African Journal of Education Science and Technology (AJEST) Vol. 8 No. 4 (May, 2026)
University of Eldoret, Kenya, Mount Kenya University, Kenya, Chukwuemeka Odumegwu Ojukwu University, Nigeria, Kyambogo University, Uganda and University of Makeni, Sierra Leone

# BiCHAT: A Hybrid CNN-BiLSTM Attention Transformer for Hate Speech Detection in Kenyan Multilingual Political Discourse

***Kimani Julia Waceke, Kennedy Ndenga Malanga & Josphat Mwai Karani***
*School of Computing and Informatics, Karatina University, Kenya*

***Corresponding Email: kimjulie54@gmail.com***

## Abstract

*Detecting hate speech on social media presents significant challenges in multilingual, low-resource contexts. In Kenya, political discourse frequently alternates between English and Swahili, emphasizes ethnic identities, and suffers from insufficient annotated training data for hate speech detection. This study introduces BiCHAT (Bidirectional CNN-LSTM with Hierarchical Attention Transformer), a hybrid deep learning model that integrates a pre-trained cross-lingual transformer (XLM-RoBERTa), a convolutional block, a bidirectional LSTM encoder, and a linear attention mechanism. A lexicon-based labeling system was applied to 121,149 political tweets, resulting in a highly imbalanced binary training set (HATE: 1.45%). Counterfactual Data Augmentation (CDA) was implemented to mitigate identity-group bias. Class-weighted cross-entropy loss helped tackle the class imbalance issue. When tested on a separate set of 11,994 samples, BiCHAT achieved an overall accuracy of 99.10%. Its hate-class F1 score was 0.6250, and it had a ROC AUC of 0.8358. By optimizing the threshold at 0.43 and applying temperature scaling (T = 1.053), the study increased the validation-set hate-class F1 to 0.6548. We have set up infrastructure for language debiasing and systematic ablation studies, ready for use. This study shows that hybrid transformer models, combined with thoughtful data preparation and fair training, can effectively detect hate speech in multilingual low-resource contexts.*

***Keywords:*** *Hate Speech Detection, Transformer, CNN, BiLSTM, Attention, Multilingual NLP, Kenyan Political Discourse, Code-Switching, Counterfactual Data Augmentation, Class Imbalance.*

## Introduction

Hate speech refers to communication spoken, written, visual, or behavioral—that attacks, demeans, insults, threatens, or discriminates against an individual or group because of characteristics associated with their identity, such as ethnicity, race, nationality, religion, gender, or other identity factors. It can occur through face-to-face communication, traditional media, or digital platforms such as Facebook, X, TikTok, WhatsApp and other social media. The rise of social media has increased civic participation as well as the spread of harmful content. In Kenya, political discussions intensify around election time, often referring to ethnic identities and at times involving direct incitement. The violence following the 2007-2008 elections in Kenya, which was partly attributed to the dissemination of inflammatory messages, underscores the severe consequences of unregulated hate speech (Wamuyu, 2017). Automated hate speech detection offers a scalable solution for identifying and flagging harmful content, enabling platforms to review and restrict such material at scale, a task that manual moderation cannot accomplish efficiently. However, most existing models are trained exclusively on English-language data, resulting in reduced effectiveness when applied to low-resource or code-switched languages. For example, Kenyan political content often mixes English and Swahili and frequently uses Sheng and ethnic words that may be offensive depending on the context. To address these issues, this study proposes BiCHAT (Bidirectional Convolutional Neural Network–Long Short Term Memory with Hierarchical Attention Transformer), a hybrid model based on XLM-RoBERTa, a pre-trained transformer encoder. It processes the output obtained in the CNN and BiLSTM layers before an attention-based classification optimized for the Kenyan political landscape. The model is an ensemble of cross-lingual token embeddings generated with XLM-RoBERTa input into specific convolutional and recurrent layers to capture local phrase patterns, long-term dependencies, and prominent tokens through the attention mechanism. In this setup, XLM-RoBERTa serves only as the embedding layer, while the CNN and BiLSTM components are used to extract all feature representations and sequence modeling outputs. Since no publicly available Kenyan hate speech dataset has been annotated by humans, this study presents a reusable lexicon-based labeling pipeline. In doing so, we address fairness through Counterfactual Data Augmentation (CDA) and a framework for adversarial debiasing. This paper is structured as follows: Section 2 reviews existing literature, and Section 3 presents the methodology. Section 4 presents the results and discussion, and Section 5 offers conclusions and recommendations.

