http://dx.doi.org/10.21622/ACE.2025.05.2.1524

MACHINE LEARNING APPROACHES FOR DETECTING HATE SPEECH IN AFRICAN LANGUAGES ON SOCIAL MEDIA: A SYSTEMATIC LITERATURE REVIEW

Banchale Adhi Gufu[1], Audrey Mbogho[2] and Edward Ombui[3]

[1]United States International University, School of Science and Technology, Kenya, Africa

[2]United States International University, School of Science and Technology, Machine Learning, Kenya, Africa

[3]United States International University, School of Science and Technology, Artificial Intelligence, Kenya, Africa

Emails: {agufu@usiu.ac.ke, ambogho@usiu.ac.ke, eombui@usiu.ac.ke}

Received on, 30 July 2025 - Accepted on, 18 October 2025 - Published on, 26 November 2025

ABSTRACT

Significant research efforts have been made towards the development of machine learning models to detect hate speech worldwide. However, for Africa, which is home to over 2,000 languages with diverse dialects, there is an urgent need for inclusive natural language processing (NLP) tools tailored to the continent's linguistic diversity. More specifically, the literature reveals that limited research has been conducted on hate speech detection in African languages, thus providing a strong justification for this systematic literature review. Whereas hate speech has continued to intrigue African communities, detection has been hampered by the complexity of multiple languages, thus calling for a localised approach to solving the problem. This review aims to discover these localized approaches and identify remaining gaps.

The study adopted the PRISMA guidelines for a systematic literature review (SLR), synthesising findings from research published between 2019 and 2024, focusing on machine learning detection techniques in African low-resourced languages. This study advances the Natural Language Processing (NLP) of African languages by comprehensively reviewing the machine learning models and datasets, pinpointing crucial research gaps. It highlights the need for diverse, multi-faceted approaches and collaborative, community-based efforts to effectively combat social media hate speech. The findings reveal that machine learning models, including SVM, BiLSTM, mBERT, and XLM-RoBERTa, show significant potential in detecting hate speech in African languages. However, their performance is often constrained by the scarcity and limitations of available datasets. These findings provide valuable insights into the current state of hate speech detection for African languages and underscore the need to develop more comprehensive machine learning models and datasets for widely spoken African languages.

Keywords: African languages, Dataset, Hate speech detection, Low-resource languages, Machine learning, Multilingual model, Natural Language Processing (NLP).

1. INTRODUCTION

The pervasive reach of social media has interconnected societies globally, making the spread of hate speech a universal concern that affects both developed and developing countries, including those in Africa [1], [2]. Hate speech is any form of expression that seeks to degrade, attack, or vilify individuals or groups based on their identities, such as ethnicity, religion, nationality, race, gender, descent, and colour, among other identity factors, thus violating human rights [3], [4], [5], [6], [7]. The African continent has over 2,000 languages [8] that remain underrepresented in Natural Language Processing (NLP) research, often due to a lack of digital resources such as annotated datasets [9].

Developing annotated datasets for African languages has been a significant milestone in addressing these challenges. For instance, Muhammad et al. [10] described the creation of NaijaSenti, a sentiment analysis dataset for four major Nigerian languages (Hausa, Igbo, Yoruba, and Nigerian Pidgin), which serves as a foundational resource for multilingual hate speech analysis in the region. Such efforts demonstrate the potential for leveraging community-driven initiatives and pre-trained language models to bridge the resource gap in underrepresented languages. Similarly, Vargas et al. [9] introduced HausaHate1, an annotated dataset for Hausa hate speech detection, emphasising the importance of culturally informed classifiers.

Despite these advancements, the challenges of multilingualism and code-switching remain significant barriers. Multilingual societies, such as those in Africa, often use code-switching in spoken and written communication, complicating the task of NLP tools. Ababu et al. [11] explored bilingual hate speech detection in Amharic and Afaan Oromo, utilising deep learning techniques to address the complexities of code-mixed language usage. Another critical area of focus is the adaptation of pre-trained language models to low-resource settings. Pre-trained models, such as AfroXLMR and InkubaLM, have shown promise in addressing the linguistic challenges unique to Africa. Tonja et al. [12] introduced InkubaLM, a small yet efficient language model tailored for African languages, including Hausa, Yoruba, Swahili, isiZulu, and isiXhosa, achieving competitive performance despite limited computational resources. This approach highlights the feasibility of developing scalable and efficient models for low-resource settings, a crucial step in democratising access to NLP technologies. Integrating multimodal approaches has also been a key development in addressing hate speech. Debele et al. [13] demonstrated the potential of combining acoustic and textual features for detecting hate speech in Amharic, achieving notable accuracy with deep learning models. This multimodal strategy not only enhances detection accuracy but also opens new avenues for addressing the spread of harmful content across various social media platforms.

The socio-political implications of hate speech and sentiment analysis in African contexts cannot be overstated. Studies have highlighted the prevalence of hate speech during politically sensitive periods, such as elections in Kenya and Nigeria, and the role of social media in amplifying these issues [14], [15]. Similarly, Kotzé and Senekal [16] analysed the discourse around minority communities in South Africa, revealing how online platforms can both reflect and exacerbate societal divisions. These findings underline the urgent need for targeted interventions and policy frameworks to mitigate the impact of online hate speech. Gashe et al. [17] emphasised the importance of community involvement in annotating datasets for Amharic hate speech detection, achieving high performance with the Bidirectional Long Short-Term Memory (Bi-LSTM) model. This collaborative approach ensures NLP tools' cultural and linguistic relevance while boosting local capacity-building and knowledge-sharing. The work by Adelani et al. [18] on adapting pre-trained models to low-resource languages provides a roadmap for future research, demonstrating that even small amounts of high-quality data can significantly improve model performance. Additionally, efforts to create domain-specific corpora, such as those discussed by Oriola et al. [19], highlight the potential for customised solutions to address specific challenges, such as hate speech during elections or in particular social contexts.

Several studies have examined machine learning (ML) and hate speech detection; however, existing findings remain fragmented and lack a cohesive framework, particularly within the context of African languages. Despite the growing adoption of ML in automated hate speech detection, a systematic review reveals that research in this area has been limited and regionally skewed, as most of the studies are concentrated in West Africa, East Africa, and the Horn of Africa. This underrepresentation poses a serious gap in the literature, especially given the rich linguistic diversity of the African continent, where low-resource languages constitute the majority. Further, the lack of annotated datasets, limited availability of pre-trained language models, and minimal efforts to tailor ML approaches to local linguistic and cultural nuances further compound the challenge. As a result, current systems often fail to generalise across different African languages and social media contexts, reducing their effectiveness and fairness.

In light of this, a systematic literature review (SLR) is essential to consolidate the dispersed body of knowledge, assess methodological approaches, and uncover the limitations in current research. This review is particularly necessary to identify where and how ML techniques have been applied to African language hate speech detection, what datasets and evaluation metrics have been used, and where significant gaps persist.

Therefore, the purpose of this SLR is to synthesise existing research, identify critical challenges, especially the lack of datasets and suitable ML models for low-resource African languages, and propose directions for future research. Through this process, the review aims to support the development of inclusive, context-aware, and scalable ML-based hate speech detection systems that reflect the linguistic realities of African social media spaces.

This study is guided by the following research questions:

RQ1: What machine learning approaches have been used to detect hate speech in African languages on social media, and how can they be categorised?

RQ2: What types of datasets and annotation techniques have been applied in existing studies on hate speech detection in African languages?

RQ3: What are the current gaps and challenges in hate speech detection research for African languages, and what future directions can be proposed to improve these efforts using machine learning?

The study found that despite major efforts towards research on the development of machine learning in Africa, very little research is directed at hate speech detection. While the findings provide valuable insights into the state of hate speech detection for African languages, most research has been conducted in West Africa, East Africa and the Horn of Africa and this underscores the need for development of more comprehensive machine learning models for common languages that could form the basis for designing automated tools capable of identifying a large variety of social media hate speech that is fast growing in all African regions.

The remainder of this paper is structured as follows. Section 2 provides background and related literature on hate speech detection, with a focus on machine learning approaches in low-resource African languages. Section 3 details the methodology adopted for this systematic literature review, including the search strategy, search criteria, inclusion and exclusion criteria, and data extraction procedures. Section 4 presents the key findings and synthesises thematic patterns from the reviewed studies. Section 5 discusses the identified research gaps, practical challenges, and implications for future work. Finally, Section 6 concludes the paper by summarising the main insights and offering recommendations for further research in this domain.

