*Science World Journal Vol. 20(No 4) 2025* https://dx.doi.org/10.4314/swj.v20i4.21
*www.scienceworldjournal.org*
*ISSN: 1597-6343 (Online), ISSN: 2756-391X (Print)*
*Published by Faculty of Science, Kaduna State University*

Full Length Research Article

# HATE SPEECH DETECTION IN HAUSA CODE-MIXED TWEETS USING MACHINE LEARNING

*Bashir Idris Sulaiman, Muhammad Aminu Ahmad

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

*Corresponding Author Email Address: bashir.idris@kasu.edu.ng

**ABSTRACT**
Code-mixed communication in Nigeria, involving English, Nigerian Pidgin, Hausa, Yoruba, and Igbo, poses challenges for Natural Language Processing (NLP) systems, especially in detecting hate speech. Existing research typically focuses on high-resource languages, leaving code-mixed African data underexplored. This study applies logistic regression and random forest algorithms to identify hate speech in Hausa code-mixed tweets, utilizing two datasets of annotated posts. Key preprocessing steps included text normalization and feature extraction using TF-IDF and Bag-of-Words. Results showed that the Logistic Regression model with TF-IDF features outperformed Random Forest in accuracy, recall, and F1-score, while the optimized Random Forest model demonstrated notable performance improvements. The results demonstrate the effectiveness of machine learning for hate speech detection in low-resource languages like Hausa.

**Keywords:** Hausa code-mixing, hate speech detection, logistic regression, random forest, TF-IDF, bag-of-words, multilingual NLP

## INTRODUCTION

Hate speech is the use of expression for violence, discrimination, or animosity toward people or groups based on ethnicity, religion, gender, nationality, or political affiliation (Waseem & Hovy, 2016; Davidson et al., 2017). Digital communication and social media have facilitated the rapid dissemination of online hate speech (Vidgen & Derczynski, 2020). In Nigeria, social media chats frequently combine English, Nigerian Pidgin, Hausa, Yoruba, and Igbo (Nwafor & Nguyen, 2025). Therefore, code-mixing has become prevalent in Nigerian social media interactions, conveying different forms of expressions, including hostility and mockery. A social media post that comprises English with a Hausa, Yoruba, or Igbo proverb may carry an offensive meaning that can be recognised by native speakers only, which makes hate speech detection more critical in Nigeria (Tonneau et al., 2024).

Speech formats exist in the form of text, pictures, audio, and videos, but the majority of online hate speech is text-based (Mozafari et al., 2020; Röttger et al., 2021). Researchers used machine learning (ML) and natural language processing (NLP) techniques like Term Frequency-Inverse Document Frequency (TF-IDF) and Bag-of-Words (BoW) to detect hate speech (Davidson et al., 2017; Zhang et al., 2018; Founta et al., 2018) using annotated English hate speech datasets (Waseem and Hovy, 2016; Davidson et al., 2017). Additionally, transformer-based models like BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and XLM-R (Conneau et al., 2022) were used to enhance hate speech detection. However, limited availability of data for low-resource languages and code-mixed text hinders hate speech detection in the Nigerian context (Onyebuchi et al., 2023). Additionally, English-Hindi code-mixed datasets demonstrated that monolingual models did not generalise successfully, resulting in lower F1-scores (Jain et al., 2025). The challenges of tokenisation, morphological diversity, and mixed syntactic structure have also been emphasised by studies in Spanish-English and Arabic-English code-mix (Mubarak et al., 2022; Mubarak et al., 2023; Zaghouani et al., 2024; Al Yousef et al., 2025).

Furthermore, in the African setting, the field of hate speech detection is still relatively new but expanding quickly. While AFRIHATE (Muhammad et al., 2025) is a multilingual corpus covering 15 African languages, including Hausa, Yoruba, and Swahili, projects like Masakhane (Adelani et al., 2022) and Lacuna Fund have started creating datasets and tools for African languages. The West African Twitter Hate Speech Dataset (Sosimi et al., 2024) further contributed region-specific data, covering Nigeria and Ghana. However, the majority of these datasets are either monolingual or bilingual, which restricts their applicability to Nigerian situations with code-mixing, where users frequently switch between Pidgin, English, and native languages in a single sentence (Adewumi et al., 2022). Additionally, studies on Nigerian social media have indicated intensified linguistic creativity, with users mixing English, Pidgin, and Hausa or Yoruba to express solidarity or anger during the EndSARS movement (Akpan & Targema et al., 2023).

