FUDMA Journal of Sciences (FJS)
ISSN online: 2616-1370
ISSN print: 2645 - 2944
Vol. 10 No. 1, January, 2026, pp 249 – 252
DOI: https://doi.org/10.33003/fjs-2026-1001-4541

HAUSA HATE SPEECH DETECTION USING LOGISTIC REGRESSION

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

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

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

ABSTRACT

Hate speech on social media poses serious risks to social cohesion, particularly in multilingual and politically sensitive regions such as West Africa. While Natural Language Processing techniques have achieved strong performance for high-resource languages, African languages remain under-represented due to limited annotated data and linguistic complexity. This study investigates hate speech detection in Hausa using traditional machine learning approaches, focusing on interpretability and efficiency in low-resource settings. Experiments are conducted on the AFRIHATE Hausa corpus using Logistic Regression as the primary classifier and Random Forest as a comparative model. Text is represented using Term Frequency-Inverse Document Frequency (TF-IDF) and Bag-of-Words features. Model performance is evaluated using accuracy, precision, recall, and F1-score under stratified cross-validation. Results show that Logistic Regression with TF-IDF features achieves the best overall performance, with an accuracy of 94% and an F1-score of 93%, outperforming Random Forest across feature representations. The findings indicate that simple, interpretable models remain strong baselines for Hausa hate speech detection and offer practical value for content moderation in low-resource African language contexts.

Keywords: Hausa, Hate Speech Detection, Logistic Regression, Random Forest, TF-IDF, Low-resource Languages, African NLP

INTRODUCTION

Hate speech refers to expressions that promote hostility, discrimination, or violence against individuals or groups based on attributes such as ethnicity, religion, gender, or political identity (Davidson et al., 2017). The rapid growth of social media platforms has amplified the reach and impact of such expressions, allowing harmful narratives to spread quickly and widely (Vidgen & Derczynski, 2020; Maikano, 2024). In West Africa, hate speech on platforms such as Twitter, Facebook, and WhatsApp often intensifies during election periods and political crises, frequently targeting ethnic or religious groups and contributing to social tension (Sosimi et al., 2024).

Research on automatic hate speech detection has expanded significantly over the past decade. Early computational studies primarily focused on English-language data and relied on traditional machine learning models trained on lexical features such as n-grams and TF-IDF representations (Waseem & Hovy, 2016; Davidson et al., 2017). Subsequent work introduced deep learning architectures, including convolutional and recurrent neural networks, which improved performance by capturing sequential and semantic patterns in text (Zhang et al., 2018; Mozafari et al., 2020). More recently, transformer-based models such as BERT and its multilingual variants have become dominant due to their ability to model contextual representations (Devlin et al., 2019; Conneau et al., 2022). Although these approaches achieve strong results, they typically require large annotated datasets and substantial computational resources, limiting their applicability in low-resource language settings (Adelani et al., 2022; Vidgen & Derczynski, 2020).

In contrast, African languages remain severely under-represented in Natural Language Processing research due to structural and resource-related constraints (Adelani et al., 2022; Masakhane NLP Community, 2021). Key challenges include the scarcity of labeled data, orthographic variation, dialectal diversity, and frequent code-switching in online communication (Adelani et al., 2022). Hausa, a major Chadic language spoken by tens of millions of people across West Africa, exemplifies these challenges. The language is written in both Latin (Boko) and Arabic (Ajami) scripts, which complicates text normalization and processing (Muhammad et al., 2025). In social media contexts, Hausa content frequently includes informal spellings and code-mixed expressions involving English and Nigerian Pidgin, further increasing linguistic variability (Sosimi et al., 2024; Adelani et al., 2022).

Recent initiatives such as the Masakhane project and the release of multilingual datasets like AFRIHATE have begun to address data scarcity for African languages by providing open, community-driven resources (Masakhane NLP Community, 2021; Adelani et al., 2022; Muhammad et al., 2025). However, empirical studies evaluating baseline machine learning models specifically for Hausa hate speech detection using these newly released resources remain limited, with most existing work focusing either on multilingual benchmarks or broader West African language groupings rather than Hausa in isolation (Adewumi et al., 2022; Sosimi et al., 2024).

This study addresses this gap by systematically evaluating Logistic Regression and Random Forest classifiers for Hausa hate speech detection using the AFRIHATE corpus. The objectives of the are as follows:

1. Assess the effectiveness of Logistic Regression and Random Forest for Hausa hate speech detection.
2. Compare TF-IDF and Bag-of-Words feature representations.
3. Examine whether simple, interpretable models can provide reliable baselines in low-resource African language settings.

