JOURNAL OF SCIENCE, TECHNOLOGY AND EDUCATION (JSTE)

A PUBLICATION OF THE DEPARTMENT OF SCIENCE, TECHNOLOGY AND MATHEMATICS EDUCATION (STME), NASARAWA STATE UNIVERSITY, KEFFI, NIGERIA

VOLUME 10

ISSN: 2651-5539

JOURNAL OF SCIENCE, TECHNOLOGY AND EDUCATION (JSTE) VOL. 10. NO. 11

JSTE PUBLICATIONS, 2026 www.nsukjste.com/

ISSN: 2651-5539

MULTILABEL BINARY CLASSIFICATION HATE SPEECH DETECTION IN HAUSA TWITTER DATA USING PRE-TRAINED TRANSFORMER MODELS

[1]Aliyu M., [2]Garko A. B., [3]Awwalu J., and [4]Rabiu A. M.
[1'2'3'4]Department of Computer Science, Faculty of Computing, Federal University Dutse, Jigawa State
Corresponding author: abgarko@gmail.com

Citation: Aliyu M., Garko A. B., Awwalu J., & Rabiu A. M. (2026). Multilabel binary classification hate speech detection in Hausa twitter data using pre-trained transformer models. *Journal of Science, Technology, and Education (JSTE);* www.nsukjste.com/. *10*(11), 139-153.

Abstract

This study presents the development of a Hausa-specific hate speech detection model, HausaBERT, designed to address the limitations of existing multilingual transformers in understanding low-resource African languages. Data were collected from Twitter using language and keyword filters to extract Hausa tweets containing ethnoreligious, political, and personal insults. After expert-guided annotation and validation, the dataset underwent automated preprocessing, removing duplicates, cleaning text, and normalizing linguistic structures. Data augmentation was employed to balance classes at approximately 36,000 instances per category. Four baseline transformer models namely BERT, mBERT, RoBERTa, and DistilBERT were fine-tuned and compared with the proposed HausaBERT ensemble model using weighted precision, recall, and F1-score as evaluation metrics. The results revealed that HausaBERT outperformed some of the base models with a Weighted F1-score of 0.96 and an ensemble accuracy of 0.78, demonstrating superior adaptability to Hausa linguistic patterns. The ensemble approach also achieved a ROC-AUC of 0.9611, indicating robust classification capability across hate speech subcategories. These findings confirm the effectiveness of transformer-based transfer learning for hate speech detection in low-resource languages. The study contributes a reproducible Hausa hate-speech corpus and modeling framework, providing a foundation for future research in African language NLP, responsible AI, and culturally aware content moderation.

Keywords: Hate Speech, Transformer Models, Hausa, BERT, NLP, Machine Learning

Introduction

Hate speech on social media platforms like Twitter has become a growing concern globally, with significant implications for social cohesion and individual well-being. Hate speech refers to forms of

communication that are offensive, inflammatory, and promote intolerance, often leading to division and the demeaning of others. Although there is no universally accepted legal definition, the term is commonly understood to include speech, actions, or written content that expresses prejudice or discrimination against individuals or groups. This discrimination is typically based on characteristics such as religion, ethnicity, nationality, origin, race, skin color, gender, or other aspects of identity (Vashistha & Zubiaga, 2021). While substantial research exists for high-resource languages like English, low-resource Nigerian languages (such as Hausa, Yoruba, and Igbo) remain understudied despite their widespread use in online communications (Tonneau et al., 2024; Mutanga et al., 2020). This research addressed this gap by developing hate speech detection models for Hausa language using pre-trained transformer models.

Statement of the Problem

Despite Nigeria's linguistic diversity, most hate speeches detection systems are built for English, leaving indigenous languages underrepresented. As a result, harmful speeches in these languages often spread unchecked on platforms like Twitter. This poses risks to national cohesion and public safety (Tonneau et al., 2024; Mutanga et al., 2020). The following are some of the challenges:

I. Lack of effective tools for hate speech detection in Hausa language, lacks datasets and language-specific NLP tools. A study conducted by Adam et al. (2025) This makes it difficult to train reliable hate speech detection systems. Most social media content in the Hausa language remains unmoderated. As a result, harmful speech often spreads without detection.

II. Existing models focus on high-resource languages: Most pre-trained models are developed for high-resource languages like English (Adam et al., 2025). When applied to Hausa language, their performance drops significantly due to vocabulary gaps and linguistic differences. This creates a major barrier to accurate detection. It limits the applicability of global models to local contexts.

III. Socio-political risks of unmoderated hate speech: Nigeria's ethnic and religious diversity

makes it highly sensitive to divisive language. Unchecked hate speech can incite violence, deepen mistrust, and destabilize communities. Social media amplifies these messages rapidly. Addressing this threat is crucial for national peace and security (Mutanga et al., 2020).