According to Adewumi et al. (2022) and Sosimi et al. (2024), a robust feature extraction technique such as GloVe embeddings or Term Frequency-Inverse Document Frequency (TF-IDF) can improve machine learning and deep learning models for hate speech detection. Additionally, transfer learning models such as AfriBERTa (Ogueji et al., 2023) and mBERT (Devlin et al., 2019) were refined for African language problems and demonstrated significant improvements in Hausa and Yoruba text categorisation. However, deep learning models need a lot of training data and are computationally costly. Additionally, transfer learning algorithms are less effective at detecting hate speech when dealing with noisy and code-mixed social media text. On the other hand, logistic regression and SVM have interpretability and reduced processing demands, which are suitable for low-resource languages. In order to identify hate speech in Nigerian tweets, this study employs logistic regression and random forest with TF-IDF and bag-of-words as feature extraction.

The objective of this study is to develop a machine learning model to detect hate speech in Hausa code-mixed tweets, addressing the gap in existing research on low-resource languages. Specifically, we aim to: (1) investigate the performance of logistic regression and random forest algorithms, and (2) evaluate the effectiveness of TF-IDF and Bag-of-Words features for hate speech detection.

## MATERIALS AND METHODS

This section presents the method used for detecting hate speech in Nigerian code-mixed tweets. As illustrated in Figure 1, the

*Science World Journal Vol. 20(No 4) 2025* https://dx.doi.org/10.4314/swj.v20i4.21
*www.scienceworldjournal.org*
*ISSN: 1597-6343 (Online), ISSN: 2756-391X (Print)*
*Published by Faculty of Science, Kaduna State University*

approach began by obtaining a publicly available Hausa code-mixed dataset. The datasets were then pre-processed with TF-IDF and Bag-of-Words for were methods for feature extraction. The pre-processed dataset was then divided into training and testing sets. Two machine learning models were developed using logistic regression and random forest algorithms, with and without optimization, for hate speech detection. Finally, a comparative analysis was conducted to determine the best model.

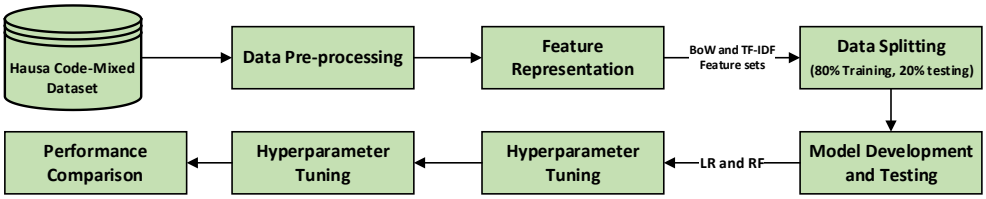


**Figure 1:** Workflow diagram of the research method

**Datasets**

The dataset used comprises tweets from the Herdphobia (Aliyu et al., 2022) and AfriSenti (Muhammad et al., 2025) Hausa code-mixed datasets in the English, Pidgin, and Hausa languages. Tweets from the two datasets were manually annotated into hateful and non-hateful speeches. Hate speech comprises explicit or implicit attacks, slurs, or incitement toward individuals or groups based on ethnicity, religion, or politics, whereas non-hate speech comprises factual or neutral discussions. The annotation was conducted by experts in English, Pidgin, and Hausa languages based on the Waseem & Hovy (2016) and Davidson et al. (2017) guidelines to ensure consistency with international hate speech labelling standards.