Hate speech detection refers to the automatic identification of online content that targets individuals or groups based on protected characteristics, including race, ethnicity, religion, or national origin (Fortuna and Nunes, 2018). Early lexicon-

based methods, which identify hate speech by matching text to predefined lists of offensive or hateful terms, exhibit low recall for implicit hate speech. Machine learning approaches, such as support vector machines (SVMs) and naive Bayes classifiers, have demonstrated improvements over rule-based systems (Schmidt and Wiegand, 2017). The introduction of transformer architectures, initiated by Vaswani et al. (2017) and further advanced by Devlin et al. (2019) with BERT, has fundamentally transformed natural language processing (NLP). Fine-tuned transformer models consistently outperform previous architectures on hate speech detection tasks (Mozafari et al., 2020). However, in low-resource, code-switched, and noisy social media environments, transformer models may underperform due to limited training data and domain mismatch. Hybrid architectures that integrate transformer encoders with convolutional neural networks (CNNs) and recurrent layers have been shown to capture complementary feature types, specifically local n-gram relationships and longer-range sequential dependencies, which pure transformer models may not effectively exploit in these contexts (Ranasinghe and Zampieri, 2021).

**Multilingual and Low-Resource Challenges**

Code-switching frequently occurs in multilingual communities. The prevalence of Swahili-English code-switching on Kenyan social media platforms has been extensively documented (Ngugi et al., 2021). The Cross-lingual Language Model – Robustly Optimized BERT Approach (XLM-RoBERTa) (Conneau et al., 2020), which is pre-trained on 2.5TB of filtered CommonCrawl text spanning 100 languages including Swahili, provides effective subword representations for code-switched Kenyan text without requiring language-specific pre-training. However, multilingualism and limited linguistic resources pose significant challenges to the development and adoption of inclusive digital technologies. Many African languages remain underrepresented in digital systems due to insufficient language datasets, limited digitized texts, speech corpora, and linguistic resources. As a result, technologies such as artificial intelligence, machine translation, speech recognition, and natural language processing tend to perform better in high-resource languages than in indigenous languages. These challenges contribute to digital exclusion, particularly among communities that primarily communicate in indigenous languages. Limited access to appropriate language technologies may restrict participation in digital education, healthcare information, government services, and knowledge-sharing platforms. Addressing these challenges necessitates the development of indigenous-language datasets, active community involvement in language documentation, culturally responsive AI systems, and multilingual digital platforms. Such interventions are crucial for promoting equitable access to digital information and ensuring that technological development reflects Africa's linguistic and cultural diversity.

**Class Imbalance and Fairness**

Hate speech in political Twitter datasets typically accounts for less than 5% of total posts (Founta et al., 2018), resulting in a pronounced class imbalance that biases classifiers trained on unprocessed data toward the majority non-hate class. To address this imbalance, researchers commonly employ class-weighted loss functions, which assign greater penalties to misclassification of the minority class, as well as oversampling techniques that enhance minority-class representation in training data. Dixon et al. (2018) reported substantially higher false positive rates for comments concerning specific identity groups in the Civil Comments dataset. Counterfactual Data Augmentation (CDA), introduced by Lu et al. (2020), generates training pairs by swapping identity terms, thereby discouraging models from using group membership as a predictive shortcut.

**Hybrid Architectures for Text Classification**

Although pure transformer models achieve strong performance on standard benchmarks, hybrid architectures that integrate transformer encoders with convolutional neural networks (CNNs) and recurrent layers offer advantages in specific contexts. Kim (2014) demonstrated that one-dimensional CNNs are effective for detecting local n-gram patterns. Bidirectional long short-term memory (BiLSTM) networks perform well in conjunction with global self-attention, enabling the capture of bidirectional structures. Recent research indicates that transformer-CNN-BiLSTM hybrid models surpass pure transformer models in hate speech detection tasks, particularly when processing short, noisy, or code-switched text (Ranasinghe and Zampieri, 2021).

