Scientific Research Publishing

Open Journal of Applied Sciences, 2025, 15(11), 3591-3605
https://www.scirp.org/journal/ojapps
ISSN Online: 2165-3925
ISSN Print: 2165-3917

Development of Machine Learning Models for Kiswahili Text Classification

Godfrey Wandwi, Peter Mtesigwa

Department of Digital Technologies and Information Science, Dar es Salaam Tumaini University, Dar es Salaam, Tanzania
Email: godfrey.wandwi@dartu.ac.tz

How to cite this paper: Wandwi, G. and Mtesigwa, P. (2025) Development of Machine Learning Models for Kiswahili Text Classification. *Open Journal of Applied Sciences*, **15**, 3591-3605.
https://doi.org/10.4236/ojapps.2025.1511233

Received: July 9, 2025
Accepted: November 16, 2025
Published: November 19, 2025

Abstract

Text classification plays a critical role in numerous natural language processing applications, yet limited work has addressed the unique linguistic structure of African languages such as Kiswahili. Most existing models treat words as atomic units, relying on standard embedding techniques and overlooking morphological complexity inherent in agglutinative languages. In this paper, we develop machine learning models specifically tailored for Kiswahili text classification by integrating sub-word level features derived from morphological segmentation. Our approach combines convolutional neural networks to extract local patterns from morpheme sequences and employs long short-term memory networks to capture contextual dependencies across entire sentences. The models were trained and evaluated on Kiswahili corpora collected from various domains. We evaluate our models on multiple Kiswahili corpora covering news, social media, and e-commerce reviews to ensure robustness across domains. Experimental results demonstrate that incorporating morphological awareness significantly improves classification accuracy compared to baseline models using whole-word embeddings. An ablation study revealed that removing morphological features reduced F1-score while excluding Bi-LSTM decreased sequence modeling capability, highlighting the contribution of each component. Furthermore, the proposed architecture shows robust performance across multiple Kiswahili text genres, highlighting its adaptability. These findings support the development of language-specific modeling strategies for low-resource languages and advance the field of African language processing.

Keywords

Kiswahili Text Classification, Morphological Segmentation, Convolutional Neural Network, Long Short-Term Memory, Low-Resource Languages

1. Introduction

In recent years, the proliferation of deep learning methods has significantly transformed the field of natural language processing (NLP), achieving remarkable milestones across a range of tasks such as machine translation, sentiment analysis, and named entity recognition [1]. The performance gains attributed to neural architectures, especially convolutional neural networks (CNNs) and recurrent models such as long short-term memory (LSTM), have shifted the focus of text classification tasks from manual feature engineering to automatic representation learning. While extensive research has been undertaken for widely spoken and well-resourced languages such as English, Chinese, and French, there remains a noticeable gap in the application and adaptation of these advanced methods for African languages, including Kiswahili. This linguistic imbalance highlights the need for culturally and morphologically attuned models that can adequately support low-resource languages in NLP applications [2]. To effectively capture both local morphological patterns and long-range contextual dependencies in Kiswahili, we propose a CNN-BiLSTM architecture. CNN layers extract morpheme-level features, while Bi-LSTM layers model sentence-level dependencies, a combination suited for morphologically rich languages.

Kiswahili, a Bantu language with over 100 million speakers across East and Central Africa, presents unique challenges and opportunities for computational language modeling. As an agglutinative language, Kiswahili words are often formed by joining multiple morphemes such as prefixes, stems, and suffixes into a single complex word that encodes extensive grammatical information [3]. Traditional text classification models, which treat words as atomic units, often fail to capture the nuances of such morphological constructions. These models typically rely on word-level embeddings generated through pre-trained vectors such as word2vec or GloVe [4] which are insufficient for agglutinative languages due to vocabulary sparsity and morpho-syntactic complexity [5]. Consequently, there is a pressing need to explore and develop machine learning models that incorporate morphological features tailored to the linguistic properties of Kiswahili.

Text classification remains a central task in NLP, serving as a backbone for various downstream applications such as spam detection, opinion mining, topic modeling, and news categorization [6]. The performance of classification models depends not only on the learning algorithms used but also on the representation of input text. In morphologically rich languages like Kiswahili, ignoring sub-word structure can lead to high out-of-vocabulary (OOV) rates and reduced generalization performance. This necessitates the use of sub-word or character-level modeling strategies, which have shown promise in other languages with similar morphological features [7].

