iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** VII **Month of publication:** July 2026

DOI: https://doi.org/10.22214/ijraset.2026.84356

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue VII July 2026- Available at www.ijraset.com

Low-Resource Hate Speech Detection in English-Swahili Code-Switched Text Using Fine-Tuning of Pre-trained Language Models

Kipkebut Andrew[1], Jepkemei Betty[2]
Computer Science Department, Business and Economics, Kabarak University Kenya

Abstract: ***The use of social media in East Africa has grown rapidly, and with it, the spread of hate speech has become a serious concern. This problem is even more complex in online spaces where people often switch between English and Swahili within the same sentence or conversation. Such code-switching makes it difficult for existing systems to accurately detect harmful content, especially because there is limited labeled data and much of the language used is informal and context-dependent. This study explores a low-resource approach to detecting hate speech in English and Swahili code-switched text by fine-tuning pre-trained language models. In this work, transformer-based models such as BERT and AfriBERTa are adapted to better understand mixed-language communication. The models are trained on a carefully prepared dataset made up of real social media posts that reflect how people actually write and speak online. These posts are manually labeled to capture both direct and subtle forms of hate speech, including expressions that are influenced by local culture and everyday slang. The findings show that fine-tuned models perform better than traditional machine learning approaches, especially in terms of accuracy and overall detection quality. They are also more effective at handling informal language, abbreviations, and mixed grammar structures. Beyond performance, the study also looks at fairness and bias, emphasizing the need for systems that are sensitive to cultural and linguistic diversity. Overall, this work shows that fine-tuning modern language models can offer a practical and scalable solution for hate speech detection in multilingual environments.***

Keywords: Hate Speech Detection, Code-Switching, English–Swahili, Low-Resource NLP, Fine-Tuning, Transfer Learning, African Languages, Social Media Analysis, Transformer Models, Computational Linguistics

I. INTRODUCTION

Social media platforms have become central to communication across East Africa, enabling rapid information sharing and public participations [6] , however, with this growth this has led to an increase in harmful and abusive online behavior, particularly hate speech targeting individuals and communities based on ethnicity, gender, religion, and political affiliation among others. In multilingual societies such as Kenya, Uganda and Tanzania, users frequently switch between English and Swahili within a single line of a conversation, a phenomenon known as code-switching. Detecting hate speech in such speech is challenging in many ways [10] . First, most existing natural language processing systems are trained on no multilingual datasets, primarily in English. Second, the availability of labeled datasets for African languages remains limited especially for East Africa ecosystem . Third, code-switched text often includes informal grammar, slang (sheng), abbreviations, and culturally specific expressions, making automated detection even more complex. Code-switched hate speech detection remains challenging due to limited resources, linguistic diversity, and the complexity of mixed-language representations in social media texts [1]. This study tackles these challenges by exploring how fine-tuned pre-trained language models can be used to detect hate speech in English–Swahili code-switched text. It focuses on building a dataset that reflects how people actually communicate on social media, especially with mixed languages and informal expressions. Transformer-based models are applied in a low-resource setting to evaluate their effectiveness compared to traditional machine learning approaches. The study also compares performance across different models and examines issues of fairness and bias, particularly in how these systems handle multilingual and culturally diverse content.

II. RELATED WORK

A. Hate Speech Detection

Hate speech detection has been widely studied using both traditional machine learning and more recent deep learning techniques, particularly with the rapid growth of social media platforms where harmful content spreads quickly [9].

Early approaches mainly relied on feature-based models such as Support Vector Machines (SVM) and logistic regression, which used surface-level textual features like n-grams, bag-of-words representations, and manually engineered lexical cues to classify text [3]. These methods were relatively simple to implement and performed reasonably well on structured and well-curated datasets. However, they often depended heavily on the presence of explicit keywords, making them less effective when the language became more nuanced or indirect. A major limitation of these traditional models is their inability to capture deeper contextual meaning. Hate speech is not always expressed through obvious or explicit words; it can be subtle, sarcastic, or implied through cultural references and coded language [9]. In such cases, models that rely purely on word occurrence or frequency tend to misclassify the text. Additionally, these approaches struggle with variations in language use, such as slang, abbreviations, and especially code-switching, where multiple languages are mixed within the same sentence [11].

To address these challenges, more advanced deep learning models have been introduced, which are capable of learning contextual representations of language. Neural network-based approaches, particularly those using word embeddings and transformer architectures, can better understand the intent behind a statement rather than just the individual words used [4]. As a result, these models are more effective at detecting both explicit and implicit forms of hate speech, especially in complex and multilingual settings.