The class distribution of the dataset is as follows:

3 Hateful (Positive): 751 tweets (22%)
4 Not-Hate (Negative): 2,644 tweets (78%)

The final corpus comprises 3,395 tweets. The dataset thus reflects a realistic representation of online discourse while maintaining analytical focus on hate and non-hate discrimination.

**Data Pre-processing and Feature Extraction**

Pre-processing plays a crucial role in transforming raw text into structured, machine-readable input (Founta et al., 2018; Röttger et al., 2021). Nigerian tweets often contain informal expressions, emojis, hashtags, and spelling variations that can distort vector representations if not standardised. The pre-processing pipeline consisted of the following sequential steps:

1 Lowercasing all characters to ensure consistency.
2 Removing URLs, hashtags, mentions, and punctuation using regular expressions.
3 Filtering out stop words from both English and Nigerian Pidgin.
4 Expanding contractions (e.g., “don’t” is expanded to “do not”).
5 Tokenising the text using NLTK for word-level segmentation.
6 Stemming and lemmatisation to reduce words to their root forms (e.g., from “hateful” to “hate”).
7 Whitespace normalisation to eliminate redundant spacing.

This pipeline ensured that linguistic features were standardised across all code-mixed languages.

**Feature Representation**

After pre-processing, tweets were converted into numerical vectors using two feature extraction techniques:

1. Bag-of-Words (BoW): This represents the frequency of each token within a document, ignoring word order but preserving term importance.
2. Term Frequency-Inverse Document Frequency (TF-IDF): This scales the BoW representation by penalising common words and amplifying rare, discriminative terms. TF-IDF is effective for textual classification tasks such as hate speech detection (Zhang et al., 2018; Mozafari et al., 2020).

The vectorisation process was implemented in scikit-learn using *TfidfVectorizer* with *ngram_range = (1,2)* to capture both unigram and bigram patterns. The resulting feature matrix was then split into 80% training and 20% testing subsets using stratified sampling to preserve class proportions.

**Model Development**

Machine learning models were developed using logistic regression and random forest algorithms. Logistic regression is a probabilistic linear classifier that estimates the likelihood of a sample belonging to a specific class. It applies a sigmoid activation function that maps the input vector to a probability value between 0 and 1. Logistic regression was chosen as the baseline model due to its interpretability, computational efficiency, and resilience against overfitting in small datasets (Davidson et al., 2017; Mozafari et al., 2020). Its coefficients can be interpreted to reveal the most predictive words contributing to hate or non-hate classifications. Furthermore, random forest is an ensemble learning technique that constructs multiple decision trees and aggregates their predictions for robust classification. Each tree is trained on random subsets of features and samples, thereby improving generalisation and reducing variance. The strength of the random forest classifier is handling high-dimensional data efficiently, reducing overfitting through averaging and minimising preprocessing requirements (Zhang et al., 2018).

The two models were implemented in Python (Scikit-learn) in the Google Colab cloud GPU environment. Hyperparameter tuning was conducted to ensure optimal performance and model stability. The “C” hyperparameter for regularisation strength was used for the logistic regression model. Additionally, the number of trees and maximum depth were used for the random forest model. The optimisation targeted key parameters affecting model depth and generalisation, as shown in Table 1.

*Science World Journal Vol. 20(No 4) 2025* https://dx.doi.org/10.4314/swj.v20i4.21
*www.scienceworldjournal.org*
*ISSN: 1597-6343 (Online), ISSN: 2756-391X (Print)*
*Published by Faculty of Science, Kaduna State University*

**Table 1.** Hyperparameter Options Considered for RF Optimisation

| S/N | Hyperparameter | Value Options |
|---|---|---|
| 1 | max_depth | [None, 10, 20, 30] |
| 2 | n_estimators | [50, 100, 200] |
| 3 | min_samples_split | [2, 5, 10] |
| 4 | min_samples_leaf | [1, 2, 4] |
| 5 | max_features | ['auto', 'sqrt'] |

**Evaluation**

This study used experiments on the Google Colab cloud GPU environment to evaluate the two machine learning models. Table 2 displays the number of experiments conducted to evaluate the performance of the developed machine learning models. Both models were tested with the TF-IDF and Bag-of-Words feature representations, resulting in four experiments. The four experiments were repeated using the optimised models, making a total of 8 experiments.

