140, no. 4, pp. 1073–1137, 2014, doi: 10.1037/a0035618.

[15] R. Kumar, A. K. Ojha, S. Malmasi, and M. Zampieri, "Benchmarking aggression identification in social media," in Proc. 1st Workshop Trolling, Aggression and Cyberbullying (TRAC-2018), 2018, pp. 1–11. [Online]. Available: https://aclanthology.org/W18-4401/

[16] L. Liu, D. Xu, P. Zhao, D. D. Zeng, P. J.-H. Hu, Q. Zhang, Y. Luo, and Z. Cao, "A cross-lingual transfer learning method for online COVID-19-related hate speech detection," Expert Systems with Applications, vol. 234, Art. no. 121031, 2023, doi: 10.1016/j.eswa.2023.121031.

[17] F. A. Maikano, "Machine learning approaches for cyber bullying detection in Hausa language social media: A comprehensive review and analysis," FUDMA Journal of Sciences, vol. 8, no. 3, pp. 344–348, 2024, doi: 10.33003/fjs-2024-0803-2517.

[18] National Bureau of Statistics, "Telecoms data: Active voice and internet per state, porting and tariff information (Q4 2023)," 2024. [Online]. Available: https://www.nigerianstat.gov.ng/elibrary/read/1241461

[19] C. L. Nixon, "Current perspectives: The impact of cyberbullying on adolescent health," Adolescent Health, Medicine and Therapeutics, vol. 5, pp. 143–158, 2014, doi: 10.2147/AHMT.S36456.

[20] I. A. Owoade, A. A. Adeomi, P. A. Akinyemi, and A. F. Owoade, "Prevalence and determinants of cyberbullying perpetration among adolescents in rural and urban secondary schools in Osun State, South-Western Nigeria," *Journal of Community Medicine & Primary Health Care*, vol. 35, no. 1, pp. 24–39, 2023, doi: 10.4314/jcmphc. v35i1.3.

[21] A. G. Philipo, D. S. Sarwatt, J. Ding, M. Daneshmand, and H. Ning, "Cyberbullying detection: Exploring datasets, technologies, and approaches on social media platforms," *ACM Computing Surveys*, vol. 58, no. 7, Art. no. 186, 2026, doi: 10.1145/3785654.

[22] M. Sap, D. Card, S. Gabriel, Y. Choi, and N. A. Smith, "The risk of racial bias in hate speech detection," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 1668–1678, doi: 10.18653/v1/P19-1163.

[23] M. Sihab-Us-Sakib, M. M. Rahman, M. S. A. Forhad, and M. T. Aziz, "Cyberbullying detection of resource constrained language from social media using transformer-based approach," *Natural Language Processing Journal*, vol. 9, Art. no. 100104, 2024, doi: 10.1016/j.nlp.2024.100104.

[24] R. Slonje, P. K. Smith, and A. Frisén, "The nature of cyberbullying, and strategies for prevention," *Computers* in *Human Behavior*, vol. 29, no. 1, pp. 26–32, 2013, doi: 10.1016/j.chb.2012.05.024.

[25] P. K. Smith, J. Mahdavi, M. Carvalho, S. Fisher, S. Russell, and N. Tippett, "Cyberbullying: Its nature and impact in secondary school pupils," *Journal of Child Psychology and Psychiatry*, vol. 49, no. 4, pp. 376–385, 2008, doi: 10.1111/j.1469-7610.2007.01846. x.