2. BACKGROUND

Whereas there isn't a single globally standardized definition for "hate speech", an often referenced version by the United Nations defines hate speech as "Any kind of communication in speech, writing or behaviour, that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, colour, descent, gender or other identity factor".[20] The current body of literature demonstrates that hate speech detection has been extensively explored in high-resource languages but remains underdeveloped for low-resource languages, particularly in Africa, mainly due to the multilingual nature of the diverse languages spoken, worsened by code switching [8]. Ridwanullah et al. [15] examined hate speech trends during election periods in Nigeria, revealing the prevalence of political, ethnic, and regional biases. These findings align with those of Vargas et al. [9], who highlighted the dominance of racial hate in West African hate speech datasets. Such research findings underscore the need for context-sensitive approaches that address the specific socio-political dimensions of hate speech in Africa. Researchers have designed different machine learning algorithms that are used to automatically detect and monitor hate speech to compliment manual processes to identify and remove hate speech online perpetrated through high level resource languages such as English, French or Chinese among others [20], [21], [22], [23], [24], [25], [26], [27], [28]. Furthermore, multimodal models that integrate text, audio, and visual features are gaining attention. Debele et al. [13] achieved 88.15% accuracy by combining acoustic and textual features for Amharic hate speech detection. These approaches provide promising solutions for addressing the rich, multimodal communication styles prevalent in African languages. One of the most influential breakthroughs in machine learning is the development of the Transformer architecture by [29]. Unlike earlier models such as Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks, the Transformer architecture introduced a self-attention mechanism that allows for parallel processing of input data and more efficient handling of long-range dependencies in sequences. This innovation laid the foundation for subsequent advancements in deep learning and has become the dominant architecture for many natural language processing (NLP) tasks. Building upon this foundation, models such as BERT (Bidirectional Encoder Representations from Transformers) by [30] and the evolution of Large Language Models (LLMs) represent a significant leap in generative Artificial Intelligence. Models like GPT (Generative Pre-trained Transformer) by Radford et al. [31], [32] have demonstrated remarkable capabilities in text understanding and generation. These models are pre-trained on massive corpora and fine-tuned for specific downstream tasks, enabling them to generalise well even with minimal supervision.

Oriola and Kotzé [33] underscore the importance of domain-specific models in enhancing classification accuracy. Similarly, Aliyu et al. [34] released an annotated hate speech corpus focused on Fulani herders in Nigeria, covering English, Hausa, and Nigerian Pidgin. These contributions provide foundational resources for developing machine-learning models tailored to specific linguistic and cultural contexts. Muhammad et al. [10] emphasised that hate speech research prioritises high-resource languages, leaving underrepresented languages with limited tools and datasets. Developing annotated datasets has been a critical step in addressing this gap. For example, AfriSenti, introduced by [8], is a benchmark dataset containing over 110,000 tweets in 14 African languages, enabling sentiment analysis research in previously neglected languages. Sosimi et al. [35] studied the application of Long Short-Term Memory (LSTM) networks, Gated Recurrent Units (GRUs), and hybrid models to analyse West African languages, with a specific focus on Pidgin English. Their approach highlights how these models effectively capture contextual meanings and linguistic nuances that emerge in African languages, especially in code-switched contexts. The authors argue that LSTMs and GRUs can maintain contextual memory across sequences of words, making them well-suited for managing the complexity of mixed (coded) linguistic inputs. Tesfaye and Kakeba [36] trained an LSTM and GRU model using a manually annotated dataset of 30,000 Amharic hate speech texts (race, religion, ethnicity) collected from Facebook; the LSTM model outperformed GRU, achieving a test accuracy of 97.9%.

The complexity of multilingualism and bilingual communication significantly impacts hate speech detection on social media platforms. Ababu and Woldeyohannis [37] tackled this issue by developing a corpus and baseline models for Afaan Oromo hate speech detection, achieving an 84% accuracy using BiLSTM models with word2vec embeddings. Their dataset of 12,812 Facebook posts focused on hate speech themes such as gender, religion, and race. These efforts underscore the necessity of tailoring NLP tools to accommodate linguistic diversity and the unique challenges of bilingual communication. Despite significant progress, imbalances in data representation remain a challenge for African languages. Aliyu et al. [38] observed that their annotated corpus predominantly featured English (97.2%), with Hausa and Nigerian Pidgin comprising only a tiny fraction. This disparity reflects broader issues in developing comprehensive datasets for low-resource languages. Efforts like the creation of AfriSenti aim to address these gaps but are still far from encompassing Africa's vast linguistic diversity.

The study by [39] evaluated nine models for hate speech detection in English-Kiswahili code-switched text, including traditional machine learning and deep learning approaches. Traditional models using features like character-level Term Frequency-Inverse Document Frequency (TF-IDF) with a Support Vector Machine (SVM) classifier yielded the best performance, achieving 82.5% accuracy, compared with the deep learning algorithms, which showed lower classification accuracy because they require larger datasets to achieve optimal performance. The study highlights the limitations of static embeddings and the need for more advanced contextual models for code-switched data.

2.1. RECENT CONTRIBUTIONS TO MACHINE LEARNING DATASETS

Significant research efforts have been made in curating hate speech datasets for African languages. The success of machine learning (ML) based hate speech detection initiatives like Masakhane, Lacuna Fund, AfriNLP, and the creation of multilingual corpora such as KenCorpus reflect a growing, community-driven movement to bridge linguistic gaps in ML, especially for underrepresented African languages.

Masakhane is a grassroots NLP project involving African researchers and linguists in developing ML models focused on African languages, aiming to build and democratize AI through open collaboration. The Lacuna Fund, now closed, provided funding for creating labelled datasets, especially in languages and regions that are historically underserved. The fund supported several hate speech detection and classification efforts by financing dataset creation in multiple African languages. The fund supported the building of KenCorpus, a Swahili-English multilingual dataset, instrumental for tasks like sentiment analysis and hate speech detection. AfriNLP is a community initiative focused on African Natural Language Processing research and regularly organizes events to share tools and advances in African NLP. It also offers open-access corpora and pretrained models tailored for African languages.

Vargas et al. [9] introduced HausaHate1, the first expert-annotated corpus for detecting hate speech in Hausa, including labels for offensive speech and specific hate targets such as race and gender. In the bilingual space, Ababu et al. [11], [37] compiled hate speech datasets in both monolingual Afaan Oromo and bilingual Amharic-Afaan Oromo. These datasets, which include over 30,000 social media posts, were used to train deep learning classifiers such as BiLSTM and CNN-BiGRU, supported by FastText and word2vec embeddings, achieving classification accuracies above 78%. Complementing this, Gashe et al. [17] developed a dataset of 5,000 Amharic social media posts, annotated into categories such as racial, religious, gender-based, and non-hate speech. Their work utilised a stacked BiLSTM (SBi-LSTM) architecture, yielding a 94.8% F1-score, demonstrating the effectiveness of deep learning approaches in low-resource settings.

Further, the majority of hate speech detection automated tools have been developed in high-resource languages, such as English, which have an enormous amount of publicly available datasets to train the models. However, for low-resource languages, little machine learning based research has been conducted, due to limited available annotated datasets, which in turn has nurtured an environment for perpetuating and propagating hate speech in these languages. Further, most recent studies have focused on the detection of textual hate speech, giving little attention to other forms of multimodal modes such as images, audio, and video. This review underscores the value of leveraging multiple data modalities, combining text and image to enhance the precision and robustness of hate speech detection systems, particularly for low-resourced languages.

3. METHODOLOGY

This study used the Systematic Literature Review (SLR) informed by the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines [40]. The review process was planned by formulating research questions, identifying a search strategy, and formulating search strings. Further, inclusion and exclusion criteria and an analysis process were defined as outlined in the following subsections.

3.1. SEARCH STRATEGY

First, a set of electronic databases was selected, consisting of the following: Google Scholar, Emerald, IEEE, ACM Digital Library, and African Journals Online. This selection was guided by the study's focus, as these databases published content related to the use of machine learning algorithms to detect and monitor hate speech on social media and are also subscribed to by the authors' institution. For example, the Google Scholar database indexes research articles and abstracts from most major academic publishers and repositories worldwide, including both free and subscription sources, making it easy to uncover details of the topic under study. The noted search phrases were used to create search strings employing the Boolean operators 'AND' and 'OR.' Consequently, the formulated search strings were executed on the selected databases, and the resulting output was recorded for each run. The scope of this study covers the period between 2019 and 2024, as illustrated in Figure 1. Furthermore, inclusion and exclusion criteria were applied to the output to guide the study analysis.

