Kasu Journal of Computer Science Vol.2 No.4 [December, 2025], pp. 347-360
https://doi.org//10.47514/kjcs/2025.2.4.002

Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

A Comparative Study of Embedding Models for Offensive Language Detection in Nigeria

Saminu Mohammad Aliyu[1] and Shamsuddeen Hassan Muhammad[2]

[1] Department of Computer Science, Bayero University, Kano

[2] Department of Software Engineering, Bayero University, Kano

smaliyu.cs@buk.edu.ng[1], shamsuddeen2004@gmail.com[2]

Abstract

Offensive language detection has become a pivotal challenge in Natural Language Processing (NLP), particularly in the context of low-resource languages. Existing models, predominantly trained on high-resource languages such as English, often struggle to generalize across linguistically and culturally diverse contexts. This study addresses the limitation by developing and introducing novel datasets for the detection of offensive and hate speech in three major Nigerian languages, Hausa, Yoruba, and Igbo. Data was collected from Twitter and manually annotated using native speakers. The study leverages Afro-XLM-R embedding pre-trained on African language corpora combined with a Convolutional Neural Network (CNN) architecture for robust classification. To evaluate its effectiveness, the proposed hybrid model was benchmarked against three alternative CNN-based models utilizing FastText, word2Vec, and mBERT embedding. Experimental results demonstrate the superiority of the Afro-XLM-R + CNN approach in capturing nuanced linguistic features across the selected languages, highlighting its potential for scalable offensive language detection in low-resource settings.

Keywords: Natural language processing, offensive language, pre-trained language models, low-resource languages, embedding

1. Introduction

Offensive and Hate speech remain a threat in the digital era, particularly within multilingual societies like Nigeria, where linguistic diversity intersects with deep-seated cultural and ethnic identities (Ikuelogbon et al., 2025). Online platforms such as Facebook and Twitter (now X) have amplified this challenge by enabling rapid dissemination of offensive and hateful contents across languages (Ridwanullah et al., 2025; Ubong, 2024). Offensive language is a complex, context-dependent phenomenon shaped by the author's intent and societal norms (Alshalabi et al., 2024; Husain & Uzuner, 2021). It encompasses expressions that belittle, insult, mock, or disparage individuals or

Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

groups (Antonio Garcia-Diaz et al., 2023; Fieri & Suhartono, 2023), manifesting in various forms such as cyberbullying, sexism, and abusive language, with hate speech being one of the most severe and extensively studied categories (Caselli et al., 2020; Davidson et al., 2019; Dorris et al., 2020). Although it lacks a universally accepted definition, hate speech is widely understood to involve communication that attacks, demeans, or incites discrimination against people based on protected attributes such as race, religion, gender, ethnicity, or political affiliation (Alkomah et al., 2022; Negi Advocate, 2024; Patil et al., 2023). The impact on victims is profound, often causing significant emotional distress and, in severe cases, escalating into real-world physical conflict (Mahmood & Mohammad, 2024; Saha et al., 2019). This challenge is acutely felt in Nigeria, a richly multicultural nation in West Africa home to over 522 languages. The three most widely spoken indigenous languages are: Hausa in the north, Yoruba in the west, and Igbo in the east, alongside Nigerian Pidgin English, dominate everyday communication, especially on social media (Heaton, 2024; Odeyemi, 2014). Platforms like Twitter are vibrant spaces for conversation in these native languages, yet these interactions are frequently marred by offensive and hateful expressions (Ndabula et al., 2023). The country's history of communal, tribal, and religious conflicts is often exacerbated by the unchecked spread of online hate speech and fake news (Pate & Ibrahim, 2020, Arinzechi et al., 2025), underscoring an urgent need for effective countermeasures. While research on hate speech detection has advanced in English and other high-resource languages, African indigenous languages especially Hausa, Yoruba, and Igbo remain significantly underrepresented in Natural Language Processing research (Inuwa-Dutse, 2025; Usman et al., 2025). This scarcity of resources and tools creates a critical gap in safeguarding digital spaces for millions of speakers. Therefore, this study creates datasets for offensive and hate speech detection in Hausa, Yoruba and Igbo languages. It also conducts a comparative analysis of embedding models in the task of offensive language detection.