## Materials and Methods

### Research Design and Rationale

The study employed an applied natural language processing (NLP) experimental system, specifically designed to address practical NLP challenges. The methodology consisted of three steps: (i) data preparation, which entailed transforming an unlabeled political corpus into a supervised binary training set; (ii) model construction, which involved developing and training a hybrid neural architecture; and (iii) evaluation and fairness analysis. Table 1 presents the rationale for each component of BiCHAT.

**Table 1. Rationale for Each BiCHAT Architectural Component**

| Component | Primary Function | Linguistic Motivation | Key Parameter |
|---|---|---|---|
| XLM-RoBERTa | Multilingual contextual embeddings | Handles English-Swahili code-switching; Swahili in pre-training data | MAX_LEN=128 |
| 1D-CNN | Local n-gram feature extraction | Detects compound slurs and hate-bearing trigrams | filters=128, kernel=3 |
| BiLSTM | Bidirectional sequence modelling | Captures long-range dependencies across code-switches | hidden=256 total |
| Attention | Weighted token aggregation | Focuses classifier on most discriminative tokens | Linear proj + softmax |

The BiCHAT architecture integrates four sequential components, each addressing a distinct linguistic challenge in code-switched hate speech detection. XLM-RoBERTa produces multilingual contextual embeddings (MAX_LEN=128), utilizing its Swahili pre-training to manage English-Swahili code-switching. Subsequently, a one-dimensional convolutional neural network (1D-CNN; filters=128, kernel=3) extracts local n-gram features, identifying compound slurs and hate-bearing tri-grams. A bidirectional long short-term memory network (BiLSTM; hidden=256) captures long-range dependencies across code-switch points, where semantic meaning frequently spans both languages. An attention layer (linear projection and softmax) then assigns weights to tokens based on relevance, directing the classifier toward the most discriminative sequence elements. This pipeline progresses from multilingual comprehension, to local pattern recognition, to long-range contextualization, and finally to selective focus, reflecting the layered complexity of implicit and code-switched hate speech.

### Data Source and Lexicon-Based Labeling

The primary dataset utilized is NEW_Politico_Dataset.csv, which comprises unlabeled Kenyan public Twitter political tweets. Due to the absence of human-annotated labels, a rule-based lexicon classifier was developed to automatically assign hate-speech labels using four semantic categories: identity seeds (references to ethnic and national groups), violent seeds (threats or incitement), offensive seeds (derogatory language), and neutral seeds (common non-hateful terms). Tokenization relies on Unicode-aware regular expressions to accurately separate English, Swahili, and other African languages. A minimum frequency threshold of five was applied to exclude rare tokens.The three-class output is denoted as binary (HATE = 1, NOT_HATE = 0).

### Text Cleaning and Normalisation

A deterministic five-step pipeline produces the normalised text_clean field from the raw Tweet column, ensuring full reproducibility. The pipeline is summarised in Table 2.

**Table 2. Text Cleaning Pipeline**

| Step | Operation | Description | Example |
|---|---|---|---|
| 1 | URL removal | Replaces all hyperlinks with URL token | http://t.co/xyz to URL |
| 2 | Mention masking | Replaces @handles with USER token | @Wanjiku to USER |
| 3 | Punctuation normalisation | Strips excess punctuation; retains sentence-boundary markers | !!!!! to ! |

| Step | Operation | Description | Example |
|---|---|---|---|
| 4 | Lowercasing | Converts all text to lowercase for uniform token matching | HATE to hate |
| 5 | Whitespace collapsing | Removal of repeated spaces and newlines | Double spacing to single space |

Table 2 presents the five sequential procedures used to transform the raw Twitter data into a standardized text_clean variable. The pipeline begins with URL removal, which replaces hyperlinks with the standardized token URL. This step reduces unnecessary variation caused by different web addresses while retaining an indication that a hyperlink was present in the original tweet. This is important because URLs may contain lengthy strings of characters that do not contribute directly to the linguistic analysis.

The second step involves mention masking, whereby Twitter user handles are replaced with the token USER. This procedure minimizes the influence of individual usernames on subsequent text analysis and ensures that tweets mentioning different users are treated consistently. It also helps reduce unnecessary variability in the vocabulary used for classification or statistical analysis.

