Swahili News Classification: Performance, Challenges, and Explainability Across ML, DL, and Transformers

Manas Pandya, Avinash Kumar Sharma, Arpit Shukla
{zda23b019, zda23m011, zda23m007}@iitmz.ac.in
Indian Institute of Technology Madras, Zanzibar Campus

Abstract

In this paper, we propose a comprehensive framework for the classification of Swahili news articles using a combination of classical machine learning techniques, deep neural networks, and transformer-based models. By balancing two diverse datasets sourced from Harvard Dataverse and Kaggle, our approach addresses the inherent challenges of imbalanced data in low-resource languages. Our experiments demonstrate the effectiveness of the proposed methodology and set the stage for further advances in Swahili natural language processing.

1 Introduction

The rapid growth of digital news platforms has intensified the need for automated text classification systems. Although substantial progress has been made in natural language processing (NLP) for high-resource languages, low-resource languages such as Swahili remain significantly underrepresented. Swahili, spoken by millions across East Africa, is essential for disseminating information; however, the scarcity of balanced and annotated datasets poses a major challenge for developing robust NLP models.

This study addresses these challenges by leveraging two prominent Swahili news datasets - one from Harvard Dataverse and another from Kaggle. By applying advanced data balancing techniques, we mitigate class imbalances and enhance the reliability of our models. Furthermore, we explore a diverse set of classification methodologies, ranging from traditional machine learning algorithms to deep neural networks and transformer-based architectures. To promote transparency and trust in automated decisions, explainability tools such as LIME and SHAP are suggested as promising avenues for future work, to shed light on the inner workings of these classifiers.

2 Related Work

Text classification has long been a core task in Natural Language Processing (NLP), with early work relying on classical machine learning techniques such as Support Vector Machines (SVM), Naïve Bayes, and Random Forests (Joachims, 1998; McCallum and Nigam, 1998). These methods, despite their simplicity, have shown considerable success in various domains. With the advent of deep learning, models such as Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks have been increasingly applied to capture complex sequential dependencies in text data (Kim, 2014; Hochreiter and Schmidhuber, 1997).

In recent years, transformer-based models have revolutionized NLP by leveraging self-attention mechanisms to learn contextual representations at scale (Vaswani et al., 2017; Devlin et al., 2019). These models have not only improved overall performance on benchmark tasks but have also enabled more effective handling of nuanced language phenomena. However, while substantial progress has been made for high-resource languages, low-resource languages like Swahili continue to receive limited attention.

Prior research on Swahili text processing has predominantly utilized traditional machine learning techniques for tasks such as sentiment analysis and named entity recognition (Nyoni et al., 2020). Only recently have deep learning and transformer-based approaches been explored for Swahili. The introduction of models such as AfriBERTa (Ogueji et al., 2021a) and SwahBERTa (Martin et al., 2022) marks a significant step forward, as these pretrained models provide richer contextual embeddings tailored for African languages. Despite these advancements, the application of state-of-the-art transformers to Swahili news classification remains underexplored.

Our work builds upon this diverse body of re-

Proceedings of the Sixth Workshop on African Natural Language Processing (AfricaNLP 2025), pages 203–209
July 31, 2025 ©2025 Association for Computational Linguistics

search by integrating classical machine learning, deep learning, and transformer-based models for Swahili news classification. By leveraging multiple model architectures and employing advanced explainability techniques, we aim to bridge the gap in low-resource NLP and provide a comprehensive evaluation framework that not only improves classification performance but also enhances model transparency.

3 Data

We use two datasets for Swahili news classification:

Swahili News Classification Dataset The Swahili News Classification Dataset was obtained from Kaggle (Antudre, 2020). It contains Swahili news articles categorized into five classes: *kitaifa* (national), *michezo* (sports), *burudani* (entertainment), *uchumi* (economy), and *kimataifa* (international). Initially, the dataset consists of 22,409 samples across three features. To mitigate class imbalance, undersampling was applied by taking 1,906 samples from each remaining category, resulting in a balanced dataset of 9,530 samples. The data was then split into 7,624 training samples and 1,906 testing samples.

Harvard Swahili News Dataset The Harvard Swahili News Dataset was obtained from Harvard Dataverse (Harvard Dataverse, 2020). This dataset comprises news articles from various Swahili media sources and includes six categories: *kitaifa* (national), *michezo* (sports), *kimataifa* (international), *burudani* (entertainment), *afya* (health), and *biashara* (business). The original dataset contains 31,044 samples across two features. To address class imbalance, undersampling was performed by taking 2,611 samples from each category, yielding a balanced dataset of 15,666 samples. This dataset was partitioned into 12,532 training samples and 3,134 testing samples.