**Table 2**: Experiments

| Model | Experiments | |
|---|---|---|
| | TF-IDF | Bag-of-Words |
| LR | 1 | 2 |
| RF | 3 | 4 |
| Optimized LR | 5 | 6 |
| Optimized RF | 7 | 8 |

A confusion matrix was used to determine the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) for each model. True positives are the number of hate speeches that are correctly flagged as hateful. True negatives are normal speeches that are flagged as non-hateful. False Positives (FP) are non-hate speeches that are wrongly flagged as hateful. False Negatives (FN) are harmful speeches that are wrongly flagged as non-hateful. The models were then evaluated using accuracy, precision, recall and F1-score (Davidson et al., 2017; Mozafari et al., 2020; Vidgen & Derczynski, 2020). The derivation of the four metrics is based on the confusion matrix results as illustrated in equations 1, 2, 3, and 4, respectively.

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

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

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

$$F1 - Score = 2 * \frac{precision * recall}{precision + recall} \quad (4)$$

These metrics collectively provide a comprehensive understanding of model effectiveness, balancing the trade-off between identifying hate speech (recall) and avoiding false accusations (precision) (Röttger et al., 2021; Vidgen & Derczynski, 2020).

**RESULTS**

This section presents the results of the experiments and performance of the evaluated models for hate speech detection in code-mixed Hausa tweets. The two models were trained and evaluated using two text representation techniques, namely TF-IDF and Bag-of-Words (BoW). Therefore, the results of the confusion matrix for each model based on TF-IDF and Bag-of-Words (BoW) are presented for each model. Then, the accuracy, precision, recall, and F1-score values are presented to determine the performance of the models. Table 3 displays the confusion matrix results of the logistic regression and random forest models with TF-IDF and Bag-of-Words (BoW).

**Table 3**: Confusion Matrix Results

| Model | Feature Rep. | TP | TN | FP | FN |
|---|---|---|---|---|---|
| Logistic Regression | TF-IDF | 1,512 | 1,668 | 83 | 132 |
| | BoW | 1,485 | 1,645 | 106 | 159 |
| Random Forest | TF-IDF | 1,470 | 1,615 | 136 | 174 |
| | BoW | 1,272 | 1,388 | 363 | 372 |
| Optimized Logistic Regression | TF-IDF | 1,520 | 1,678 | 73 | 124 |
| | BoW | 1,495 | 1,655 | 96 | 149 |
| Optimized Random Forest | TF-IDF | 1,500 | 1,640 | 111 | 144 |
| | BoW | 1,425 | 1,545 | 206 | 219 |

The confusion matrix results provide detailed insights into how each model classified hate and non-hate tweets. The results show that logistic regression with TF-IDF features correctly classified 1,512 hate speeches (TP) and 1,668 non-hate speeches (TN), but misclassified 83 non-hate speeches (FP) and 132 hate speeches (FN) as hateful and non-hateful speeches, respectively. In contrast with the BoW features, the misclassification of logistic regression increased to 106 false positives and 159 false negatives. The results of random forest show that the classifier correctly identified 1,470 hate speeches (TP) and 1,615 non-hate speeches (TN) with the TF-IDF features but misclassified 136 non-hate speeches (FP) and 174 hate speeches (FN) as hateful and non-hateful, respectively. The misclassification of speeches by random forest increased substantially with the BoW features; false positives increased from 136 to 363, whereas false negatives increased from 174 to 372. This shows that the model's complexity and reliance on feature diversity made it more sensitive to sparse BoW representations.

Furthermore, hyperparameter tuning was used to optimise the performance of the two models by adjusting regularisation strength and solver for logistic regression and using grid-search tuning (adjusting the values of parameters in Table 1) for random forest. The performance of the optimised models increased due to the reduction in the number of false positives and false negatives, as summarised in Table 4.