The third step is punctuation normalization. Excessive punctuation, such as repeated exclamation marks, is reduced while important sentence-boundary markers are retained. This process reduces textual noise while preserving useful structural information. For example, converting !!!!! to ! prevents repeated punctuation from being interpreted as multiple independent features.

The fourth step is lowercasing, which converts all textual content into lowercase. This ensures that words such as HATE, Hate, and hate are recognized as the same lexical item during token matching. Consequently, the procedure reduces duplication in the vocabulary and improves consistency during subsequent text processing and analysis.

The final step involves whitespace collapsing, where repeated spaces, tabs, and line breaks are converted into a single space. This produces a clean and consistently formatted textual representation, facilitating reliable tokenization and subsequent computational analysis.

**Dataset Distribution and Class Imbalance**

Table 3 shows the class distribution in stratified training, validation, and test splits. HATE examples represent only 1.45% of the total samples. Class-weighted cross-entropy loss was applied during training, making misclassification of a HATE example approximately 68 times more costly than a NOT_HATE misclassification. The 1.45% representation of HATE examples implies severe class imbalance, which may cause the classification model to favor the majority NOT_HATE class and overlook hate-speech cases. Consequently, class-weighted cross-entropy loss was necessary to give greater importance to the minority HATE class. This improves the model's sensitivity to hate-speech instances but may also increase false-positive classifications. Therefore, accuracy alone may not adequately reflect model performance, making precision, recall, F1-score, and the confusion matrix particularly important for evaluating the classifier.

**Table 3. Dataset Class Distribution (* after CDA augmentation)**

| Split | Total Samples | HATE (label=1) | NOT_HATE (label=0) |
|---|---|---|---|
| Training | 84,804 to 85,382* | 1,230 | 83,574 |
| Validation | 24,229 | 352 | 23,877 |
| Test | 11,994 | 174 | 11,820 |
| TOTAL | 121,149 | 1,761 (1.45%) | 119,388 (98.55%) |

The distribution in Table 3 confirms that class imbalance persists at a near-identical ratio (approximately 1.45% HATE) across the training, validation, and test splits. This consistency indicates that the stratified split preserved the original label distribution rather than distorting it, so evaluation on the test set is not biased by an artificially easier or harder class balance relative to training conditions. It also confirms that the minority class remains small in absolute terms even after stratification, which is the main driver of the class-weighting strategy adopted during training.

**Counterfactual Data Augmentation (CDA)**

CDA was used to mitigate identity-group bias, a fairness concern identified in previous research (Dixon et al., 2018) where, during training, the model did not associate certain ethnic or national group labels with hate speech as such, whereas it would

associate hate speech with the language used. CDA creates paired counterfactual training examples: every time a target identity term appears in a post, a paralleled version is produced by replacing the specific identity term with the corresponding pair and receiving the same label for both. Two sets of swaps were set up: nationality (Kenyan vs Ugandan) and ethnicity (Kikuyu vs Luo). The 578 augmented samples generated expanded the training set from 84,804 to 85,382 examples.

### Model Architecture: BiCHAT

BiCHAT integrates four components in a sequential stack. First, an XLM-RoBERTa encoder produces 768-dimensional contextual token embeddings for sequences up to 128 tokens. Second, a 1D-CNN block with 128 filters and kernel size 3 extracts local trigram-level features. Third, a Bidirectional LSTM produces concatenated hidden states of dimension 256. Fourth, a linear attention mechanism computes a weighted average over BiLSTM hidden states, allowing the classifier to focus on the most discriminative tokens in each post.

### Training Configuration

Training hyperparameters are summarized in Table 4. The model was trained using AdamW optimisation with learning rate 2e-5 and early stopping on validation loss with patience 2. A post-hoc threshold search on the validation set replaced the default 0.5-classification threshold with a value optimising hate-class F1. Temperature scaling was applied to correct systematic overconfidence in raw model outputs.