Objective of the Study

To compare the performance of existing transformer models with the proposed ensemble model on Hausa Multi-label hate speech dataset.

Research Question

What are the comparative performance outcomes of transformer-based models such as BERT, mBERT, DistilBERT, and RoBERTa compared with the proposed ensemble model in Multi-label Hausa hate speech detection?

Literature Review

Traditional ML and Deep Learning Methods for Hate speech detection

Antypas and Camacho-Collados (2023) identified that Traditional ML methods for hate speech detection often involve feature engineering and classification algorithms such as Naive Bayes, Logistic Regression (LR), Support Vector Machines (SVM), and Random Forest (RF). For instance, one study derived classification options from tweet content and grammatical dependencies to identify "othering" phrases and incitement to antagonistic action. Another work utilized SVM with N-gram features for toxic tweet detection Jahan and Oussalah (2023) With the advent of deep learning (DL), more sophisticated models like Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Long Short-Term Memory (LSTM) networks have been employed (Hate Speech Detection Using LSTM and Machine Learning Models. Deep learning technologies have increasingly gained prominence since 2017, with a preference for various word embeddings combined with architectures like Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN). Comparative studies by Jahan and Oussalah (2023) have demonstrated the merits of deep learning models, including CNN, RNN, and GRU, using embedding like Word2Vec, GloVe, and FastText, over traditional ML models. Furthermore, the concatenation of two or more deep learning models, such as CNN + LSTM and CNN + GRU, has shown improved performance compared to single models. Deep neural network structures have also been proposed as effective feature extractors for capturing

the semantics of hate speech, outperforming other methods on Twitter datasets.

Transformer-based Models for Hate Speech Detection

Transformer models like Multilingual BERT (mBERT) and RoBERTa have demonstrated strong performance across a wide range of languages, including some low-resource ones, due to their pre-training on vast amounts of text in multiple languages (Insights into Low-Resource Language Modelling: Improving Model Performances for South African Languages, Cross-lingual transfer of multilingual models on low resource African Languages, Transfer Learning and Distant Supervision for Multilingual Transformer Models: A Study on African Languages(Narula & Chaudhary, 2024)

The transformer architecture, introduced in 2017, has revolutionized NLP and serves as the foundation for models used in hate speech detection. After the introduction of contextual relations-based models like BERT in 2018, several works claimed BERT's superior performance over earlier models such as ELMO, CNN, and RNN. BERT-based models have also been ranked among the top deep-learning models in hate speech detection competitions. Fine-tuning language models on different hate speech detection datasets, including transformer-based models, can contribute to the development of robust hate speech detection models Fredric Ng'ang'an et al. (2024)

Fetahi et al. (2025) explored hate speech detection in Albanian, a language with significant dialectal variation, using XLM-RoBERTa combined with explainable AI tools such as SHAP and LIME. Their model achieved an F1-score of 86%, setting a new benchmark for low-resource hate speech detection.

Similarly in multilingual settings, pre-trained language models have been used in research conducted by Aliyu et al. (2025) to evaluate their efficacy in detecting offensive language in newly created datasets for Nigerian languages like Hausa, Yoruba, and Igbo, achieving an accuracy of up to 90%. Transformer-based models have also been explored for hate speech detection in code-switched social media text, such as English-Kiswahili. These models learn shared linguistic features, enabling cross-lingual transfer capabilities where knowledge from resource-rich languages can be leveraged for low-resource language(Sreeja & Bharathi, 2025)

The most dominant methodological trend is the application of transformer-based models, such as BERT, RoBERTa, and DistilBERT, often fine-tuned for specific tasks or languages Ahmed et al. (2022). These models consistently outperform traditional machine learning methods like Naïve Bayes and Decision Trees(Omran et al., 2023; Sosimi et al., 2024), especially in terms of F1-score and contextual understanding. A few studies explore hybrid models combining CNNs, LSTMs, and attention mechanisms to process multimodal data (Prabhu & Seethalakshmi, 2025; Sreeja K & Bharathi B, 2025), which show promising results but are still largely experimental.

Recent research in natural language processing by Ramos et al. (2024) shows that transformer-based models have become the dominant approach for automated hate speech detection because of their ability to learn deep contextual representations of text. Unlike traditional machine learning models that rely on handcrafted features, transformers use self-attention mechanisms to capture relationships between all words in a sentence simultaneously, improving detection of implicit, context-dependent, or subtle insult patterns common in social media text. A comprehensive review of automatic hate speech detection methods highlights that transformer models such as BERT and its variants significantly outperform classical machine learning and earlier deep learning approaches on a wide range of hate speech datasets and tasks (Solanki et al., 2025).