**Table 4**: Summary of Reduction in False Detection Due to Optimization

| Optimised Model | Feature Rep. | FP | FN |
|---|---|---|---|
| Logistic Regression | TF-IDF | 10 | 8 |
| | BoW | 10 | 10 |
| Random Forest | TF-IDF | 25 | 30 |
| | BoW | 157 | 153 |

The optimised logistic regression model reduced the false positives and false negatives by 10 and 8 with the TF-IDF features,

respectively, and by 10 with the BoW features. The small but consistent improvement illustrates the stability of LR and its robustness for code-mixed text. For the optimised random forest, the model reduced the false positives and false negatives by 25 and 30 with the TF-IDF features and 157 and 153 for the BoW features.

Table 5 displays the performance of the logistic regression and random forest models based on accuracy, precision, recall and F1-score values, whereas Figure 2 illustrates the performance of the models for visual comparison.

**Table 5**: Hate Speech Detection Performance of the LR and RF Models

| Model | Featur e Rep. | Accura cy | Precisi on | Recal l | F1-score |
|---|---|---|---|---|---|
| LR | TF-IDF | 93.67% | 94.80% | 91.97 % | 93.36 % |
| | BoW | 92.19% | 93.34% | 90.33 % | 91.81 % |
| RF | TF-IDF | 90.87% | 91.53% | 89.42 % | 90.46 % |
| | BoW | 78.35% | 77.80% | 77.37 % | 77.58 % |
| Optimiz ed LR | TF-IDF | 94.20% | 95.42% | 92.46 % | 93.91 % |
| Optimiz ed RF | BoW | 92.78% | 93.97% | 90.94 % | 92.43 % |
| | TF-IDF | 92.49% | 93.11% | 91.24 % | 92.17 % |
| | BoW | 87.48% | 87.37% | 86.68 % | 87.02 % |

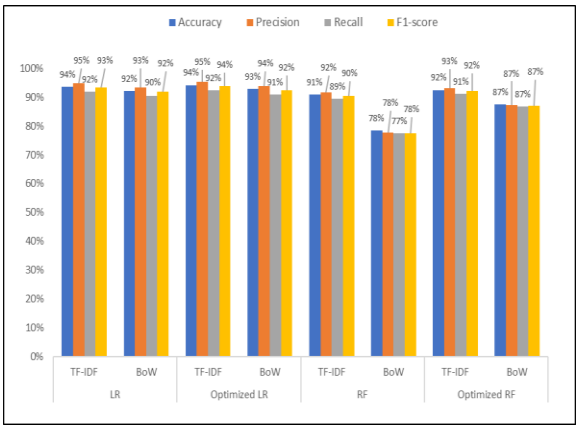


**Figure 2**: Comparative performance of the LR and RF Models

The Logistic Regression (LR) classifier demonstrated excellent predictive performance, particularly when trained on TF-IDF features. Using TF-IDF, the model achieved an accuracy of 93.67%, a precision of 94.80%, a recall of 91.97%, and an F1-score of 93.36%. When trained on BoW features, its accuracy slightly decreased to 92.19%, with precision, recall, and F1-score also marginally lower. This is due to the increase in the number of false detections with the BoW features. The superior performance of LR with TF-IDF aligns with previous research (Zhang et al., 2018; Waseem & Hovy, 2016), which highlights the ability of TF-IDF to capture contextual relevance by penalising frequent but uninformative words. This suggests that term weighting plays a crucial role in differentiating hateful and non-hateful expressions in multilingual settings.

On the other hand, the Random Forest (RF) classifier also performed strongly but exhibited greater sensitivity to feature representation. When tested on TF-IDF features, RF achieved 90.87% accuracy, 91.53% precision, 89.42% recall**,** and an F1-score of 90.46%**.** However, the performance dropped substantially with the BoW features, where accuracy reduced to 78.35%**,** precision reduced to 78.80%**,** recall reduced to 77.37%**,** and the F1-score reduced to 77.58%**.** The decline in performance with BoW features reflects the dependency of random forest on informative features, which overfit when dealing with sparse and high-dimensional data. Nonetheless, the random forest model maintained high precision, demonstrating its robustness in minimizing false positives, which is a valuable property in moderation systems where over-censorship must be avoided. After optimisation, the classification performance increased as displayed in Table 6. The improvement in performance ranges from 0.49% to 9.56%.