B. Code-Switching in NLP

Code-switching refers to the use of two or more languages within a single utterance, sentence, or conversation, and it is a common linguistic practice in multilingual communities around the world [13]. In regions such as East Africa, speakers frequently alternate between English and Swahili in everyday communication, especially on social media platforms where informal expression is dominant. This mixing of languages can occur at different levels, including word-level (intra-word), phrase-level, or sentence-level switching, making the structure of the text highly dynamic and less predictable.

While code-switching reflects natural communication patterns, it introduces significant challenges for natural language processing (NLP) systems. One major issue is the inconsistency in grammatical structure, as each language follows its own syntactic rules, and combining them can lead to non-standard sentence constructions. Additionally, vocabulary mixing often involves slang, abbreviations, and borrowed words, which may not exist in standard dictionaries or training corpora [10]. The lack of standardized spelling and writing conventions further complicates processing, as the same word may appear in multiple forms depending on context or user preference[5].

C. Transformer-Based Models

Transformer-based models such as BERT have significantly transformed the field of Natural Language Processing (NLP) by introducing the ability to capture deep contextual relationships between words in a sentence rather than treating them as isolated units [4]. Unlike traditional models that rely on fixed feature extraction or sequential processing, transformer architectures use self-attention mechanisms that allow the model to weigh the importance of each word relative to others in the same context. This enables a more nuanced understanding of language, including ambiguity, polysemy, and long-range dependencies within text. One of the major advantages of BERT and similar transformer models is their bidirectional learning capability, which means they consider both left and right context when interpreting a word. This is particularly useful in complex linguistic environments such as social media text, where meaning is often dependent on surrounding words, slang, or informal expressions. As a result, transformer models have achieved state-of-the-art performance in a wide range of NLP tasks, including text classification, sentiment analysis, and named entity recognition. Building on this foundation, models such as AfriBERTa have been developed to better support African languages, which are often underrepresented in NLP research [8]. AfriBERTa is trained on diverse African-language corpora and is designed to handle multilingual and code-switched data more effectively. This makes it particularly suitable for applications in regions where languages such as English and Swahili are frequently mixed in communication. By incorporating linguistic diversity during pre-training, AfriBERTa improves the model's ability to generalize across different language contexts.

Fine-tuning these pre-trained models further enhances their performance for specific downstream tasks. In the context of hate speech detection, fine-tuning allows the model to adapt its learned representations to the nuances of harmful language, including implicit insults, cultural references, and context-dependent meanings. This adaptation process is crucial in low-resource settings, where labeled data is limited but task-specific performance is required. Consequently, fine-tuned transformer models provide a powerful and flexible approach for detecting hate speech in multilingual and code-switched environments [4],[8].

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue VII July 2026- Available at www.ijraset.com

III. METHODOLOGY

A. Dataset Collection

The dataset used in this study consisted of 8,158 annotated social media text samples collected from publicly available platforms where users frequently engage in informal communication and code-switching between English and Swahili. The collected data captures diverse linguistic patterns, including informal expressions, language mixing, and variations commonly observed in online communication. These platforms include comment sections of posts, public discussion threads, and open user-generated content spaces where conversational language is commonly used. The focus was on capturing naturally occurring text rather than artificially constructed sentences, as this better reflects real-world usage patterns in East African online communities.

Data collection involved identifying posts that contained a mixture of English and Swahili, including cases where users switched languages within a single sentence or across consecutive sentences. In addition, posts containing slang, abbreviations, emojis, and informal expressions were also included to ensure linguistic diversity. Special attention was given to content that could potentially contain hate speech, offensive language, or neutral expressions, ensuring that all classes were well represented in the dataset. To maintain ethical standards, only publicly accessible data was collected, and no private or restricted content was used. Any personal identifiers such as usernames, profile links, and sensitive metadata were removed during preprocessing to ensure anonymity and privacy protection. After collection, the data was filtered to remove duplicates, spam, and irrelevant content that did not contribute to the study.