**Table 4. Training Hyperparameter Configuration**

| Hyperparameter | Value | Rationale |
|---|---|---|
| Base model | xlm-roberta-base | Multilingual; includes Swahili |
| MAX_LEN | 128 tokens | Covers typical tweet length |
| CNN filters | 128 | Sufficient n-gram capacity |
| CNN kernel size | 3 | Trigram window |
| LSTM hidden size | 128 per direction | 256 total after concatenation |
| Dropout | 0.1 | Regularisation |
| Learning rate | 2e-5 | Selected from grid search over {1e-5, 2e-5, 5e-5}; 2e-5 yielded the highest validation F1 and is the standard recommendation for transformer fine-tuning |
| Batch size | 32 | Fits GPU memory |
| Max epochs | 5 (early stopping) | Patience = 2 on validation loss |
| Class weights | approx. 1:68 (HATE:NOT_HATE) | Inverse frequency weighting |
| Optimiser | AdamW | Weight decay = 0.01 |

The configuration in Table 4 reflects a balance between adequate model capacity and the constraints of fine-tuning a large multilingual embedding model on a small positive class. The relatively conservative learning rate (2e-5) and short epoch budget with early stopping (patience = 2) were chosen to limit overfitting on the 1,761 HATE-labelled examples, while the heavily weighted loss function (approximately 1:68) directly compensates for the 1.45% positive-class prevalence documented in Table 3.

## Results and Discussion

### Training Dynamics

The model was trained for five epochs with early stopping. During epochs 1-3, training loss declined from 0.5404 to 0.4069 and validation loss from 0.4846 to 0.3773, while hate-class validation F1 improved from 0.6159 to a peak of 0.6524. Beginning at epoch 4, training loss rose approximately 49% and validation loss 53%; validation F1 fell to 0.4913. By epoch 5, the model ceased to identify any hate-class instance. The early-stopping mechanism correctly restored the epoch 3 checkpoint. Total training time was 46 minutes and 45 seconds.

**Final Evaluation on the Held-Out Test Set**

The restored best-checkpoint model was evaluated on 11,994 held-out test samples. Table 5 presents the full classification report.

**Table 5. Classification Report on Held-Out Test Set**

| Class / Metric | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| HATE | 0.7895 | 0.5172 | 0.6250 | 174 |
| NOT_HATE | 0.9929 | 0.9980 | 0.9954 | 11,820 |
| Weighted Average | 0.9901 | 0.9910 | 0.9901 | 11,994 |
| Overall Accuracy | — | 0.9910 | — | 11,994 |
| ROC AUC (HATE) | — | — | 0.8358 | — |

The overall accuracy of 99.10% must be contextualised against the 98.55% achievable by a naive always-NOT_HATE classifier. For NOT_HATE, the model achieves near-perfect precision (0.9929), recall (0.9980), and F1 (0.9954). For the hate class, precision of 0.7895 indicates that approximately 79% of hate-flagged instances are genuine, while recall of 0.5172 reveals that roughly 48% of actual hate speech goes undetected. The hate-class F1 of 0.6250 reflects the inherent difficulty of minority-class detection.

The ROC AUC of 0.8358 provides a threshold-independent measure: the model ranks a randomly selected hate instance above a non-hate instance with probability 83.6%. The confusion matrix yields 90 true positives, 84 false negatives (the principal limitation), 24 false positives, and 11,796 true negatives.

**Threshold Optimization and Probability Calibration**

A threshold of 0.4300 was identified through the validation-driven threshold search, resulting in a validation F1 of 0.6548 with precision of 0.7632 and recall of 0.5734. The reduction of the threshold from the default 0.5 to 0.43 means that the model classifies hate speech when its predicted probability of being hate is at least 43%, rather than 50%, which improves sensitivity to the minority class as well as recall at an acceptable cost to precision. Temperature scaling (T = 1.0530) is a post-hoc probability calibration procedure in which the model's raw logits are divided by a scalar T prior to softmax being applied. A value slightly above 1.0 indicates that the model was somewhat overconfident — meaning that, on average, its predicted probabilities were consistently too high — and the calibration step corrects this so that a predicted probability of, say, 80% actually corresponds to the model being correct 80% of the time.

**Ablation Study Design**

The BiCHAT architecture allows for selective component disabling through constructor flags, enabling five configurations to be evaluated under identical conditions. Table 6 presents the ablation configuration matrix. The full-configuration run serves as the upper-bound reference; comparative ablation runs are specified for subsequent sessions. In Table 6, Yes indicates that a particular BiCHAT component or feature is enabled and included in the model configuration, whereas No indicates that the component is disabled and excluded from that configuration. The columns represent the specific architectural components under investigation, while the rows represent the different experimental configurations.