2. Related Works

Offensive and hate speech detection has emerged as a critical challenge in the age of social media and online communication platforms. The increasing prevalence of offensive and hate speech has prompted researchers to investigate diverse approaches, including machine learning, deep learning, and transformer-based models, to address the problem effectively. However, most existing studies focus primarily on widely spoken languages such as English. In contrast, African languages like Hausa, Yoruba, and Igbo remain underexplored due to the lack of large annotated corpora, which has significantly limited the advancement of NLP models for these languages.

While English serves as the official language, these three indigenous languages alongside Nigerian

Kasu Journal of Computer Science Vol.2 No.4 [December, 2025], pp. 347-360
https://doi.org//10.47514/kjcs/2025.2.4.002
Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

Pidgin English dominate everyday communication, particularly on social media platforms. The country has witnessed numerous communal, tribal, and religious conflicts, many of which have been linked to the unchecked spread of hate speech online (Pate & Ibrahim, 2020). Twitter remains one of the most widely used social networking platforms in Nigeria, serving as a space where users actively engage in conversations using native languages. These interactions frequently include offensive and hateful expressions (Ndabula et al., 2023).

Recently, there has been a significant increase in research on offensive and hate speech detection in low resource languages like Arabic (Husain & Uzuner, 2021), Indonesia (Ibrohim & Budi, 2019) and India (Bohra et al., 2018). Low-resource African languages are also being explored. These include Amharic: Mossie & Wang, (2020) and Yimam et al., (2019); Afan-Oromo: Kanessa & Tulu, (2021); Ganfure, (2022); Defersha et al., (2022); Defersha & Tune, (2021); Nigeria Pidgin: Ilevbare et al., (2024); Aliyu et al., (2022); Hausa: Gbeyonron, (2024); Adam et al., (2023), Yoruba and Igbo: Ndabula et al., (2023); Tonneau et al., (2024). Some authors treated the problem as a binary classification task: Risch et al., (2020); Pelicon et al., (2021); Mozafari et al., (2022), multiclass: Djandji et al., (2020); Plaza-del-Arco et al., (2022) and multi-label: Ibrohim & Budi, (2019); Omar et al., (2021); Azzi & Zribi, (2023). In terms of approach, many have experimented with classical machine learning algorithms: Abdrakhmanov et al., (2024); Vaidyanathan et al., (2025); Das et al., (2024) and Swain et al., (2022), deep learning models: Hussain et al., (2025) Wei et al., (2021); Roy et al., (2022); and the state-of-the-art transformer models: Naib et al., (2025); Aodhora et al., (2025); Molero et al., (2023); Elmadany et al., (2020); Ranasinghe & Zampieri, (2021); Subramanian et al., (2022).

3. Methodology

This study employed a structured methodology to detect offensive language in Hausa, Yoruba, and Igbo tweets. The design followed a systematic pipeline: datasets creation, preprocessing, annotation, feature representation, models training and then evaluation as illustrated in figure 1.

Kasu Journal of Computer Science Vol.2 No.4 [December, 2025], pp. 347-360
https://doi.org//10.47514/kjcs/2025.2.4.002
Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