MATERIALS AND METHODS

Method

The approach used began with data collection and annotation, followed by preprocessing, feature extraction using TF-IDF and Bag-of-Words (BoW), and model development using logistic regression as the primary classifier. Random Forest was also used as a comparative model to evaluate

performance consistency and validate the robustness of logistic regression.

Datasets

The study uses the Hausa subset of the AFRIHATE corpus (Muhammad et al., 2025), a publicly available multilingual dataset for hate speech and abusive language detection in African languages. The Hausa subset contains 6,644 annotated tweets collected from social media platforms. Annotations were performed by native speakers, with reported inter-annotator agreement between 0.75 and 0.80. Although the dataset includes hate, abusive, and normal categories, this study formulates the task as a binary classification problem by merging the hate and abusive instances into a single hate class. This formulation aligns with common baseline approaches and reduces class sparsity and facilitate model comparison (Adewumi et al., 2022).

Data Preprocessing

Hausa social media text is highly informal and often contains mixed scripts, slang, emojis, and code-switching. The following preprocessing steps were applied (Röttger et al., 2021):

1. Conversion of all text to lowercase.
2. Removal of URLs, usernames, hashtags, punctuation, and numbers.
3. Stop-word removal using a custom Hausa stop-word list
4. Expansion of common informal contractions. Examples include "lfy" → "lafiya".
5. Tokenization using standard word tokenization
6. Stemming and lemmatization to reduce lexical variation
7. Normalization of whitespace
8. Manual normalization of frequent Hausa–English and Hausa-Pidgin code-switched expressions. Example "wallahi bro" to "wallahi dan'uwa".

This cleaning process removed irrelevant characters while preserving linguistically meaningful contents.

Feature Representation

Feature extraction converts text into numerical vectors for model training. Two feature extraction methods were employed:

1. Bag-of-Words (BoW): This represents documents as vectors of token frequencies without considering term importance across the corpus. Although it ignores word order, BoW is effective for identifying frequent hate-related terms (Zhang et al., 2018).
2. Term Frequency–Inverse Document Frequency (TF-IDF): This weighs tokens by their frequency within a document relative to their distribution across the corpus, emphasizing discriminative terms (Mozafari et al., 2020).

Model Development

Machine learning models were developed using logistic regression and random forest algorithms. Logistic regression was chosen as the baseline model due to its interpretability, computational efficiency, and suitability for high-dimensional data (Mozafari et al., 2020).

The models were implemented using the scikit-learn library. Hyperparameter tuning was performed to assess performance stability. For Logistic Regression, the regularization parameter was adjusted, while Random Forest tuning focused on the number of trees, maximum depth, and minimum sample constraints. The optimization targeted key parameters affecting model depth and generalization, as shown in Table 1.

Table 1: Hyperparameter Options Considered for RF Optimization

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

Model performance was evaluated using stratified five-fold cross-validation. Accuracy, precision, recall, and F1-score were computed for each fold and averaged as displayed in Equations 1 to 4.

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

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

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

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

Confusion matrices were used to analyze correct identifications (True Positives and True Negatives) and classification errors (False Positives and False Negatives), focusing on the latter due to their implications for content moderation (Röttger et al., 2021). True Positives (TP) are the number of correctly classified instances of hate speech. True Negatives (TN) are the number of correctly classified instances of normal speech False Positives (FP) are the number of incorrectly classified instances of hate speech. False Negatives (FN) are the number of incorrectly classified instances of normal speech.

RESULTS AND DISCUSSION

Table 2 presents the confusion matrix results for all model configurations. Logistic Regression with TF-IDF features produced the highest number of correct classifications (TP=590, TN=654) and the lowest misclassification rates (FP=33, FN=52). Performance declined slightly when Bag-of-Words features were used due to higher number of classification errors (FP=42, FN=61). Random Forest with TF-IDF features also produced higher number of correct classifications (TP=575, TN=633) than misclassification rates (FP=53, FN=68), but the performance also declined when Bag-of-Words features were used due to higher number of classification errors (FP=142, FN=146). Hyperparameter tuning led to modest improvements for both models, with more noticeable gains for Random Forest under sparse feature representations.

Table 2: Confusion Matrices Results

Model	Feature Rep.	TP	TN	FP	FN	Total
		Before Optimization				
Logistic Regression	TF-IDF	590	654	33	52	1329
	Bag-of-Words	580	645	42	61	1329
Random Forest	TF-IDF	575	633	53	68	1329
	Bag-of-Words	497	544	142	146	1329
		After Optimization				
Logistic Regression	TF-IDF	593	658	29	49	1329
	Bag-of-Words	584	649	38	58	1329
Forest	TF-IDF	586	643	44	56	1329
	Bag-of-Words	556	606	81	86	1329

Table 3 summarizes performance metrics. Logistic Regression with TF-IDF achieved an accuracy of 94%, precision of 95%, recall of 92% and an F1-score of 93%, with slight decline in performance when Bag-of-Words features were used. Random Forest achieved an accuracy of 91%, precision of 92%, recall of 89% and an F1-score of 90%, with sharp decline in performance when Bag-of-Words features were used. Random Forest achieved a maximum F1-score of 92% after optimization. Hyperparameter tuning resulted in larger improvements observed for Random Forest when using Bag-of-Words features. Across all experiments, TF-IDF consistently outperformed Bag-of-Words.