**Table 6. Ablation Study Configuration Matrix**

| Configuration | CNN | BiLSTM | Attention | Description |
|---|---|---|---|---|
| Full BiCHAT (reported) | Yes | Yes | Yes | Complete architecture — upper-bound reference |
| No CNN | No | Yes | Yes | Removes local n-gram features |
| No BiLSTM | Yes | No | Yes | Removes sequence-level encoding |
| No Attention | Yes | Yes | No | Mean-pool instead of weighted aggregation |
| BiLSTM only | No | Yes | No | Minimal recurrent baseline |

As shown in Table 6, each ablation configuration isolates the contribution of a single architectural component by disabling it while holding the others constant, with the BiLSTM-only configuration serving as a minimal recurrent baseline against which the added value of the CNN and attention components can be assessed. Comparative results for these configurations are reported in a subsequent experimental session.

**Fairness and Adversarial Debiasing**

Language-group distribution confirms correct population of the lang column across splits: training 72,367 non-Swahili and 13,015 Swahili; validation 20,564 and 3,787; test 10,157 and 1,837. Adversarial debiasing infrastructure, including data-loader objects, an adversarial classifier, and a gradient reversal layer, was confirmed operational but not activated in the primary run. Per-language-subgroup metrics await a subsequent adversarial run.

**Discussion**

BiCHAT achieved a hate-class ROC AUC of 0.8358 and an F1 score of 0.6250 on the held-out test set, despite being trained without human-annotated labels and with only 1,761 positive training examples. These findings indicate that the model reliably ranks hate speech above non-hate content in the majority of cases, even though the minority class remains difficult to identify with high recall. Relating these findings to prior work, this performance is broadly in line with results reported for comparable low-resource, code-switched settings: Ranasinghe and Zampieri (2021) report AUC values of 0.76–0.84 for Tamil-English code-switched content, and Nkemdirim et al. (2022) report a hate-class F1 of 0.58 on a Nigerian Pidgin corpus. BiCHAT's results sit within this range, notably without relying on the larger human-annotated corpora used in those studies. The primary limitation observed in this study — a recall of 0.5172 — reflects training on a small lexicon-labeled minority class. Lexicon-based labels under-represent implicit hate speech where offensiveness arises from pragmatic implicature rather than explicit seed words. Human annotation of ambiguous, borderline cases is the most direct path to improving recall. The deployment of CDA as the primary bias-mitigation strategy is consistent with findings that identity-aware augmentation produces more equitable error distributions (Dixon et al., 2018; Lu et al., 2020).

## Conclusion and Recommendations

**Conclusion**

This study presented BiCHAT (Bidirectional CNN-LSTM with Hierarchical Attention Transformer), a hybrid deep learning model in which XLM-RoBERTa serves as a multilingual embedding layer. It utilizes 768-dimensional token representations to carry out sequential processing using a 1D-CNN block for local n-gram feature extraction, a Bidirectional LSTM encoder for long-range sequence modeling, and a linear attention mechanism to highlight the most discriminative tokens before binary classification to detect hate speech in Kenyan multilingual political discourse. The results confirm the effectiveness of the proposed model, achieving a hate-class ROC AUC of 0.8358 and an F1 score of 0.6250 on a held-out test set. The study makes four main contributions. First, it provides a practical lexicon-based labeling pipeline that generates supervised training data, without human annotation, for unlabeled political tweets. Second, it presents a hybrid transformer architecture suited to code-switched East African political discourse, suggesting that CNN and BiLSTM layers enhance transformer embeddings for this low-resource setting. Third, it employs an identity-aware counterfactual augmentation strategy that reduces ethnic and national group bias. Fourth, it includes infrastructure established for adversarial language debiasing and systematic ablation studies, allowing subsequent researchers to directly build upon this work.
The results confirm that hybrid transformer architectures, combined with principled data preparation and fairness-aware training strategies, can deliver competitive hate speech detection performance in low-resource multilingual settings. The study offers critical insights for AI ethics researchers, social media platforms, and policymakers engaged in the design of content moderation systems for African multilingual contexts.