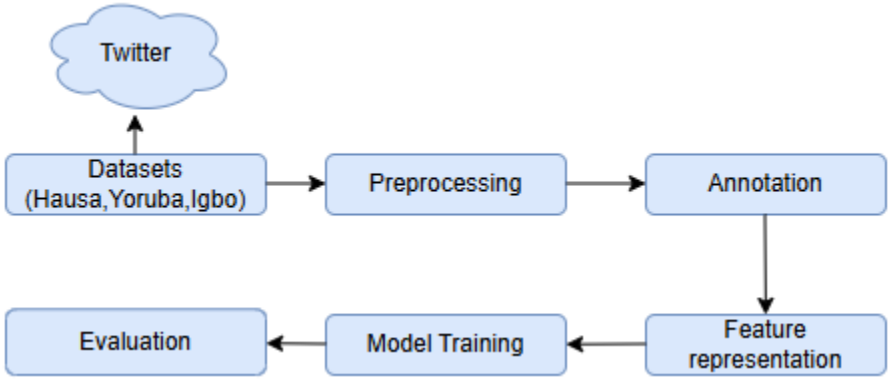

Figure 1. Methodology overview

3.1 Dataset

To gather data relevant for offensive and hate speech detection, this study utilized a keyword-driven strategy targeting three Nigerian languages: Hausa, Yoruba, and Igbo. Keywords associated with offensive and hateful expressions were initially sourced through crowd-sourcing and subsequently vetted by native speakers to ensure both linguistic and contextual validity. These validated keywords informed the construction of search queries, which were used to extract tweets via the Twitter Academic API. For each language, 20,000 tweets were collected. The keyword-based approach was chosen to increase the probability of retrieving tweets containing offensive or hateful content, aligning with methodologies adopted in prior research (Davidson et al., 2019; Roy et al., 2023). All tweets were retained in their original language to preserve authenticity and respect linguistic nuances.

3.2 Pre-processing and Annotation

The collected tweets underwent a comprehensive pre-processing phase, which involved the removal of duplicates and exclusion of tweets written in other languages or deemed unintelligible. To standardize the data, all username mentions, email addresses, and URLs were replaced with placeholders: "@USER," "@EMAIL," and "@URL," respectively. For the annotation process, three native speakers were recruited and trained for each language. An annotation guideline adapted from the framework proposed by Sigurbergsson & Derczynski, (2019) was developed to ensure consistency and clarity. The annotation was structured into two levels.

Level 1: Tweet Categorization

Each tweet was labeled as offensive, hate, indeterminate, or normal. A tweet was considered

Kasu Journal of Computer Science Vol.2 No.4 [December, 2025], pp. 347-360
https://doi.org//10.47514/kjcs/2025.2.4.002
Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

offensive if it contained derogatory language targeting an individual or group, while hate tweets were those that were offensive and based on protected characteristics such as religion, race, or ethnicity. Tweets were marked indeterminate if they were unintelligible or written in a different language, and normal if they were intelligible and free of offensive content.

Level 2: Hate Target Identification

Tweets labeled as hate were further annotated to specify the target of hate, including categories such as religion, ethnicity, gender, disability, and politics. Annotators selecting the "others" category were prompted to manually specify the target type. The dataset is publicly available on github: https://github.com/smaliyu/Offensive_Naija.

Following the training phase, each annotator was assigned a set of 100 tweets for annotation. Inter-annotator agreement (IAA) was calculated using Fleiss' kappa as shown in table 1(Fleiss et al., 1979).

Table 1: Inter-annotator agreement

Language	Score
Hausa	77.39%
Yoruba	82.72%
Igbo	85.42%

Preliminary analysis of the annotated samples revealed that the majority of tweets across all three languages were labeled as normal. To improve the representation of offensive and hate speech categories, a pre-annotation filtering step was introduced, wherein potentially harmful tweets were selectively sampled prior to the main annotation exercise. To determine the final label for each tweet, a simple majority voting strategy was employed, whereby a label was accepted if agreed upon by at least two out of the three annotators. After excluding tweets labeled as indeterminate, the final datasets comprised 6,476 tweets in Hausa, 4,926 in Yoruba, and 4,435 in Igbo as shown in table 2. Each dataset was evaluated independently.

Table 2: Annotated tweets by language

Language	Normal	Offensive	Hate	Total
Hausa	4008	2384	75	6467
Igbo	1079	3019	337	4435
Yoruba	2221	2484	221	4926

Kasu Journal of Computer Science Vol.2 No.4 [December, 2025], pp. 347-360
https://doi.org//10.47514/kjcs/2025.2.4.002
Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

3.3 Feature extraction

The feature extraction consists of two stages: (1) tokenization of raw text into subword units, and (2) the generation of contextual or static word embeddings. Tokenization is the process of segmenting raw text into discrete linguistic units for numerical representation. Given the morphological richness of the target languages and the prevalence of creative or coded spellings in offensive online speech, the subword tokenization was employed for all models. For mBERT and Afro-XLM-R, their respective WordPiece tokenizers, which are designed to handle the models' multilingual vocabularies were used. For fastText and Word2Vec models, the SentencePiece tokenizer with a unigram language model were employed to build a vocabulary of 20,000 subword units directly from our combined training corpora. This ensures the tokenizer is optimized for the specific lexical and morphological characteristics of Hausa, Igbo, and Yoruba text.