Many studies emphasize overall accuracy as the primary evaluation metric, neglecting performance indicators that are more suitable for imbalanced data such as Precision, Recall, and F1-score (Ranjan et al., 2025; Fetahi et al., 2025). This practice obscures poor model sensitivity to minority hates classes and gives a misleading impression of performance, especially in datasets where non-hate content dominates.

Methodology

The methodological framework of this study followed a systematic sequence of processes designed to achieve accurate and reliable detection of Hausa hate speech. Below are the processes involved in the development of dataset, see figure 3.1 for diagrammatical workflow

Publicly available Hausa-language tweets were collected using the X (formerly Twitter) API v2 through the official X Developer Platform. Authentication was carried out using OAuth 2.0 Bearer Token access. Data retrieval employed the API

search endpoints with carefully designed keyword and hashtag queries reflecting Hausa social, religious, ethnic, and political discourse.

Preprocessing process was automated through embedded Python functions that systematically cleaned and prepared the Hausa text data for analysis. Functions such as cleanHtml(), cleanPunc(), and keepAlpha() were implemented to automatically remove HTML tags, punctuation marks, and non-alphabetic symbols from the tweets.

A manual annotation process was carried out by six (6) native Hausa language students from the department of Hausa Language, Jigawa State College of Education and Legal Studies, Ringim, under the supervision of three (3) subject experts Hausa Language Course lecturers who also served as validators, the annotators labeled each tweet according to predefined categories such as:

i. Hate Speech (HS)
ii. Abusive Language
iii. Hate Target (Individual, Group, Religion, Ethnicity, Gender, etc.)

To guarantee annotation reliability, a label verification procedure was conducted. The annotations were reviewed by same subject lecturers who oversees the annotations to ensure consistency and reduce human bias. Disagreements were resolved through consensus, and the final dataset was affirmed by Hausa Language Lecturers as accurate and valid for model training.

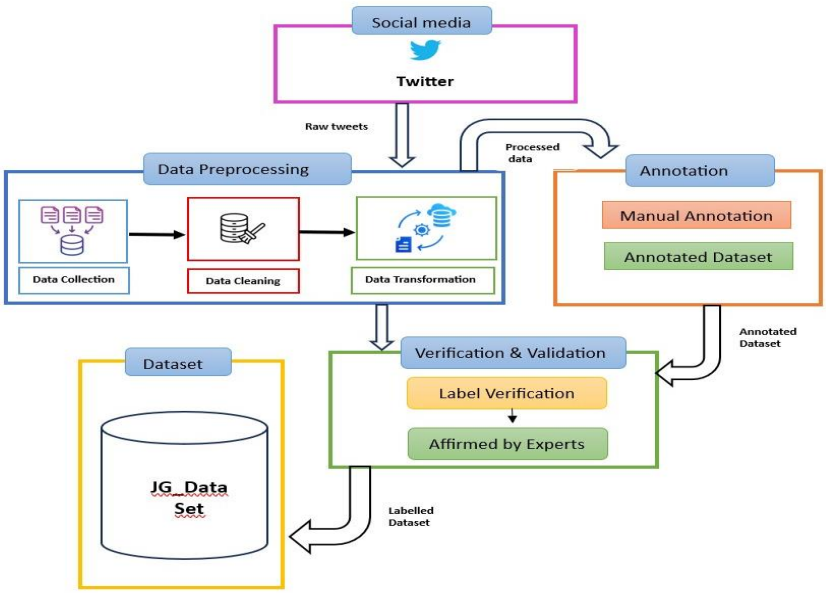

Figure 3.1: Dataset Development Work Flow

Dataset Augmentation

The original 2,794 Hausa Hate Speech dataset was highly imbalanced across its nine annotation categories (HS, Abusive, HS_Individual, HS_Group, HS_Religion, HS_Ethnic, HS_Physical, HS_Gender and HS_Other). To mitigate class imbalance in the JG_Dataset, an automated data augmentation module was embedded within the preprocessing pipeline. Easy Data Augmentation (EDA) techniques random deletion, random swap, and random insertion was employed to ensure sufficient representation and improved model learning stability. These augmented samples simulated natural language variations by reintroducing linguistic diversity and underrepresented hate categories. Figure 3.2 shows augmented representation of the dataset by positive and negative samples per label

Although the design goal was to ensure per-label balance (2,000 positives + 2,000 negatives), the implemented pipeline operated in a label-isolated manner:

1. For each label column *independently*, a 4,000-row subset was constructed consisting of:
    - Up to 2,000 positive samples (using oversampling + text augmentation where necessary)
    - Up to 2,000 negative samples (using under sampling where necessary)
2. Each of these nine balanced subsets was then concatenated vertically into a final dataset containing 36,000 total rows (9 × 4,000).