**Table 6**: Improvement in Classification Performance After Optimisation

| Model | Feature Rep. | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|---|
| LR | TF-IDF | 0.53% | 0.62% | 0.49% | 0.55% |
| | BoW | 0.59% | 0.63% | 0.61% | 0.62% |
| RF | TF-IDF | 1.62% | 1.58% | 1.82% | 1.71% |
| | BoW | 9.13% | 9.57% | 9.31% | 9.44% |

The accuracy, recall, and F1-score values increased substantially for random forest, particularly with the BoW features. The 9% minimum gain across the metrics shows the importance of parameter tuning, especially for ensemble models in high-dimensional text tasks. The improvement suggests that even with simpler feature spaces such as BoW, model generalisation can be significantly enhanced through controlled hyperparameter adjustment. Table 7 shows the classification performance increase when using TF-IDF over BoW.

**Table 7**: Improvement in Classification Performance with TF-IDF

| Model | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|
| LR | 1.48% | 1.46% | 1.64% | 1.55% |
| RF | 12.52% | 13.73% | 12.05% | 12.88% |
| Optimized LR | 1.42% | 1.45% | 1.52% | 1.48% |
| Optimized RF | 5.01% | 5.74% | 4.56% | 5.15% |

The results indicate that logistic regression with TF-IDF has a better balance of precision and recall by 1.64% and 1.55% respectively. For the optimised logistic regression model, the recall and F1-score increased by 1.52% and 1.48% respectively. This makes the classifier a strong baseline for detecting hate speech in Hausa code-mixed tweets. The performance of random forest increased substantially with the TF-IDF by at least 12.05% and 12.88% for recall and F1-score, respectively. For both classifiers, the results indicate that weighted feature representations (TF-IDF) and model tuning are key to handling code-mixed text effectively.

**DISCUSSION**

The results indicate that the logistic regression model outperformed the random forest model on both feature sets, reaffirming the model's strength in sparse text data with high dimensionality (Davidson et al., 2017; Mozafari et al., 2020). The use of TF-IDF also consistently outperformed BoW across all models, emphasising the significance of frequency weighting and contextual term importance. TF-IDF captures rare but semantically meaningful tokens (e.g., slurs, local proverbs, or political nicknames), while BoW treats all terms equally. Given that Nigerian tweets often include code-switching and orthographic variation, frequency-based weighting helps mitigate noise from repetitive neutral words such as "government", "people", or "country". Finally, the performance margin between logistic regression and random forest with TF-IDF is modest but consistent across folds. Such closeness indicates that both models are stable baselines for low-resource NLP tasks where computational simplicity and interpretability are prioritised.

The findings align with prior literature emphasising the robustness of classical machine learning models in hate speech detection. For instance:

1. Waseem & Hovy (2016) and Davidson et al. (2017) established that linear classifiers like logistic regression and support vector machines provide strong baselines, particularly on social media text.
2. Founta et al. (2018) demonstrated that TF-IDF enhances classifier precision by encoding contextual salience.
3. In African code-mixed settings, Adelani et al. (2022) and Muhammad et al. (2025) emphasise the scarcity of annotated multilingual corpora, validating the use of lightweight, interpretable models like logistic regression for practical deployment.

Thus, using logistic regression with TF-IDF features is both effective and computationally efficient for hate speech detection in code-mixed data. Its interpretability supports transparent moderation practices. This is crucial in multilingual societies where ethical implications of misclassification can be significant (Vidgen & Derczynski, 2020).

**Conclusion and Future Work**

This study developed hate speech detection models in code-mixed Hausa tweets using logistic regression and random forest classifiers. The study used Bag-of-Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF) for feature representation. The models were evaluated using a hate speech dataset that comprises tweets from the HERDPhobia (Aliyu et al., 2022) and AfriSenti (Muhammad et al., 2025) Hausa code-mixed datasets. Both models performed strongly, with logistic regression (TF-IDF) emerging as the most effective model for Hausa code-mixed hate speech detection. The study also used hyperparameter tuning for optimisation, which enhanced classification performance, particularly for random forest. The overall results show that using TF-IDF and hyperparameter tuning can effectively enhance hate speech classification in Hausa code-mixed text.