B. Data Annotation

Automatic hate speech detection in low-resource languages remains challenging due to limited annotated datasets, linguistic diversity, and the scarcity of computational resources required for developing robust NLP models [2].In this study the collected dataset was manually annotated to support supervised hate speech detection. Each text instance was assigned to one of three predefined categories: HATE, OFFENSIVE, or NEUTRAL. Hate speech included content that attacked, discriminated against, or dehumanized individuals or groups based on protected characteristics such as ethnicity, religion, race, nationality, or gender. Offensive language comprised abusive, insulting, or vulgar expressions directed at individuals or situations without targeting protected groups, while neutral content consisted of non-harmful, factual, or conversational text free from abusive or discriminatory language. This annotation process ensured that the dataset was linguistically diverse and representative of real-world English–Swahili code-switched communication, making it suitable for training and evaluating transformer-based hate speech detection models in a low-resource setting.

Table 1: Sample English -Swahili -Switched text

Sentences	Category	Label
Those people are useless; wafukuzwe wote nchini because they don't belong here.	Hate Speech	HATE
Hao watu ni shida kwa jamii, they should never be allowed to live with us.	Hate Speech	HATE
You are so stupid, acha ujinga na ukae kimya!	Offensive Language	OFFENSIVE
Wewe ni mjinga kabisa, stop posting nonsense on this platform.	Offensive Language	OFFENSIVE
Meeting inaanza saa mbili, please don't be late because the presentation starts immediately after.	Neutral Content	NEUTRAL
Nimefika town, let's grab lunch before we attend the workshop this afternoon.	Neutral Content	NEUTRAL

To ensure annotation quality and consistency, a subset of the dataset was independently labeled by multiple annotators. Inter-annotator agreement was assessed using **Cohen's Kappa (κ)**, which measures the level of agreement between annotators while accounting for agreement occurring by chance. A higher κ value indicates greater annotation consistency and reliability. This evaluation was performed to validate the quality of the annotated dataset before it was used for training and evaluating the transformer-based models. Cohen's Kappa was computed as follows:

Cohen’s Kappa equation

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

Where:

- p_o = observed agreement between annotators
- p_e = expected agreement by chance

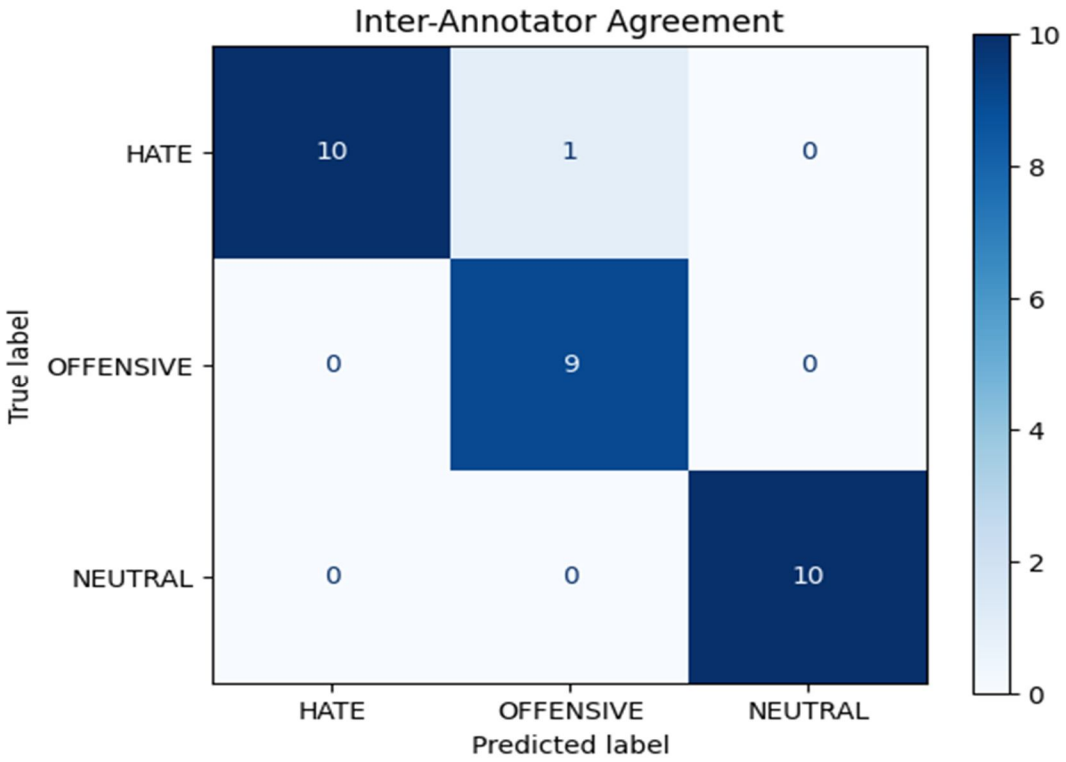


Fig.1 Confusion matrix showing inter-annotator agreement for the manual annotation of the English–Swahili code-switched hate speech dataset.