**Recommendations**

It is beneficial to add transformer-based classifiers incrementally to the Kenyan social media moderation pipeline, starting with cases where the model assigns a calibrated probability above the optimized threshold of 0.43, thus ensuring that only the most robust instances are acted upon when deploying new systems. The most viable short-term strategy for improving hate-class recall is to manually annotate a stratified sample of borderline lexicon-tagged cases, since recognition of implicit

hate speech from a lexicon-only training set is constrained to a recall ceiling of around 0.52. The adversarial debiasing training loop should be set up and verified prior to production deployment to ensure that detection performance is comparable across both predominantly Swahili and heavily English-dominant content. Ultimately, future research should extend the BiCHAT pipeline to a wider range of East African political corpora, including data from Uganda and Tanzania, which share comparable multilingual, code-switching dynamics with Kenya, to assess cross-national transferability.

**Limitations of the Study**

The absence of human-annotated ground truth means evaluation metrics reflect performance against lexicon-derived labels rather than human judgement. The small HATE training set (1,761 examples) limits exposure to the full distributional diversity of hate speech. Generalizability to non-political or non-Twitter domains has not been assessed. Adversarial debiasing and ablation comparative results are pending activation in future experimental sessions.

**Suggestions for Further Research**

- Integrate human annotation for stratified groups of borderline, ambiguous instances to improve label quality beyond the lexicon-based approach.
- Activate and test the adversarial debiasing training branch — a gradient-reversal-based module already implemented that penalises the model for predicting language group (Swahili vs. non-Swahili) from the learned representations, thereby encouraging language-invariant hate-speech features — and report per-language-subgroup metrics.
- Complete the ablation study to quantify the individual contribution of the CNN, BiLSTM, and attention components.
- Run the pipeline over several random seeds to capture metric variance and establish confidence intervals for reported results.
- Test BiCHAT on Ugandan and Tanzanian political Twitter corpora — chosen for their comparable English-Swahili code-switching dynamics and similar political social-media cultures — to evaluate cross-national transferability.

**References**

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzman, F., & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In Proceedings of ACL 2020 (pp. 8440-8451).

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of NAACL-HLT 2019 (pp. 4171-4186).

Dixon, L., Li, J., Sorensen, J., Thain, N., & Vasserman, L. (2018). Measuring and mitigating unintended bias in text classification. In Proceedings AAAI/ACM AIES 2018 (pp. 67-73).

Fortuna, P., & Nunes, S. (2018). A survey on automatic detection of hate speech in text. ACM Computing Surveys, 51(4), 1-30.

Founta, A. M., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., & Kourtellis, N. (2018). Large scale crowdsourcing and characterization of Twitter abusive behavior. In Proceedings of ICWSM 2018 (pp. 491-500).

Kim, Y. (2014). Convolutional neural networks for sentence classification. In Proceedings of EMNLP 2014 (pp. 1746-1751).

Lu, K., Mardziel, P., Wu, F., Amancharla, P., & Datta, A. (2020). Gender bias in neural natural language processing. In Logic, Language, Information, and Computation (pp. 189-202). Springer.

Mozafari, M., Farahbakhsh, R., & Crespi, N. (2020). A BERT-based transfer learning approach for hate speech detection. In Complex Networks Conference (pp. 928-940).

Ngugi, J. M., Fleisch, E., & Afande, F. O. (2021). Swahili-English code-switching in Kenyan digital media. Journal of African Languages and Linguistics, 42(1), 89-117.

Nkemdirim, C., Obiora, O., & Adegoke, T. (2022). Hate speech detection in Nigerian Pidgin. African Language Technology, 3(2), 45-58.

Patwa, P., Aguilar, G., Kar, S., Pandya, S., Pykl, S., Gelbukh, A., & Solorio, T. (2020). SemEval-2020 task 9: Sentiment analysis of code-mixed tweets. In SemEval Workshop (pp. 774-790).

Ranasinghe, T., & Zampieri, M. (2021). MUDES: Multilingual detection of offensive spans. In NAACL-HLT 2021 (pp. 3659-3666).

Schmidt, A., & Wiegand, M. (2017). A survey on hate speech detection using natural language processing. In SocialNLP Workshop (pp. 1-10).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30.

Wamuyu, P. K. (2017). Bridging the digital divide among low income urban communities. Telematics and Informatics, 34(8), 1709-1720.