http://dx.doi.org/10.21622/ACE.2026.06.1.2076

MULTIMODAL MACHINE LEARNING MODEL FOR DETECTING KISWAHILI HATE SPEECH ON SOCIAL MEDIA

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

[1] United States International University, School of Science and Technology, Kenya, Africa.
[2] United States International University, Artificial Intelligence , Kenya, Africa.
[3] United States International University, Machine Learning , Kenya, Africa.

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

Received 10 April 2026 - Accepted 01 June 2026 - Published 08 June 2026

ABSTRACT

The rapid growth of social media platforms has transformed communication and information sharing across the world. However, it has also increased the spread of harmful online content, particularly hate speech. This problem is more critical in low-resource languages such as Kiswahili, where limited annotated datasets, language resources, and computational tools make it difficult to develop effective automated detection systems. Although many previous studies have focused on text-based hate speech detection in high-resource languages, hate speech on social media is often multimodal, combining both textual and visual elements such as images and memes. This study, therefore, solved this problem by developing a multimodal machine learning model for detecting hate speech in Kiswahili social media content using both text and image data. The study was guided by the research question: How can multimodal machine learning improve hate speech detection in Kiswahili using text and image data from social media platforms? To address this question, the research focused on several objectives, including the collection and annotation of a multimodal dataset, the development and evaluation of machine learning models, and the validation of the developed system using quantitative evaluation metrics and expert feedback. A dataset containing 115,204 Kiswahili text samples and 2,607 image samples was collected from social media platforms such as X and Facebook. Annotation was carried out on the preprocessed 112,571 texts and 2607 images, with the support of Kiswahili language experts to ensure linguistic and contextual accuracy. The study adopted a mixed research design, combining qualitative insights into hate speech interpretation with quantitative machine learning techniques. To address class imbalance in the training data, the Synthetic Minority Oversampling Technique (SMOTE) was applied only to the training set, while the validation and test sets were retained as representations of real-world data.

The dataset was divided using an 80:10:10 ratio, and hold-out validation was used to ensure reliable evaluation results. Several machine learning and deep learning models were developed and evaluated. The Bidirectional Long Short-Term Memory (BiLSTM) model achieved the best performance in text classification with an F1-score of 97.33%, while the ResNet50V2 model achieved 89.76% F1-score in image classification. The developed multimodal fusion model achieved an overall F1-score of 92.11%, demonstrating the effectiveness of combining textual and visual features in improving contextual understanding of hate speech. Although the text-only model achieved a slightly higher F1-score, the multimodal model provided a more robust and context-aware approach, especially in identifying implicit and visually encoded hate speech. The study contributes to the field by providing one of the first large-scale multimodal Kiswahili hate speech datasets and demonstrating the applicability of deep learning techniques in low-resource language settings. In addition, the study developed an interactive hate speech detection interface called ChujaHate, which was validated by experts. The findings support the development of more inclusive and culturally relevant content moderation systems while highlighting the importance of ethical AI considerations such as fairness, bias mitigation, and data privacy. Future research should explore

http://dx.doi.org/10.21622/ACE.2026.06.1.2076

transformer-based multimodal architectures, real-time deployment, and the integration of additional modalities such as audio and video in dynamic online environments.

Keywords: *Hate speech detection, Late fusion, Low-resource languages, Machine learning, Multimodal data, Kiswahili, Natural Language Processing (NLP)*

1. INTRODUCTION

The rapid expansion of social media platforms has significantly transformed global communication, enabling individuals, communities, and institutions to interact and share information in real time across geographical boundaries. Platforms such as Facebook and X (formerly Twitter) have created open and participatory digital spaces that support collaboration and information exchange. However, alongside these benefits, social media has also facilitated the rapid spread of harmful content, particularly hate speech [1][2][3][4]

Hate speech has become a growing concern in online environments due to its potential to incite discrimination, hostility, and violence. The United Nations defines hate speech as any form of communication that attacks or uses discriminatory language toward a person or group based on identity factors such as Religion, ethnicity, nationality, race, or gender [5]. The viral nature of social media amplifies the reach and impact of such content, often intensifying social divisions and contributing to real-world conflict.

In many regions, including East Africa, hate speech has been linked to politically sensitive events such as elections, where it can fuel polarization and violence [6][7][8][9]. The increasing volume and speed of content generation make manual moderation ineffective, highlighting the urgent need for automated and scalable detection mechanisms.

To address the challenges associated with online social media hate speech, researchers have increasingly turned to automated detection systems based on machine learning techniques [10]. Machine learning enables systems to learn patterns from large datasets and make predictions with minimal human intervention [11][12]. Over time, hate speech detection has evolved from rule-based approaches to more sophisticated machine learning, deep learning models, transformer-based, and large language models.

Early studies relied on traditional machine learning algorithms such as Support Vector Machines (SVM) and logistic regression, which utilized manually engineered features like Bag-of-Words and Term Frequency Inverse Document Frequency (TF-IDF) representations [13][14]. While these approaches provided a foundation, they often struggled to capture contextual nuances and implicit forms of hate speech.

The introduction of deep learning techniques, including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Long Short-Term Memory (LSTM) networks, significantly improved the ability to model sequential and contextual information in text [15]. More recently, transformer-based architectures such as BERT and its multilingual variants, such as SwahBERT [16], have further enhanced performance by capturing deeper semantic relationships [17][18]. Despite these advancements, most research has focused on high-resource languages such as English, where large annotated datasets are readily available, leaving low-resource languages underrepresented.

Low-resource languages face significant challenges in the development of natural language processing systems due to the limited availability of annotated datasets, linguistic resources, and computational tools [19]. Despite Kiswahili being spoken by over 200 million people across Africa, it remains under-represented in machine learning research [20]. Existing studies on Kiswahili hate speech detection have primarily focused on text-based approaches using machine learning and deep learning models such as SVM, CNN, and LSTM [9][21]. While these studies demonstrate promising results, they are limited by small datasets and a narrow focus on textual data.

http://dx.doi.org/10.21622/ACE.2026.06.1.2076

Additionally, Kiswahili presents unique linguistic challenges, including dialectal variation, code-switching with English, and evolving slang, which complicate the development of robust models [6][22]. These factors often lead to reduced model generalization and performance, particularly when applied to diverse real-world contexts.

The lack of large-scale, diverse, and well-annotated datasets further exacerbates these challenges, underscoring the need for more comprehensive resources tailored to the Kiswahili language.

A key limitation of existing hate speech detection systems is their reliance on unimodal data, particularly text. However, modern social media communication is inherently multimodal, combining text, images, videos, and audio to convey meaning [23].

Hate speech is often expressed implicitly through the interaction of multiple modalities. For instance, an image may contain symbolic or contextual cues that, when combined with accompanying text, convey a harmful message that cannot be detected through text analysis alone. This makes unimodal detection approaches insufficient for capturing the full spectrum of online hate speech.

Recent studies have demonstrated the effectiveness of multimodal approaches in improving detection accuracy [24][23]. For instance, [25] combined text and image features using BERT and ResNet models, achieving high performance. Similarly, [26] showed that fusion strategies integrating multiple modalities outperform unimodal baselines. Despite these advancements, multimodal hate speech detection remains largely unexplored in low-resource languages such as Kiswahili, where most existing research is still limited to text-based analysis [27][28]. Datasets play a critical role in the development and evaluation of machine learning models. However, the availability of high-quality datasets for hate speech detection in Kiswahili remains limited. Existing datasets are often small, domain-specific, and predominantly text-based [1][9]. The process of creating annotated datasets in low-resource languages is complex and resource-intensive. It requires linguistic expertise and cultural understanding to ensure accurate labeling, particularly for implicit and context-dependent forms of hate speech [29][30][31]. Recent efforts such as AfriSenti and AfriHate have contributed to the development of multilingual datasets for African languages, but these remain largely unimodal and do not capture the multimodal nature of social media content [32][33].

The absence of publicly available multimodal datasets for Kiswahili significantly limits the development of robust detection systems and hinders comparative evaluation across studies. This gap highlights the need for large-scale, culturally contextualized datasets that integrate both text and image data.

Given the limitations of existing approaches, there is a growing need for advanced solutions that integrate multimodal data and address the challenges associated with low-resource languages. Multimodal machine learning provides a promising framework for combining textual and visual information to improve detection accuracy and contextual understanding [34].

This study addressed these gaps by curating a multimodal Kiswahili hate speech dataset and developing a machine learning model that integrated both text and image data from social media platforms such as Facebook and X. The approach leveraged deep learning architectures, including Bidirectional LSTM for text processing and convolutional neural networks for image analysis, combined through a late fusion strategy to generate a unified prediction.