Preprocessing

Prior to model training, both datasets underwent the following preprocessing steps: removal of special characters; conversion of text to lowercase to ensure uniformity; tokenization and stopword removal using Swahili-specific NLP libraries; splitting the data into 80% training and 20% testing sets; and balancing the datasets using undersampling to ensure equal distribution across categories. Tables 1 and 2 summarize the balanced datasets.

Attribute	**Swahili News Classification Dataset**
Total Samples	9,530
Training Samples	7,624
Testing Samples	1,906
Categories	*kitaifa, michezo, burudani, uchumi, kimataifa*

Table 1: Summary of Swahili News Classification Dataset statistics after preprocessing and balancing.

Attribute	**Harvard Swahili News Dataset**
Total Samples	15,666
Training Samples	12,532
Testing Samples	3,134
Categories	*kitaifa, michezo, kimataifa, burudani, afya, biashara*

Table 2: Summary of Harvard Swahili News Dataset statistics after preprocessing and balancing.

4 Methodology

4.1 Data Characteristics

Although our primary focus is on the classification of Swahili news, we first analyze important properties of the data that may influence model performance. In particular, we observe the distribution of text length across the different categories in both datasets. Figure 1 present the box-and-whisker plots, illustrating the minimum, first quartile, median, third quartile, and maximum text lengths for each category.

From these plots, a few notable patterns emerge:

Certain categories (e.g., *burudani*) tend to have lower median text lengths, potentially impacting the richness of vocabulary captured and affecting classification performance.

Outliers reaching beyond 20,000 characters in categories such as *kitaifa* may contain in-depth or repeated text, possibly influencing classifier decisions if not handled properly.

Categories with fewer words or shorter articles on average (e.g., *afya* in the Harvard dataset) tend to exhibit slightly lower performance, likely due to less contextual information per sample.

In the subsequent sections, we detail the modeling approaches used to address these challenges.

4.2 Machine Learning Approach[1]

We begin our methodology with classical machine learning algorithms, leveraging `scikit-learn` pipelines. The process involves:

[1] Kindly look at the appendix ’A’ for more details about exact implementation of our models

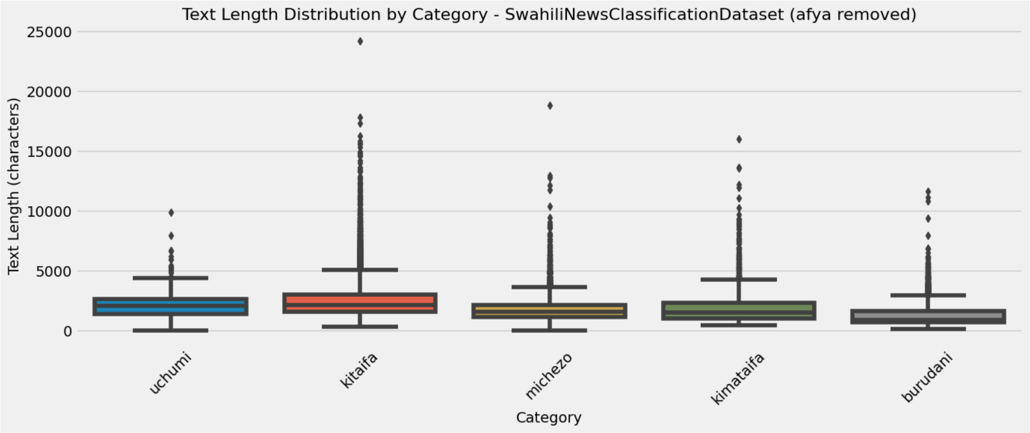

Figure 1: Text Length Distribution by Category — SwahiliNewsClassificationDataset (afya removed).

Text Representation: We apply a `TfidfVectorizer` to convert text into numerical feature vectors, setting `max_features=5000` to limit dimensionality.

Model Training: We train **four models: SVM, Logistic Regression, Random Forest, and XGBoost.** Each model is embedded in a `Pipeline` to ensure reproducible and streamlined experimentation.

Evaluation: We track metrics like accuracy, F1-score, precision, recall, training time, and inference time. Additionally, we save each trained pipeline for later analysis and potential use in explainability methods.

This approach provides initial baselines to compare against the more complex deep learning architectures.