3.4 Models training

The task was formulated as a binary text categorization problem because of the low prevalence of tweets explicitly categorized as hate speech. To mitigate the challenges of severe class imbalance and data sparsity for the hate speech class, initial offensive and hateful categories were merged into a single offensive class. As a result, each dataset consists of two classes: offensive and normal. To ensure rigorous model development and evaluation, each dataset was partitioned using a stratified random split into three subsets: 70% for training, 15% for validation, and 15% for testing. Stratification was applied to maintain the original class distribution within each subset, preserving the representativeness of the data. Four distinct hybrid models were developed each integrating a different pre-trained embedding strategy: Afro-XLM-R, mBERT, Word2Vec, and FastText, combined with a parallel multi-branch CNN classifier. This is to enable the capturing of multi-scale linguistic patterns indicative of offensive speech, from local n-gram sequences to broader contextual semantics. At the core of each model is a parallel multi-scale convolutional block, for capturing features at varying granularities. Following the embedding layer, the sequence of token vectors is processed simultaneously by three separate one-dimensional convolutional (Conv1D) layers. These layers employ kernel sizes of 2, 3, and 4, respectively, allowing the model to specialize in detecting discriminative local patterns corresponding to bigrams, trigrams, and 4-grams. This multi-scale approach is particularly advantageous for the task, as offensive constructs can manifest at different syntactic lengths from short abusive slurs to longer, contextually dependent phrases. Each convolutional operation applies a set of learnable filters (64 filters per branch) that slide across the sequence, generating feature maps that highlight the presence of specific local patterns. The outputs

Kasu Journal of Computer Science Vol.2 No.4 [December, 2025], pp. 347-360
https://doi.org//10.47514/kjcs/2025.2.4.002
Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

from all three parallel convolutional branches are subsequently concatenated along the feature dimension. the concatenated feature vector is subjected to a dropout regularization layer with a rate of 0.4. This technique randomly nullifies 40% of the activations during training, forcing the network to develop robust, distributed feature representations rather than relying on any small set of neurons. The regularized feature vector is then fed into a fully connected layer of 64 units with a ReLU activation function. This layer performs a non-linear transformation, integrating the diverse local features extracted by the convolutional branches into a higher-level abstract representation. Finally, a softmax classification layer projects this representation onto the two output classes: offensive and normal. All models were trained using a consistent hyperparameter framework to ensure a fair comparison. The training employed the AdamW optimizer with a learning rate of 2e-5 and a weight decay of 0.01 to promote stable convergence and mitigate overfitting. The training also used a batch size of 12 over 4 epochs, utilizing a cross-entropy loss function. Input sequences were standardized to a maximum length of 128 tokens, determined by the 95th percentile of our training data. The Afro-XLM-R and mBERT leveraged the Trainer API from the HuggingFace transformers library, while the Word2Vec and FastText-based CNNs were implemented in PyTorch, inheriting default library settings for other parameters. A dedicated validation set was used for early stopping and hyperparameter tuning.

3.5 Evaluation metrics

The performances of the developed hybrid models were evaluated using F1-score which is the harmonic mean of precision and recall Accuracy.

Precision: This measures the fraction of the correctly predicted positive instances out of all instances predicted as positive.

$$Precision = \frac{TP}{TP + FP} \quad (1)$$

Recall: This measures the proportion of actual positive instances that the model correctly identified.

$$Recall: \frac{TP}{TP + FN} \quad (2)$$

F1-score: The F1-score is the harmonic mean of precision and recall.

$$F1 - score = \frac{Precision * Recall}{Precision + Recall} \quad (3)$$

Kasu Journal of Computer Science Vol.2 No.4 [December, 2025], pp. 347-360
https://doi.org//10.47514/kjcs/2025.2.4.002
Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

4. Results and discussion

This study examined the performance of four different embedding models combined with a CNN. The architecture was designed to assess whether integrating semantic embeddings with the CNN's ability to capture local n-gram patterns could improve effectiveness in analyzing short, informal social media text. The results in table 3 present performances of the four hybrid models evaluated using the metrics explained in the evaluation metrics section.

Table 3: F1-score for the four hybrid models

Model	Hausa	Igbo	Yoruba
mBERT + CNN	0.848	0.920	**0.903**
Word2Vec + CNN	0.690	0.816	0.802
FastText + CNN	0.588	0.556	0.774
Afro-XLM-R + CNN	**0.871**	**0.936**	0.882

The results prove that contextual embedding models significantly and consistently outperform static embedding models across all languages. This performance gap provides strong evidence that understanding word sense in context is paramount for this task. Static embeddings, which assign a single vector per word, fail to disambiguate polysemous words often used in offensive speech.