By capturing complementary information across modalities, the model can detect both explicit and implicit forms of hate speech more effectively than unimodal systems. In addition, the study incorporated ethical considerations such as bias mitigation, data privacy, and fairness in model development [35][36].

http://dx.doi.org/10.21622/ACE.2026.06.1.2076

This study does not primarily introduce a new deep learning architecture but rather addresses an important research and resource gap in low-resource African language processing through multimodal integration. The main contribution of the study lies in the development of a large-scale multimodal Kiswahili hate speech dataset and the implementation of a multimodal framework that integrates textual and visual information for hate speech detection in social media contexts. Unlike many prior Kiswahili hate speech studies that rely solely on textual analysis, this work demonstrates the practical applicability of multimodal machine learning in capturing implicit and context-dependent hate expressions communicated through both language and imagery. The study further contributes to comparative benchmarking across classical machine learning, deep learning, and multimodal approaches within a low-resource setting.

2. RELATED WORK

In Africa, hate speech has emerged as a significant challenge, contributing to ethnic polarization, social unrest, and threats to democratic processes [6][7][8]. The rapid growth of social media platforms has further intensified this problem by enabling the widespread dissemination of inflammatory and harmful content in real time. Consequently, the development of automated hate speech detection systems has become increasingly important in supporting peaceful ethnic discourse and safeguarding social cohesion.

Early studies on hate speech detection primarily focused on text classification approaches using both traditional machine learning and deep learning techniques. Commonly applied methods include Support Vector Machines (SVMs), Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Long Short-Term Memory (LSTM) networks. For instance, [37] demonstrated the effectiveness of feature-engineering approaches, particularly TF-IDF combined with psychosocial features, in detecting hate speech in Kiswahili text. Similarly, [1] reported that transformer-based architectures such as mBERT and XLM-RoBERTa outperform traditional models in code-switched English-Swahili datasets. In another notable contribution, [6] developed a Swahili and code-switched English-Swahili political hate speech dataset by integrating the Politikweli, AfriSenti, and Hate_Speech_Kenya datasets. Their comparative evaluation showed that non-linear SVM, Bi-LSTM with fastText embeddings, and SwahBERT achieved strong classification performance.

Despite the valuable contributions of these studies, most existing approaches remain limited to textual analysis, overlooking other modalities through which hate speech is communicated. Recent advancements in artificial intelligence and natural language processing indicate a growing transition from unimodal text-based systems to multimodal approaches that integrate text, images, audio, and video data [27][23]. This transition is motivated by the observation that hate speech on social media is often implicit, contextual, and distributed across multiple modalities, especially in memes, edited images, and multimedia posts where harmful intent may not be directly expressed in text alone.

Multimodal approaches seek to capture complementary information from different data sources [38], thereby improving the ability of models to identify hidden or implicit hate speech patterns [39]. For example,[25] proposed a multimodal framework that combined BERT for textual representation and ResNet for image feature extraction, achieving high classification performance with an accuracy of 97.0% and an F1-score of 94.7%. Similarly, [40] developed a transfer learning-based multimodal framework known as TCAM using the Facebook Hateful Memes dataset and demonstrated the importance of modeling weak semantic relationships between textual and visual modalities.

Further studies have continued to demonstrate the effectiveness of multimodal learning strategies in hate speech detection. For instance, [41] evaluated several multimodal datasets and models and reported that early fusion techniques combined with transformer-based architectures such as CLIP, UNITER, and BERT consistently outperform unimodal baselines on benchmark datasets, including MMHS150K and MAMI. Likewise, [38] introduced a transformer-based multimodal framework incorporating multi-head attention mechanisms,

http://dx.doi.org/10.21622/ACE.2026.06.1.2076

achieving strong generalization performance across multiple hate speech datasets. In addition, [42] emphasized the importance of explainability and interpretability in multimodal hate speech detection systems, particularly for enhancing transparency and trustworthiness in automated moderation systems.

Although multimodal hate speech detection has gained considerable attention in high-resource languages, research in African and other low-resource languages remains limited and underexplored. Most studies involving Kiswahili and related African languages continue to focus primarily on text-based approaches. Examples include [37], which employed SVM-based classification using psychosocial features, and [1], which applied transformer-based models to code-switched English–Swahili text classification. More recently,[33] introduced the AFRIHATE dataset covering 15 African languages, including Kiswahili, where transformer-based models such as AfroXLMR achieved a strong performance of up to 91.72% in Swahili hate speech detection. However, the AFRIHATE dataset remains predominantly text-based, highlighting a critical limitation in current African hate speech research.

Only a few studies have attempted multimodal hate speech detection within low-resource African contexts [39]. For example, [27] developed a multimodal hate speech detection model for Amharic memes using CNN-based text and image fusion techniques, achieving approximately 81% classification accuracy. Similarly, [28] introduced a multimodal Urdu-English hate speech dataset and demonstrated that early fusion techniques significantly improve detection performance. Additionally, [22] and [43] explored audio-based hate speech detection in Kiswahili and English contexts, respectively, further demonstrating the potential of multimodal methods in low-resource environments.

Despite these advancements, research on multimodal hate speech detection in African languages remains fragmented, limited in scope, and underdeveloped. A review of recent studies, as summarized in Table 1, reveals that the majority of existing works are predominantly text-based, with very limited multimodal research focusing on African languages, particularly Kiswahili. This study, therefore, addresses this critical gap by developing a multimodal Kiswahili hate speech dataset sourced from X (formerly Twitter) and Facebook and a multimodal machine learning framework that integrates both textual and visual information. In doing so, the study contributes to the advancement of multimodal natural language processing research for low-resource African languages and provides a foundation for more robust hate speech detection systems in multilingual and multicultural online environments.

Table 1: Summary of key findings from the literature

Study	Dataset size and language focus	Best model	Metrics	Findings	Gaps
Transformer-based multi-task learning for multimodal hate speech detection (memes) By [26]	Four English benchmark meme datasets: MMHS150K, Facebook Hateful Memes, MultiOFF, SemEval MAMI	Multi-task learning framework combining UNITER + CLIP + BERT	AUC-ROC 0.817, Acc 77.35; MultiOFF: F1 79.82, Acc 81.54 (selected headline results)	Shows that training related meme tasks jointly can transfer useful multimodal knowledge and outperform several unimodal, fusion-based, and pretrained V+L baselines on multiple benchmarks.	Limited exploration of culturally diverse and low-resource contexts in multimodal settings; lacks real-world validation.
Swahili and code-switched English-Swahili political hate speech detection textual dataset by [6]	Merged Politikweli, AfriSenti, Hate_Speech_Kenya -101,014 tweets; includes hate/offensive/neither subset (19,369 labeled as hate/offensive/neither reported) Focused on Swahili + English–Swahili code-switching (Kenyan social media)	SwahBERT fine-tuning (plus BiLSTM and SVM baselines)	SwahBERT F1 = 0.89; BiLSTM (fastText) Acc/Prec/Rec/F1 = 0.80; recall for hate improved 0.57→0.76 with SMOTE	Contributes a larger, better-annotated Swahili/code-switched resource with language and target labels, and demonstrates that Swahili-specific pretraining (SwahBERT) can separate hateful from non-hateful content effectively despite imbalance.	Does not incorporate multimodal data, such as images, video, audio, among others, and is limited to political text.