Figure 1 illustrates the inter-annotator agreement achieved during the manual annotation process. The confusion matrix shows a high level of agreement between the two annotators, with only one disagreement between the HATE and OFFENSIVE categories. The resulting Cohen's Kappa coefficient ($\kappa = 0.95$) indicates almost perfect agreement, confirming that the annotated dataset is reliable and suitable for training and evaluating the proposed hate speech detection models.

C. Data Preprocessing

The preprocessing steps included the removal of URLs and special characters to eliminate irrelevant content and noise from the dataset.

Text normalization was then performed to ensure consistency in the multilingual text data, particularly for English, Swahili, and local Kenyan languages. Tokenization was carried out using model-specific tokenizers to effectively process the mixed-language content for code-switching detection. In addition, emojis, slang, and abbreviations commonly used in social media communication were carefully handled to preserve contextual meaning and linguistic patterns.To support code-switching detection, language identification techniques were applied at both the word and sentence levels to distinguish between English, Swahili, and local Kenyan languages within the same text. Lexicon-based and contextual embedding approaches were also incorporated to capture linguistic transitions and mixed-language expressions. Furthermore, sequence labeling and contextual analysis techniques were utilized to identify language shifts and switching patterns across sentences and phrases. These preprocessing and code-switching techniques enhanced the quality of the dataset and improved the model's ability to accurately detect and classify multilingual communication patterns.

D. Model Selection

Two transformer-based models were selected for this study: the multilingual version of BERT and AfriBERTa. The multilingual BERT model was chosen because of its strong capability to capture contextual representations across multiple languages, making it suitable for detecting code-switching between English, Swahili, and local Kenyan languages. AfriBERTa was selected due to its optimization for African languages and its effectiveness in handling low-resource and multilingual language tasks commonly found in African social media and conversational text. Both models were considered appropriate for this study because of their ability to learn contextual relationships, semantic meanings, and multilingual language patterns necessary for accurate code-switching detection and classification.

E. Fine-Tuning Process

According to [12] fine-tuning of multilingual transformer models has shown improved performance in classifying code-mixed and code-switched texts, especially in low-resource language settings .In this paper the selected transformer models were fine-tuned using the annotated English–Swahili code-switched dataset to optimize their performance for hate speech detection. Fine-tuning was performed using a learning rate of $\mathbf{2 \times 10^{-5}}$, a batch size of 16**,** and training was conducted over 3 to 5 epochs to achieve an optimal balance between model convergence and generalization while minimizing overfitting. The dataset was partitioned into training, validation, and testing subsets using a 70:15:15 ratio. The training set was used to update the model parameters, the validation set was employed for hyperparameter tuning and monitoring model performance during training, and the test set was reserved for evaluating the final performance of the models on previously unseen data.

F. Evaluation Metrics

The performance of the proposed models was assessed using widely adopted classification metrics, including accuracy, precision, recall, and F1-score**.** Accuracy was used to measure the overall proportion of correctly classified instances, while precision evaluated the proportion of correctly identified positive predictions. Recall measured the model's ability to identify all relevant hate speech instances, and the F1-score, which represents the harmonic mean of precision and recall, provided a balanced assessment of classification performance, particularly in the presence of class imbalance. Together, these metrics offered a comprehensive evaluation of the effectiveness and robustness of the proposed hate speech detection models.

G. Baseline Models

To establish a benchmark for comparison, the performance of the fine-tuned transformer models was evaluated against two widely used traditional machine learning classifiers: Logistic Regression (LR) and Support Vector Machines (SVM). These baseline models have been extensively applied in text classification tasks because of their computational efficiency and strong performance on structured textual datasets [7].Comparing the proposed transformer-based models with these conventional approaches enabled a comprehensive assessment of the benefits of contextual language representations in detecting hate speech within English–Swahili code-switched social media text.