The Afro-XLM-R + CNN model achieved the highest F1-score in Hausa (0.871) and Igbo (0.936) languages. Its superior performance, particularly over the more general mBERT in these languages, directly proves the concept that domain-specific pretraining on African languages is beneficial. Afro-XLM-R's vocabulary and training data include these languages, allowing it to better capture morphologically rich structures and culturally nuanced insults that are opaque to models without African language pretraining. The mBERT + CNN model also performed well, even outperforming Afro-XLM-R + CNN in Yoruba language because of the high number of code-mixed texts in the dataset. This supports the alternative concept that a model with massively broad multilingual pretraining can develop robust cross-lingual representations that transfer effectively, even to languages not heavily represented in its pretraining corpus. This underscores the strength of the transformer architecture and its ability to learn generalizable contextual patterns. The results from the static hybrid model shows a consistent performance of Word2Vec over FastText which is a critical analytical finding. Our hypothesis is that this relates to dataset characteristics. The texts are short and informal, with many misspellings and code-mixed words. While FastText's subword information is theoretically advantageous for handling out-of-vocabulary words, in this specific case, it may have introduced ambiguity by generating vectors for non-morphemic character n-grams. Word2Vec's

Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

word-level embeddings, trained directly on the available corpus vocabulary, appear to have provided a more stable and discriminative feature set for the subsequent CNN to process, thus proving more effective for this particular data domain.

5. Conclusion

This research developed hybrid CNN models for offensive language detection in Hausa, Igbo, and Yoruba, establishing crucial benchmarks for these low-resource Nigerian languages. Our findings demonstrate that model performance is directly tied to linguistic suitability, with the Afro-XLM-R + CNN model achieving superior results due to its African-centric pre-training. The notable exception where mBERT + CNN outperformed on Yoruba reflects that dataset's higher English code-mixing, illustrating how sociolinguistic factors critically impact model effectiveness and challenging assumptions about general multilingual models' universal superiority. This study acknowledges some limitations such as: (1) The crowd-sourced keyword method for data collection may introduce sampling bias. (2) Merging the hate and offensive classes, though necessary for model stability, obscures the critical distinction between general offensive and targeted hate speech, which is essential for appropriate real-world intervention. (3) Models trained and evaluated exclusively on data from Twitter (X) may not perform reliably on text from other social platforms such as WhatsApp and Facebook due to differences in spelling conventions and discourse styles. (4) Focusing only on Hausa, Igbo, and Yoruba, and excluding Nigeria's hundreds of other languages and dialects, limits the broader applicability of the findings across the nation's full linguistic diversity.

Acknowledgement

This work was supported by a Research Grant from the Nigeria Artificial Intelligence Research (NAIR) Scheme, administered by the National Information Technology Development Agency (NITDA).

Reference

Abdrakhmanov, R., Kenesbayev, S. M., Berkimbayev, K., Toikenov, G., Abdrashova, E., Alchinbayeva, O., & Ydyrys, A. (2024). Offensive Language Detection on Social Media using Machine Learning. *International Journal of Advanced Computer Science & Applications*, *15*(5).

Adam, F. M., Zandam, A. Y., & Inuwa-Dutse, I. (2023). Detection of offensive and threatening online content in a low resource language. *arXiv Preprint arXiv:2311.10541*.

Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

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*.

Alkomah, F., Salati, S., & Ma, X. (2022). A New Hate Speech Detection System based on Textual and Psychological Features. *International Journal of Advanced Computer Science and Applications*, *13*(8). https://doi.org/10.14569/IJACSA.2022.01308100

Alshalabi, N., Lahiani, H., & Yasin, A. (2024). The Role of Culture in Abusive Language on Social Media: Examining the Use of English and Arabic Derogatory Terms. *Theory and Practice in Language Studies*, *14*(10), 3057–3066.