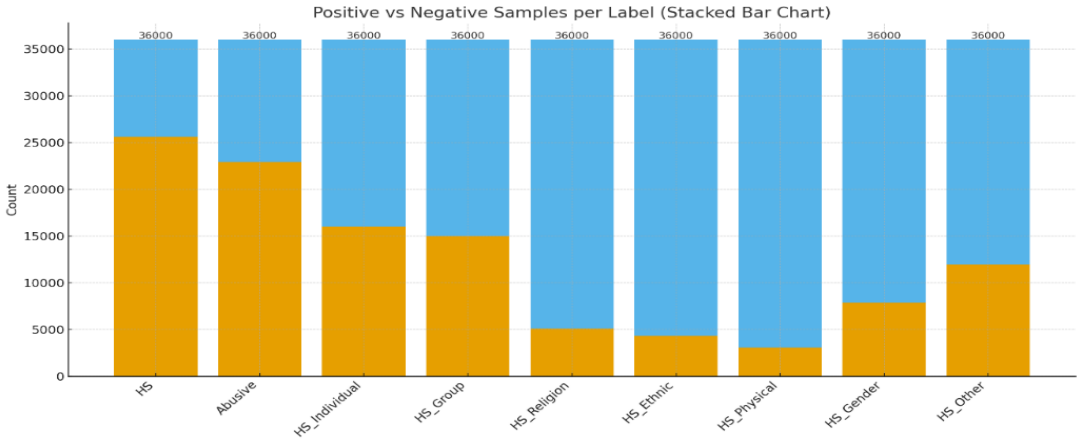

Figure 3.2 Augmented dataset distribution by Positive and Negative Samples per label

Model Selection and Training

In this study, BERT, mBERT, RoBERTa, and DistilBERT were adopted as selected transformer architectures for transfer learning in detecting Hausa hate speech. These models were chosen due to their strong contextual understanding, adaptability, and proven success in various natural language processing tasks (Solanki et al., 2025). Almaliki (2023) DistilBERT, a lighter and faster version of BERT, was selected to reduce computational cost while maintaining competitive performance. However, Solanki et al. (2025) emphasized that transformer models are selected due to their proven effectiveness in low-resource settings through transfer learning, and their ability to capture contextual meaning crucial for hate speech detection

Ensemble Model: HausaBERT (Proposed Model)

To further improve classification performance and leverage the complementary strengths of individual transformer models, an ensemble model named HausaBERT was developed. HausaBERT integrates the learned representations and probabilistic outputs from the four fine-tuned transformer models using a meta-learning approach. Figure 3.3 represents HausaBERT Ensemble Framework

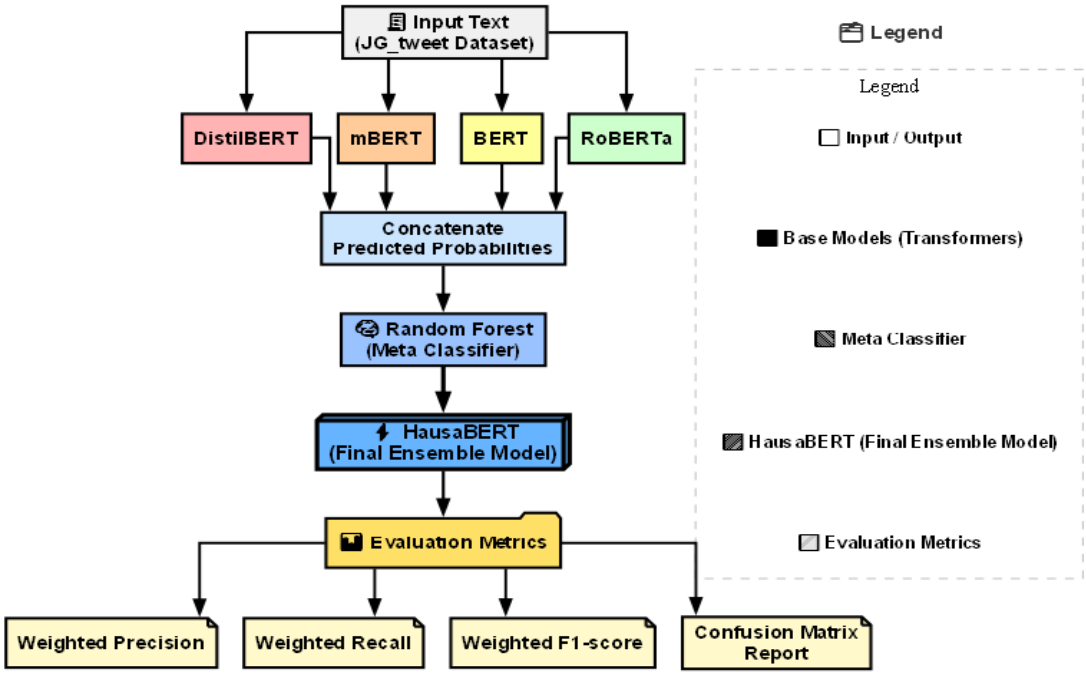

Figure 3.3: HausaBERT Transformer Ensemble Framework

Model Training and Evaluation

Following data preprocessing and augmentation, optimized dataset of 36,000 samples was divided into training (70%), validation (15%), and testing (15%) subsets using the train-test-split function from the *scikit-learn* library. Model training was conducted using the Hugging Face Transformers library within a Python environment Table 3.1 illustrate the hyper-parameters for fin-tuning the baseline models.