3.1.1 Search Criteria

The search criteria used for this review consisted of three parts:

Str1: ("hate speech" OR "offensive language" OR "abusive content") AND ("machine learning" OR "artificial intelligence" OR "deep learning") AND ("African languages" OR "low resource African languages") AND ("social media" OR "online platforms")"

Str2: "hate speech detection" AND ("machine learning") AND ("African languages" OR "low resource African languages") AND (social media)

Str3: ("hate speech detection") AND ("machine learning") AND ("African languages") OR ("low resource African languages")

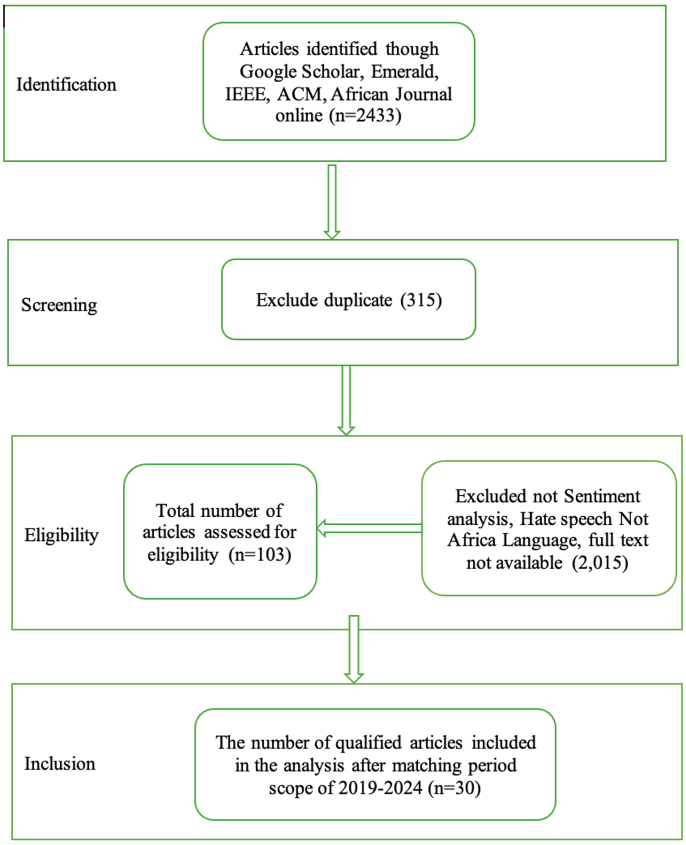
From the above search criteria, a total of 2,433 articles were identified (Google Scholar= 668, Emerald= 293, IEEE= 1241, ACM Digital Library= 117, African Journals Online= = 114) as demonstrated in Table 1.

3.1.2 Inclusion and Exclusion Criteria

The following criteria were used for inclusion and exclusion in the study. Figure 1 below provides a summary of the identification and selection of relevant articles used in this review.

http://dx.doi.org/10.21622/ACE.2025.05.2.1524

Figure 1: Inclusion and exclusion criteria flowchart


Identification
Articles identified though Google Scholar, Emerald, IEEE, ACM, African Journal online (n=2433)
Screening
Exclude duplicate (315)
Eligibility
Total number of articles assessed for eligibility (n=103)
Excluded not Sentiment analysis, Hate speech Not Africa Language, full text not available (2,015)
Inclusion
The number of qualified articles included in the analysis after matching period scope of 2019-2024 (n=30)

3.1.2.1 Inclusion Criteria

The inclusion criteria were based on the following:

i. The study is a peer-reviewed publication and is either a journal article or a conference paper, and
ii. The publication language is English, and
iii. The publication is relevant to the defined search phrases, and
iv. The study was published between 2019 and 2024.

3.1.2.2 Exclusion Criteria

The exclusion criteria were based on the following:

i. Studies that did not focus on machine learning algorithms for detecting and monitoring hate speech in African languages.
ii. Studies that did not meet the inclusion criteria or;
iii. Studies where the target language was not African, or
iv. Studies that were already published by one of the targeted publishers and were therefore a duplication to a large extent of other studies within the period 2019 to 2024.

3.1.3 Search Coding and Analysis Process

All publications that met the defined search strategy and criteria were documented using an Excel sheet, where all output was coded for further analysis. From the selected databases, output from conference papers was minimal and was later dropped from the analysis. Therefore, the studies that were analysed were peer-reviewed journal articles from the four databases.

At the start, a total of 2,433 publications, as shown in Table 1, were recorded. The study inclusion and exclusion criteria were then applied, as demonstrated in Figure 1, resulting in 30 publications indicated in Table 2. The 30 publications were then used to inform the research questions for further analysis and reporting.

Table 1: Articles initially coded for study analysis

Online database	Number of publications (initial coding)
IEEE	1241
Google scholar	668
Emerald	293
African Journals Online	114
ACM Digital Library	117
Total Publications	2433

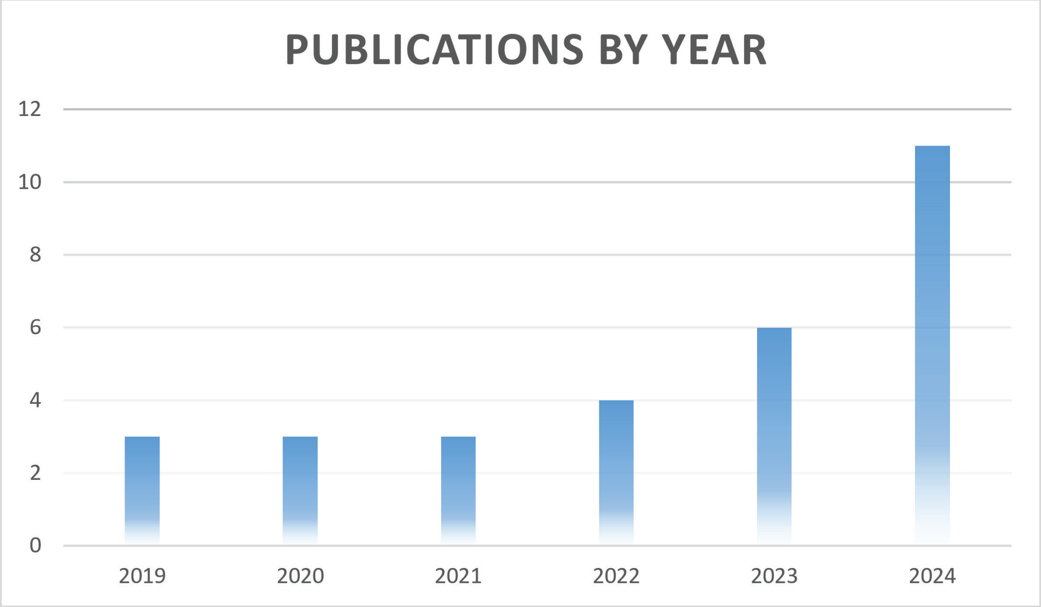
Table 2: Articles finally considered for study analysis

Online database	Number of publications (for final study analysis)
IEEE	3
Google scholar	21
Emerald	2
African Journals Online	2
ACM Digital Library	2
Total Publications	30

4. RESULTS

The research was conducted using search sources, and the search string was as described under the methodology section. A total of 2,433 articles related to the application of machine learning models, the availability and limitations of datasets, and identifying key gaps in research were found. Out of these, 30 articles published within the study period of 2019 to 2024 were identified and retained. The number of publications fluctuated annually throughout the observed timeframe. The highest publication volumes occurred in 2019, 2022, and 2024, as shown in Figure 1 below.

Figure 2: Publication by year



The findings of this SLR are organised according to the three research questions as outlined below:

4.1. MACHINE LEARNING APPROACHES FOR HATE SPEECH DETECTION IN AFRICAN LANGUAGES

RQ1 sought to review machine learning models applied to hate speech detection in African languages employing Classical Machine Learning (ML), also known as traditional ML models, Deep Learning (DL), and Transformer-Based Models. The review findings indicate that traditional ML models, such as Support Vector Machines (SVM), have demonstrated high effectiveness on smaller datasets, as shown by [41] and [14], where SVM outperformed other models on Bambara language and Kiswahili-English (code-switched) datasets. In contrast, deep learning models, such as CNN-BiLSTM, excel on larger datasets due to their ability to capture more complex patterns [42]. Transformer-based models like XLM-RoBERTa and mBERT have demonstrated superior performance, particularly in handling multilingual or code-switched data Nganga et al. [43]. For example, Nganga et al. [43] found that XLM-RoBERTa surpassed traditional models like SVM in detecting hate speech in Kiswahili-English datasets, achieving an 88% F1 score, highlighting the significance of dataset structure and language characteristics in model performance. Studies such as [37] have utilized BiLSTM models with Word2Vec embeddings for Afaan Oromo hate speech detection, achieving an 84% accuracy rate. Similarly, [33] explored multilingual BERT (mBERT) for South African social media texts, showcasing how transformer-based models can adapt to African contexts. The study by [44] trained a language-specific model called TwiBERT from scratch to detect hate speech in the Twi language, achieving an F1 score of 64.29%.