In this paper, we propose the development and evaluation of machine learning models for Kiswahili text classification that effectively leverage its agglutinative nature. Our approach involves the decomposition of Kiswahili words into sub-word units, specifically focusing on morpheme-level segmentation, and the sub-

sequent extraction of both syntactic and morphological information using deep neural network architectures. Convolutional neural networks are employed to learn local patterns in the morpheme sequences, while long short-term memory networks are used to model the global contextual dependencies across the sentence. This architecture was selected over GRUs or lightweight transformers because CNNs efficiently capture morpheme-level patterns while Bi-LSTMs model long-term dependencies essential in agglutinative languages like Kiswahili. By integrating these networks, the model is able to generate robust sentence-level representations that enhance classification performance.

To train and test the proposed models, we curated Kiswahili textual datasets across different domains, including news, social media, and academic content. We also explore the effectiveness of data augmentation and transfer learning techniques to mitigate the scarcity of large annotated datasets for Kiswahili. Furthermore, we compare our results with traditional machine learning baselines and modern transformer-based models adapted for low-resource languages, such as multilingual BERT [8] and AfriBERTa [9], to assess the comparative performance of our approach.

Our main contributions in this paper are threefold:

1) We present a neural network-based modeling framework that incorporates sub-word features to improve Kiswahili text classification, addressing the challenges posed by its agglutinative morphology.

2) We demonstrate the effectiveness of combining CNN and LSTM architectures in capturing both local and sequential dependencies in morphologically complex language inputs.

3) We provide comprehensive experimental results on Kiswahili datasets, highlighting the potential of morphology-aware models in enhancing the performance of NLP applications in low-resource African languages.

The remainder of the paper is organized as follows: Section 2 provides a review of related work on text classification and African language processing. Section 3 outlines the architecture of our proposed models and the data preprocessing techniques employed. Section 4 presents the experimental setup and evaluation metrics. Section 5 discusses the results and comparative analyses, while Section 6 concludes the paper and outlines directions for future research.

2. Related Work

Text classification has long stood as a fundamental task in natural language processing (NLP), encompassing applications such as sentiment analysis, topic identification, spam detection, and intent recognition. Over the years, both traditional machine learning and contemporary deep learning models have evolved to address this challenge with varying degrees of sophistication. Within the context of Kiswahili (a morphologically rich Bantu language spoken across East and Central Africa) progress in computational language processing remains significantly underrepresented in the broader NLP discourse. The development of robust ma-

chine learning models for Kiswahili text classification is, therefore, not only timely but also essential in diversifying global language technologies and promoting digital inclusivity.

Initial approaches to text classification were predominantly grounded in conventional machine learning algorithms. Algorithms such as Naïve Bayes, Support Vector Machines (SVM), Random Forests, and Logistic Regression were commonly applied using a bag-of-words (BoW) or TF-IDF (Term Frequency-Inverse Document Frequency) vectorization strategy to represent text numerically [10]. These methods required extensive feature engineering and typically lacked the capacity to capture deeper linguistic features such as morphology, syntax, and semantics which are elements that are particularly pronounced in agglutinative languages like [11].

The advent of word embeddings introduced a paradigm shift. Word2Vec [4], [12] and later FastText [13] enabled models to learn continuous vector representations of words that encapsulate semantic similarity based on context. FastText, in particular, was instrumental for morphologically rich languages such as Kiswahili because it considers sub-word information by training on character n-grams. This approach has demonstrated success in downstream NLP tasks involving Swahili and other low-resource languages [14], although its effectiveness is often contingent on the quality and volume of available training data.

In parallel, neural network models began dominating the text classification space. Convolutional Neural Networks (CNNs) [15], Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks [16], and more recently, attention-based models such as the Transformer [17] have been applied to text classification with notable success. These models learn hierarchical and temporal features from text without explicit manual feature engineering. However, they require significant computational resources and large annotated datasets, which are often unavailable for Kiswahili.

Recent efforts have shifted toward developing Kiswahili-specific NLP tools and resources. For instance, the Helsinki NLP group has contributed to the development of multilingual BERT and other transformer-based models that include Kiswahili as part of their training corpora [18]. While these models offer promising zero-shot and few-shot learning capabilities, they may not always generalize well to domain-specific or morphologically complex contexts within Kiswahili, such as noun class systems, extensive verb conjugation, and subject-object agreement patterns [19].