Table 3.1: Hyper-parameters used for Fine-tuning

Model	Learning Rate	Training Batch Size	Validation Batch Size	Test batch size	Epochs	Max Sequence Length	Optimizer
BERT	1e-05	8	8	8	7	256	AdamW
mBERT	1e-05	8	8	8	7	256	AdamW
DistilBERT	1e-05	8	8	8	5	256	AdamW
RoBERTa	2e-05	8	8	8	7	256	AdamW

Model Training

Figures 4.4 and 4.5 show that all transformer models steadily improved in training and validation accuracy, indicating effective learning and generalization. DistilBERT and BERT converged fastest, while mBERT and RoBERTa had slightly slower starts but reached similar final accuracy. Consistently decreasing training and validation losses confirm model stability, with all models learning meaningful hate speech representations for the HausaBERT ensemble.

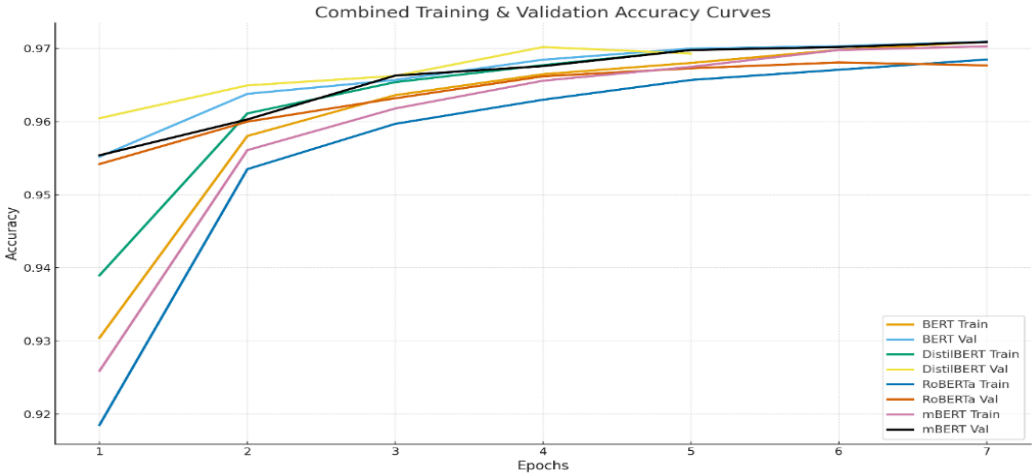

Figure: 3.4: Training and Validation Accuracy of the baseline Transformer Models

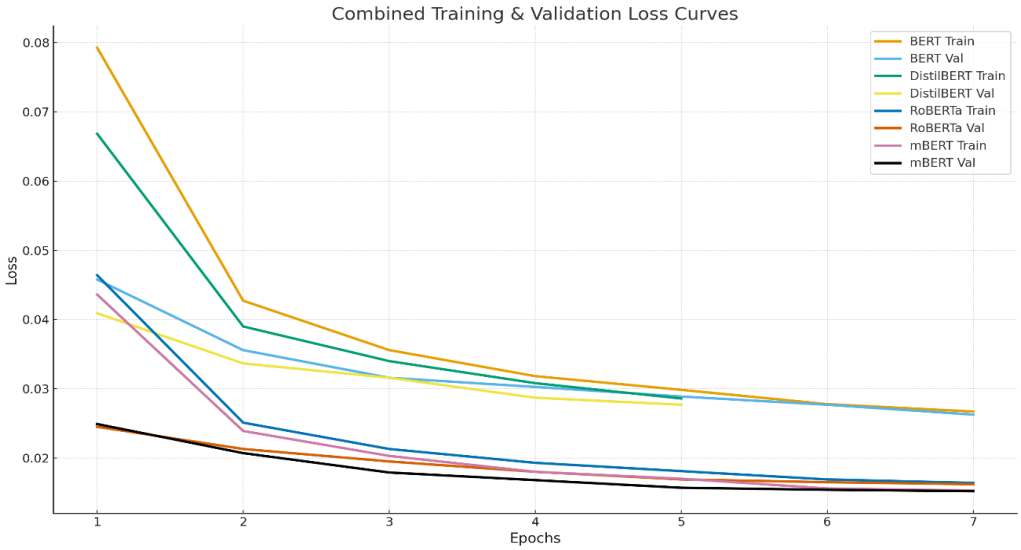

Figure: 3.5: Training and Validation loss of the baseline Transformer Models

Software Configuration

The experiments were implemented in Python 3.10 using Google Colaboratory, a cloud-based environment with GPU support. Transformer models were accessed via the Hugging Face Transformers library and trained using PyTorch, while scikit-learn was used for evaluation metrics. Pandas, NumPy, Matplotlib, and Seaborn supported data processing, analysis, and visualization, with datasets and model artifacts stored via Google Drive integration.