4.1.1 Impact of Domain-Specific Fine-Tuning

Domain adaptation remains crucial for model success. Ahmed et al. [45] and Tonneau et al. [46] demonstrated that fine-tuning transformer models on regional dialects, such as Egyptian-Arabic and Nigerian Pidgin, significantly boosted performance. The study by [11] further emphasized the importance of fine-tuning when working with bilingual datasets, such as Amharic and Afaan Oromo. The study found that BiLSTM models with FastText feature extraction outperformed other models with an accuracy of 78.05% for bilingual hate speech detection. The success of fine-tuning strategies underscores that while multilingual models are powerful, their effectiveness can be limited without domain-specific pretraining.

In contrast, Ilevbare et al. [47] showed that ensemble models could also enhance cross-domain performance. This suggests that pre-training on domain-specific data, coupled with ensemble methods, offers a comprehensive approach to improving model performance across various contexts.

4.1.2 Class Imbalance

Class imbalance remains a prevalent challenge in hate speech detection, particularly when hate content is underrepresented in training datasets. Arlim et al. [48] used the Synthetic Minority Over-sampling Technique (SMOTE) to rebalance datasets, improving SVM and AfriBERT performance, respectively. However, Ayele et al. [49] showed that transformer models, such as Afro-XLMR-large, can handle class imbalance more effectively than traditional ML models without requiring oversampling. Therefore, the deep learning models benefit more from larger, diverse datasets that reflect real-world complexities rather than synthetic data augmentation.

4.1.3 Bias and Ethical Concerns

Bias detection and mitigation are central concerns in hate speech detection research, carrying significant ethical implications for real-world applications. Davidson et al. [50] found that classifiers trained on existing datasets were more likely to flag African-American English (AAVE) tweets as abusive compared to white-aligned tweets, revealing a troubling pattern of algorithmic discrimination. This bias arises not only from the technical aspects of model training but also from the way datasets are annotated, which can lead to the unjust censorship of marginalized voices. Ombui et al. [6] observed similar issues in code-switched datasets and demonstrated that incorporating psychosocial features into the model could help capture cultural nuances that traditional lexicon-based methods often overlook. The ethical challenge, as discussed by [50] and [51], extends beyond technical limitations to the broader societal impact of these systems, underscoring the importance of culturally informed annotation and feature engineering.

When hate speech detection systems are deployed, they introduce two major ethical risks. The first is the risk of false positives, where legitimate speech is incorrectly flagged and censored. This is particularly concerning for marginalized groups, whose voices may be disproportionately silenced. In politically sensitive contexts, such errors can suppress dissent or minority perspectives, deepening social divides and eroding trust in automated moderation. The second risk is that of false negatives, where actual hate speech goes undetected. This allows harmful content to spread unchecked, potentially inciting violence, perpetuating discrimination, and causing psychological harm to those targeted. The consequences of these errors are not merely technical but have real-world implications for individuals and communities.

Balancing these risks requires careful calibration of detection thresholds, ongoing evaluation, and transparent reporting of error rates. Responsibility for these errors is distributed across several stakeholders. Developers and researchers must prioritize fairness and inclusivity in model design, dataset selection, and feature engineering, drawing on diverse expert input and conducting regular bias audits. Platform operators are responsible for implementing robust appeal mechanisms, providing context-sensitive moderation, and ensuring transparency in their decision-making processes. Annotators and domain experts play a crucial role in shaping culturally sensitive datasets and guiding ethical annotation practices.

To mitigate these risks, several concrete strategies can be employed. Engaging a diverse group of annotators and domain experts helps ensure that datasets reflect a wide range of cultural and linguistic contexts, reducing the likelihood of annotation bias. Feature engineering should move beyond simple lexicon-based approaches to incorporate psychosocial and sociolinguistic insights, allowing models to better understand nuanced expressions. Regular bias audits are essential to evaluate models for disparate impact across demographic groups, using metrics that reveal differences in false positive and false negative rates. Transparent reporting of model limitations, error rates, and mitigation efforts fosters accountability and builds trust with users. Finally, providing clear pathways for users to contest moderation decisions ensures that legitimate speech is not unjustly censored. By integrating these strategies, hate speech detection systems can better address the ethical complexities of deployment, minimizing harm while promoting fairness and inclusivity.

4.1.4 Feature engineering

Feature engineering has been a major theme across the studies. Most of the studies utilized the Word2Vec (Word to Vector) and Term Frequency-Inverse Document Frequency (TF-IDF) features are used in the preprocessing of the data. Ombui et al. [14] introduced a psychosocial feature set with the TF-IDF, which achieved the highest classification accuracy when used with the Support Vector Machine (SVM) algorithm to detect hate speech in code-switched text messages, outperforming deep learning algorithms such as Convolutional Neural Networks (CNN). On the other hand, Ilevbare et al. [47] focused on ensemble methods for political hate speech detection, achieving superior performance on multi-class tasks. While both studies emphasized feature engineering, each approached it differently. Ombui's work relied on theoretical constructs for feature design, whereas the study [47] focused on algorithmic approaches through ensemble methods. This highlights the complementary nature of innovations in feature and model design.

4.1.5 Emerging trends

While these advancements demonstrate the potential of machine learning for African languages, challenges persist. Many existing models designed for high-resource languages fail to incorporate linguistic and cultural nuances essential for accurately detecting hate speech in low-resource languages.

Hate speech detection in African languages is influenced by various factors, including dataset size, language specificity, and domain adaptation. Traditional models like SVM are efficient on smaller datasets, while deep learning and transformer-based models excel in multilingual and complex contexts. Oversampling techniques can enhance the performance of ML models, but transformer models often mitigate class imbalance more effectively. Ethical concerns around bias are prevalent, requiring both inclusive datasets and culturally informed feature design. Finally, cross-lingual transfer and domain-adaptive pretrained language models offer promising solutions for low-resource languages, but more research is needed to address regional linguistic complexities. The future of hate speech detection in African languages lies in the integration of ensemble methods (hybrid methods) with culturally aware feature sets and localized training data to reduce errors and improve generalization compared to any individual model [33]

4.2. DATASETS AND ANNOTATION TECHNIQUES IN EXISTING STUDIES

Hate speech detection in African languages has seen an area of growing research, and various datasets have been developed to aid in the advancement of this field. Datasets remain a cornerstone of machine learning research, yet their scarcity and limitations severely impact progress in hate speech detection for African languages. Small datasets are particularly insufficient for training data-hungry deep learning models [10], [46], [47], [52]. The study revealed that many existing datasets tend to focus on specific languages or regional dialects, which limits the generalizability of the trained models.

For example, datasets for Amharic Text-Image Data [53], HausaHate1[49], and Afaan Oromo [37] represent major African languages but often fail to capture the full range of dialects and regional variations that require local expertise for accurate annotation. Languages like Hausa have several dialects across West and Central Africa, and Afaan Oromo is spoken across multiple regions with different dialects and writing systems. However, these datasets, while valuable, primarily represent one version of each language, meaning they may not be fully reflective of the variations found in real-world scenarios. The lack of dialectal diversity limits the ability of machine learning models to generalize across different linguistic variations, making it harder for models to detect hate speech effectively in regions where dialectal differences are prominent. Another critical theme in African hate speech datasets is the issue of code-switching and multilingualism, which is common in many African communities. African languages are frequently spoken alongside global languages, such as English and French, among others, creating multilingual environments in which speakers alternate between languages and dialects, making it difficult to effectively detect hate speech. For example, the English-Swahili dataset [33], [39], [43], [47] was used to train models to detect hate speech in code-switched speech in multilingual contexts. However, the studies lack the breadth needed to cover the full range of linguistic diversity across the continent. Combining these code-switched datasets enables existing high-resource English-language models, such as multilingual BERT (mBERT), to be trained on local languages. This approach bridges the gap caused by the limited availability of annotated code-switched datasets, enhances hate speech detection, and better represents the complex and diverse multilingual landscape of many African regions, while addressing the issue of dataset scarcity.

The reviewed studies show significant efforts in curating datasets for African languages, emphasizing their critical role in advancing hate speech detection, as tabulated in Table 3 below. Multiple datasets have been developed across various African languages, such as NaijaSenti for Nigerian languages [8], AfriSenti covering 14 African languages [10], 16 African Languages, MAFAND-MT covering 11 African languages, and a bilingual Amharic-Afaan Oromo dataset, English, isiZulu, English-Kiswahili language [6], [37], [43]. Despite these efforts, data scarcity, limited language coverage, and lack of multimodal datasets remain major challenges because most of the studies focused on textual data, with limited exploration of multimodal approaches such as combining text, images, audio, and video, among others

[13], [54]. Additionally, issues such as dialectal variations [41], inconsistent annotations, and the absence of standardized benchmarks [50] hinder model generalization and comparison across studies. Overall, while the diversity of datasets is expanding, significant gaps remain in language inclusivity, multimodal integration, and dataset standardization, particularly for underrepresented African languages.