4.3 Deep Learning Approach

To capture rich semantic and syntactic features, we develop PyTorch-based models that utilize embedding layers and sequence-processing components. Specifically, we examine:

BiLSTM: A bidirectional LSTM that can process text from left to right and right to left, capturing long-term dependencies.

CNN: A text-based convolutional neural network that extracts local features via sliding filters.

BiLSTM+CNN: A hybrid model that first uses BiLSTM to glean temporal context, followed by a 1D convolution to capture local n-gram features.

4.3.1 Model Architecture Visualization

Figure 2 illustrates two of our core deep learning architectures side by side. We train all deep models for a fixed number of epochs (e.g., 5), track training and validation losses, and then evaluate on held-out test data to assess generalization.

4.4 Transformer-Based Approach[2]

Transformers leverage self-attention to learn contextual embeddings and have shown state-of-the-art performance in various NLP tasks. We finetune the following models: AfriBERTa (Ogueji et al., 2021b) , XLM-RoBERTa (Conneau et al., 2019), and RoBERTa Swahili (Minixhofer et al., 2022) on our datasets, enabling them to adapt to domain-specific Swahili news content.

4.4.1 Transformer Architecture Visualization

Figure 2 shows a schematic of two representative transformer models used in our experiments. We tokenize the input text using each model's recommended tokenizer and then feed it through the pretrained layers. Finally, a simple classification head produces the output probabilities. We finetune for a small number of epochs (e.g., 3) on our training sets with an early stopping criterion to avoid overfitting.

Implementation Details. We employ the Hugging Face Transformers library for loading and fine-tuning models. Training arguments (`TrainingArguments`) are set with a small batch size (e.g., 4), a learning rate of $2e{-}5$, and a maximum sequence length of 256. F

5 Experimental Results and Discussion

Table 3 summarizes the experimental results across three modeling paradigms: classical machine learning (ML) models, deep learning (DL) models, and transformer-based models. For each dataset, the