Hardware Configuration

All experiments were conducted on Google Colab's virtualized cloud infrastructure, utilizing an Intel Xeon CPU (2 cores) and an NVIDIA Tesla T4 GPU with 16 GB VRAM. The runtime provided 12–16 GB RAM, approximately 68 GB of temporary disk storage, and operated on Ubuntu 20.04 LTS within a Google Cloud–hosted environment.

Results

Model performance was evaluated using both quantitative metrics and diagnostic visualizations to assess classification accuracy and generalization capability across models. The evaluation compared the four baseline transformer models

Table 4.1: Comparative Evaluation of Model Performance

Model	Weighted Precision	Weighted Recall	Weighted F1-Score	Accuracy
BERT	0.96	0.96	0.96	0.9729
DistilBERT	0.96	0.96	0.96	0.9708
RoBERTa	0.97	0.94	0.95	0.9695
mBERT	0.97	0.94	0.95	0.9708
HausaBERT (Proposed)	**0.96**	**0.95**	**0.96**	**0.7826**

The model achieved a high ROC AUC of 0.9611, showing strong discrimination across hate speech categories. Its overall accuracy of 0.7826, with an ensemble error rate of 0.2174, indicates most samples were correctly classified despite some disagreement among base models. Weighted precision, recall, and F1 scores around 0.95–0.96 reflect a strong balance between accuracy and capturing true hate speech cases. The ensemble effectively combines multiple transformers to enhance robustness and generalization. Figure 4.1 shows the confusion matrices of the HausaBERT ensemble metaclassifier per label.

From per-label confusion matrices shown in figure 4.1, high true positive and true negative rates across most categories, particularly HS, HS_Religion, and HS_Physical, with increased misclassification in more context-dependent labels such as HS_Ethnic and HS_Other. This pattern explains the observed disparity between moderate accuracy and high ROC AUC, as the model demonstrates strong discriminative ability across decision thresholds while fixed-threshold classification remains sensitive to subtle and ambiguous hate expressions.