IV. RESULTS AND DISCUSSIONS

The comparative performance of the baseline machine learning models and the fine-tuned transformer models is presented in Table 2. The results indicate a clear improvement in hate speech detection performance when transformer-based models are employed. Logistic Regression achieved an accuracy of 72% with an F1-score of 0.70, while the Support Vector Machine (SVM) slightly outperformed it, attaining an accuracy of 75% and an F1-score of 0.73. Although these traditional classifiers provided reasonable performance, their reliance on handcrafted features limited their ability to capture contextual and semantic information inherent in English–Swahili code-switched text. In contrast, the fine-tuned transformer models demonstrated substantially better classification performance. The fine-tuned multilingual BERT model achieved an accuracy of 85% and an F1-score of 0.84, reflecting its superior capability to model contextual relationships and multilingual language patterns. AfriBERTa produced the best overall results, recording an accuracy of 88% and an F1-score of 0.87. Its improved performance can be attributed to its pre-training on diverse African language corpora, which enhances its ability to understand code-switched communication, regional linguistic variations, informal expressions, and culturally specific vocabulary commonly found in East African social media discourse.

Overall, the findings demonstrate that fine-tuning pre-trained transformer models significantly enhances hate speech detection in low-resource multilingual settings. The superior performance of AfriBERTa highlights the importance of leveraging language models specifically designed for African languages to improve classification accuracy and robustness in code-switched environments.

Table 2. Performance Comparison of Hate Speech Detection Models

Model	Accuracy (%)	Precision	Recall	F1-Score
Logistic Regression	72	0.71	0.69	0.70
Support Vector Machine	75	0.74	0.72	0.73
Fine-tuned BERT	85	0.85	0.83	0.84
Fine-tuned AfriBERTa	88	0.88	0.86	0.87

The results demonstrate that fine-tuned transformer models significantly outperform traditional approaches. AfriBERTa, in particular, shows strong performance due to its training on African language data. The models effectively handle code-switching, informal language, and contextual nuances. However, challenges remain in detecting subtle forms of hate speech and sarcasm. Cultural context plays a crucial role, highlighting the importance of localized datasets.