The rise of social media in Africa has played a significant role in the development of datasets for hate speech detection, as social media platforms are key sources of hate speech. Datasets like HausaHate1. [9], EKOHATE [47], Naijahate [46], AfriSenti [8], Naijahate [37], TArC [55], and HateSpeech_Kenya [39] collected data from various social media platforms, providing insights into how hate speech is expressed in informal and real-world online interactions. Social media content is inherently informal, which introduces a host of challenges for dataset creation and annotation, as detailed in Table 3 below. The majority of studies used Twitter as the primary data source for hate speech detection, followed by Facebook and YouTube, respectively, as illustrated in Table 3 below.

Social media posts often contain slang, emojis, abbreviations, and cultural references, making it difficult to label data consistently. Additionally, the language used on these platforms is often highly context-dependent, with meanings shifting based on the speaker's intention, audience, and platform. These features make data quality a major concern for hate speech detection, as inconsistencies in annotation can lead to errors in model predictions. While datasets like Naijahate [37] have attempted to tackle these issues, they remain limited in scope and size. Furthermore, social media datasets tend to be platform-specific, focusing on specific social networks or user bases. As a result, they may fail to represent the diversity of online interactions across different African countries. Cultural and regional sensitivity is another critical factor influencing the quality and relevance of datasets. Many African languages are spoken in politically, ethnically, and culturally diverse regions, and hate speech often targets specific groups based on ethnicity, religion, or political affiliation. Datasets such as Fulani Herdsmen Corpus (HERDPhobia) by [38] and Algerian Dialect (AlgD) [56] attempt to capture these culturally sensitive issues but face several limitations.

The main challenge lies in the size of these datasets. Due to the sensitive nature of topics such as ethnic conflict or political violence, collecting large-scale data becomes difficult. Moreover, these datasets may not fully represent the diversity of hate speech within these cultural contexts. For instance, the HERDPhobia dataset [38] primarily focuses on hate speech related to Fulani herders, a specific ethnic group in Nigeria. This narrow focus makes it difficult to develop models that can generalize to other types of hate speech, such as those targeting urban migrants, religious minorities, or women.

In addition, datasets focused on culturally sensitive topics may suffer from annotation biases. Annotators may interpret the same statements differently based on their own cultural or political beliefs, leading to inconsistencies in the data. This challenge requires more robust annotation strategies that incorporate diverse perspectives and ensure that all forms of hate speech are adequately captured. [41] emphasized the challenges of annotating datasets for African languages, noting that dialectal variations and limited linguistic expertise often led to inconsistencies. Additionally, the limited size and scope of existing datasets restrict the development of robust models. The lack of standard benchmarks further complicates model evaluation and comparison, making it difficult to assess progress in the field comprehensively. Additionally, no one-model-fits-all solution exists for a per-language evaluation [50].

Table 3 presents an overview of the datasets for hate speech detection in African languages. The EthioHate dataset by [20] is the largest, comprising approximately 139 million posts from Facebook and X (formerly Twitter) in Amharic, Ge'ez, Afan Oromo, Somali, and Tigrinya. It is followed by AfriSenti [8], which contains 110,000 annotated tweets spanning 14 African languages, including Amharic, Algerian Arabic, Hausa, Igbo, Kinyarwanda, Moroccan Arabic, Mozambican Portuguese, Nigerian Pidgin, Oromo, Swahili, Tigrinya, Twi, Xitsonga, and Yoruba. These tweets were annotated by native speakers and used in the AfriSenti and SemEval shared tasks. Similarly, Belete et al. [54] curated a dataset of 70,703 Facebook messages

http://dx.doi.org/10.21622/ACE.2025.05.2.1524

to detect Amharic text-image data. Meanwhile, [37] compiled a dataset of 12,812 Facebook posts and comments covering thematic areas such as gender, religion, race, and offensive speech, culminating in 20,000 annotated Amharic and Afaan Oromo bilingual datasets. The HausaHate1 corpus by [9] consists of 2,000 manually annotated comments extracted from West African Facebook pages. Additionally, [39] developed a dataset of 25,000 tweets collected during Kenya's 2017 elections, specifically focusing on Kiswahili-English code-switching.

Table 3: Machine learning dataset summary

No	Focus Area	Key Datasets & Authors	Techniques / Models Used	Goals / Future Directions
1	Multilingual & Code-Switched	HateSpeech_Kenya [39], [43] EKOHATE [47], AfriSenti [8] , SemEval [57]	Multilingual ML, XLM-R, RoBERTa, supervised learning	The need to improve code-switching handling and enhance culturally nuanced hate-speech detection across African languages. Researchers should develop unified multilingual benchmarks to evaluate cross-lingual transfer and model robustness. There is also a need to expand code-switched datasets and advance transformer-based, context-aware models integrating text and speech for resilient detection in African digital spaces.
2	Language-Specific	Afaan Oromo Hate Speech [37] HausaHate1 [9], Naijahate [46]	Supervised ML, BERT-based models	Address ethnic and regional biases by incorporating diverse and dialectal datasets across African languages. Research should expand supervised and transformer-based models for low-resource contexts, enhancing data quality and cultural nuance. Building collaborative language resource hubs and benchmarks will strengthen inclusivity and reproducibility in hate-speech detection.
3	Sentiment Analysis	NaijaSenti [10], Bambara V2 [41]	Transformer-based models, supervised classifiers	Bridge sentiment and hate-speech detection by developing multi-task models that jointly capture polarity, intent, and toxicity. Leveraging AfriSenti and NaijaSenti corpora as pretraining resources can enhance cross-domain learning and contextual understanding. There is also a need to build emotion-aware and zero-shot models capable of detecting sarcasm, implicit hate, and sentiment nuances.
4	Multimodal & Dialectal	Amharic Text-Image [54], AlgD [52]	Deep learning, multimodal fusion, dialect-aware NLP	Expand multimodal datasets by integrating culturally grounded text–image and speech data from diverse social platforms. Researchers should develop dialect-aware deep learning models that capture linguistic variation and visual–textual context. Emphasis should be placed on cross-modal fusion techniques to improve detection accuracy across dialects.
5	Emerging Trends	Hatebase [50], Amharic Hate Speech Corpus; [49], [58] Masakhane Hate Speech, [18], AfriHate, AfriBERTa,	Transformer fine-tuning, lightweight model dev., multimodal expansion (BERT, RoBERTa, XLM-RoBERTa, AfriBERTa, mBERT); multimodal fusion models (ResNet-50 + BiLSTM)	Focus on advancing transformer fine-tuning and domain adaptation to enhance multilingual generalisation across African languages. Researchers should develop lightweight, efficient multilingual models suitable for low-resource settings. There is also a need to create cross-lingual and multimodal tools that support scalable and inclusive hate-speech detection across diverse digital ecosystems.

4.3. CURRENT GAPS AND CHALLENGES IN HATE SPEECH DETECTION RESEARCH FOR AFRICAN LANGUAGES

The literature on hate speech detection in African languages has progressed significantly, but several research gaps remain, inhibiting the development of more robust models. These gaps emerge from analyzing the datasets and models used across thirty studies (30) studies, which cover a wide variety of African languages and dialects. The key research gaps identified are as follows:

4.3.1 Inadequate Representation of Dialectal and Regional Variations

Although various languages like Hausa, Swahili, and Afaan Oromo have been represented in datasets, the existing models fail to adequately account for the dialectal variations and regional differences within African languages [10], [42], [43], [46], [59]. Many of the models, such as those by [9] and [37], focus on major languages, leading to reduced accuracy in detecting hate speech across diverse African contexts.

4.3.2 Lack of Multimodal Hate Speech Detection

Current datasets are predominantly text-based, neglecting the growing prevalence of images, videos, emojis, and memes as mediums for expressing hate speech. As pointed out by studies [54] and [11], incorporating multimodal data is crucial for effective hate speech detection. The absence of these additional forms of communication limits the capacity of models to recognize nuanced and complex expressions of hate speech, especially in platforms like Instagram, Facebook, and YouTube, where multimodal hate speech is increasingly common.

4.3.3 Data Imbalance and Limited Coverage of Hate Speech Categories

Many African datasets are biased towards specific categories of hate speech, primarily focusing on ethnic or religious tensions, as seen in datasets like HausaHate1 [9] and NaijaSenti [10]. There is a lack of sufficient representation for other types of hate speech, such as gender-based or political hate speech. This imbalance restricts the models' ability to generalize across different forms of hate speech and makes them less effective in detecting hate speech outside of the categories they were trained on.

4.3.4 Insufficient Focus on Low-Resource Languages

While datasets such as NaijaSenti [10] and AfriSenti [8] have made strides in collecting data for widely spoken African languages, many low-resource languages remain underrepresented. Dialects and smaller African languages like Bambara, Fulani, and Mande lack sufficient datasets for effective hate speech detection. As highlighted by [10] and [34], authors, these languages often lack the necessary digital content and linguistic resources, making it difficult to develop robust hate speech detection models for these communities.