http://dx.doi.org/10.21622/ACE.2026.06.1.2076

AfriHate: multilingual hate/abusive datasets for 15 African languages (incl. Swahili) By [33]	Train/dev/test counts per language (e.g., Swahili: train 14,760; dev 3,164; test 3,168). Labels: hate, abusive, neutral; targets annotated Focused on 15 African languages (incl. Swahili)	Baseline comparisons: Africa-centric PLMs, SetFit few-shot, and prompted LLMs	Average macro-F1: AfroXLMR-76L ≈ 78.16; GPT-4o: ≈61.89 (zero-shot) and ≈70.79 (20-shot)	Provides a comparatively large, standardized multilingual benchmark with native-speaker annotation and explicit ethical guidance, enabling more realistic evaluation across African languages.	Focused only on text and lacks multimodal datasets and unified benchmarks.
Transfer learning + cross-attention/ cross-mask for hateful meme classification (FHMC) By [39]	FHMC: 12,140 unique meme images; official splits used (train 8,500; dev_seen 500; test_seen 1,500) with additional unseen splits reported. Focused on English	TCAM (with CLIP + TweetEval encoders; cross-attention + cross-mask fusion)	AUROC 0.8362; Accuracy 0.764 on FHMC test_seen	Shows gains from explicitly modeling weak/indirect relationships between meme text and imagery, outperforming several reported baselines and approaching top competition systems while remaining simpler than ensemble-heavy approaches.	Evaluated mainly on a single dataset; it lacks generalization and robustness testing.
Comparative study of transformer models for English–Kiswahili code-switched hate speech detection by [1]	HateSpeech_Kenya (Kaggle): 48,057 instances (Hate 3,181; Offensive 8,543; Neither 36,333). Focused on English–Kiswahili code-switched texts.	XLM-RoBERTa (best on balanced data); also mBERT/ mDistilBERT as feature extractors and end-to-end	Balanced: XLM-R Acc 0.6069; Macro-F1 0.49. Feature extractor best: mBERT-SVM Acc 0.5461; Macro-F1 0.40. Imbalanced: mBERT Acc 0.7820; Macro-F1 0.53	Highlights that end-to-end fine-tuning generally beats using transformers only as feature generators, but results are sensitive to class balancing choices, underscoring the need to report macro metrics and dataset conditions clearly.	Performance remains relatively low and shows unresolved issues in code-switching understanding.
ASR vs acoustic word embeddings for hate-speech keyword spotting in Wolof/Swahili radio By [22]	Scraped Swahili radio broadcasts from three Kenyan stations (out-of-domain); hate keywords labeled by native Swahili experts Focused on: Wolof + Swahili (speech)	Keyword spotting via fine-tuned multilingual ASR (XLS-R) vs ASR-free multilingual AWE (query-by-example)	In-the-wild precision (top-100 retrievals): ASR 30h=52%, 1h=42%, 5min=36%; AWE=45% (music retrievals far lower for AWE: 2% vs 17–30%)	Shows that even when ASR is sample-efficient in controlled settings, ASR-free embeddings can be more robust under real broadcast conditions, making them attractive when labeled in-domain data are scarce.	Keyword-based detection lacks deep semantic/contextual understanding of hate.
Amharic hate speech detection from Facebook memes using deep learning (OCR-based) By [27]	5,000 Facebook text-images (memes); split using StratifiedKFold with 80/10/10 (train/val/ test) Focused on Ahmaric	BiLSTM + Dense (two hidden layers)	Best accuracy ≈ 81%	Demonstrates a workable pipeline for a low-resource language by extracting meme text via OCR and training sequence models, indicating that reasonable performance is possible with modest data sizes.	Relies on OCR text only; ignores visual features in memes.
Multimodal hate speech detection in Greek social media (rendered tweets) By [25]	Collected ~126,000 tweets; labelled sample n=4,004 (1,040 toxic; 2,964 non-toxic); rendered tweet screenshots used for vision input Focused on Greek (with some English tweets retained)	Early fusion approach: fine-tuned BERT (text) + ResNet (image) jointly	Accuracy 0.970; F1 0.947 (best model)	Introduces a practical way to treat posts as they appear to users (rendered screenshots), capturing visual cues that plain text pipelines can miss, and reports strong performance on their Greek xenophobic/ racist dataset.	Potential dataset bias and lack of cross-domain validation.
Psychosocial features for hate speech detection in code-switched texts (Kenya elections) By [37]	≈50k human-annotated tweets (Kenya 2012 & 2017 elections); each tweet annotated by ≥3 annotators; 80/20 split with cross-validation Focused on English/ Swahili/Sheng + some local-language tokens	SVM with psychosocial (PDC) features (plus lexical/PoS/app features explored)	SVM overall accuracy ~76.2%; hate-class precision 0.81, recall 0.85, F1 0.83; also notes uniform P/R/F1 ≈0.77 for SVM in one evaluation summary	Shows that engineered psychosocial/topic-informed features can improve detection of camouflaged hate in code-switched settings, where purely lexical keyword approaches often miss nuanced attacks.	Focused only on English, Kiswahili conde switched text and lacks multimodal datasets and unified benchmarks for Kiswahili, feature engineering not.

http://dx.doi.org/10.21622/ACE.2026.06.1.2076

3. METHODOLOGY

A mixed research design approach was used in which qualitative socio-cultural insights were integrated into the quantitative machine learning pipeline to enhance construct validity and annotation consistency. The architectural model presented in Figure 1 provides a structured framework for detecting Kiswahili hate speech by integrating both textual and visual information within a unified multimodal system. The process begins with the collection of 115,204 texts and 2,607 image data from social media platforms such as X (formerly Twitter) and Facebook, followed by annotation to label content as hate or non-hate speech. The collected data is then preprocessed separately according to modality. Text preprocessing involved cleaning, normalization, tokenization, stop-word removal, and conversion into numerical representations using Word2Vec embeddings. Similarly, image preprocessing included resizing, rescaling, augmentation, and feature extraction to prepare visual data for learning.

A structured annotation process of cleaned 112,571 texts and 2,607 images was conducted to ensure the reliability and validity of the multimodal dataset used in this study. Seven annotators, who were native Kiswahili speakers and active social media users, participated in the annotation exercise over a one-month period. Prior to annotation, the annotators underwent training and orientation sessions where they were introduced to the study objectives, labeling categories, and detailed annotation guidelines. The annotation process involved labeling both text and image data, with overlapping subsets assigned to multiple annotators to assess consistency and reliability. Regular discussions and review sessions were conducted to resolve disagreements and clarify ambiguous cases, ensuring a consistent interpretation of hate speech within the Kiswahili social media context. The study adopted a dual annotation scheme consisting of a nine-class labeling system for text data and a binary hate and non-hate classification system for image data. The text categories included not hate, offensive, sexual, gender, disability, race, chronic_disease, Religion, and Tribe, enabling fine-grained classification of different hate speech forms. Annotation guidelines emphasized context sensitivity, target identification, distinction between offensive and hateful content, handling ambiguity, and multimodal interpretation of text-image relationships. To evaluate annotation reliability, fifty text samples were independently labeled by multiple annotators and analyzed using Randolph's free-marginal multirater Kappa. The obtained Kappa score of 0.567 indicated substantial agreement among annotators, demonstrating that the annotation process was systematic, reproducible, and guided by well-defined labeling criteria.

The raw text was then preprocessed through the removal of special characters, URLs, and user mentions, among others, followed by tokenization and text normalization. Further, keywords were organized into categories reflecting common forms of hate expression observed in online discourse, such as *mshenzi* (uncivilized person), *shenzi* (barbaric), *malaya* (prostitute), *mbwa* (dog), *kumbafu* (idiot), and *mshirikina* (pagan), among others.

Feature extraction was subsequently performed using Term Frequency–Inverse Document Frequency (TF-IDF) and Word2Vec to obtain numerical vector representations. Text modeling was conducted iteratively using data partitioning into training (90,056), validation (11,258), and test (11,257) splits representing a ratio of 80:10:10, respectively. Benchmarking was done across traditional machine-learning models such as Naïve Bayes, Logistic Regression, SVM, Decision Tree, and deep-learning models such as RNN, GRU, BiLSTM, CNN, with final model selection guided by validation performance. Text inputs for the deep learning models were processed using TensorFlow's text vectorization layer. The vocabulary size was limited to 20,000 tokens, and each text instance was converted into a fixed-length sequence of 20 token indices. These sequences were then mapped into dense 128-dimensional vectors using a trainable embedding layer. Unlike static embedding methods, this trainable representation was learned jointly with the classifier, enabling the model to adapt to the linguistic patterns of Kiswahili hate-speech data. While this approach is computationally efficient and task-specific, it does not model contextual meaning as effectively as more recent transformer-based embeddings.

http://dx.doi.org/10.21622/ACE.2026.06.1.2076

The image hate detection model was trained on a binary hate speech dataset labeled as hate or non-hate. Additionally, for the image modality, the qualitative strand similarly informed the image annotation guidelines to support consistent labeling of hate and not hate and further guided image screening and selection prior to model training. Image preprocessing was broken down into steps, such as resizing and normalization, to ensure compatibility with the deep learning architecture. The pre-trained weights of ResNet50V2 were used for initialization so that the model could benefit from its experience of large-scale feature learning from the ImageNet dataset. This drastically reduced computation and training time while enhancing the performance since the model is already good at recognizing general patterns in images. The pre-trained models were adapted for binary classification on hate and non-hateful content by discarding the last fully connected layers of this model, and they were replaced by custom layers such as Global Average Pooling (GAP) to reduce the spatial dimensions of feature maps into one feature vector in a computationally efficient manner. The fully connected classification head comprised a dense layer with 256 neurons and ReLU activation, followed by a dropout layer with a rate of 0.5 to reduce overfitting and improve model generalization. One neuron with a sigmoid activation function was used to output the probability of the image belonging to the hateful class. The model was fine-tuned on the collected dataset with the Adam optimizer, with a learning rate of 0.001 and beta values for adaptive learning. Binary cross-entropy loss was used to calculate the error in classification, and the batch size was kept at 32 to strike a balance between computational efficiency and learning stability. Further, the models were also trained for 50 epochs to achieve robustness. Early stopping based on validation loss was applied to avoid overfitting of the model. The earning rate scheduler decreased the learning rate dynamically in case of a plateau in validation performance. Eventually, to make the models robust and avoid overfitting, they were trained using data augmentation such as random horizontal flipping, rotation up to 15 degrees, zooming, and brightness adjustment. Data augmentation created training data variations, enabling the model to generalize to unseen samples better.