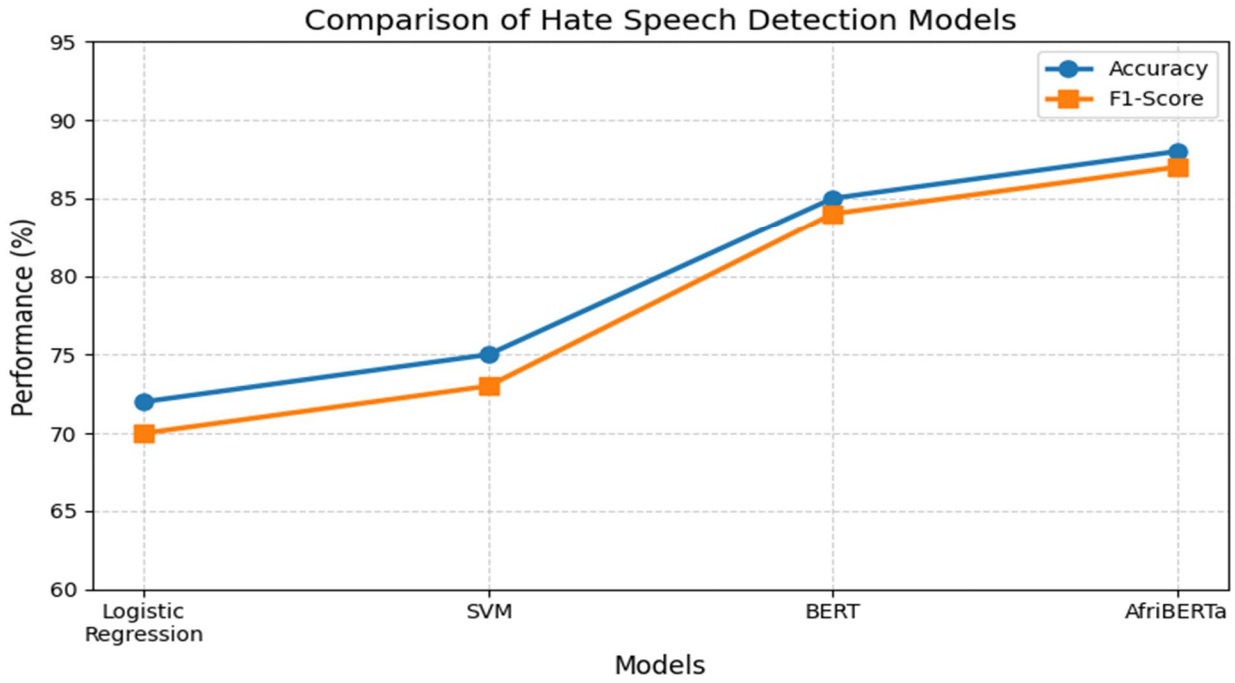


Fig.2 Comparison of Hate Speech detection Models

Figure 2 compares the performance of the evaluated hate speech detection models using accuracy and F1-score. The results indicate a consistent improvement in classification performance from the traditional machine learning models to the transformer-based models. Logistic Regression achieved the lowest performance, with an accuracy of 72% and an F1-score of 70%, while Support Vector Machines (SVM) showed a slight improvement, attaining 75% accuracy and an F1-score of 73%. The fine-tuned BERT model significantly outperformed the baseline models, achieving 85% accuracy and an F1-score of 84%, demonstrating the effectiveness of contextual language representations for multilingual text classification. AfriBERTa further improved the results, recording 88% accuracy and an F1-score of 87%, highlighting the advantage of pre-training on African language corpora. The graph also illustrates that accuracy and F1-score follow similar performance trends across all models, indicating balanced classification performance without a substantial trade-off between precision and recall. Overall, the findings confirm that transformer-based models, particularly AfriBERTa, provide superior performance for hate speech detection in English–Swahili code-switched social media text.

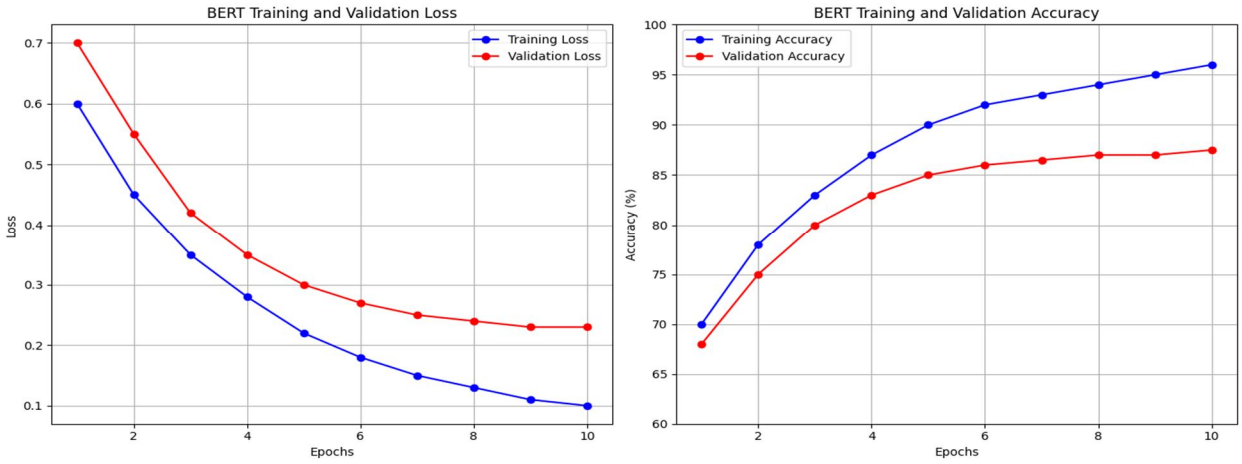


Fig.3 BERT training and Validation Loss

Figure 3 presents the training and validation accuracy of the fine-tuned BERT model across ten epochs. Training accuracy improved consistently from 70% at the first epoch to 96% at the final epoch, indicating that the model progressively learned discriminative features for hate speech classification. Similarly, validation accuracy increased from 68% to approximately 87.5%, confirming improved predictive performance on unseen data.