4.3.5 Challenges in Code-Switching and Multilingual Contexts

Given the high prevalence of code-switching and multilingualism in African online discourse, existing models often struggle to detect hate speech in mixed-language contexts. Datasets such as KenyaHateSpeech [39] and NaijaSenti [10] have tackled some aspects of code-switched datasets, but there is still much room for improvement. The difficulty in parsing mixed-language text hinders effective detection, and more advanced methods are required to handle such complex linguistic features.

4.3.6 Actionable Future Directions

Based on the identified gaps above, the following future directions are proposed to advance the field of hate speech detection for African languages:

4.3.7 Expansion of Datasets to Include More Languages and Dialects

To bridge the regional and dialectal variations gap, researchers should focus on expanding datasets to include a broader range of African languages and dialects. Furthermore, incorporating diverse hate speech categories, including racism, gender-based hate speech, and political discourse, will better reflect the complexities of online hate speech in Africa [9], [38]. Researchers should institutionalise community-driven annotation pipelines by partnering with universities and NGOs to recruit native speakers, running mobile-friendly micro-tasks, maintaining living lexicons, and embedding robust quality-control mechanisms such as scenario-based training, calibration rounds, multi-annotator labelling with adjudication, and

http://dx.doi.org/10.21622/ACE.2025.05.2.1524

explicit inter-annotator-agreement targets to raise reliability and curb "teacher noise" [6], [10], [14] In parallel, teams should publish reusable, culturally-aware annotation guidelines that clearly distinguish hate from offensiveness, mark targets and intensity, and set explicit rules for identity terms and code-switching to reduce over-blocking and improve cross-regional portability [49], [50]. The MasakhaneNER 2.0 project demonstrated that effective community-driven annotation relies on recruiting native speakers from African AI and NLP communities, providing structured training workshops, and ensuring fair remuneration for their contributions. Each language team was led by a coordinator who supervised annotation, resolved disagreements, and maintained high inter-annotator agreement through the ELISA annotation tool. This participatory approach, grounded in ethical collaboration, empowered local experts to co-create high-quality and culturally representative NLP datasets for African languages [18]. There is also a pressing need to co-develop operational definitions of hate speech for African languages grounded in local sociolinguistic realities and validated with community stakeholders so that labelling captures intent, target, and severity across code-switched and dialectal discourse [1], [14], [37], [49]. For sustainable practice, corpora should be domain aware, deliberately sampling election and conflict periods and refreshing periodically to capture distribution shifts typical of West and Southern African platforms [1], [35], [47], [60]. Dataset releases ought to include reproducible splits, rich metadata, and reliability statistics (e.g., Fleiss' κ or Krippendorff's α), alongside ethical safeguards for annotator well-being and clear licences to enable reuse and cross-country comparability [46], [47].

4.3.7.1 Incorporating Multimodal Data for Enhanced Detection

Future datasets should incorporate multimodal data to enhance hate speech detection in varied communication forms [54]. Expanding the research to images, videos, audio, and memes will improve the models' ability to detect hate speech in contexts where text alone is insufficient. Additionally, researchers like [11], [54] suggest incorporating emoji-based hate speech to enhance detection across social media platforms like WhatsApp, Instagram, and Facebook.

4.3.7.2 Addressing Data Imbalance and Bias in Datasets

Given the imbalance in the types of hate speech represented in current datasets, future work should focus on annotating various hate speech categories. More attention should be paid to detecting hate speech related to gender, political discourse, and inter-ethnic tensions [9]. Additionally, addressing data imbalance by ensuring equal representation across different categories will help improve the accuracy and robustness of detection models.

4.3.7.3 Utilisation of Transfer Learning and Low-Resource NLP Techniques

Researchers should explore transfer learning methods to help tackle the problem of low-resource languages. Researchers can overcome the scarcity of annotated data by leveraging large pre-trained models like BERT or GPT-3 and fine-tuning them on smaller African language datasets. Models that employ hyperparameter optimisation and transfer learning techniques could significantly improve the performance of hate speech detection models in low-resource languages [44], [61].

4.3.7.4 Improving Handling of Code-Switching in African Social Media Data

Due to the common practice of code-switching in African languages, future models must be specifically designed to handle mixed-language text [13], [39]. This can be achieved by exploring more sophisticated word segmentation and syntax parsing methods, which can enhance the ability of models to detect hate speech in code-switched text. Improved linguistic feature extraction will be key to dealing with the complexities of multilingual social media data [62].

5. DISCUSSION

The review reveals that hate speech detection in African languages using machine learning is a growing research area, though facing numerous challenges [57]. The dynamic nature of social media platforms poses challenges to dataset quality and annotation, which remains a significant barrier, especially for low-resource languages and dialects across the continent [9], [35], [63]. Most researchers employed X (formerly Twitter) as the primary data source for hate speech detection, followed by Facebook and YouTube, reflecting the dominant social media platforms in Africa where online discourse often occurs. Of the 30 models reviewed, only a few are custom-built, while most adopt pre-trained models designed to detect hate speech in high-resource languages. The evolution of machine learning models, particularly transformer architectures, has significantly enhanced hate speech detection capabilities. For example, Custom-built models such as AfroXLMR [57], InkubaLM [12], NaijaXLM-T [46], and AfriBERTa were effective in the detection of hate speech but still lack extensively annotated datasets from the majority of African languages, which still limits their effectiveness and performance.

While transformer-based models like XLM-RoBERTa have shown promise, their reliance on large-scale pretraining datasets that exclude many African languages limits their applicability. This gap underscores the importance of creating pretraining datasets that include diverse African languages and dialects [33], [37], [39], ensuring broader representation and improved model performance.

The lack of multimodal datasets is particularly problematic [11]. Hate speech often combines text with visual elements, such as images or memes, to convey harmful messages. Current datasets fail to capture this complexity, limiting the applicability of models in real-world scenarios [9], [11], [54]. Addressing this gap requires the creation of large-scale, multimodal datasets that integrate text, images, and even audio data. Collaborative efforts involving local communities and linguists are essential for ensuring these resources' accuracy and cultural relevance.

In examining research gaps and future directions, this review identifies several critical shortcomings in hate speech detection for African languages. Firstly, the challenges posed by code-switching and bilingual communication remain largely unaddressed. Future models should incorporate multilingual training and context-aware embeddings to effectively process mixed-language inputs [39], [43]. Secondly, cultural adaptation in existing models is insufficient. To enhance detection accuracy, models must recognise and integrate cultural nuances, such as idiomatic expressions and societal norms [53].

Thirdly, addressing data imbalance and bias requires annotating a broader range of hate speech categories to improve model generalisation. Lastly, findings indicate an inadequate representation of dialectal and regional variations within African languages. Most models predominantly focus on widely spoken languages, neglecting minority dialects, which limits their applicability and effectiveness across diverse linguistic communities [9], [18].

6. CONCLUSION

The analysis of thirty (30) machine learning models for detecting hate speech in African languages demonstrates important findings. Despite advances in machine learning, the lack of datasets remains a significant obstacle. Efforts such as Naijaseti [10] and the Afaan Oromo project [37] provide valuable resources. However, most available datasets are small, single-language, and text-oriented model reverberation. The findings reveal that transformer models such as AfriBERT, BiLSTM, mBERT, and XLM-RoBERTa have shown significant potential in detecting hate speech in African languages like Afaan Oromo, Amharic, Swahili, and Hausa. These models demonstrate the power of advanced algorithms to adapt to linguistic contexts that lack extensive computational resources. However, their performance is often hindered by the scarcity and limitations of datasets. Existing datasets, such as AfriSenti, HausaHate1, InkubaLM, and BambaraV2, provide valuable contributions but remain constrained in scope, size, and linguistic diversity. Multimodal datasets that integrate text and images, essential for capturing the complexity of hate speech in real-world scenarios, are largely absent.

The challenges in hate speech detection extend beyond technical considerations to encompass linguistic, cultural, and social dimensions. Many African languages exhibit dialectal variations, code-switching, and cultural nuances that are difficult to model with traditional NLP techniques. Furthermore, the lack of community-driven approaches in data annotation and resource creation has limited hate speech detection tools' contextual accuracy and inclusivity. These gaps highlight the need for interdisciplinary collaboration among linguists, technologists, and local communities to address the unique requirements of African languages in NLP research. This study highlights that advancing hate-speech detection for African languages requires community-driven, ethically grounded data creation and transfer-learning practices that centre linguistic diversity and local participation. The findings can inform social media platforms to improve moderation by integrating region-specific datasets, recruiting native-speaker annotators, and prioritising contextual understanding of multilingual and code-switched discourse rather than relying on English-centric models. Such strategies would reduce false positives and negatives and enhance cultural sensitivity in automated moderation systems. Similarly, NGOs and civil society organisations can leverage these insights to advocate for stronger platform accountability through transparent reporting of model coverage, equitable data governance, and fair remuneration of annotators. By supporting participatory annotation initiatives and monitoring algorithmic fairness, these stakeholders can ensure that technological interventions in African digital spaces are both inclusive and socially responsible [18].