Hold-out validation was employed as part of the model validation process. This approach provided a more robust and generalizable estimate of model performance, particularly important in settings where dataset composition can influence results. In this method, the multimodal dataset was partitioned into 80:10:10 for the train, validation, and testing sets, respectively. To address class imbalance in the training data, the Synthetic Minority Oversampling Technique (SMOTE) was applied only to the training set, while the validation and test sets were retained as representations of real-world data. This way, we prevented the data leakage, which could have led to overfitting of the models. This increases confidence that the observed results reflect the model's true ability to generalize to unseen data.

http://dx.doi.org/10.21622/ACE.2026.06.1.2076

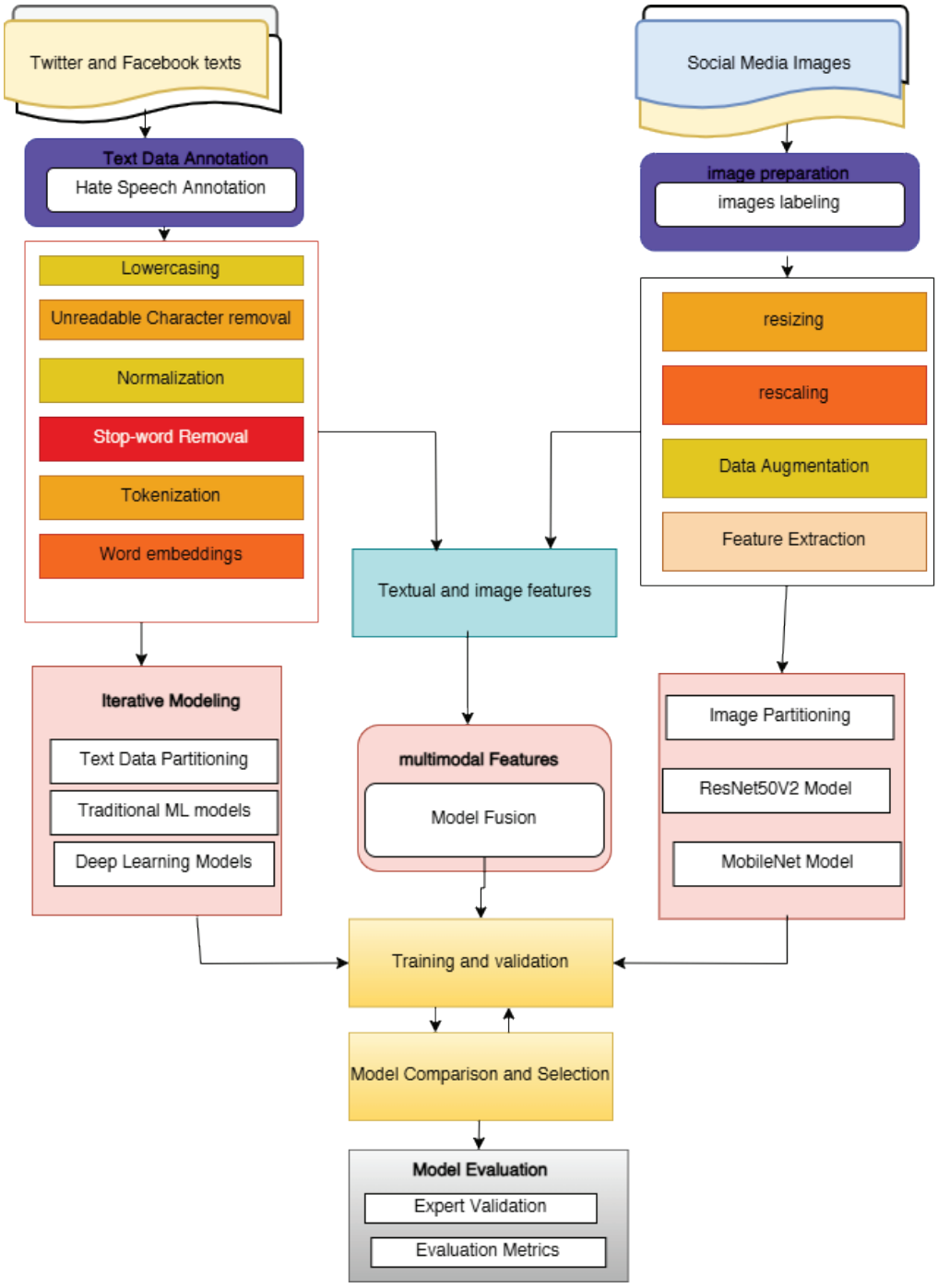

Figure 1: Multimodal architectural model

The multimodal hate speech detection model, ChujaHate, was implemented using a late fusion strategy since the image dataset with 2,607 samples was much smaller than the text dataset with 112,571 samples, creating a clear modality imbalance. The collection of hate speech image datasets was increasingly constrained by recent Application Programming Interfaces (APIs) and policy changes on Facebook and X (formerly Twitter), which limited the availability of multimodal data for this study, and further, image data required more extensive preprocessing than textual data. Given these differences in dataset size and modality characteristics, late fusion was considered more appropriate than early fusion. In late fusion, the probability outputs generated by the SoftMax layer of the BiLSTM model and the sigmoid probability outputs from the ResNet50V2 model were concatenated into a unified feature representation. This avoided the need for early alignment of heterogeneous features while still enabling effective multimodal integration [44]. Prior studies have shown

http://dx.doi.org/10.21622/ACE.2026.06.1.2076

that late fusion could outperform early fusion across key performance metrics in multimodal settings, where textual and image embeddings are concatenated before classification [40]. The fused model was evaluated using metrics such as accuracy, precision, Recall, and F1-score, while hyperparameter tuning helps identify the best-performing configuration. Finally, a web-based user interface enables users to input text, images, or both and receive hate speech classification results in real time.

4. EXPERIMENTAL RESULTS

The experimental results of the unimodal and multimodal models developed for hate speech detection are presented in this section. The evaluation was conducted in two phases, where text and image models were trained independently on an annotated Kiswahili dataset and integrated at inference time. For the text modality, both traditional machine-learning baselines and deep learning model results are indicated in this section.

4.1. TRADITIONAL ML MODELS:

The comparative evaluation of traditional machine learning algorithms for Kiswahili hate speech detection, given in Table 2, reveals clear performance differences across models, both in baseline and optimized configurations.

The Multinomial Naïve Bayes model recorded the lowest performance across all metrics, with an accuracy of 78.50% and an F1-score of 77.46%. While its relatively higher precision of 79.37% suggests it can reasonably identify hate speech when predicted, the low Recall indicates that it fails to capture a significant portion of actual hate speech instances. This limitation can be attributed to its strong independence assumption, which is often unsuitable for complex linguistic patterns such as those found in Kiswahili and code-switched text. The Support Vector Machine (SVM) model emerged as one of the best-performing models, achieving an accuracy of 87.95%, a precision of 87.76%, a recall of 87.95%, and an F1 score of 87.63%. The balanced performance across all metrics indicates that SVM effectively handles both positive and negative classes. Its robustness can be linked to its ability to find optimal decision boundaries in high-dimensional feature spaces, especially when combined with TF-IDF features.

The Logistic Regression model demonstrated competitive performance, with an initial accuracy of 86.27% and an F1 score of 85.59%. However, after hyperparameter tuning, the model improved significantly, achieving an accuracy of 88.76% and an F1 score of 88.46%, surpassing the SVM model slightly. This improvement highlights the importance of parameter optimization, particularly in linear models where regularization plays a critical role in balancing bias and variance.

The Decision Tree model also performed well, with an accuracy of 87.99% and an F1 score of 87.87%. Its performance suggests that it is capable of capturing non-linear relationships in the data. However, decision trees are prone to overfitting, which may limit their generalization capability despite strong training performance.