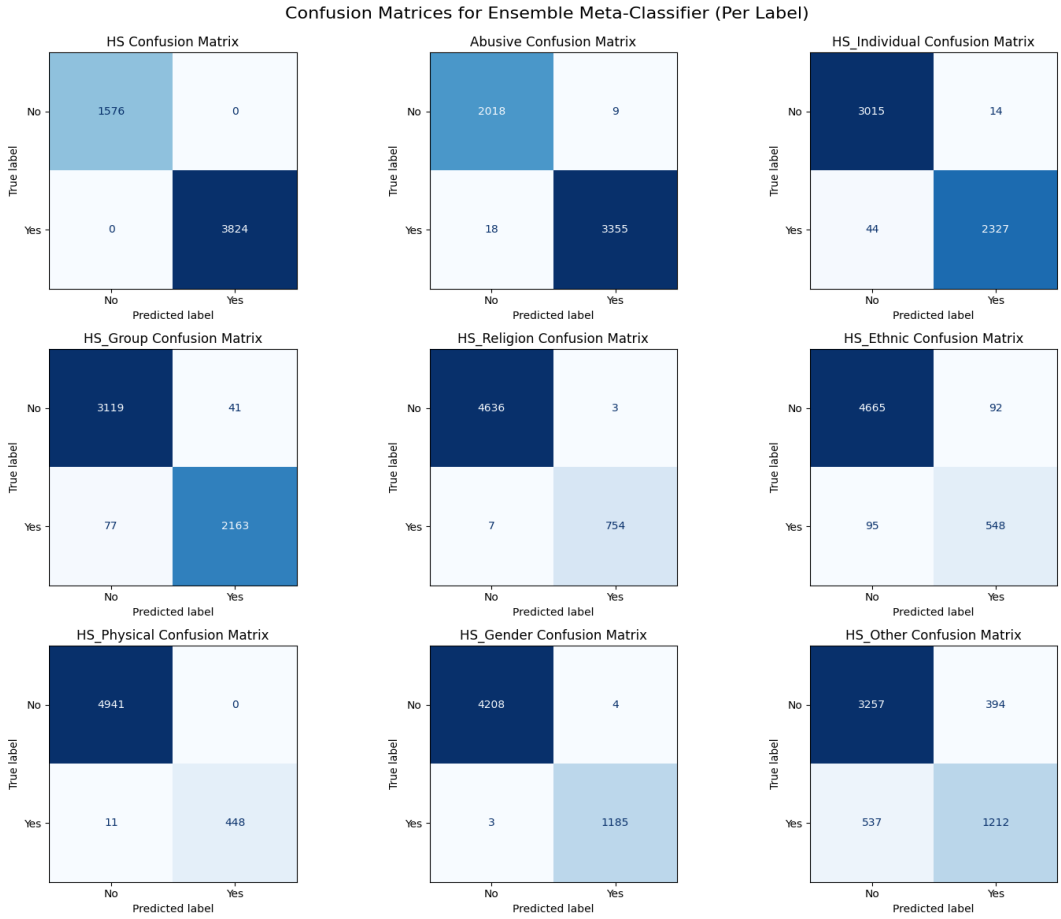

Figure. 4.1: Confusion matrices of the HausaBERT ensemble meta-classifier across individual hate speech labels.

Discussion

The proposed HausaBERT ensemble model, which integrates the outputs of the four base transformers through a Random Forest meta-classifier, achieved a weighted F1-score of 96%, when compared with roBERTa and mBERT with weighted F1 scores of 95% equally, this indicates that HausaBERT produced a stable prediction across all classes, effectively capturing complementary contextual information from multilingual and monolingual models.

HausaBERT benefits from linguistic and cultural alignment that the larger English- and multilingual-based transformers lack. Its ensemble structure further enhances its robustness by combining multiple prediction sources, enabling it to achieve strong weighted performance precision of 0.96, recall of 0.95, and F1-score of 0.96. However, the most impressive metric is the ROC AUC score of 0.9611, which demonstrates the ensemble model's strong ability to distinguish between classes across different thresholds. An AUC near

1.0 indicates excellent discriminative power, meaning the model is highly effective at ranking positive instances higher than negative ones, even if it does not always select the correct decision boundary.

Figure 4.2 reveals the visual comparative performance of the five transformer-based models across major hate speech classes. HausaBERT demonstrates consistently balanced scores across all three metrics, particularly excelling in minority categories such as HS_Ethnic and HS_Others. While baseline BERT and DistilBERT show strong precision, their recall lags in nuanced categories, indicating less adaptability to code-switched Hausa-English expressions. In a nutshell, the figure confirms HausaBERT's superiority in harmonizing precision and recall making it the most reliable model for Hausa hate speech detection.

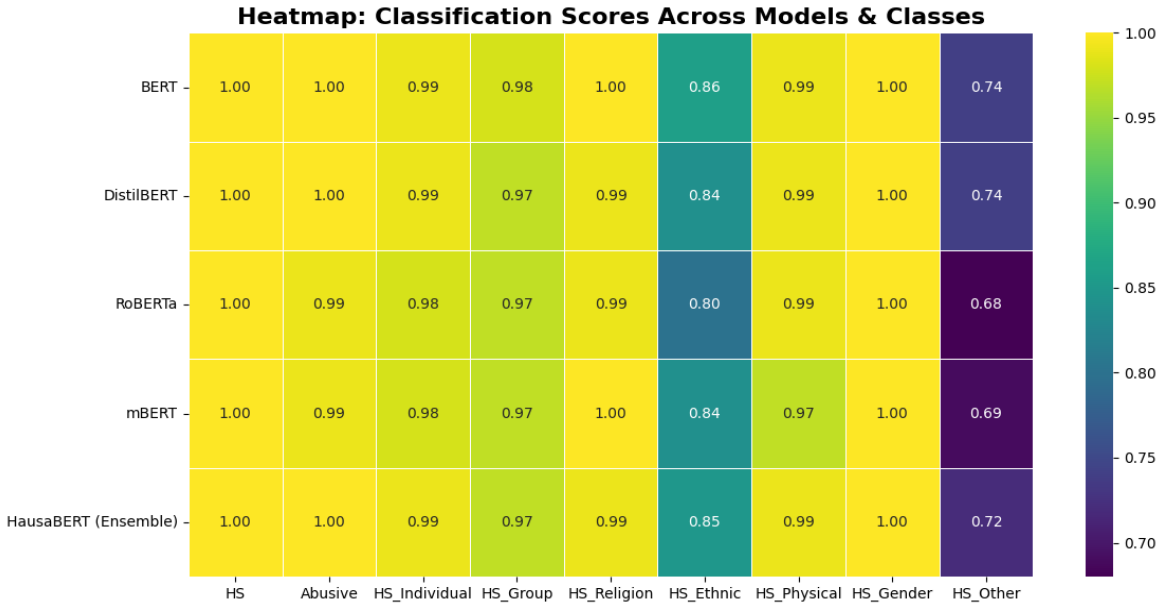

Figure 4.2: Heatmap Models Performance Comparison based on Classes

Conclusion

The analysis demonstrates that HausaBERT is a strong and practically advantageous model, particularly for nuanced hate speech categories where broader contextual awareness is required. While it may not claim the top score in every category, it delivers consistently high performance across all classes and avoids major weaknesses, making it the most well-rounded and robust model among the tested architectures.