Future research directions should focus on the expansion of datasets to include more languages and dialects, incorporating multimodal data for enhanced detection in both text and images. In addition, they should address data imbalance and bias in datasets and increase the use of transfer learning and other low-resource NLP techniques to train accurate models with moderate amounts of data. Improving the handling of code-switching in African social media data should also be prioritised because no one-model-fits-all solution exists for a per-language evaluation [50].

REFERENCES

[1] C. Sinyangwe, D. Kunda, and W. P. Abwino, "Detecting Hate Speech and Offensive Language using Machine Learning in Published Online Content," Zambia ICT Journal, vol. 7, no. 1, pp. 79–84, Mar. 2023, doi: 10.33260/zictjournal.v7i1.143.

[2] F. M. Ndahinda and A. S. Mugabe, "Streaming Hate: Exploring the Harm of Anti-Banyamulenge and Anti-Tutsi Hate Speech on Congolese Social Media," J Genocide Res, vol. 26, no. 1, pp. 48–72, Jan. 2024, doi: 10.1080/14623528.2022.2078578.

[3] U. K. Schmid, A. S. Kümpel, and D. Rieger, "How social media users perceive different forms of online hate speech: A qualitative multi-method study," New Media Soc, vol. 26, no. 5, pp. 2614–2632, May 2024, doi: 10.1177/14614448221091185.

[4] C. Silva and P. Carvalho, "When Can Compliments and Humour Be Considered Hate Speech? A Perspective From Target Groups in Portugal," Comunicacao e Sociedade, vol. 43, 2023, doi: 10.17231/COMSOC.43(2023).4135.

[5] Y. Musa and G. Asuquo, "HATE SPEECH AND HUMAN SOCIETY: A CRITICAL ANALYSIS," Humanities and Education Journal (SHE Journal), vol. 2, no. 3, 2021.

[6] E. Ombui, L. Muchemi, and P. Wagacha, "Building and Annotating a Codeswitched Hate Speech Corpora," International Journal of Information Technology and Computer Science, vol. 13, no. 3, 2021, doi: 10.5815/ijitcs.2021.03.03.

[7] G. Njovangwa and G. Justo, "Automated Detection of Bilingual Obfuscated Abusive Words on Social Media Forums: A Case of Swahili and English Texts," Tanzania Journal of Science, vol. 47, no. 4, 2021, doi: 10.4314/tjs.v47i4.2.

http://dx.doi.org/10.21622/ACE.2025.05.2.1524

[8] S. H. Muhammad et al., "AfriSenti: A Twitter Sentiment Analysis Benchmark for African Languages," in EMNLP 2023 - 2023 Conference on Empirical Methods in Natural Language Processing, Proceedings, 2023. doi: 10.18653/v1/2023.emnlp-main.862.

[9] F. Vargas et al., "HausaHate: An Expert Annotated Corpus for Hausa Hate Speech Detection," in Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024), Stroudsburg, PA, USA: Association for Computational Linguistics, 2024, pp. 52–58. doi: 10.18653/v1/2024.woah-1.5.

[10] S. H. Muhammad et al., "NaijaSenti: A Nigerian Twitter Sentiment Corpus for Multilingual Sentiment Analysis," in 2022 Language Resources and Evaluation Conference, LREC 2022, 2022.

[11] T. M. Ababu, M. M. Woldeyohannis, and E. B. Getaneh, "Bilingual hate speech detection on social media: Amharic and Afaan Oromo," J Big Data, vol. 12, no. 1, p. 30, Feb. 2025, doi: 10.1186/s40537-024-01044-y.

[12] A. L. Tonja et al., "InkubaLM: A small language model for low-resource African languages," arXiv preprint arXiv:2408.17024 , 2024.

[13] A. G. Debele and M. M. Woldeyohannis, "Multimodal Amharic Hate Speech Detection Using Deep Learning," in 2022 International Conference on Information and Communication Technology for Development for Africa, ICT4DA 2022, 2022. doi: 10.1109/ICT4DA56482.2022.9971436.

[14] E. Ombui, L. Muchemi, and P. Wagacha, "Psychosocial Features for Hate Speech Detection in Code-switched Texts," International Journal of Information Technology and Computer Science, vol. 13, no. 6, 2021, doi: 10.5815/ijitcs.2021.06.03.

[15] A. O. Ridwanullah, S. Y. Sule, B. Usman, and L. U. Abdulsalam, "Politicization of Hate and Weaponization of Twitter/X in a Polarized Digital Space in Nigeria," J Asian Afr Stud, vol. 60, no. 5, 2025, doi: 10.1177/00219096241230500.

[16] E. Kotzé and B. Senekal, "Employing sentiment analysis for gauging perceptions of minorities in multicultural societies: An analysis of Twitter feeds on the Afrikaner community of Orania in South Africa," The Journal for Transdisciplinary Research in Southern Africa, vol. 14, no. 1, 2018, doi: 10.4102/td.v14i1.564.

[17] S. M. Gashe, S. M. Yimam, and Y. Assabie, "Hate Speech Detection and Classification in Amharic Text with Deep Learning," 2024, doi: https://doi.org/10.48550/ARXIV.2408.03849.

[18] D. I. Adelani et al., "MasakhaNER 2.0: Africa-centric Transfer Learning for Named Entity Recognition," in Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, 2022. doi: 10.18653/v1/2022.emnlp-main.298.

[19] O. Oriola and E. Kotzé, "Automatic detection of toxic south african tweets using support vector machines with N-gram features," in 2019 6th International Conference on Soft Computing and Machine Intelligence, ISCMI 2019, 2019. doi: 10.1109/ISCMI47871.2019.9004298.

[20] A. L. Tonja et al., "EthioLLM: Multilingual Large Language Models for Ethiopian Languages with Task Evaluation," arXiv:2403.13737, 2024, doi: https://doi.org/10.48550/arXiv.2403.13737.

[21] S. Ranathunga, E. S. A. Lee, M. Prifti Skenduli, R. Shekhar, M. Alam, and R. Kaur, "Neural Machine Translation for Low-resource Languages: A Survey," ACM Comput Surv, vol. 55, no. 11, 2023, doi: 10.1145/3567592.

[22] M. S. Jahan and M. Oussalah, "A systematic review of hate speech automatic detection using natural language processing," 2023. doi: 10.1016/j.neucom.2023.126232.

[23] A. M. U. D. Khanday, B. Bhushan, R. H. Jhaveri, Q. R. Khan, R. Raut, and S. T. Rabani, "NNPCov19: Artificial Neural Network-Based Propaganda Identification on Social Media in COVID-19 Era," Mobile Information Systems, vol. 2022, 2022, doi: 10.1155/2022/3412992.

http://dx.doi.org/10.21622/ACE.2025.05.2.1524

[24] N. S. Mullah and W. M. N. W. Zainon, "Advances in Machine Learning Algorithms for Hate Speech Detection in Social Media: A Review," 2021. doi: 10.1109/ACCESS.2021.3089515.

[25] R. Duwairi, A. Hayajneh, and M. Quwaider, "A Deep Learning Framework for Automatic Detection of Hate Speech Embedded in Arabic Tweets," Arab J Sci Eng, vol. 46, no. 4, 2021, doi: 10.1007/s13369-021-05383-3.

[26] A. Tontodimamma, E. Nissi, A. Sarra, and L. Fontanella, "Thirty years of research into hate speech: topics of interest and their evolution," Scientometrics, vol. 126, no. 1, 2021, doi: 10.1007/s11192-020-03737-6.

[27] F. E. Ayo, O. Folorunso, F. T. Ibharalu, and I. A. Osinuga, "Machine learning techniques for hate speech classification of twitter data: State-of-The-Art, future challenges and research directions," 2020. doi: 10.1016/j.cosrev.2020.100311.

[28] A. S. Alammary, "BERT Models for Arabic Text Classification: A Systematic Review," 2022. doi: 10.3390/app12115720.

[29] A. Vaswani et al., "Attention Is All You Need," 2017, Advances in Neural Information Processing Systems (NeurIPS 2017), Curran Associates, Inc. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[30] J. Devlin, M.-W. Chang, K. Lee, K. T. Google, and A. I. Language, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," Naacl-Hlt 2019, no. Mlm, 2018.

[31] A. Radford, K. Narasimhan, I. Sutskever, and T. Salimans, "Improving Language Understanding by Generative Pre-Training," 2018. [Online]. Available: https://cdn.openai.com/research-covers/language

[32] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language Models are Unsupervised Multitask Learners," 2019. [Online]. Available: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

[33] O. Oriola and E. Kotzé, "Exploring Neural Embeddings and Transformers for Isolation of Offensive and Hate Speech in South African Social Media Space," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2022. doi: 10.1007/978-3-031-10522-7_44.