The Random Forest model, an ensemble extension of decision trees, showed stable and slightly improved results compared to a single decision tree. The model achieved an accuracy of 87.63% and an F1-score of 84.04% after tuning. The marginal improvement during training of the random forest indicates that ensemble learning helped reduce overfitting and improved generalization, although the gains were not substantially higher than those of SVM or tuned Logistic Regression.

The results indicate that SVM and tuned Logistic Regression models provide the best balance between precision and Recall, making them highly suitable for hate speech detection tasks. While ensemble methods such as Random Forest offer robustness, their performance gains in this case are moderate. The findings also emphasize the importance of hyperparameter tuning, as seen in the significant improvement of Logistic Regression after optimization.

http://dx.doi.org/10.21622/ACE.2026.06.1.2076

This finding is consistent with prior studies by researchers such as [9], who reported SVM as the top classical baseline with strong overall effectiveness and particularly strong hate-class separation, while [6] similarly demonstrated competitive SVM results on Kiswahili and English hate speech benchmarks.

Table 2: Traditional ML result comparison

ML ALGORITHM	ACCURACY	PRECISION	RECALL	F1-SCORE
Naïve Bayes	78.50	79.37	78.50	77.46
Support Vector Machine (SVM)	87.63	87.76	87.95	87.63
Logistic regression	88.76	88.51	88.76	88.46
Decision tree	87.99	87.87	87.99	87.87
Random Forest	87.63	87.75	87.64	84.84

4.2. DEEP LEARNING MODELS

i) LSTM Model Results

The Long Short-Term Memory (LSTM) model achieved strong performance, with an accuracy of 90.44%, demonstrating its ability to capture sequential dependencies in Kiswahili text. This confirms that recurrent architectures are well-suited for modeling linguistic structures in hate speech detection tasks. However, compared to traditional machine learning models such as SVM and LR, which were the best among classical approaches, the LSTM clearly surpasses them, highlighting the advantage of deep learning in capturing contextual relationships. Despite its strengths, the LSTM shows a slight imbalance between macro and weighted metrics, suggesting reduced effectiveness on minority classes. This indicates that while it learns general patterns well, it may miss subtle or less frequent hate expressions. Nevertheless, its performance is significantly higher than that of Logistic Regression, Decision Trees, and Random Forest models. The LSTM model, therefore, provides a strong foundation for deep learning-based text classification. It demonstrates that moving beyond feature-engineered approaches to sequence modeling leads to substantial gains in performance.

ii) GRU Model Results

The GRU model improves upon the LSTM by achieving an accuracy of 94.86%, indicating more efficient learning of contextual dependencies in the dataset. Compared to traditional machine learning models such as SVM, which already performed well, the GRU further demonstrates the superiority of deep learning approaches for this task. Its simpler architecture allows faster convergence while still maintaining high predictive power. However, the lower macro recall of 88.88% suggests that the model may not generalize equally well across all hate speech categories, particularly minority classes. Despite this, its weighted metrics remain very high, indicating strong overall classification capability. The GRU strikes a balance between computational efficiency and performance, making it practical for large-scale applications. It clearly outperforms classical approaches like Random Forest and Decision Trees. This reinforces the idea that neural architectures are better suited for capturing semantic and contextual nuances in Kiswahili text.

iii) Bidirectional Long Short-Term Memory Model (BiLSTM) Results

The Bidirectional LSTM (BiLSTM) achieved the best overall performance, with an accuracy of 97.33% and consistently high precision of 97.34%, Recall of 97.32%, and F1-score of 97.32%. This significantly surpasses both traditional machine learning models and other deep learning architectures such as LSTM and GRU. The superior performance of BiLSTM confirms that bidirectional context is critical for Kiswahili hate speech detection, where meaning often depends on both preceding and succeeding words. Unlike unidirectional models, BiLSTM captures richer semantic relationships, leading to improved classification accuracy across all classes. The close alignment between macro and weighted metrics further indicates

http://dx.doi.org/10.21622/ACE.2026.06.1.2076

strong generalization and balanced learning. This model effectively minimizes both false positives and false negatives. Its dominance across all evaluation metrics demonstrates that it is the most suitable model for this task. Therefore, BiLSTM stands out as the most effective approach due to its ability to fully exploit contextual information.

iv) CNN Model Results

The CNN model achieved a strong accuracy of 93.42%, confirming its effectiveness in capturing local textual patterns such as key phrases associated with hate speech. Compared to traditional machine learning models like SVM, CNN still demonstrates superior performance, reinforcing the advantage of deep learning methods. However, when compared to recurrent architectures such as LSTM, GRU, and especially BiLSTM, its performance is slightly lower. This is mainly due to its limitation in capturing long-range dependencies in text. The relatively low macro recall 83.02% suggests that CNN struggles more with minority classes, focusing primarily on dominant patterns. Despite this, its high weighted scores indicate strong overall predictive ability. CNN remains computationally efficient and suitable for scenarios where speed is critical. While it does not outperform BiLSTM, it provides a solid alternative within deep learning approaches. Overall, it highlights the trade-off between efficiency and deep contextual understanding.

Among the traditional machine learning models, the Support Vector Machine (SVM) outperformed Logistic Regression when logistic regression is not fine-tuned, Decision Tree, and Random Forest, suggesting that margin-based classifiers are more effective for the curated hate speech dataset. Despite this strong performance, the deep learning models consistently achieved higher results than all traditional approaches. In particular, the Bidirectional Long Short-Term Memory (BiLSTM) model delivered the best overall performance, with an accuracy of 97.33%, precision of 97.34%, recall of 97.32%, and an F1-score of 97.32%. These consistently high metrics indicate that the model is both accurate and well-balanced in its predictions. The superior performance of the BiLSTM can be attributed to its ability to process text in both forward and backward directions, allowing it to capture richer contextual information. This is especially important in Kiswahili hate speech detection, where meaning often depends on surrounding words and sentence structure. As a result, the BiLSTM emerged as the most effective model for textual hate speech detection, making it more suitable for integration into a multimodal fusion model.

The summary showing the comparison of results is shown in Figure 2.

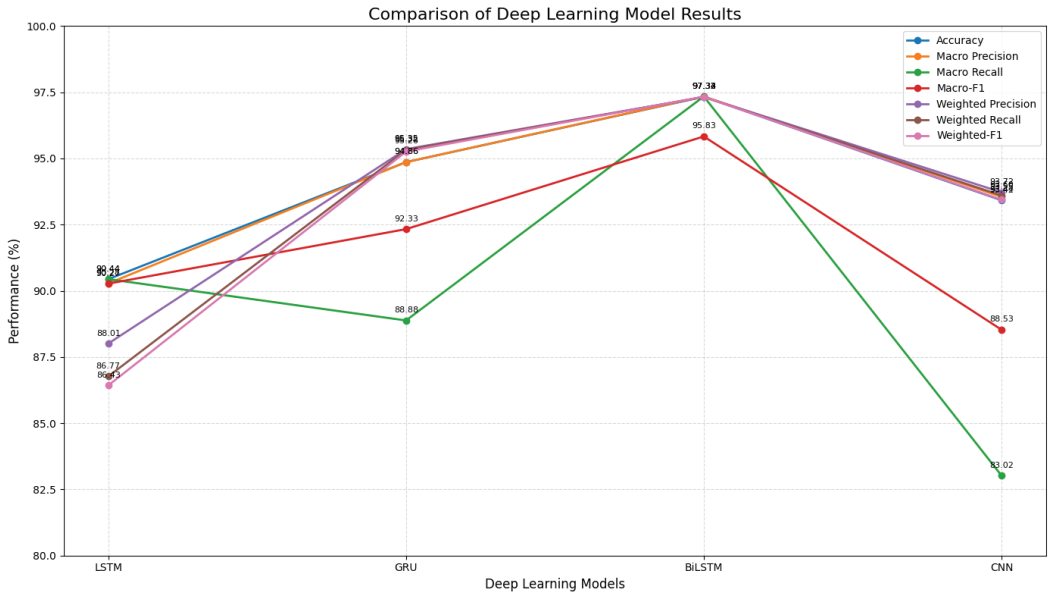

Figure 2: Deep Learning result comparison

http://dx.doi.org/10.21622/ACE.2026.06.1.2076

4.3. CLASS-LEVEL EVALUATION (BILSTM)

Table 4 illustrates the class-wise performance of the BiLSTM model. The model achieved the highest F1-score in the Gender category with 98.43%, indicating very strong discrimination for the most represented class. The high performance was also observed for Not hate, Offensive, Sexual, Disability, Race, and Chronic disease, all of which recorded F1-scores above 95%. Compared to a lower performance was observed for Religion (92.16%) and Tribe (93.56%), suggesting that these categories were relatively more difficult to distinguish. This may be attributed to contextual overlap with other abusive or identity-based expressions, fewer representative training examples, or semantic ambiguity in the dataset. The overall class-level results confirm that BiLSTM maintained high and relatively balanced performance across the nine hate-speech categories.