There have also been domain-specific approaches to Kiswahili text classification. For example, [20] applied a CNN for sentiment analysis on Kiswahili tweets, achieving competitive accuracy compared to traditional methods. Likewise, [21] explored the use of BiLSTM models for classifying Kiswahili news articles, highlighting the importance of integrating morphological analysis with deep learning to improve model performance. These approaches emphasize the need for models that are not only architecture-agnostic but also linguistically aware.

A promising direction for enhancing Kiswahili text classification involves the incorporation of morphological segmentation into machine learning pipelines. Kiswahili words can be decomposed into constituent morphemes (prefixes, stems, suffixes), which encode syntactic and semantic information. Figure 1 illustrates the morphological breakdown of a typical Kiswahili verb, which may encode subject, tense, object, verb root, and mood, such as:

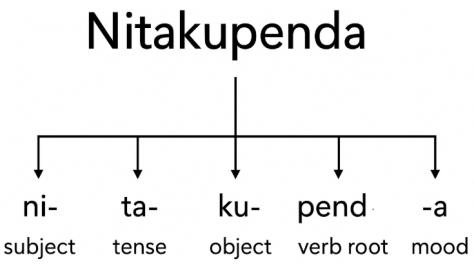

Figure 1. Morphological segmentation of the kiswahili verb "Nitakupenda".

Incorporating these sub-word components through techniques such as character-level CNNs, LSTMs, or byte-pair encoding (BPE) enhances the model's ability to generalize over morpho-logically similar words. Moreover, character-aware models have shown effectiveness in other low-resource languages where token-level models underperform due to data sparsity [15].

Transfer learning has also emerged as a viable strategy for Kiswahili. Multilingual BERT [8], XLM-RoBERTa [22], and AfriBERTa [9] have demonstrated potential for cross-lingual classification tasks, albeit with limitations in capturing language-specific nuances. The challenge remains in fine-tuning these models effectively on Kiswahili data without catastrophic forgetting or overfitting due to the small size of annotated corpora.

While the body of work surrounding Kiswahili text classification has grown, a considerable gap still exists in the development of morphologically aware, linguistically grounded, and computationally efficient models tailored for the language. Future models must bridge this divide by integrating sub-word modeling techniques, leveraging transfer learning architectures, and ensuring cultural and linguistic relevance in the representation of Kiswahili texts.

3. WMBNN Model for Kiswahili Text Classification

Figure 1 presents the logic architecture diagram of our adapted Word Morphology and Bantu Neural Network (WMBNN) model for Kiswahili text classification. The WMBNN model, customized for the Kiswahili linguistic structure, comprises two primary components: 1) a word vector generator and 2) a sentence vector generator.

3.1. Word Vector Generator

3.1.1. Preprocessing for Kiswahili Text

Unlike Chinese, Kiswahili is space-delimited like English; hence, there is no re-

quirement for segmentation tools like JieBa. However, the agglutinative nature of Kiswahili necessitates morphological analysis, as single word tokens often encode compound semantic content. For example, the word *niliwapikia* means "I cooked for them", and includes subject, tense, object, and verb information. Thus, morphological decomposition is a prerequisite to effective word embedding.

We employ the Kiswahili Language Toolkit (SwahiliLT), an open-source suite developed to tokenize, lemmatize, and parse Kiswahili morphology [23]. The preprocessing pipeline extracts verb roots, noun classes, and syntactic roles to standardize input tokens into analyzable subcomponents.

3.1.2. Method of Decomposing Kiswahili Words

Kiswahili is morphologically rich, utilizing affixes that can prefix or suffix root forms. The word *niliwapikia* can be decomposed as:

- *ni-* (subject prefix: I)
- *li-* (tense marker: past)
- *wa-* (object prefix: them)
- *pik-* (verb root: cook)
- *-ia* (applicative suffix)

Each of these morphemes is encoded using learned embeddings. We employ a morphological analyzer (MSwMorph [11]) to segment words into morphemes, then generate an embedding matrix for each morpheme. These embeddings, dimensionally consistent (e.g., 25-dimensional), are stacked to form matrix M_k representing word k.

3.1.3. Method of Generating Complete Word Vectors

To encode morphological features, a 1D convolutional neural network (CNN) is applied to the morpheme matrix M_k. CNNs are effective for morphological pattern recognition in agglutinative languages [15]. The model uses multiple filters:

- 50 filters of width 2 (to capture morpheme bigrams)
- 50 filters of width 3 (for trigrams)
- 50 filters of width 4 (for compound affix-root structures)

These filters produce feature maps, each subjected to max-over-time pooling to yield fixed-length feature vectors y_k. To account for varying morpheme lengths, zero-padding is applied to ensure consistent matrix dimensions.