Antonio Garcia-Diaz, J., Maria Jimenez-Zafra, S., Angel Garcia-Cumbreras, M., & Valencia-Garcia, R. (2023). Evaluating feature combination strategies for hate-speech detection in Spanish using linguistic features and transformers. In *COMPLEX & INTELLIGENT SYSTEMS* (Vol. 9, Issue 3, pp. 2893–2914). SPRINGER HEIDELBERG. https://doi.org/10.1007/s40747-022-00693-x

Aodhora, S. R., Ahsan, S., & Hoque, M. M. (2025). *Cuet_hateshield@ nlu of devanagari script languages 2025: Transformer-based hate speech detection in devanagari script languages*. 260–266.

Arinzechi, M. M., Ibrahim, M., Haruna, K., Daniel, M. S., Ahmed, A., Mustapha, R., & Abdulkari, S. (2025). A scalable machine learning model for enhancement of fake news detection. *Kasu Journal of Computer Science*, *2*(2), 144–159. https://doi.org/10.47514/kjcs/2025.2.2.002

Azzi, S. A., & Zribi, C. B. O. (2023). A new classifier chain method of bert models for multi-label classification of arabic abusive language on social media. *Procedia Computer Science*, *225*, 476–485.

Bohra, A., Vijay, D., Singh, V., Akhtar, S. S., & Shrivastava, M. (2018). A Dataset of Hindi-English Code-Mixed Social Media Text for Hate Speech Detection. *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, 36–41. https://doi.org/10.18653/v1/W18-1105

Caselli, T., Basile, V., Mitrović, J., Kartoziya, I., & Granitzer, M. (2020). *I Feel Offended, Don't Be Abusive! Implicit/Explicit Messages in Offensive and Abusive Language*. 6193–6202.

Das, A., Rahgouy, M., Feng, D., Zhang, Z., Bhattacharya, T., Raychawdhary, N., Jamshidi, F., Jain, V., Chadha, A., & Sandage, M. J. (2024). OffensiveLang: A Community Based Implicit Offensive Language Dataset. *IEEE Access*.

Davidson, T., Bhattacharya, D., & Weber, I. (2019). Racial Bias in Hate Speech and Abusive Language Detection Datasets. *Proceedings of the Third Workshop on Abusive Language Online*, 25–35. https://doi.org/10.18653/v1/W19-3504

Defersha, N. B., Abawajy, J., & Kekeba, K. (2022). Deep Learning based Multilabel Hateful Speech Text Comments Recognition and Classification Model for Resource Scarce Ethiopian Language: The

Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

case of Afaan Oromo. *2022 IEEE International Conference on Current Development in Engineering and Technology (CCET)*, 1–11. https://doi.org/10.1109/CCET56606.2022.10080837

Defersha, N. B., & Tune, K. K. (2021). Detection of Hate Speech Text in Afan Oromo Social Media using Machine Learning Approach. *Indian Journal of Science and Technology*, *14*(31), 2567–2578. https://doi.org/10.17485/IJST/v14i31.1019

Djandji, M., Baly, F., Antoun, W., & Hajj, H. (2020). *Multi-task learning using AraBert for offensive language detection*. 97–101.

Dorris, W., Hu, R., Vishwamitra, N., Luo, F., & Costello, M. (2020). *Towards automatic detection and explanation of hate speech and offensive language*. 23–29.

Elmadany, A., Zhang, C., Abdul-Mageed, M., & Hashemi, A. (2020). Leveraging affective bidirectional transformers for offensive language detection. *arXiv Preprint arXiv:2006.01266*.

Fieri, B., & Suhartono, D. (2023). Offensive Language Detection Using Soft Voting Ensemble Model. *MENDEL*, *29*(1), 1–6. https://doi.org/10.13164/mendel.2023.1.001

Fleiss, J. L., Nee, J. C., & Landis, J. R. (1979). Large sample variance of kappa in the case of different sets of raters. *Psychological Bulletin*, *86*(5), 974.

Ganfure, G. O. (2022). Comparative analysis of deep learning based Afaan Oromo hate speech detection. *Journal of Big Data*, *9*(1), 76. https://doi.org/10.1186/s40537-022-00628-w

Gbeyonron, C. I. (2024). Analysis of Hate Speech in Responses to Two Hausa Online Media Outlets on the Spread of the COVID-19 Pandemic. *Indonesian Journal of English Language Studies (IJELS)*, *10*(1), 56–64.