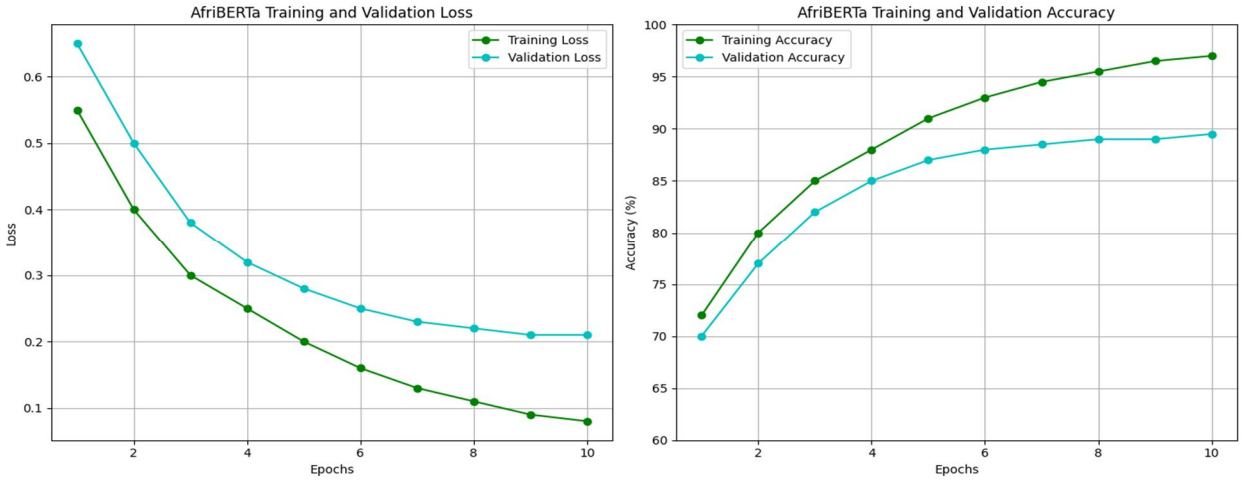


Fig.4 Afri-BERTa training and Validation Loss

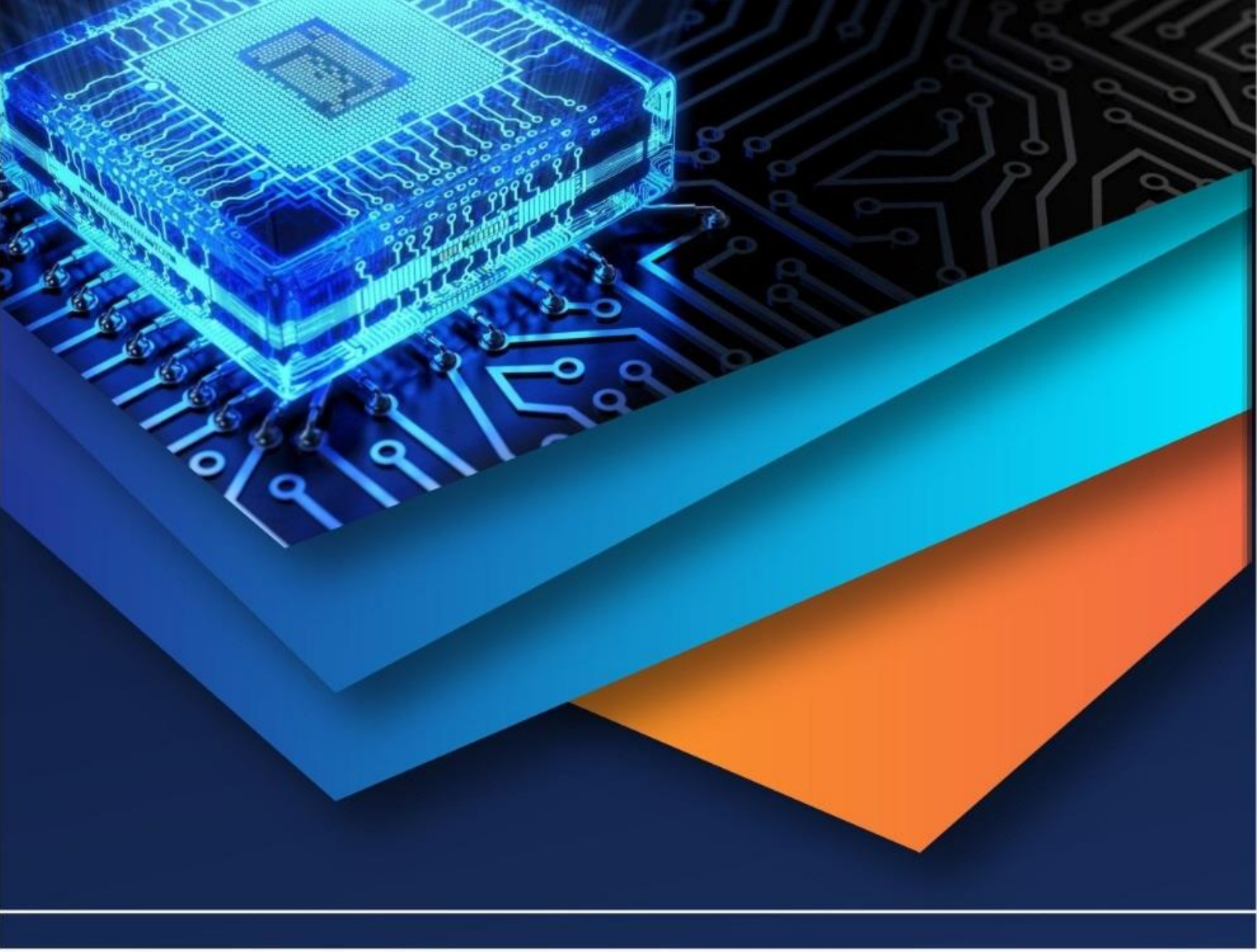
Figure 4 show the training and validation performance of the fine-tuned AfriBERTa model over ten epochs. Both training and validation loss decreased steadily, while training and validation accuracy increased consistently, indicating effective learning and stable model convergence. The training accuracy reached 97%, with validation accuracy improving to approximately 89.5% by the final epoch. The relatively small gap between the training and validation curves suggests that the model generalized well to unseen data with minimal overfitting, demonstrating the effectiveness of AfriBERTa for hate speech detection in English–Swahili code-switched text.