Table 3: BiLSTM model Class-level evaluation

Class	Precision (%)	Recall (%)	F1-score (%)	Support (%)
Not hate	96.87	96.56	96.71	320
Offensive	98.90	94.50	96.65	382
Sexual	98.22	93.80	95.96	823
Gender	97.89	98.98	98.43	6455
Disability	97.45	95.89	96.66	2508
Race	98.88	93.62	96.17	94
Chronic disease	94.74	97.67	96.18	129
Religion	88.92	95.64	92.16	344
Tribe	94.03	93.10	93.56	203

4.4. IMAGE-BASED MODELS' PERFORMANCE

The image classification pipeline based on the ResNet50V2 architecture achieved an overall accuracy of 89.77%, outperforming MobileNetV2 with an accuracy of 75.67. This demonstrates the ResNet50V2 model's ability to identify visual indicators of hate speech within social media images. Although this performance is slightly lower than that of the text-based BiLSTM model, the results remain significant given the inherent complexity of visual data and the contextual ambiguity often associated with images. Unlike text, where meaning is explicitly encoded in language, images frequently require contextual interpretation, making hate-related cues more subtle and sometimes difficult to generalize.

The model effectively learned to recognize key visual patterns commonly associated with hateful content, including offensive symbols, provocative imagery, and contextually charged visuals. The confusion matrix analysis indicated a high number of correctly classified instances, particularly for clearly distinguishable hate-related images. This suggests that the ResNet50V2 model was successful in extracting meaningful spatial features and leveraging pre-trained representations to identify patterns relevant to the classification task.

However, some misclassifications were observed, particularly in cases where images lacked clear semantic indicators or required external context for interpretation. False Positives occurred when non-hateful images contained visually similar elements to known hate symbols, leading the model to misinterpret them. Conversely, False Negatives were noted in instances where hate speech was implied through subtle or culturally specific imagery that the model could not fully capture. These limitations highlight a key challenge in image-based hate speech detection, where meaning is often implicit and dependent on contextual or cultural knowledge beyond pixel-level features.

Despite these challenges, the ResNet50V2 model demonstrated robustness and strong

http://dx.doi.org/10.21622/ACE.2026.06.1.2076

generalization capability, benefiting from transfer learning and deep feature extraction. Its performance confirms that visual data provides valuable complementary information in hate speech detection, even though it may not be sufficient when used in isolation. These findings underscore the importance of integrating image analysis with textual understanding to improve overall detection accuracy. The comparison of the results is summarized in Table 5.

Table 4: MobileNetV2 and ResNet50V2 Models Performance"

MODELS	ACCURACY(%)	PRECISION(%)	RECALL(%)	F1(%)
MobileNetV2	75.67	78.95	75.67	73.05
ResNet50V2	89.77	89.76	89.76	89.76

4.5. MULTIMODAL FUSION RESULTS

The multimodal fusion framework adopted a decision-level late fusion strategy. In this approach, the BiLSTM text classifier and the ResNet50V2 image classifier were first trained independently on their respective modalities. During inference, the probability outputs generated by the SoftMax layer of the BiLSTM model and the sigmoid probability outputs from the ResNet50V2 model were concatenated into a unified feature representation. The combined prediction vectors were then passed through a fully connected neural fusion layer followed by SoftMax activation to generate the final hate speech classification. This late fusion approach was selected because the textual and image datasets differed substantially in size and structure, making direct feature-level alignment less suitable for the current study. The approach also reduced the complexity associated with heterogeneous multimodal feature synchronization while still enabling complementary contextual learning across modalities.

The performance of the multimodal model, as illustrated in Figure 3, demonstrates strong and balanced results across all evaluation metrics. The model achieved an accuracy of 92.15%, a precision of 91.87%, a recall of 92.35%, and an F1-score of 92.11%. The close alignment of these metrics indicates that the model maintains a good balance between correctly identifying hate speech and minimizing false classifications. In particular, the slightly higher Recall suggests that the model is highly effective in capturing hate speech instances, reducing the likelihood of missed detections.

The multimodal approach significantly reduced misclassification rates by leveraging the strengths of both text and image models. For instance, it addressed challenges such as sarcasm or implicit hate that may not be easily detected in text-only models, as well as contextual nuances in memes that image-only models might fail to interpret. By combining these complementary perspectives, the system achieved more robust and reliable predictions.

Furthermore, the fusion of the best-performing text model (BiLSTM) and image model (ResNet50V2) resulted in superior performance compared to individual unimodal approaches. While the BiLSTM alone achieved higher standalone accuracy, the multimodal model provided a more balanced and generalized performance across different types of data. This demonstrates that integrating textual and visual features enhances the overall effectiveness of Kiswahili hate speech detection systems. From the results, we confirm that multimodal fusion is a valuable strategy for improving classification performance in complex, real-world scenarios.

Although transformer-based models such as SwahBERT are widely regarded as state-of-the-art in NLP tasks, the results obtained from the BiLSTM model in this study demonstrate that competitive performance can still be achieved using less computationally intensive architectures. This is particularly important in real-world deployment scenarios where computational efficiency and latency are critical factors. The findings suggest that while transformer-based models such as SwahBERT may provide superior contextual

http://dx.doi.org/10.21622/ACE.2026.06.1.2076

representations, the performance gap may not always justify the significantly higher computational cost, especially in applications targeting resource-constrained environments. The BiLSTM model, when properly tuned and combined with effective preprocessing techniques, provides a viable alternative that balances performance and efficiency. Furthermore, the integration of ResNet50v2 for image classification within the multimodal framework enhances the overall system's capability to detect hate speech beyond text alone. This multimodal approach represents a key contribution of the study, distinguishing it from purely text-based transformer models. Based on the multimodal approach, we developed an interactive hate speech detection interface called *ChujaHate*, which was validated by experts, presenting another key contribution of the study.

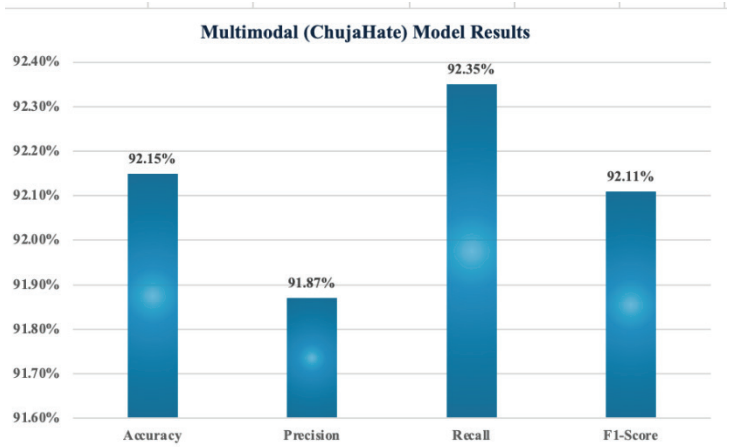

Figure 3: Multimodal Hate Speech Model results

4.6. COMPARATIVE ANALYSIS BETWEEN THE REVIEWED AND THE DEVELOPED MULTIMODAL FUSION MODEL

The findings of this study demonstrate that multimodal hate speech detection in low-resource African languages can achieve performance levels comparable to, and in some cases exceeding, those reported in prior multilingual hate speech studies. While previous studies, such as [1], achieved strong results using transformer-based architectures on code-switched English-Kiswahili text, the current study extends this body of knowledge by integrating both textual and visual modalities within a unified multimodal framework. Unlike earlier approaches that relied solely on textual analysis, the ChujaHate framework demonstrated that combining BiLSTM and ResNet50V2 models significantly improves contextual understanding, particularly for implicit and image-supported hate speech. This finding aligns with studies by [27] and [45], which observed that multimodal fusion provides richer semantic representation than unimodal systems.

The study further contributes to low-resource language research by demonstrating that computationally efficient architectures such as BiLSTM can still achieve highly competitive performance despite the increasing dominance of transformer-based models.