Table 3: Classification Performance

Model	Feature Rep.	Accuracy	Precision	Recall	F1-score
		Before Optimization			
Logistic Regression	TF-IDF	94%	95%	92%	93%
Logistic Regression	Bag-of-Words	92%	93%	90%	92%
Random Forest	TF-IDF	91%	92%	89%	90%
Random Forest	Bag-of-Words	78%	78%	77%	78%
		After Optimization			
Logistic Regression	TF-IDF	94%	95%	92%	93%
Logistic Regression	Bag-of-Words	93%	94%	91%	92%
Random Forest	TF-IDF	92%	93%	91%	92%
Random Forest	Bag-of-Words	87%	87%	87%	87%

The results demonstrate that traditional machine learning models remain effective for hate speech detection in Hausa, despite linguistic complexity and limited resources. Logistic Regression consistently outperformed Random Forest across feature representations, particularly when combined with TF-IDF features. This outcome reflects the capability of TF-IDF to emphasize linguistically informative terms while reducing the impact of frequent but uninformative tokens. Random Forest showed higher sensitivity to feature sparsity, especially with raw frequency representations. While ensemble methods can capture nonlinear relationships, their effectiveness is limited when training data is small and feature spaces are highly sparse, as is common in short social media texts. An important advantage of Logistic Regression is interpretability. Model coefficients can be inspected to identify tokens strongly associated with hateful content, which is critical for transparency, accountability, and trust in moderation systems. In sociopolitical contexts such as Nigeria, explainable models are particularly important for policy adoption and ethical deployment.

CONCLUSION

This study evaluated traditional machine learning approaches for Hausa hate speech detection using the AFRIHATE corpus. Logistic Regression combined with TF-IDF features achieved the best overall performance, demonstrating that simple, interpretable models remain strong baselines in low-resource African language contexts. The findings contribute empirical evidence supporting the continued relevance of lightweight models for practical hate speech moderation.

The limitations of this study are the use of a small dataset compared to high-resource benchmarks, which may limit generalizability and the preprocessing pipeline relies partly on English-centric tools that may not fully capture Hausa morphology or dialectal variation. Future work will expand the Hausa hate speech dataset, explore multilingual transformer-based models, improved Hausa-specific preprocessing, and finer-grained classification schemes to better capture linguistic nuance and sociocultural context.

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). MasakhaNER 2.0: Africa-centric transfer learning for named entity recognition. Transactions of the Association for Computational Linguistics, 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. In Proceedings of LREC (pp. 1678–1686).

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*.

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of

offensive language. In Proceedings of ICWSM (pp. 512–515). https://doi.org/10.1609/icwsm.v11i1.14955

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 (pp. 4171–4186). https://doi.org/10.48550/arXiv.1810.04805

Maikano, F. A. (2024). Machine Learning Approaches for Cyber Bullying Detection In Hausa Language Social Media: A Comprehensive Review And Analysis. *FUDMA Journal of Sciences*, *8*(3), 344-348.

Masakhane NLP Community. (2021). Building open, community-driven resources for African languages. In Proceedings of the ACL Workshop on African NLP.

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

Muhammad, S. H., Abdulmumin, I., Ayele, A. A., et al. (2025). AFRIHATE: A multilingual collection of hate speech and abusive language datasets for African languages. In Proceedings of NAACL (pp. 1705–1720).

PeaceTech Lab. (2017). Social media hate speech lexicons: Nigeria. Washington, DC: PeaceTech Lab.

Röttger, P., Vidgen, B., Nguyen, D., & Derczynski, L. (2021). HateCheck: Functional tests for hate speech detection models. Proceedings of 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. Arid Zone Journal of Engineering, Technology & Environment, 20(7), 55–68.

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

Waseem, Z., & Hovy, D. (2016). Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In Proceedings of NAACL-HLT (pp. 88–93). https://doi.org/10.18653/v1/N16-2013

Zhang, Z., Robinson, D., & Tepper, J. (2018). Detecting hate speech on Twitter using a convolution-GRU based deep neural network. In Proceedings of the European Semantic Web Conference (pp. 745–760). https://doi.org/10.1007/978-3-319-93417-4_48