[2] See appendix 'A' for more details about exact implementation of our models

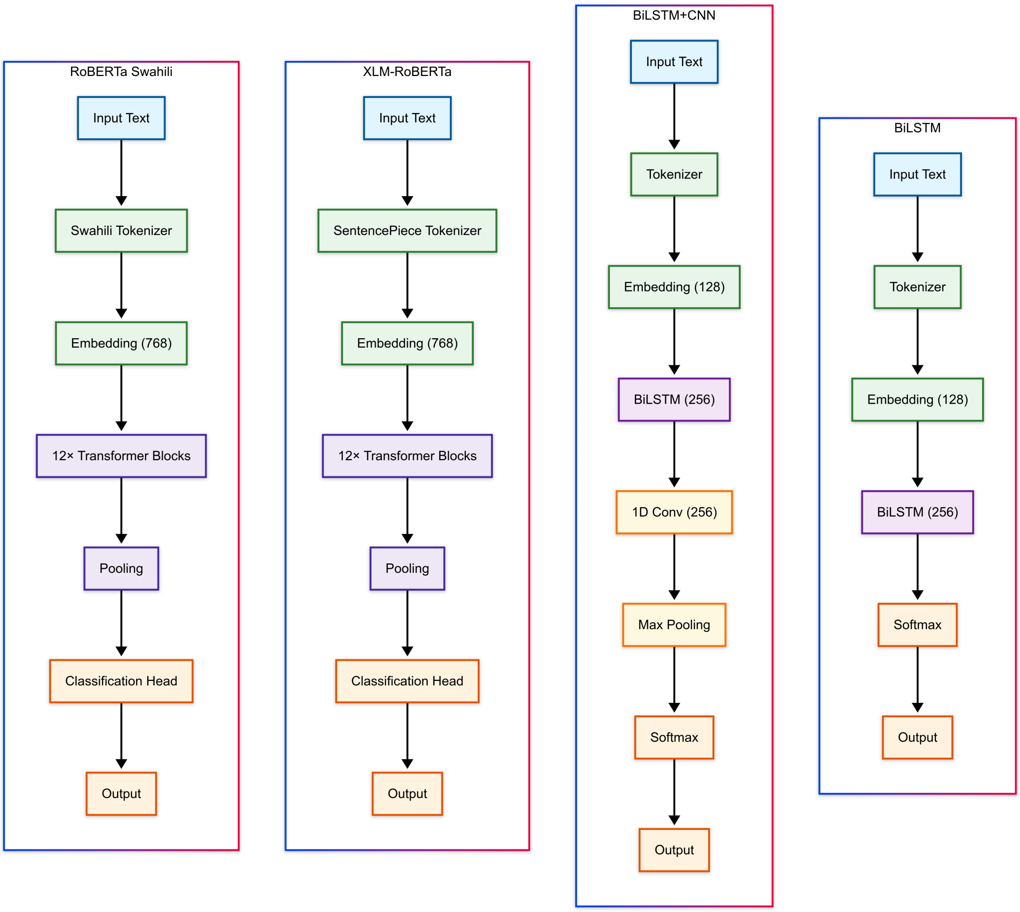

Figure 2: Transformer based and Deep Learning based model Architecture Examples

best performing metric values are highlighted in bold.

Our experimental evaluation reveals several noteworthy insights:

Classical Machine Learning Models: For the SwahiliNews dataset, the SVM model achieved the highest accuracy, F1-score, precision, and recall, indicating robustness in handling the textual features extracted via TF-IDF. Logistic Regression offered faster training and inference times, which may be advantageous in real-time or resource-constrained scenarios. On the Harvard Dataset, SVM again outperformed other ML models in terms of classification metrics, while Logistic Regression maintained computational efficiency.

Deep Learning Models: Among DL models, the CNN architecture outperformed both the BiLSTM and the hybrid BiLSTM+CNN model on the SwahiliNews dataset. In the Harvard Dataset, BiLSTM+CNN and CNN models showed similar effectiveness. CNNs were especially valuable in distinguishing closely related news categories by capturing local features.

Transformer-Based Models: Transformer models, leveraging self-attention, consistently yielded the highest performance across both datasets. Notably, the RoBERTa Base Wechsel Swahili model achieved the best accuracy, F1-score, precision, and recall. While transformers incur longer training and inference times, their ability to capture contextual nuances in Swahili news articles leads to significant performance gains.

Additional Interpretations: Shorter text lengths in some categories, such as entertainment and health, correlated with slightly reduced performance. ML models provide computational efficiency but are generally outperformed by DL and transformer-based models, which offer better predictive robustness. Variations in dataset size and composition emphasize the importance of tailored preprocessing and model fine-tuning; transformer models in particular demonstrated strong adaptability.

Comparison with Prior Work: Compared to the results reported by (Murindanyi et al., 2023), where the best SVM achieved 83% and their CNN-

Dataset	**Model**	**Accuracy**	**F1-Score**	**Precision**	**Recall**	**Train Time (s)**	**Inference Time (s)**
ML Models							
SwahiliNews	SVM	**0.8898**	**0.8897**	**0.8897**	**0.8898**	142.0781	7.6992
SwahiliNews	Logistic Regression	0.8814	0.8816	0.8820	0.8814	**3.0558**	**0.4052**
SwahiliNews	Random Forest	0.8683	0.8691	0.8718	0.8683	26.3243	0.6207
SwahiliNews	XGBoost	0.8788	0.8791	0.8799	0.8788	80.3573	0.4206
Harvard Dataset	SVM	**0.8535**	**0.8532**	**0.8536**	**0.8535**	299.9927	17.3810
Harvard Dataset	Logistic Regression	0.8462	0.8461	0.8465	0.8462	**4.5977**	**0.5304**
Harvard Dataset	Random Forest	0.8287	0.8285	0.8303	0.8287	47.8084	0.8566
Harvard Dataset	XGBoost	0.8481	0.8480	0.8491	0.8481	127.2364	0.5829
DL Models							
SwahiliNews	BiLSTM	0.5315	0.5007	0.5767	0.5315	44.7930	0.7452
SwahiliNews	CNN	**0.8620**	**0.8630**	**0.8662**	**0.8620**	**3.7755**	**0.0899**
SwahiliNews	BiLSTM_CNN	0.8421	0.8422	0.8515	0.8421	48.2156	0.8496
Harvard Dataset	BiLSTM	0.7128	0.7152	0.7306	0.7128	67.9060	1.2125
Harvard Dataset	CNN	0.8293	0.8292	**0.8330**	0.8293	**5.3939**	**0.1006**
Harvard Dataset	BiLSTM_CNN	**0.8325**	**0.8304**	0.8316	**0.8325**	79.5473	1.3700
Transformer Models							
SwahiliNews	AfriBERTa	0.9355	0.9354	0.9355	0.9355	**709.0182**	**18.5460**
SwahiliNews	XLM-RoBERTa	0.9344	0.9342	0.9344	0.9344	876.8414	22.4779
SwahiliNews	RoBERTa Wechsel sw	**0.9391**	**0.9391**	**0.9393**	**0.9391**	779.2871	20.8529
Harvard Dataset	AfriBERTa	0.9148	0.9141	0.9142	0.9148	**1142.9654**	**29.2730**
Harvard Dataset	XLM-RoBERTa	0.9065	0.9060	0.9064	0.9065	1393.1364	35.0627
Harvard Dataset	RoBERTa Weschel sw	**0.9167**	**0.9165**	**0.9166**	**0.9167**	1248.5649	33.3698