Husain, F., & Uzuner, O. (2021). A Survey of Offensive Language Detection for the Arabic Language. *ACM Transactions on Asian and Low-Resource Language Information Processing*, *20*(1), 1–44. https://doi.org/10.1145/3421504

Hussain, N., Qasim, A., Mehak, G., Kolesnikova, O., Gelbukh, A., & Sidorov, G. (2025). ORUD-Detect: A Comprehensive Approach to Offensive Language Detection in Roman Urdu Using Hybrid Machine Learning–Deep Learning Models with Embedding Techniques. *Information*, *16*(2), 139.

Ibrohim, M. O., & Budi, I. (2019). Multi-label Hate Speech and Abusive Language Detection in Indonesian Twitter. *THIRD WORKSHOP ON ABUSIVE LANGUAGE ONLINE*, 46–57.

IKUELOGBON, K., IDOWU, O. A., & ALOBA, A. M. (2025). A Critical Discourse Analysis of Selected Hate Comments on X Social Media Platform in the 2023 Presidential Election Nigeria. *Àgídìgbo: ABUAD Journal of the Humanities*, *13*(1), 245–264.

Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

Ilevbare, C. E., Alabi, J. O., Adelani, D. I., Bakare, F. D., Abiola, O. B., & Adeyemo, O. A. (2024). Ekohate: Abusive language and hate speech detection for code-switched political discussions on nigerian twitter. *arXiv Preprint arXiv:2404.18180*.

Inuwa-Dutse, I. (2025). Naijanlp: A survey of nigerian low-resource languages. *arXiv Preprint arXiv:2502.19784*.

Kanessa, L. G., & Tulu, S. G. (2021). Automatic Hate and Offensive speech detection framework from social media: The case of Afaan Oromoo language. *2021 International Conference on Information and Communication Technology for Development for Africa (ICT4DA)*, 42–47. https://doi.org/10.1109/ICT4DA53266.2021.9672232

Mahmood, R. J., & Mohammad, M. H. (2024). Verbal Violence and Hate Speech Related to Political Conflict: Facebook Comments as Example. *Koya University Journal of Humanities and Social Sciences*, *7*(2), 338–349.

Molero, J. M., Pérez-Martín, J., Rodrigo, A., & Peñas, A. (2023). Offensive language detection in spanish social media: Testing from bag-of-words to transformers models. *IEEE Access*, *11*, 95639–95652.

Mossie, Z., & Wang, J.-H. (2020). Vulnerable community identification using hate speech detection on social media. *Information Processing & Management*, *57*(3), 102087. https://doi.org/10.1016/j.ipm.2019.102087

Mozafari, M., Farahbakhsh, R., & Crespi, N. (2022). Cross-Lingual Few-Shot Hate Speech and Offensive Language Detection Using Meta Learning. *IEEE ACCESS*, *10*, 14880–14896. https://doi.org/10.1109/ACCESS.2022.3147588

Naib, M. M., Shohag, M. S. H., Hossain, A., Hossain, J., & Hoque, M. M. (2025). *cuetRaptors@ DravidianLangTech 2025: Transformer-Based Approaches for Detecting Abusive Tamil Text Targeting Women on Social Media*. 739–745.

Ndabula, J. N., Olanrewaju, O. M., & Echobu, F. O. (2023). Detection of Hate Speech Code Mix Involving English and Other Nigerian Languages. *Journal of Information Systems and Informatics*, *5*(4), 1416–1431. https://doi.org/10.51519/journalisi.v5i4.595

Negi Advocate, C. (2024). The Rise of Hate Speech Around the World. *Available at SSRN 4719266*.

Omar, A., Mahmoud, T. M., Abd-El-Hafeez, T., & Mahfouz, A. (2021). Multi-label arabic text classification in online social networks. *Information Systems*, *100*, 101785.

Pate, U. A., & Ibrahim, A. M. (2020). Fake News, Hate Speech and Nigeria's Struggle for Democratic Consolidation: A Conceptual Review. In A. M. G. Solo (Ed.), *Advances in Human and Social Aspects of Technology* (pp. 89–112). IGI Global. https://doi.org/10.4018/978-1-7998-0377-5.ch006

Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

Patil, P., Raul, S., Raut, D., & Nagarhalli, T. (2023). Hate Speech Detection using Deep Learning and Text Analysis. *2023 7th International Conference on Intelligent Computing and Control Systems (ICICCS)*, 322–330. https://doi.org/10.1109/ICICCS56967.2023.10142895

Pelicon, A., Shekhar, R., Škrlj, B., Purver, M., & Pollak, S. (2021). Investigating cross-lingual training for offensive language detection. *PeerJ Computer Science*, *7*, e559.