Although transformer-based architectures such as mBERT, AfroXLMR, CLIP, and multimodal vision-language transformers have demonstrated strong performance in multilingual hate speech detection tasks, they were not implemented in the current study due to computational and dataset-related constraints. Training and fine-tuning such models require substantially larger multimodal datasets, higher GPU memory requirements, and longer training durations. The present study, therefore, prioritized computationally efficient architectures, including BiLSTM and ResNet50V2, which remain suitable for deployment in resource-constrained African NLP environments. Nevertheless, transformer-based multimodal models remain an important direction for future research.

http://dx.doi.org/10.21622/ACE.2026.06.1.2076

5. ETHICS STATEMENT

Data collection adhered to social media platform guidelines, ensuring compliance with privacy policies. User data was anonymized to protect personal information, and only publicly available datasets were used. Further, this study was undertaken under a data collection license issued to the researcher by the Kenya National Commission for Science, Technology, and Innovation (NACOSTI). Ethical approval was also obtained from the United States International University - Africa prior to the commencement of the study.

6. LIMITATIONS

The text dataset was substantially larger than the image dataset, creating a modality imbalance that may have influenced multimodal learning and fusion performance. The hate speech dataset may also not have fully captured all forms of hate speech, particularly sarcasm, coded Kiswahili expressions, and culturally specific references.

Although the study evaluated the performance of text-only, image-only, and multimodal models, a comprehensive ablation analysis involving multiple fusion strategies and feature contribution experiments was not conducted. Specifically, the study did not compare early fusion, attention-based fusion, and transformer-based multimodal architectures due to computational limitations and the relatively small size of the image dataset. Future work should therefore include expansion of the dataset to ensure balance, extensive ablation experiments to better quantify modality contribution, feature interaction, and the effectiveness of alternative multimodal fusion strategies.

7. RECOMMENDATIONS AND FUTURE WORK

i) Community-Driven Annotation and Local Collaboration:

Establishing community-based annotation hubs through partnerships with universities, research institutions, and civic technology organizations would also provide a sustainable approach to developing high-quality datasets for low-resource languages such as Kiswahili.

ii) Expansion to Additional Modalities:

Future research should extend beyond text and image analysis by incorporating additional modalities such as audio, video, emojis, and other forms of digital expression commonly used on platforms like TikTok and WhatsApp. Hate speech communication on modern social media platforms is increasingly multimodal, and integrating these additional data forms would provide a more comprehensive understanding of how harmful content is created, shared, and interpreted online.

iii) Institutional Integration and Policy Alignment:

Government institutions, digital rights organizations, and social media platforms should work collaboratively to integrate multimodal hate speech detection systems into existing policy enforcement and content moderation frameworks. The adoption of AI-assisted moderation tools can support faster and more proactive detection of harmful online content while reducing the workload placed on human moderators.

iv) Public Awareness and Digital Literacy:

There is also a need to strengthen public awareness on the role of artificial intelligence and machine learning in combating online hate speech. Governments, civil society organizations, educational institutions, and private sector stakeholders should develop digital literacy programs that educate users on how automated hate speech detection systems operate, their benefits, and their limitations. Such initiatives would promote responsible online

http://dx.doi.org/10.21622/ACE.2026.06.1.2076

behavior, encourage informed use of digital platforms, and foster a culture of respectful and inclusive digital communication.

5. CONCLUSION

This study introduced *ChujaHate*, which, to the best of our knowledge, is the first multimodal hate speech detection framework developed specifically for Kiswahili social media content. The study addressed an important gap in low-resource African language NLP research by integrating both textual and visual modalities in hate speech detection. The findings showed that the text-based model achieved the strongest overall performance across the nine fine-grained hate speech categories, while the image-based model also demonstrated promising capability in detecting hateful visual content. When both modalities were combined using a late fusion approach, the multimodal framework showed complementary benefits, confirming that hateful meaning in social media is often distributed across both text and images. However, the improvement achieved through multimodal fusion was moderated by the relatively smaller image dataset, which created a modality imbalance during model training.

Beyond the technical findings, the study has practical significance for content moderation in Kiswahili-speaking online communities, where hate speech is frequently communicated through culturally contextualized combinations of language, memes, and imagery. By incorporating both text and image information, *ChujaHate* provides a more context-aware and inclusive framework that can support social media platforms, policymakers, researchers, and civil society organizations in developing more effective moderation systems for low-resource multilingual environments. The study therefore contributes not only to the advancement of multimodal machine learning research but also to ongoing efforts aimed at promoting safer and more responsible digital communication spaces in Africa.

Despite these contributions, the study faced several limitations, including the relatively small image dataset, possible subjectivity and cultural bias during manual annotation, and the use of non-contextual embeddings in parts of the modeling pipeline, which may have limited the capture of nuanced contextual meanings in Kiswahili hate speech.

Nevertheless, the findings of this study demonstrate that multimodal machine learning can provide more context-aware hate speech detection than unimodal systems, particularly in low-resource multilingual environments where hateful meaning is often distributed across both text and imagery. While the text-only BiLSTM achieved the highest standalone accuracy, the multimodal framework improved contextual robustness and practical applicability in real-world social media scenarios. The study, therefore, highlights the importance of balancing predictive performance, computational efficiency, and deployment feasibility when designing AI systems for African language technologies.

Future research should therefore focus on developing larger and more balanced multimodal datasets, adopting transformer-based models such as mBERT fine-tuned for Kiswahili, exploring advanced fusion techniques such as attention-based and cross-modal learning, and extending analysis to audio and video modalities to further improve multimodal hate speech detection systems in African languages.

REFERENCES

[1] 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: https://doi.org/10.30534/ijatcse/2024/011352024.

[2] G. Arya *et al.*, "Multimodal Hate Speech Detection in Memes using Contrastive

Language-Image Pre-training," *IEEE access*, vol. 12, pp. 22359–22375, Jan. 2024, doi: https://doi.org/10.1109/access.2024.3361322.

[3] A. Marshan, F. Nasreen, A. Ioannou, and K. Spanaki, "Comparing Machine Learning and Deep Learning Techniques for Text Analytics: Detecting the Severity of Hate Comments Online," *Information Systems Frontiers*, pp. 1–19, Nov. 2023, doi: https://doi.org/10.1007/s10796-023-10446-x.

[4] S. MacAvaney, H.-R. Yao, E. Yang, K. Russell, N. Goharian, and O. Frieder, "Hate speech detection: Challenges and solutions," *PLOS ONE*, vol. 14, no. 8, p. e0221152, Aug. 2019, doi: https://doi.org/10.1371/journal.pone.0221152.

[5] United Nations, "United Nations Strategy and Plan of Action on Hate Speech," *United Nations Digital Library System*, 2019. https://digitallibrary.un.org/record/3889290

[6] N. Onyango, L. Wanzare, and J. I. Obuhuma, "Swahili and Code-Switched English-Swahili Political Hate Speech Detection Textual Dataset," *Data Intelligence*, vol. 7, no. 3, pp. 819–850, 2025, doi: https://doi.org/10.3724/2096-7004.di.2025.0053.

[7] B. C. Solomon, Matthew, A. Hemmen, and J. N. Druckman, "Illusory interparty disagreement: Partisans agree on what hate speech to censor but do not know it," *Proceedings of the National Academy of Sciences*, vol. 121, no. 39, Sep. 2024, doi: https://doi.org/10.1073/pnas.2402428121.

[8] F. M. Ndahinda and A. S. Mugabe, "Streaming Hate: Exploring the Harm of Anti-Banyamulenge and Anti-Tutsi Hate Speech on Congolese Social Media," *Journal of Genocide Research*, vol. 26, no. 1, pp. 1–25, May 2022, doi: https://doi.org/10.1080/14623528.2022.2078578.

[9] E. Ombui, M. Karani, and L. Muchemi, "Annotation Framework for Hate Speech Identification in Tweets: Case Study of Tweets During Kenyan Elections," *IEEE Xplore*, May 01, 2019. https://ieeexplore.ieee.org/document/8764868 (accessed Apr. 14, 2022).

[10] T. M. Ababu and M. M. Woldeyohannis, "Afaan Oromo Hate Speech Detection and Classification on Social Media," in *ACL Anthology*, Proc. 13th Language Resources and Evaluation Conf, Jun. 2022, pp. 6612–6619.

[11] O. Oriola and Kotze E., "Improved semi-supervised learning technique for automatic detection of South African abusive language on Twitter," *South African Computer Journal*, vol. 32, no. 2, pp. 56–79, Dec. 2020, doi: https://doi.org/10.18489/sacj.v32i2.847.

[12] S. Badillo *et al.*, "An Introduction to Machine Learning," *Clinical Pharmacology & Therapeutics*, vol. 107, no. 4, pp. 871–885, Mar. 2020, doi: https://doi.org/10.1002/cpt.1796.

[13] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated Hate Speech Detection and the Problem of Offensive Language," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 11, no. 1, pp. 512–515, May 2017, doi: https://doi.org/10.1609/icwsm.v11i1.14955.