To enhance hate speech detection in code-mixed speech, future work should:

i. Develop a large-scale, balanced, and diverse dataset: Collect more balanced samples across Hausa, Yoruba, Igbo, and Pidgin tweets to reduce skewness in hate speech detection. Resources like AfriHate (Muhammad et al., 2025) can be expanded and integrated with other hate speech datasets to form a large-scale hate speech dataset.
ii. Develop Code-Mix-Aware Tokenisers: Standard tokenisation often fails to separate mixed-language constructs. Hybrid tokenisers trained on Nigerian social media text would improve preprocessing accuracy.
iii. Enhance Contextual Understanding: TF-IDF and BoW fail to capture contextual meaning, sarcasm, or implicit hate (e.g., "Shebi you sabi their tribe?"). Unlike transformer-based models, LR and RF cannot infer meaning beyond term frequency. Therefore, combining TF-IDF with semantic embeddings (e.g., fastText, AfriBERTa) or fine-tuning mBERT (Devlin et al., 2019), XLM-R (Conneau et al., 2022), or AfriBERTa (Ogueji et al., 2023) on code-mixed data to capture deeper contextual and multilingual embeddings is necessary.
iv. Deep learning models: Several hateful posts on Nigerian social media include images, memes, or audio clips. Future systems should integrate text and image analysis to detect hate expressed visually or symbolically using deep learning models to enhance the identification of implicit or sarcastic hate content in text, images, audio, and videos, although this will require larger datasets and higher computational power.

**REFERENCES**

Adelani, D. I., Abbott, J., Neubig, G., Dossou, B. F. P., Kreutzer, J., Lignos, C., Palen-Michel, C., Buzaaba, H., Rijhwani, S., Ruder, S., & Adewumi, T. (2022). Masakha NER 2.0: Africa-centric transfer learning for named entity recognition. Transactions of the Association for Computational Linguistics (TACL), 10(1), 1467–1484. https://doi.org/10.1162/tacl_a_00524

Adewumi, T. O., Adebara, I., & Adelani, D. I. (2022). Towards benchmark datasets for African language hate speech detection. Language Resources and Evaluation Conference (LREC), 1678–1686.

Akpan, E. O. B., & Targema, T. S. (2022). Social Media, Mass Mobilisation and National Development in Nigeria: Lessons from the #EndSARS Protest. *ASEAN Journal of Community Engagement*, *6*(2), 228-243. https://doi.org/10.7454/ajce.v6i2.1166

Aliyu, S. M., Wajiga, G. M., Murtala, M., Muhammad, S. H., Abdulmumin, I., & Ahmad, I. S. (2022). Herdphobia: A dataset for hate speech against Fulani in Nigeria. *arXiv preprint arXiv:2211.15262*. https://doi.org/10.48550/arXiv.2211.15262

Al Yousef, H. M., Al Jaraedah, A. R., Alomari, N. M., Al-Qawasmeh, S. I., & Al-Natour, M. M. (2025). Beyond Linguistic Gaps: Types of Code-Switching Among Jordanian Bilingual Speakers. *Journal of World Englishes and Educational Practices, 7*(1), 22-31. https://doi.org/10.32996/jweep.2025.7.1.3

Conneau, A., Bapna, A., Zhang, Y., Ma, M., von Platen, P., Lozhkov, A., ... & Johnson, M. (2022). Xtreme-s: Evaluating cross-lingual speech representations. arXiv preprint arXiv:2203.10752.