Plaza-del-Arco, F. M., Molina-González, M. D., Ureña-López, L. A., & Martín-Valdivia, M.-T. (2022). Integrating implicit and explicit linguistic phenomena via multi-task learning for offensive language detection. *Knowledge-Based Systems*, *258*, 109965. https://doi.org/10.1016/j.knosys.2022.109965

Ranasinghe, T., & Zampieri, M. (2021). An evaluation of multilingual offensive language identification methods for the languages of india. *Information*, *12*(8), 306.

Ridwanullah, A. O., Sule, S. Y., Usman, B., & Abdulsalam, L. U. (2025). Politicization of hate and weaponization of Twitter/X in a polarized digital space in Nigeria. *Journal of Asian and African Studies*, *60*(5), 3350–3370.

Risch, J., Ruff, R., & Krestel, R. (2020). Explaining offensive language detection. *Journal for Language Technology and Computational Linguistics*, *34*(1), 29–47.

Roy, P. K., Bhawal, S., & Subalalitha, C. N. (2022). Hate speech and offensive language detection in Dravidian languages using deep ensemble framework. *COMPUTER SPEECH AND LANGUAGE*, *75*. https://doi.org/10.1016/j.csl.2022.101386

Roy, S. S., Roy, A., Samui, P., Gandomi, M., & Gandomi, A. H. (2023). Hateful sentiment detection in real-time tweets: An LSTM-based comparative approach. *IEEE Transactions on Computational Social Systems*, *11*(4), 5028–5037.

Saha, P., Mathew, B., Goyal, P., & Mukherjee, A. (2019). Hatemonitors: Language agnostic abuse detection in social media. *arXiv Preprint arXiv:1909.12642*.

Sigurbergsson, G. I., & Derczynski, L. (2019). Offensive language and hate speech detection for Danish. *arXiv Preprint arXiv:1908.04531*.

Subramanian, M., Ponnusamy, R., Benhur, S., Shanmugavadivel, K., Ganesan, A., Ravi, D., Shanmugasundaram, G. K., Priyadharshini, R., & Chakravarthi, B. R. (2022). Offensive language detection in Tamil YouTube comments by adapters and cross-domain knowledge transfer. *Computer Speech & Language*, *76*, 101404. https://doi.org/10.1016/j.csl.2022.101404

Swain, M., Biswal, M., Raj, P., Kumar, A., & Mishra, D. (2022). Hate and offensive language identification from social media: A machine learning approach. In *Electronic Systems and Intelligent Computing: Proceedings of ESIC 2021* (pp. 335–342). Springer.

Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

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*.

Ubong, B. A. (2024). PUBLIC PERCEPTION OF HATE SPEECH ON SOCIAL MEDIA (FACEBOOK) DURING THE 2023 GENERAL ELECTIONS IN NIGERIA. *African Journal of Social Sciences and Humanities Research*, *Volume 7, Issue 4,* 89–103. https://doi.org/10.52589/AJSSHR-FJL4Z0WQ

Usman, M., Ahmad, M., Sidorov, G., Gelbukh, I., & Tellez, R. Q. (2025). A Large Language Model-Based Approach for Multilingual Hate Speech Detection on Social Media. *Computers*, *14*(7), 279.

Vaidyanathan, V. K., Srihari, V., & Durairaj, T. (2025). *NLP_goats@ DravidianLangTech 2025: Towards Safer Social Media: Detecting Abusive Language Directed at Women in Dravidian Languages*. 350–354.

Wei, B., Li, J., Gupta, A., Umair, H., Vovor, A., & Durzynski, N. (2021). Offensive language and hate speech detection with deep learning and transfer learning. *arXiv Preprint arXiv:2108.03305*.

Yimam, S. M., Ayele, A. A., & Biemann, C. (2019). Analysis of the Ethiopic Twitter dataset for abusive speech in Amharic. *arXiv Preprint arXiv:1912.04419*.