[34] S. M. Aliyu, "Beyond English: Offensive Language Detection in Low-Resource Nigerian Languages," in 5th Workshop on African Natural Language Processing, 2022.

[35] A. A. Sosimi, O. Ipinnimo, C. O. Folorunso, B. A. Adim, and E. Onoyom-Ita, "Hate Speech Identification in West Africa using Machine-Learning Techniques," Arid Zone Journal of Engineering, Technology and Environment (AZOJETE), vol. 20, no. 2, pp. 491–508, 2024.

[36] S. G. Tesfaye and K. Kakeba, "Automated Amharic Hate Speech Posts and Comments Detection Model Using Recurrent Neural Network," Dec. 01, 2020. doi: 10.21203/rs.3.rs-114533/v1.

[37] T. M. Ababu and M. M. Woldeyohannis, "Afaan Oromo Hate Speech Detection and Classification on Social Media," in 2022 Language Resources and Evaluation Conference, LREC 2022, 2022.

[38] A. S. Mohammad, G. M. Wajiga, M. Murtala, S. H. Muhammad, I. Abdulmumin, and I. S. Ahmad, "HERDPhobia: A Dataset for Hate Speech against Fulani in Nigeria," 2022. [Online]. Available: https://arxiv.org/abs/2211.15262

[39] E. Ombui, L. Muchemi, and P. Wagacha, "Hate Speech Detection in Code-switched Text Messages," in 3rd International Symposium on Multidisciplinary Studies and Innovative Technologies, ISMSIT 2019 - Proceedings, 2019. doi: 10.1109/ISMSIT.2019.8932845.

[40] M. J. Page et al., "The PRISMA 2020 statement: an updated guideline for reporting

systematic reviews," BMJ, p. n71, Mar. 2021, doi: 10.1136/bmj.n71.

[41] M. Diallo, C. Fourati, and H. Haddad, "Bambara Language Dataset for Sentiment Analysis," 2021. [Online]. Available: https://arxiv.org/abs/2108.02524?utm_source

[42] G. O. Ganfure, "Comparative analysis of deep learning based Afaan Oromo hate speech detection," J Big Data, vol. 9, no. 1, 2022, doi: 10.1186/s40537-022-00628-w.

[43] F. N. Njung'e, A. M. Oirere, and R. N. Ndung'u, "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, vol. 13, no. 5, pp. 181–186, Oct. 2024, doi: 10.30534/ijatcse/2024/011352024.

[44] S. Akrah and T. Pedersen, "DuluthNLP at SemEval-2023 Task 12: AfriSenti-SemEval: Sentiment Analysis for Low-resource African Languages using Twitter Dataset," in Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), Stroudsburg, PA, USA: Association for Computational Linguistics, 2023, pp. 1697–1701. doi: 10.18653/v1/2023.semeval-1.236.

[45] I. Ahmed, M. Abbas, R. Hatem, A. Ihab, and M. W. Fahkr, "Fine-Tuning Arabic Pre-Trained Transformer Models for Egyptian-Arabic Dialect Offensive Language and Hate Speech Detection and Classification," in Proceedings of the 20th Conference on Language Engineering, ESOLEC 2022, 2022. doi: 10.1109/ESOLEC54569.2022.10009167.

[46] M. Tonneau et al., "NaijaHate: Evaluating Hate Speech Detection on Nigerian Twitter Using Representative Data," 2024. [Online]. Available: https://arxiv.org/abs/2403.19260

[47] C. E. Ilevbare, J. O. Alabi, D. I. Adelani, F. D. Bakare, O. B. Abiola, and O. A. Adeyemo, "EkoHate: Abusive Language and Hate Speech Detection for Code-switched Political Discussions on Nigerian Twitter," 2024. [Online]. Available: https://arxiv.org/abs/2404.18180

[48] N. Arlim, S. Riyanto, R. Rodiah, and A. Hafiz, "GunadarmaXBRIN at SemEval-2023 Task 12: Utilization of SVM and AfriBERTa for Monolingual, Multilingual, and Zero-shot Sentiment Analysis in African Languages," Association for Computational Linguistics, pp. 869–877, Nov. 2023, doi: 10.18653/v1/2023.semeval-1.120.

[49] A. A. Ayele, E. A. Jalew, A. C. Ali, S. M. Yimam, and C. Biemann, "Exploring Boundaries and Intensities in Offensive and Hate Speech: Unveiling the Complex Spectrum of Social Media Discourse," 2024. [Online]. Available: https://arxiv.org/abs/2404.12042

[50] T. Davidson, D. Bhattacharya, and I. Weber, "Racial Bias in Hate Speech and Abusive Language Detection Datasets," arXiv:1905.12516 [cs], Nov. 2019, [Online]. Available: https://arxiv.org/abs/1905.12516

[51] N. James et al., "Limits of Language in Nigeria: The Hatred on Igbo Language," SSRN Electronic Journal, 2024, doi: 10.2139/ssrn.4946190.

[52] A. C. Mazari and H. Kheddar, "Deep Learning-based Analysis of Algerian Dialect Dataset Targeted Hate Speech, Offensive Language and Cyberbullying," International Journal of Computing and Digital Systems, vol. 13, no. 1, pp. 965–972, Apr. 2023, doi: 10.12785/ijcds/130177.

[53] F. M. Adam, A. Y. Zandam, and I. Inuwa-Dutse, "Detection and Analysis of Offensive Online Content in Hausa Language," in Proceedings of the Second IJCAI AI for Good Symposium in Africa hosted by Deep Learning Indaba, California: International Joint Conferences on Artificial Intelligence Organization, Aug. 2024, pp. 2–14. doi: 10.24963/ijcai.aai4g.2024/1.

[54] M. Degu, A. Tesfahun, and H. Takele, "Amharic Language Hate Speech Detection System from Facebook Memes Using Deep Learning System," SSRN Electronic Journal, 2023, doi: 10.2139/ssrn.4389914.

[55] C. Fourati, H. Haddad, A. Messaoudi, M. B. H. Hmida, A. B. E. Mabrouk, and M. Naski, "Introducing A large Tunisian Arabizi Dialectal Dataset for Sentiment Analysis," in

http://dx.doi.org/10.21622/ACE.2025.05.2.1524

WANLP 2021 - 6th Arabic Natural Language Processing Workshop, Proceedings of the Workshop, 2021.

[56] A. Benali, M. H. Maaloul, and L. H. Belguith, “Automatic Processing of Algerian Dialect: Corpus Construction and Segmentation,” SN Comput Sci, vol. 4, no. 5, p. 597, Aug. 2023, doi: 10.1007/s42979-023-02097-1.

[57] S. H. Muhammad et al., “SemEval-2023 Task 12: Sentiment Analysis for African Languages (AfriSenti-SemEval),” arXiv (Cornell University), Nov. 2023, doi: 10.18653/v1/2023.semeval-1.315.

[58] D. I. Adelani et al., “IrokoBench: A New Benchmark for African Languages in the Age of Large Language Models,” 2024. [Online]. Available: https://arxiv.org/abs/2406.03368

[59] E. D. Kingawa, K. Tasew, M. Sholaye, and S. Hailu, “HATE SPEECH DETECTION USING MACHINE LEARNING: A SURVEY,” Academy Journal of Science and Engineering (AJSE), vol. 17, no. 1, pp. 88–109, 2023.

[60] O. Oriola and E. Kotze, “Evaluating Machine Learning Techniques for Detecting Offensive and Hate Speech in South African Tweets,” IEEE Access, vol. 8, pp. 21496–21509, 2020, doi: 10.1109/ACCESS.2020.2968173.

[61] J. Mussandi and A. Wichert, “NLP Tools for African Languages: Overview,” in Proceedings of the 16th International Conference on Computational Processing of Portuguese (PROPOR 2024), Volume 2, Lisbon, Portugal: Association for Computational Processing of Portuguese Languages, 2024, pp. 73–82.

[62] C. Jacobs, N. C. Rakotonirina, E. A. Chimoto, B. A. Bassett, and H. Kamper, “Towards hate speech detection in low-resource languages: Comparing ASR to acoustic word embeddings on Wolof and Swahili,” in Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 2023. doi: 10.21437/Interspeech.2023-421.

[63] W. B. Demilie and A. O. Salau, “Detection of fake news and hate speech for Ethiopian languages: a systematic review of the approaches,” J Big Data, vol. 9, no. 1, p. 66, Dec. 2022, doi: 10.1186/s40537-022-00619-x.

[64] United Nations, “United Nations Strategy and Plan of Action on Hate Speech,” 2019. [Online]. Available: https://www.un.org/en/hate-speech/strategy-plan-action