[14] M. Wiegand, J. Ruppenhofer, Anna Marie Schmidt, and C. Greenberg, "Inducing a Lexicon of Abusive Words – a Feature-Based Approach," in *Publication Server of the Institute for German Language (Institute for German Language)*, Leibniz Institute for the German Language: Proc. NAACL-HLT, Jan. 2018. doi: https://doi.org/10.18653/v1/n18-1095.

[15] P. Mishra, M. D. Tredici, H. Yannakoudakis, and E. Shutova, "Author Profiling for Hate Speech Detection," 2019, doi: https://doi.org/10.48550/arXiv.1902.06734.

[16] G. Martin, M. E. Mswahili, Y.-S. Jeong, and J. Woo, "SwahBERT: Language Model of

Swahili," in *ACLWeb* , Seattle, United States: Association for Computational Linguistics, Jul. 2022, pp. 303–313. doi: https://doi.org/10.18653/v1/2022.naacl-main.23.

[17] O. F. Babu, M. Jahan, A. Faisal, Md. S. Islam, and R. Khan, "Bangla Hate Speech Detection System Using Transformer-Based NLP and Deep Learning Techniques," in *Proc. 3rd Asian Conf. Innovation in Technology (ASIANCON)*, Aug. 2023, pp. 1–6. doi: https://doi.org/10.1109/asiancon58793.2023.10269919.

[18] A. S. Alammary, "BERT Models for Arabic Text Classification: A Systematic Review," *Applied Sciences*, vol. 12, no. 11, p. 5720, Jun. 2022, doi: https://doi.org/10.3390/app12115720.

[19] D. Adelani, G. Neubig, S. Ruder, and S. Rijhwani, "MasakhaNER 2.0: Africa-centric Transfer Learning for Named Entity Recognition," *arXiv preprint arXiv:2210.12391*, 2022, doi: https://doi.org/10.48550/arXiv.2210.12391.

[20] A. Kaliba, "Performance Assessment of a New Swahili Lexicon (SWAHILILex.01) Tagged by Native Speakers for Polarity Analysis," May 2023, doi: https://doi.org/10.36227/techrxiv.22806782.

[21] M. Kandagor, "THE PLACE OF KISWAHILI IN THE TWENTY-FIRST CENTURY | Mark M. Kandagor," in *Mu.ac.ke*, Youth, Globalization, and Society in Africa and Its Diaspora, 2020, p. 114.

[22] 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," *arXiv.org*, 2023. https://arxiv.org/abs/2306.00410

[23] A. Irfan, D. Azeem, S. Narejo, and N. Kumar, "Multi-Modal Hate Speech Recognition Through Machine Learning," *Proc. IEEE 1st Karachi Section Humanitarian Technology Conf. (KHI-HTC)*, pp. 1–6, Jan. 2024, doi: https://doi.org/10.1109/khi-htc60760.2024.10482031.

[24] A. G. Debele and M. M. Woldeyohannis, "Multimodal Amharic Hate Speech Detection Using Deep Learning," *Proc. Int. Conf. Information and Communication Technology for Development for Africa (ICT4DA)*, pp. 102–107, Nov. 2022, doi: https://doi.org/10.1109/ict4da56482.2022.9971436.

[25] K. Perifanos and D. Goutsos, "Multimodal Hate Speech Detection in Greek Social Media," *Multimodal Technologies and Interaction*, vol. 5, no. 7, p. 34, Jun. 2021, doi: https://doi.org/10.3390/mti5070034.

[26] P. Kapil and A. Ekbal, "A transformer based multi task learning approach to multimodal hate detection," *Natural Language Processing Journal*, vol. 11, p. 100133, Feb. 2025, doi: https://doi.org/10.1016/j.nlp.2025.100133.

[27] M. D. Belete and Girma Kassa Alitasb, "Identification of Hateful Amharic Language Memes on Facebook using Deep Learning Algorithms," *Systems and Soft Computing*, vol. 7, pp. 200258–200258, Apr. 2025, doi: https://doi.org/10.1016/j.sasc.2025.200258.

[28] F. K. Saddozai, S. K. Badri, D. Alghazzawi, A. Khattak, and M. Z. Asghar, "Multimodal hate speech detection: a novel deep learning framework for multilingual text and images," *PeerJ Computer Science*, vol. 11, p. e2801, Apr. 2025, doi: https://doi.org/10.7717/peerj-cs.2801.

[29] F. Vargas *et al.*, "HausaHate: An Expert Annotated Corpus for Hausa Hate Speech Detection," *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024)*, pp. 52–58, 2024, doi: https://doi.org/10.18653/v1/2024.woah-1.5.

[30] B. Vidgen and T. Yasseri, "Detecting weak and strong Islamophobic hate speech on social media," *Journal of Information Technology & Politics*, vol. 17, no. 1, pp. 66–78,

Dec. 2019, doi: https://doi.org/10.1080/19331681.2019.1702607.

[31] A. K. Diallo and K. Abainia, "Offensive Language Detection in Code-Mixed Bambara-French Corpus: Evaluating machine learning and deep learning classifiers," *2023 International Conference on Decision Aid Sciences and Applications (DASA)*, pp. 121–125, Sep. 2023, doi: https://doi.org/10.1109/dasa59624.2023.10286577.

[32] S. H. Muhammad *et al.*, "AfriSenti: A Twitter Sentiment Analysis Benchmark for African Languages," Feb. 2023, doi: https://doi.org/10.48550/arxiv.2302.08956.

[33] F. Ieracitano, C. Balenzano, S. Girardi, C. G. Gemmano, and F. Comunello, "Online Hate Speech as a Moral Issue: Exploring Moral Reasoning of Young Italian Users on Social Network Sites," *Social Science Computer Review*, vol. 42, no. 1, p. 089443932311611, Mar. 2023, doi: https://doi.org/10.1177/08944393231161124.

[34] R. A. Akbar, A. Mulyana, and M. Amalia, "Legal Challenges In The Age Of Social Media: Protecting Citizens From Misuse Of Information," *Golden Ratio of Law and Social Policy Review*, vol. 3, no. 1, pp. 14–25, Dec. 2023, doi: https://doi.org/10.52970/grlspr.v3i1.328.

[35] 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, pp. 29–47, Dec. 2021, doi: https://doi.org/10.5815/ijitcs.2021.06.03.

[36] A. Chhabra and D. K. Vishwakarma, "A literature survey on multimodal and multilingual automatic hate speech identification," *Multimedia Systems*, vol. 29, no. 3, Jan. 2023, doi: https://doi.org/10.1007/s00530-023-01051-8.

[37] F. Wu, G. Chen, J. Cao, Y. Yan, and Z. Li, "Multimodal Hateful Meme Classification Based on Transfer Learning and a Cross-Mask Mechanism," *Electronics*, vol. 13, no. 14, p. 2780, Jul. 2024, doi: https://doi.org/10.3390/electronics13142780.

[38] F. Chen, X. Li, Z. Li, C. Zhou, and J. Sheng, "Multimodal Rumor Detection via Multimodal Prompt Learning," *2022 International Joint Conference on Neural Networks (IJCNN)*, vol. 7, pp. 1–8, Jun. 2024, doi: https://doi.org/10.1109/ijcnn60899.2024.10650974.

[39] P. Kapil and A. Ekbal, "A transformer based multi task learning approach to multimodal hate detection," *Natural Language Processing Journal*, vol. 11, p. 100133, Feb. 2025, doi: https://doi.org/10.1016/j.nlp.2025.100133.

[40] P. Vijayaraghavan, H. Larochelle, and D. Roy, "Interpretable Multi-Modal Hate Speech Detection," *arXiv.org*, 2021. https://arxiv.org/abs/2103.01616? (accessed Jun. 06, 2026).

[41] J. L. Imbwaga, N. B. Chittaragi, and S. G. Koolagudi, "Automatic hate speech detection in audio using machine learning algorithms," *International Journal of Speech Technology*, vol. 27, no. 2, pp. 447–469, Jun. 2024, doi: https://doi.org/10.1007/s10772-024-10116-6.

[42] Z. Zhao, Z. Zhang, and F. Hopfgartner, "Detecting Toxic Content Online and the Effect of Training Data on Classification Performance," *EasyChair Preprints*, Apr. 2019, doi: https://doi.org/10.29007/z5xk.

[43] M. Zampieri, S. Rosenthal, Preslav Nakov, Alphaeus Dmonte, and T. Ranasinghe, "OffensEval 2023: Offensive language identification in the age of Large Language Models," *Natural language engineering*, vol. 29, no. 6, pp. 1416–1435, Nov. 2023, doi: https://doi.org/10.1017/s1351324923000517.