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). https://doi.org/10.48550/arXiv.1810.04805

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

Jain, A., Jha, S., Agarwal, B., Klemen, M., & Robnik-Šikonja, M. (2025). Hate Speech Detection in Code-Mixed English-Hindi with Bilingual Large Language Models. In *Proceedings of International Conference on Recent Advancements in Artificial Intelligence* (pp. 259-267). Singapore: Springer Nature Singapore. https://doi.org/10.1007/978-981-96-7760-3_18

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimised BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

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 (ICWSM), 11(1), 512–515. https://doi.org/10.1609/icwsm.v11i1.14955

Mozafari, M., Farahbakhsh, R., & Crespi, N. (2020). A BERT-based transfer learning approach for hate speech detection in online social media. Complexity, 2020, 1–12. https://doi.org/10.1155/2020/8828421

Mubarak, H., Al-Khalifa, H., & Al-Thubaity, A. (2022, June). Overview of the OSACT5 shared task on Arabic offensive language and hate speech detection. In *Proceedings of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools with Shared Tasks on Qur'an QA and Fine-Grained Hate Speech Detection* (pp. 162-166). https://aclanthology.org/2022.osact-1.20/

Mubarak, H., Hassan, S., & Chowdhury, S. A. (2023). Emojis as anchors to detect offensive Arabic language and hate speech. *Natural Language Engineering*, *29*(6), 1436-1457. https://doi.org/10.1017/S1351324923000402

Muhammad, S. H., Abdulmumin, I., Ayele, A. A., Adelani, D. I., Dossou, B. F. P., & Kreutzer, J. (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 (NAACL) (pp. 1705–1720). https://doi.org/10.48550/arXiv.2503.09200

Nwafor, E., & Nguyen, M. P. (2025, May). Fostering Digital Inclusion for Low-Resource Nigerian Languages: A Case Study of Igbo and Nigerian Pidgin. In *Proceedings of the Eighth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2025)* (pp. 44-53). 10.18653/v1/2025.loresmt-1.6

Ogueji, K. (2022). *Afriberta: Towards viable multilingual language models for low-resource languages* (Doctoral dissertation, University of Waterloo).

Onyebuchi, C. A., Onwukwalonye, B. C., & Odoh, M. C. (2025). Social media users' awareness, knowledge, attitude and practice on monkeypox in the Southeast, Nigeria. *Nigerian Journal of Communication*, *19*(1), 16–34. https://www.ajol.info/index.php/njcomm/article/view/285548

Vidgen, B., & Derczynski, L. (2020). Directions in abusive language training data: Garbage in, garbage out. PLOS ONE, 15(12), e0243300. https://doi.org/10.1371/journal.pone.0243300

Röttger, P., Vidgen, B., Nguyen, D., & Derczynski, L. (2021). HateCheck: Functional tests for hate speech detection models. Proceedings of the Association for Computational Linguistics (ACL), 41–58. https://doi.org/10.48550/arXiv.2012.15606

Sosimi, A. A., Ipinnimo, O., Folorunso, C. O., Adim, B. A., & Onoyom-Ita, E. (2024). Hate speech identification in West Africa using machine-learning techniques. In Arid Zone Journal of Engineering, Technology & Environment, 20(7), 55–68.

Taiwo, R., Odebunmi, A., & Adetunji, A. (Eds.). (2016). *Analysing language and humour in online communication*. IGI Global.

Tonneau, M., de Castro, P. V. Q., Lasri, K., Farouq, I., Subramanian, L., Orozco-Olvera, V., & Fraiberger, S. P. (2024). NaijaHate: Evaluating hate speech detection on Nigerian Twitter using representative data. *arXiv preprint arXiv:2403.19260*.

Zhang, Z., Robinson, D., & Tepper, J. (2018). Detecting hate speech on Twitter using a convolution-GRU-based deep neural network. Journal of Semantic Web and Information Systems, 14(4), 1–17. https://doi.org/10.1007/978-3-319-93417-4_48

Waseem, Z., & Hovy, D. (2016). Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), 88–93. https://doi.org/10.18653/v1/N16-2013

Zaghouani, W., Mubarak, H., & Biswas, M. R. (2024, May). So hateful! Building a multi-label hate speech annotated Arabic dataset. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 15044-15055). https://aclanthology.org/2024.lrec-main.1308/