References

Adam, F. M., Zandam, A. Y., & Inuwa-Dutse, I. (2025). *Detection and analysis of offensive online content in hausa language*. http://arxiv.org/abs/2311.10541

Aliyu, S. M., Wajiga, G. M., & Murtala, M. (2024). *a multilingual dataset for offensive language and hate speech detection for hausa, yoruba and igbo languages*. https://doi.org/10.48550/arXiv.2406.02169

Almaliki, M. (2023). Cyberhate dissemination: a systematic literature map. *IEEE Access*, *11*, 117385–117392. https://doi.org/10.1109/ACCESS.2023.3326254

Antypas, D., & Camacho-Collados, J. (2023). *Robust hate speech detection in social media: a cross-dataset empirical evaluation*. https://huggingface.co/cardiffnlp/twitter-rober

Fetahi, E., Susuri, A., Hamiti, M., Kastrati, Z., Canhasi, E., & Misini, A. (2025). Enhancing social media hate speech detection in low-resource languages using transformers and explainable AI. *Social Network Analysis and Mining*, *15*(1). https://doi.org/10.1007/s13278-025-01497-w

Fredric Ng'ang'an, Mogeni Oirere, & Njeri Rachael Njeri Ndug'u. (2024). A comparative study of transformer-based models for hate-speech detection in english-kiswahili code-switched social media text. *International Journal of Advanced Trends in Computer Science and Engineering*, *13*(5), 181–186. https://doi.org/10.30534/ijatcse/2024/011352024

Jahan, M. S., & Oussalah, M. (2023). A systematic review of hate speech automatic detection using natural language processing. In *Neurocomputing* (Vol. 546). Elsevier B.V. https://doi.org/10.1016/j.neucom.2023.126232

Mutanga, R. T., Naicker, N., & Olugbara, O. O. (2020). Hate speech detection in twitter using transformer methods. In *IJACSA) International Journal of Advanced Computer Science and Applications* (Vol. 11, Number 9). www.ijacsa.thesai.org

Narula, R., & Chaudhary, P. (2024). Proposed framework for detecting multilingual hate speech on social media platform. *Proceedings of the 2024 3rd Edition of IEEE Delhi Section Flagship Conference, DELCON 2024*. https://doi.org/10.1109/DELCON64804.2024.10867230

Omran, E., Al Tararwah, E., & Al Qundus, J. (2023). A comparative analysis of machine learning algorithms for hate speech detection in social media. *Online Journal of Communication and Media Technologies*, *13*(4). https://doi.org/10.30935/ojcmt/13603

Prabhu, R., & Seethalakshmi, V. (2025). A comprehensive framework for multi-modal hate speech detection

in social media using deep learning. *Scientific Reports*, *15*(1), 1–21. https://doi.org/10.1038/S41598-025-94069-Z;SUBJMETA=166,301,639;KWRD=Engineering,materials+science

Ramos, G., Batista, F., Ribeiro, R., Fialho, P., Moro, S., Fonseca, A., Guerra, R., Carvalho, P., Marques, C., & Silva, C. (2024). A comprehensive review on automatic hate speech detection in the age of the transformer. In *Social Network Analysis and Mining* (Vol. 14, Number 1). Springer. https://doi.org/10.1007/s13278-024-01361-3

Ranjan, R., Ayinala, L., Vatsa, M., & Singh, R. (2025). *Multimodal zero-shot framework for deepfake hate speech detection in low-resource languages*. http://arxiv.org/abs/2506.08372

Solanki, K., Patil, S., & Pranali Patil, A. (2025). Issue 3 www.jetir.org (ISSN-2349-5162). *JETIR2503739. Journal of Emerging Technologies and Innovative Research*, *12*. www.jetir.org

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*.*20*(2), 491–508. www.azojete.com.ng

Sreeja K, & Bharathi B. (2025). *SSNCSE@DravidianLangTech 2025: multimodal hate speech detection in dravidian languages*.

Tonneau, M., Vitor, P., De Castro, Q., Lasri, K., Farouq, I., Subramanian, L., Orozco-Olvera, V., Fraiberger, S. P., & Bank, T. W. (2024). *NAIJAHATE: Evaluating hate speech detection on nigerian twitter using representative data*. https://doi.org/10.18653/v1/2024.acl-long.488

Vashistha, N., & Zubiaga, A. (2021). Online multilingual hate speech detection: Experimenting with hindi and english social media. *Information (Switzerland)*, *12*(1), 1–16. https://doi.org/10.3390/info12010005