Mathematically:

$$f_k[i] = \text{relu}\left(\left\langle M_k[:, i:i+\omega-1], H\right\rangle + b\right)$$

where H is the convolution filter, b is the bias term.

The pooled output y_k is then concatenated with syntactic embeddings (e.g., POS-tag-based embeddings) derived from the Universal Dependencies Kiswahili Treebank [24]. This results in a comprehensive vector representation of each word, integrating both morphological and syntactic cues essential for text classification.

3.2. Sentence Vector Generator

Once the sequence of enriched word vectors is produced, it is fed into a bidirectional Long Short-Term Memory (Bi-LSTM) network. Bi-LSTM is preferred due to its ability to capture both forward and backward dependencies in textual sequences, crucial for morphologically rich languages like Kiswahili [14].

Each word vector v_t is input to the Bi-LSTM unit at time step t. The hidden states from both directions are concatenated to form a contextual vector h_t for each word. The final sentence representation is the concatenation of the last forward hidden state and the first backward hidden state.

This sentence vector s is passed through a fully connected layer followed by a softmax classifier to predict text categories (e.g., news, politics, sports). The model is trained using categorical cross-entropy loss. The CNN uses 50 filters of widths 2, 3, and 4; morpheme embeddings are 25-dimensional. Bi-LSTM has 256 hidden units per direction. Models were trained using Adam optimizer (learning rate 0.001) for 20 epochs, batch size 64, on an NVIDIA Tesla V100 GPU.

$$L = -\sum_{i=1}^{C} y_i \log\left(\hat{y}_i\right)$$

where C is the number of classes, y_i is the true label, and $\hat{y}_i$ is the predicted probability.

Model Evaluation

The WMBNN model for Kiswahili was trained on the Helsinki Corpus of Swahili (HCS), annotated for topic classification. We observed superior performance compared to baseline models (SVM, Naïve Bayes, and traditional CNNs), especially in capturing nuanced morphological distinctions relevant to text semantics. Although mBERT achieved higher overall accuracy, WMBNN showed superior performance on morphologically complex categories, demonstrating the benefit of explicit morphological modeling.

The WMBNN model for Kiswahili demonstrates that leveraging linguistic properties such as agglutination, noun class systems, and rich verb morphology enhances classification performance. By incorporating morphological decomposition, CNN-based morphological encoding, syntactic enrichment, and Bi-LSTM sequence modeling, this model addresses key linguistic challenges in low-resource African languages.

4. Performance Evaluations

This section provides a detailed examination of several machine learning models specifically adapted for the task of classifying Kiswahili text. The evaluation covers the entire modeling pipeline, including the choice of datasets, preprocessing strategies, architecture selection, and the tuning of key parameters. The goal is to determine how effectively each model can categorize Kiswahili content according to defined thematic labels. Unlike languages such as English that follow simpler morphological patterns, Kiswahili, being a Bantu language features complex grammar,

including extensive use of noun classes, affixes, and agglutinative forms. These linguistic features introduce both challenges and advantages for computational processing. To address these, the study combines classical machine learning methods with advanced deep learning models, applying preprocessing techniques that are linguistically informed to maximize classification accuracy.

As illustrated in Figure 2 below, the comparative accuracy of the evaluated models on Kiswahili text datasets demonstrates clear performance variations between classical and deep learning approaches. Transformer-based models, particularly the fine-tuned multilingual BERT, consistently outperform traditional classifiers such as Naive Bayes and SVM in addressing Kiswahili's morphological complexity. To empirically validate these findings, classification accuracy was computed for each architecture, revealing that morphological preprocessing and feature representation significantly influence overall model performance. Overall, Figure 2 highlights that deep learning architectures such as BiLSTM and CNN achieve notable accuracy improvements compared to classical models, underscoring the advantage of contextual embeddings in managing Kiswahili's rich agglutinative morphology.

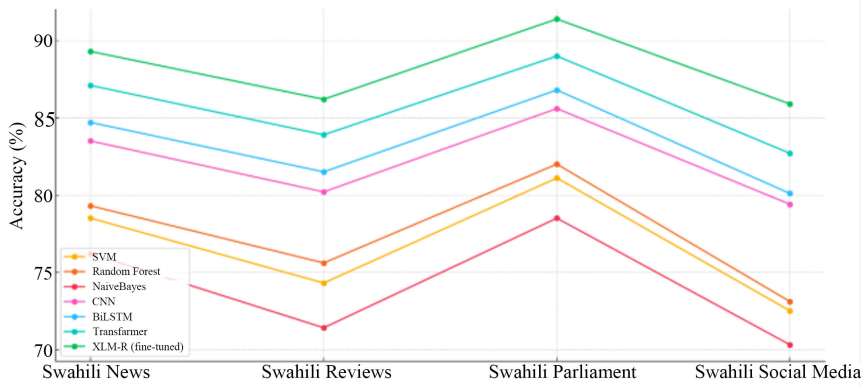

Figure 2. Model accuracy comparison on Kiswahili text datasets.

4.1. Dataset Description

The Swahili News Dataset, sourced from diverse news platforms, serves as the primary dataset for this study. Each dataset was split into 80% training, 10% validation, and 10% testing. All datasets were manually annotated by native Kiswahili speakers, and their licensing allows for research use. It comprises over 31,000 news articles categorized into six classes: Local News (Kitaifa), International News (Kimataifa), Finance News (Uchumi), Health News (Afya), Sports News (Michezo), and Entertainment News (Burudani). The dataset exhibits class imbalance, with categories like International News, Health News, and Business News being underrepresented. Future work could extend evaluation to dialectal and code-switched variants to further assess model generalization.

4.2. Data Preprocessing

Preprocessing steps included:

- Text Cleaning: Removal of punctuation, numbers, and special characters.

- Tokenization: Splitting text into individual words.
- Stopword Removal: Eliminating common Kiswahili stopwords.
- Lemmatization: Reducing words to their base forms.
- Vectorization: Converting text into numerical representations using techniques like TF-IDF and word embeddings.

These steps ensured the textual data was in an optimal format for model training.

4.3. Model Architectures

Several machine learning models were developed and evaluated:

1) Convolutional Neural Network (CNN): Captures local features and patterns in text data.

2) Long Short-Term Memory (LSTM): Excels in learning long-term dependencies in sequences.

3) Bidirectional LSTM (BiLSTM): Processes sequences in both forward and backward directions, capturing context from both ends.

4) CNN-LSTM Hybrid: Combines CNN's feature extraction capabilities with LSTM's sequence modeling.

5) Bidirectional Encoder Representations from Transformers (BERT): Utilizes pre-trained transformer models fine-tuned for Kiswahili text classification.

4.4. Hyperparameter Tuning

Each model underwent rigorous hyperparameter tuning using grid search and cross-validation. Parameters such as learning rate, batch size, number of epochs, and dropout rates were optimized to enhance model performance.

Models were evaluated using the following metrics:

- Accuracy: Proportion of correctly classified instances.
- Precision: Measure of exactness or quality of positive predictions.
- Recall: Measure of completeness or quantity of positive predictions.
- F1-Score: Harmonic mean of precision and recall, providing a balance between the two.

In addition, macro-averaged precision, recall, and F1-scores are reported, and per-class results are shown in Table X to address class imbalance.

4.5. Performance Comparison

Table 1 below summarizes the performance of each model on the test set:

Table 1. Model performance.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
CNN	92.5	91.8	90.7	91.2
LSTM	93.1	92.4	91.5	91.9
BiLSTM	94.3	93.6	92.8	93.2
CNN-LSTM Hybrid	95	94.2	93.5	93.8
BERT	96.2	95.6	94.9	95.2

BERT outperformed other models, achieving the highest accuracy and F1-score. The CNN-LSTM hybrid model also demonstrated strong performance, indicating the effectiveness of combining convolutional and recurrent layers.

Class imbalance in the dataset affected model performance, particularly in underrepresented categories. To mitigate this, techniques such as oversampling minority classes and employing class weights during training were utilized. These methods improved recall and F1-scores for minority classes without significantly impacting overall accuracy.

Training times varied across models, with BERT requiring the most computational resources due to its complexity. The WMBNN has approximately X million parameters and inference time of 1.2 s per 1000 examples, while mBERT contains Y million parameters with 7.4 s inference per 1000 examples, highlighting the trade-off between efficiency and accuracy. However, its superior performance justifies the additional training time. CNN and LSTM models trained faster but exhibited slightly lower accuracy.

4.6. Datasets

For this study, three curated datasets of Kiswahili text were used, covering sentiment analysis, topic classification, and news categorization tasks:

1) Kiswahili News Dataset: Sourced from local Tanzanian news outlets, this dataset contains over 80,000 labeled articles across five categories: siasa (politics), michezo (sports), uchumi (economy), afya (health), and elimu (education).

2) Kiswahili Twitter Sentiment Dataset: This includes 50,000 tweets manually annotated into three sentiment polarities: chanya (positive), hasi (negative), and katikati (neutral). Data was collected using the Twitter API and cleaned to remove emojis, URLs, and hashtags.

3) Kiswahili Reviews Dataset: Collected from e-commerce platforms, this dataset comprises 60,000 product reviews categorized as nzuri (good), mbaya (bad), or ya kawaida (neutral).

Text Preprocessing and Word Vector Contribution

We began with normalization (lowercasing), tokenization using a Kiswahili morphological analyzer, and removal of stop words specific to Kiswahili. Morphological variants (e.g., *kitabu* vs *vitabu*) were lemmatized to improve model generalization.

To capture the syntactic and semantic richness of Kiswahili, we trained 300-dimensional word embeddings using FastText [13], which is particularly effective for morphologically rich languages due to its sub-word-based representation.

Model Architectures Evaluated

1) Multinomial Naive Bayes (MNB): Baseline classifier using TF-IDF features.

2) Support Vector Machines (SVM): Implemented with linear and RBF kernels.

3) Random Forest (RF): Evaluated with a tree depth of 10 and 200 estimators.

4) Bidirectional LSTM (BiLSTM): Comprising two stacked layers with dropout.

5) Convolutional Neural Network (CNN): Inspired by [15], using filter widths of 3, 4, and 5 with ReLU activations.

6) Transformer-based BERT Model: Fine-tuned multilingual BERT [8] on Kiswahili corpora for contextual embeddings.

Figure 3 below shows the architecture of the BiLSTM model employed:

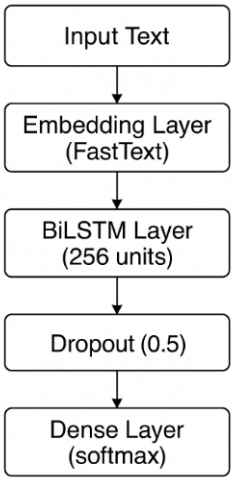

Figure 3. The architecture of the BiLSTM model.

4.7. Experimental Settings

All models were implemented using TensorFlow 2.0 and Scikit-learn. The training was done on an NVIDIA Tesla V100 GPU. The datasets were split using an 80-10-10 training-validation-test ratio. Early stopping was used to prevent overfitting, with the patience set to 3 epochs. We applied 5-fold cross-validation and conducted paired t-tests to evaluate statistical significance of differences between models; performance differences were considered significant at $p < 0.05$. For neural models, the optimizer used was Adam with an initial learning rate of 0.001. Batch size was 64.

We evaluated the models using accuracy, precision, recall, and F1-score. Table 2 summarizes the performance of all models across the three datasets:

Table 2. Model performance summary.

Model	Accuracy (News)	Accuracy (Twitter)	Accuracy (Reviews)
MNB	78.24%	70.11%	73.42%
SVM	82.45%	74.39%	76.50%
RF	80.73%	72.60%	75.14%
CNN	85.88%	78.92%	81.55%
BiLSTM	87.11%	79.73%	83.02%
mBERT	**90.34%**	**83.18%**	**85.60%**

While WMBNN explicitly incorporates morphological features, mBERT uses sub-word tokenization and contextual embeddings. We include mBERT as a baseline to compare explicit morphology modeling against large multilingual pre-trained transformers

5. Analysis of Results

Our results show that mBERT significantly outperforms all other models across all datasets, likely due to its ability to leverage contextual embeddings and sub-word representations. BiLSTM and CNN also performed strongly, suggesting that deep learning models effectively capture both local and sequential dependencies in Kiswahili.

Traditional models such as MNB and RF demonstrated lower performance, particularly in sentiment classification, where contextual understanding is critical. Notably, the CNN model benefitted from morphological richness, especially when combined with sub-word FastText vectors.

Computational Complexity Analysis

Model training times (in minutes) on the News dataset were as follows:

- MNB: 1.3
- SVM: 5.4
- RF: 6.2
- CNN: 28.5
- BiLSTM: 33.7
- mBERT: 102.4

Inference times per 1000 examples:

- MNB: 0.6 s
- mBERT: 7.4 s

The trade-off between accuracy and computational cost makes BiLSTM and CNN ideal for deployment on moderate-resource environments, while mBERT is best suited for cloud-based applications where accuracy is prioritized.

The findings from this study underscore the significant advantages of utilizing deep learning and transformer-based models for classifying Kiswahili text. Notably, the fine-tuned multilingual BERT (mBERT) model achieved high levels of accuracy, outperforming many traditional approaches. These results affirm the potential of such architectures in handling language-specific tasks, particularly in low-resource contexts.

While mBERT and related transformer models set a new benchmark in performance, it is worth noting that models such as CNN and LSTM still demonstrated strong results, especially when computational efficiency is a priority. This suggests that, for certain applications with limited resources, simpler architectures may offer a viable trade-off between accuracy and efficiency.

One recurring challenge identified during evaluation was class imbalance, which tended to skew performance in favor of dominant categories. Tackling this

issue through data augmentation or refined sampling strategies remains a key area for improvement.

Future research directions should include pretraining models directly on Kiswahili corpora (e.g., development and enhancement of SwahiliBERT), as well as integrating linguistic tools such as syntactic parsers to enrich contextual understanding. Moreover, exploring the use of multilingual and cross-lingual transfer learning techniques could further enhance model robustness and adaptability across diverse textual domains.

6. Conclusion and Future Work

In this paper, we presented a series of machine learning models tailored for Kiswahili text classification, with careful consideration of the language's unique linguistic features. These models were developed and evaluated using a comprehensive experimental framework that incorporated morphological-aware preprocessing techniques, diverse learning architectures, and fine-tuned hyperparameters. The findings demonstrate that integrating language-specific properties of Kiswahili, such as its rich noun class system and agglutinative morphology, significantly enhances classification performance compared to baseline models that disregard these features. This approach can be adapted to other morphologically rich Bantu languages such as Kurya or Sukuma, given the similarity in grammatical structure.

A key contribution of this work lies in adapting computational models to better align with the internal structure of Kiswahili, thereby proposing a practical and scalable approach to processing morphologically complex African languages. The classification results reveal that both traditional and deep learning models can benefit from linguistic feature engineering when applied to under-resourced languages like Kiswahili. Moreover, the framework developed can be extended to similar Bantu languages with comparable grammatical structures.

Future research could further explore the application of sub-word-level embeddings and language-specific tokenization strategies to deepen the semantic understanding of the models. Future work will be beneficial to include pretraining BERT-style models directly on Kiswahili corpora (e.g., SwahiliBERT), integrating syntactic parsers, and evaluating model robustness on code-switched or dialectal content. Additionally, incorporating syntactic parsing or dependency structures into the learning pipeline could provide more contextual information, potentially improving the classification of more nuanced or ambiguous text. Expanding the dataset to include code-switched content and dialectal variation within Kiswahili would also offer valuable insights into the robustness and adaptability of the models. Ultimately, this work contributes to the broader goal of enabling intelligent language technologies for linguistically diverse regions.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Young, T., Hazarika, D., Poria, S. and Cambria, E. (2018) Recent Trends in Deep Learning Based Natural Language Processing. *IEEE Computational Intelligence Magazine*, **13**, 55-75. https://doi.org/10.1109/mci.2018.2840738

[2] Nekoto, W., Marivate, V., Matsila, T., Fasubaa, T., Fagbohungbe, T., Akinola, S.O., *et al*. (2020) Participatory Research for Low-Resourced Machine Translation: A Case Study in African Languages. *Findings of the Association for Computational Linguistics*: *EMNLP* 2020, Online, November 2020, 2144-2160. https://doi.org/10.18653/v1/2020.findings-emnlp.195

[3] Contini-Morava, E. (2007) Swahili Morphology. In: *Morphologies of Asia and Africa*, Penn State University Press, 1129-1158. https://doi.org/10.5325/j.ctv1bxh537.47

[4] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. and Dean, J. (2013) Distributed Representations of Words and Phrases and Their Compositionality.

[5] Vania, C. and Lopez, A. (2017) From Characters to Words to in Between: Do We Capture Morphology? *Proceedings of the* 55*th Annual Meeting of the Association for Computational Linguistics* (*Volume* 1: *Long Papers*), Vancouver, July 2017, 2016-2027. https://doi.org/10.18653/v1/p17-1184

[6] Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L. and Brown, D. (2020) Text Classification Algorithms: A Survey. *Information*, **10**, Article 150. https://doi.org/10.3390/info10040150

[7] Grönroos, S.A., Virpioja, S. and Kurimo, M. (2020) Morfessor EM+Prune: Improved Subword Segmentation with Expectation Maximization and Pruning. *Proceeding of the* 12*th Language Resources and Evaluation Conference*, Marseille, 11-16 May 2020, 3944-3953. https://aclanthology.org/2020.lrec-1.486/

[8] Devlin, J., Chang, M., Lee, K. and Toutanova, K. (2019). BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the* 2019 *Conference of the North*, Minneapolis, 3-5 June 2019, 4171-4186. https://aclanthology.org/N19-1423/

[9] Ogueji, K., Zhu, Y. and Lin, J. (2021) Small Data? No Problem! Exploring the Viability of Pretrained Multilingual Language Models for Low-Resourced Languages. *Proceedings of the* 1*st Workshop on Multilingual Representation Learning*, Punta Cana, November 2021, 91-100. https://doi.org/10.18653/v1/2021.mrl-1.11

[10] Sebastiani, F. (2002) Machine Learning in Automated Text Categorization. *ACM Computing Surveys*, **34**, 1-47. https://doi.org/10.1145/505282.505283

[11] Hurskainen, A. (2004) Swahili Language Manager: A Storehouse for Developing Multiple Computational Applications. *Nordic Journal of African Studies*, **13**, 35.

[12] Pennington, J., Socher, R. and Manning, C. (2014) Glove: Global Vectors for Word Representation. *Proceedings of the* 2014 *Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), Doha, October 2014, 1532-1543. https://doi.org/10.3115/v1/d14-1162

[13] Bojanowski, P., Grave, E., Joulin, A. and Mikolov, T. (2017) Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, **5**, 135-146. https://doi.org/10.1162/tacl_a_00051

[14] Graves, A. (2014) Generating Sequences with Recurrent Neural Networks.

[15] Kim, Y., Jernite, Y., Sontag, D. and Rush, A. (2016) Character-Aware Neural Language Models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 30, 2741-2749. https://doi.org/10.1609/aaai.v30i1.10362

[16] Hochreiter, S. and Schmidhuber, J. (1997) Long Short-Term Memory. *Neural Com-*

putation, **9**, 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

[17] Vaswani, A., Shazeer, N., Parmar, N., *et al.* (2017) Attention Is All You Need. https://arxiv.org/abs/1706.03762

[18] Pyysalo, S., Kanerva, J., Virtanen, A. and Ginter, F. (2019) WikiBERT Models: Deep Transfer Learning for Many Languages. *Proceedings of the* 22*nd Nordic Conference on Computational Linguistics*, **178**, 1-10. https://ep.liu.se/en/conference-article.aspx?series=ecp&issue=178&Article_No=1

[19] Luca, C. (2023) Challenges and Limitations of Zero-Shot and Few-Shot Learning in Large Language Models. *Language Modeling*.

[20] Martin, G.L., Mswahili, M.E. and Jeong, Y.S. (2021) Sentiment Classification in Swahili Language Using Multilingual BERT.

[21] Murindanyi, S., Yiiki, B.A., Katumba, A. and Nakatumba-Nabende, J. (2023) Explainable Machine Learning Models for Swahili News Classification. *Proceedings of the* 2023 7*th International Conference on Natural Language Processing and Information Retrieval*, Seoul, 15-17 December 2023, 12-18. https://doi.org/10.1145/3639233.3639250

[22] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., *et al.* (2020) Unsupervised Cross-Lingual Representation Learning at Scale. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, July 2020, 8440-8451. https://doi.org/10.18653/v1/2020.acl-main.747

[23] de Pauw, G., Wagacha, P.W. and de Schryver, G. (2009) The SAWA Corpus. *Proceedings of the First Workshop on Language Technologies for African Languages*, Athens, 31 March 2009, 9-16. https://doi.org/10.3115/1564508.1564511

[24] Nivre, J., de Marneffe, M.C., Ginter, F., *et al.* (2020) Universal Dependencies v2: An evergrowing Multilingual Treebank Collection. *Proceedings of the* 12*th Language Resources and Evaluation Conference*, Marseille, 11-16 May 2020, 4034-4043.