Table 3: Experimental Results. In the table, **Bold** values indicate the best performance per metric per dataset/model and Highlights indicate best overall.

BiLSTM+Attention model achieved 84% test accuracy (with Bagging ensemble at 90%), our SVM, CNN-based, and hybrid models meet or exceed these metrics. Most notably, our transformer-based models set a new state-of-the-art, achieving over 93% test accuracy and demonstrating significant advances in Swahili news classification.

Generalizability: While our experiments are focused on Swahili, many Bantu languages share similar linguistic structures, morphological patterns, and semantic features. As a result, the methodologies and insights presented here may extend to related languages, providing cross-transfer learning and adaptation.

Overall, our experiments show that while ML and DL models offer great baselines and efficiency, transformer-based architectures, especially those fine-tuned, achieve superior classification performance.

6 Challenges

This study has demonstrated that careful model selection and preprocessing can yield robust classification results for Swahili news articles across classical machine learning, deep learning, and transformer-based approaches. Notably, transformer models, particularly the RoBERTa Base Wechsel Swahili model, have shown superior performance in capturing the nuances of Swahili language data, despite increased computational cost. However, several challenges remain that must be addressed to further improve Swahili NLP applications.

A primary challenge is the limited availability of high-quality annotated data for Swahili, which constrains both model training and generalization. The high computational demand of transformer models presents an additional barrier to efficient deployment, particularly in low-resource environments.

Models trained specifically on news data may not transfer well to other domains, such as medical or legal text, making domain adaptation an important area for future research. The complex decision-making processes of transformer models also highlight the ongoing trade-off between predictive performance and interpretability.

Furthermore, Swahili's rich morphology and regional variations continue to complicate tokenization, embedding, and model generalization, necessitating more sophisticated preprocessing strategies.

7 Future Work

In terms of future scope, enriching available datasets through new data sources and advanced data augmentation methods remains essential to mitigate class imbalances and improve representation for underrepresented categories.

A key area for future work is the systematic application and evaluation of advanced interpretability techniques, such as Local Interpretable Model-Agnostic Explanations (LIME) (Ribeiro et al., 2016) and SHAP (Lundberg and Lee, 2017), to provide transparency and insight into model predictions. Additional explorations might include attention-based interpretability in transformers, as well as resource optimization through model compression strategies like knowledge distillation and quantization to facilitate real-time deployment. Extending these models to cross-lingual or multilingual contexts could further enhance their applicability across other low-resource languages.

In future work, we plan to expand the scope of our dataset by collecting and integrating Swahili language data from a wider variety of sources, including additional news outlets, social media platforms, and blogs. By incorporating content from these diverse domains, we aim to construct a more comprehensive and representative corpus that captures the linguistic richness, topical diversity, and informal language use prevalent in real-world Swahili communication.

Such an expanded dataset would not only improve the generalizability and robustness of our models but also enable more nuanced investigations into dialectal variations, code-switching, and emerging trends within the Swahili-speaking digital ecosystem. This approach is expected to facilitate the development of more effective and inclusive NLP systems for Swahili and other low-resource languages., integration into real-world systems, such as live news aggregation platforms requiring real-time inference and continuous learning, remains a critical direction for future practical impact.

8 Conclusion:

In summary, our findings highlight the strengths and trade-offs of different NLP models for Swahili news classification. While classical machine learning models provide interpretable baselines and deep learning models offer balanced performance and efficiency, transformer-based models achieve state-of-the-art results through contextual understanding. Addressing challenges related to data availability, computational efficiency, and especially model interpretability is essential for broader adoption. By tackling these challenges and pursuing the outlined future directions, this research contributes towards advancing NLP for Swahili and other low-resource languages, promoting more inclusive and effective AI applications.

References

Waalbanny Antudre. 2020. Swahili news classification dataset. Available at: https://www.kaggle.com/datasets/waalbannyantudre/swahili-news-classification-dataset.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proc. NAACL-HLT*.

Harvard Dataverse. 2020. Swahili news dataset. Available at: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/UZHZ3I.

S. Hochreiter and J. Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

T. Joachims. 1998. Text categorization with support vector machines: Learning with many relevant features. In *Proc. European Conference on Machine Learning*.

Y. Kim. 2014. Convolutional neural networks for sentence classification. In *Proc. EMNLP*.

S. M. Lundberg and S.-I. Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*.

M. Martin et al. 2022. Swahbert: Enhancing swahili nlp with pretrained transformers. In *Proc. EACL*.

A. McCallum and K. Nigam. 1998. A comparison of event models for naïve bayes text classification. In *Proc. AAAI-98 Workshop on Learning for Text Categorization*.

Benjamin Minixhofer, Fabian Paischer, and Navid Rekabsaz. 2022. WECHSEL: Effective initialization of subword embeddings for cross-lingual transfer of monolingual language models. pages 3992–4006, Seattle, United States. Association for Computational Linguistics.

Sudi Murindanyi, Yiiki Afedra Brian, Andrew Katumba, and Joyce Nakatumba-Nabende. 2023. Explainable machine learning models for swahili news classification. In *7th International Conference on Natural Language Processing and Information Retrieval (NLPIR)*.

P. Nyoni et al. 2020. Comparative analysis of swahili text classification techniques. *Journal of African Language Technology*, 3:45–62.

A. Ogueji et al. 2021a. Afriberta: A pretrained language model for african languages. In *Proc. ACL*.

Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin. 2021b. Small data? no problem! exploring the viability of pretrained multilingual language models for low-resourced languages. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 116–126, Punta Cana, Dominican Republic. Association for Computational Linguistics.

M. T. Ribeiro, S. Singh, and C. Guestrin. 2016. Why should i trust you? explaining the predictions of any classifier. In *Proc. KDD*.

A. Vaswani et al. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*.

Appendix

A Implementation Details

This section outlines the implementation details, including model training and hyperparameter tuning for each category of models considered in our study.

A.1 Machine Learning Models

Classical ML models were trained using TF-IDF features with a vocabulary size capped at 5000 terms. For **SVM**, a linear kernel was employed with probability estimation enabled, and the regularization parameter C was selected from $\{0.01, 0.1, 1, 10\}$ using grid search with five-fold cross-validation. **Logistic Regression** was set with a maximum of 1000 iterations for convergence, and C was similarly tuned. **Random Forest** utilized 200 estimators, with maximum depth tuned between 10, 20, and None, and the random state fixed at 42 for reproducibility. **XGBoost** used the multi-log loss evaluation metric and had its number of estimators and learning rate tuned via grid search. All classical models were implemented using the `scikit-learn` pipeline, and optimal hyperparameters were chosen based on F1-score performance on the validation set.

A.2 Deep Learning Models

Deep learning models were implemented with PyTorch and TensorFlow. For the **BiLSTM** model, we used an embedding size of 128, 256 hidden units, and a bidirectional architecture, trained for five epochs. The **CNN** model for text used the same embedding size, a single 1D convolutional layer with 256 filters, and also trained for five epochs. The **BiLSTM+CNN** hybrid model first extracted features using BiLSTM and then applied CNN layers, again training for five epochs. The dataset was tokenized with a vocabulary size of 10,000 and a sequence length of 300. All models used a batch size of 32 and the Adam optimizer with a learning rate of 0.001. Hyperparameters were determined through pilot experiments and validation set performance, with early stopping applied if the validation loss did not improve for two consecutive epochs.

A.3 Transformer-Based Models

Three transformer-based models were fine-tuned using the `transformers` library. **AfriBERTa** and **XLM-RoBERTa** were trained for three epochs with a batch size of 4, a learning rate of 2×10^{-5}, and weight decay of 0.01. The **RoBERTa Base Wechsel Swahili** model was also trained for three epochs, batch size 4, and fine-tuned using gradient accumulation steps of 4. The AdamW optimizer was used for all models, and input text was tokenized to a maximum sequence length of 256 tokens. Hyperparameters were selected through small grid searches on the validation set, with early stopping based on the F1-score.

A.4 Computational Resources

All models were trained on a GPU-enabled environment. **Machine learning models** were executed on CPU, while **deep learning and transformer models** were trained using an NVIDIA Tesla V100 GPU. Training duration varied, with transformer models requiring the most time—averaging between 700 to 1400 seconds per model.