V. CONCLUSION AND RECOMMENDATIONS

This study demonstrated that fine-tuned transformer models are effective for detecting hate speech in English–Swahili code-switched text in low-resource settings. The results showed that BERT and AfriBERTa outperformed traditional machine learning models, with AfriBERTa achieving the best performance due to its strong representation of African languages. These findings highlight the potential of transformer-based models for multilingual hate speech detection in African social media. Future research should focus on developing larger and more diverse datasets, incorporating additional African languages, investigating real-time deployment for content moderation, and applying bias mitigation and explainable AI techniques to improve fairness, transparency, and model reliability.

REFERENCES

[1] A. M. Badel, T. Zhong, X. Xu, W. Tai, and F. Zhou, "Hate speech detection in Somali-English code-switched texts," in Proceedings of the CCF International Conference on Natural Language Processing and Chinese Computing, pp. 364–376, Springer Nature Singapore, 2025.
[2] S. Das, A. Dutta, K. Roy, A. Mondal, and A. Mukhopadhyay, "A survey on automatic online hate speech detection in low-resource languages," arXiv preprint arXiv:2411.19017, 2024. [Online]. Available: https://arxiv.org/abs/2411.19017
[3] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," in Proceedings of the International AAAI Conference on Web and Social Media, vol. 11, no. 1, pp. 512–515, 2017.
[4] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 4171–4186, Association for Computational Linguistics, 2019.
[5] V. Marivate et al., "PuoBERTa: Training and evaluation of a curated language model for Setswana," in Proceedings of the Southern African Conference for Artificial Intelligence Research, pp. 253–266, Springer, 2023.
[6] S. H. Muhammad, I. Abdulmumin, A. A. Ayele, D. I. Adelani, I. S. Ahmad, S. M. Aliyu, et al., "AfriHate: A multilingual collection of hate speech and abusive language datasets for African languages," in Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pp. 1854–1871, Association for Computational Linguistics, 2025.
[7] N. S. Mullah, W. M. N. W. Zainon, F. Ridzuan, and E. O. Abiodun, "Machine learning intervention on cyber-hate in code-switch texts: A systematic review with open challenges and solutions," PeerJ Computer Science, vol. 12, p. e3537, 2026, doi: 10.7717/peerj-cs.3537.
[8] K. Ogueji, Y. Zhu, and J. Lin, "Small data? No problem! Exploring the viability of pretrained multilingual language models for low-resource languages," in Proceedings of the 1st Workshop on Multilingual Representation Learning, pp. 116–126, Association for Computational Linguistics, 2021.
[9] E. Ombui, L. Muchemi, and P. Wagacha, "Hate speech detection in code-switched text messages," in Proceedings of the 2019 3rd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), pp. 1–6, IEEE, 2019.
[10] N. O. Onyango, L. Wanzare, and J. Obuhuma, "Swahili and code-switched English-Swahili political hate speech detection textual dataset," 2025.
[11] S. Rananga, A. Modupe, A. Isong, and V. Marivate, "Misinformation detection: A review for high and low resource languages," 2024.
[12] H. Rathnayake, J. Sumanapala, R. Rukshani, and S. Ranathunga, "Adapter-based fine-tuning of pre-trained multilingual language models for code-mixed and code-switched text classification," Knowledge and Information Systems, vol. 64, no. 7, pp. 1937–1966, 2022, doi: 10.1007/s10115-022-01698-1.
[13] Solorio, Thamar, et al. "Overview for the first shared task on language identification in code-switched data." Proceedings of the first workshop on computational approaches to code switching. 2014.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY