International Journal of Computer Trends and Technology
ISSN: 2231–2803 / https://doi.org/10.14445/22312803/IJCTT-V74I7P105

Volume 74 Issue 7, 52-60, July 2026

Original Article

A Distilbert Case-Based Deep Learning Model for Part-of-Speech Tagging for Under-Resourced Kenyan Language: A Case of Dholuo

Maureen Otieno[1], Lilian Wanzare[2], Calvins Otieno[3]

[1,2,3]*Department of Computer Science, Maseno University, Kisumu, Kenya.*

[1]*Corresponding Author : mscci00081023@student.maseno.ac.ke*

Received: 28 May 2026 Revised: 30 June 2026 Accepted: 17 July 2026 Published: 31 July 2026

Abstract *- One of the basic Natural Language Processing (NLP) tasks is Part-of-Speech (POS) tagging, which helps in various applications like sentiment analysis and information retrieval. However, creating accurate POS taggers for low-resource African languages continues to be difficult due to the scarcity of linguistic resources that are annotated. Its contribution is a deep learning method for POS tagging of Dholuo, a less-resourced Western Nilotic language, spoken by about four million people in Kenya and Tanzania. The suggested system uses DistilBERT, a small transformer model, in addition to FastText and Word2Vec vector representations of words that are used to capture the context and meaning of a word. The KenCorpus Dholuo POS dataset was carefully preprocessed, normalized, and standardized with the Universal POS tags and balanced using a hybrid resampling strategy to bring about class representation. The proposed method combines contextual transformer representations with complementary word representations and a training strategy that is optimized for the linguistic features of Dholuo, while previous studies primarily used multilingual transformer models or traditional sequence-labeling methods. The framework developed is a computationally efficient one that is well-suited for low-resource language processing. The results of the experiments reveal that the proposed DistilBERT-based model outperforms the baseline Conditional Random Field (CRF) and Bidirectional Long Short-Term Memory (BiLSTM) models with an accuracy 79.05%, precision 80.63%, recall 79.05% and F1-score 79.24%. To our best knowledge, these results are the best reported for Dholuo POS tagging, and for under-resourced languages in Africa in general, highlighting the suitability of lightweight transformer architectures*.

Keywords *- Deep Learning, Dholuo , DistilBERT, Natural language processing, POS Tagging.*

1. Introduction

Natural Language Processing (NLP) is a technology capable of understanding, generating, and interacting with people's language. One of the basic preprocessing steps of different NLP applications, such as machine translation, named entity recognition, sentiment analysis, and question-answering systems, is Part-of-Speech (POS) tagging, which consists of assigning a word's grammatical category to each token of a sentence. While strides have been made in multilingual transformer models, most languages in Africa still remain under-resourced due to few annotated corpora, linguistic resources and pretrained language models. The challenge is especially severe for Dholuo because of its agglutinative structure, tones and ambiguity, which render the task of determining the POS tags significantly harder than in well-resourced languages. Downstream applications like machine translation, information retrieval, speech technologies, and question answering are therefore limited in their use of NLP due to the lack of reliable linguistic preprocessing tools.

1.1. Natural language and POS Tagging

Natural Language Processing (NLP) is an AI field that empowers computers to comprehend, create, and manipulate human speech. It is one of the important tasks of NLP, in which words are labeled with grammatical tags consisting of noun, verb, adjective and adverb according to their meaning and context. POS tagging is a critical preprocessing task for many NLP systems such as machine translation, named entity recognition, question answering and summarization of text [1, 2].

POS tagging over the years became more than just manual tagging and grew into rule-based systems, statistical and deep learning. High level of accuracy in tagging has been achieved in well-resourced languages such as English, German, Chinese and Hindi because of the availability of large annotated datasets and language resources. However, there are not enough computational resources available in many African languages to make progress in NLP research [3, 4].

1.2. Dholuo Language and Challenges in POS Tagging

Dholuo is a western Nilotic language, used primarily by the Luo people in western Kenya and northern Tanzania, along the shores of Lake Victoria. It belongs to the Nilo-Saharan language group and is related to other Nilotic languages such as Acholi and Dinka. It is estimated that approximately four million people speak the language around the world [2, 5].

There are a number of language features that make automated POS tagging more challenging in the Dholuo language. First, Dholuo is a tonal language; the meaning and grammatical function of a word can change by changing its tone. Second, the language is agglutinative – that is, several aspects of the grammar (such as the pronouns, tense markers, verb forms, etc.) are combined into a single word. For "Anindo", it has both a verbal form and a first-person pronoun. Moreover, some words can be capitalized and have various meanings in different contexts, e.g., Otieno. These linguistic properties cause such ambiguities and difficulties in POS tagging.

1.3. Prior Work and Research Gap

Prior to the development of deep learning, the key approaches to POS tagging included traditional machine learning, as well as statistical methods such as Hidden Markov Models (HMM), Conditional Random Fields (CRF), Support Vector Machines (SVM), Maximum Entropy (Maxent) and memory-based learning (MBL) [6]. The strategies they employ are well-rooted in handmade linguistic features and rules.

In HMM models, POS tagging is considered a sequence prediction task, in which the most likely sequence of tags is predicted from a sentence. The results of research on the African languages have varied. For example, in [6], HMM has been applied to Igbo with 73.33% accuracy and no more than 300 words in their dataset. They used the HMM technique on the Shekki'noono language with 91.8% accuracy [7]. However, these models failed to perform well when dealing with unseen words and had limited training data.

The Conditional Random Fields (CRF) achieved better results than HMM because it considered all the sentences in the prediction process. For Yorùbá POS tagging, [3] reported a high precision and recall with the CRF model when compared to the HMM models. In addition, [2] used their CRF across 20 African languages, and they reported that their average accuracy is 85.7% and in the case of Dholuo was 85.2%, however, in Dholuo not as high because of limited data and the complicated morphology of the language.

Other, more traditional methods have been attempted as well. Setswana POS tagging was performed by [1] using the SVM, MaxEnt and combined voting model. The results showed that the ensemble methods are useful to enhance the prediction accuracy; the combined model performed the best. In order to get the highest accuracy, 90.68%, 8] adopted memory-based learning with Kamba, but the model did not work well with morphologically complex words. The situation in Dholuo is similar to that of the other groups.

Though successful in some languages, traditional models have some significant weaknesses. They rely on hand-crafted features, rather than on words they are not seeing; not very good at taking long-range context. Weaknesses are more pronounced in languages with low resources and high morphology, such as Dholuo.

Deep learning has improved the POS tagging by learning features from data. Some of the widely used deep learning models are BiLSTM, (CNN, and Transformer-based models like BERT and DistilBERT.

BiLSTM models can read text in both directions so that they can see the context around the text. BiLSTM was used by [9] in the Amazigh language, and an accuracy of 94.8% was achieved. In [4], CNN and BiLSTM were combined to perform Odia POS tagging, which achieved good performance. Although BiLSTM models work well, they usually need a large amount of data, and can have difficulties handling very long-range dependencies.

The CNN models can be helpful in extracting features at the local and character level, which is helpful in capturing morphological patterns. However, these are not as effective as recurrent models for modelling sequence relationships. Thus, researchers are inclined to combine CNN with BiLSTM to leverage the benefits of local feature extraction and contextual learning.

Transformer models have introduced a significant leap in the realm of NLP, especially with the advent of BERT and its variations. Large-scale pre-trained models that are fine-tuned for a specific task using task-specific labeled data sets. In order to address this, [10] used BERT for the Algerian dialect, with an accuracy of 83.1%, but with the challenge of overfitting since the data was limited. [2] tested the performance of multilingual and Africa-centric transformer models on 20 African languages and reported that AfroXLMR outperforms the other two models tested (CRF, multilingual BERT).

DistilBERT is a smaller and faster version of BERT, which retains the majority of the language understanding power of BERT with less computational demand. It needs fewer parameters to be trained and trains fast, appropriate for low-resource settings where data and computation resources are limited. This indicates it is a suitable model to be used for the Dholuo POS tagging.

Not much research has been done on NLP in Dholuo, and the research done in Kenya has been done in Kiswahili. In the early studies of Dholuo POS tagging, [11] reported a relatively low accuracy of 48.9% and 51.3%, respectively, with the use of annotation transfer from English and Kiswahili. Later, [2] expanded this work to cover Dholuo in multilingual African POS tagging research with advanced transformer models. However, Dholuo performance was not as robust as languages with more annotated data and morphological complexity, because of a lack of both.

Most tagging errors can be classified into two types: wrongly tagging words because of morphological ambiguity, and failing to tag words that were not seen during pretraining. They both amplify in low-resource settings where there are few annotated corpora, and it is difficult, time-consuming, and expensive to build annotated corpora, with the majority of African languages lacking the basic tools needed to create and test NLP systems. Examples of these limitations are the fact that dholuo is not well represented by a large number of digital corpora, and that it is highly inflected, with agglutination tone and a noun-class marking system making it challenging to tag the language precisely.

The release of the KenCorpus dataset by [5] was an important milestone for Dholuo NLP. It is the biggest annotated Dholuo POS dataset containing 4794 sentences and 53357 tokens. There are a number of problems with this contribution, though, such as the lack of annotation, the very uneven class distributions, and the number of detailed POS tags (more than 70).

While various studies have been conducted on the African POS tagging using CRF, BiLSTM, multilingual BERT and AfroXLMR, there are three points to take note of. First, there is a lack of research on lightweight transformer architectures (such as DistilBERT) for Dholuo. Second, the relationship between preprocessing, label standardization, and balancing between classes and how they affect transformer performance has received little attention. Third, there are not many studies that offer detailed linguistic analysis, which explains common tagging errors in morphologically rich Kenyan languages. In this study, we aim to tackle these gaps by introducing a novel framework that leverages DistilBERT with hybrid embeddings and a comprehensive preprocessing pipeline.

The reviewed studies reveal that there is a definite gap in the research. Previous studies have mainly adopted traditional machine learning models or large multilingual transformer models, and few works have investigated light transformer models such as DistilBERT for Dholuo POS tagging. Furthermore, the effect of preprocessing, feature extraction and class imbalance of Dholuo POS tagging should be investigated. Hence, this study aimed to evaluate the performance of DistilBERT and other pre-processing techniques to improve the performance of POS tagging on the KenCorpus Dholuo data.

Despite this progress, no previous work has introduced morphology-aware subword embeddings along with a light-weight transformer to solve the Dholuo agglutination problem. Previous work on Dholuo has used cross-lingual projection without any Dholuo training data [11] or only evaluated Dholuo as one of 20 languages in a large multilingual model with no direct tackling of class imbalance or agglutination [2]. The question that remains is whether a small, efficient transformer, instead of a large multilingual one, can bridge this small gap for Dholuo, the language studied in this paper.

To overcome these challenges, this study holistically addresses each problem by standardizing the data, a technique of balancing the classes, and developing an extended embedding DistilBERT model.

This study is novel in three aspects: firstly, it is the first study to incorporate the contextual features of DistilBERT with the subword and Word2Vec features of the Dholuo language in the same model; secondly, it explicitly addresses the issue of agglutination by breaking it down into subwords instead of assuming it is an inherent limitation; thirdly, it introduces a two-stage tag-standardization and mixed-resampling pipeline that transforms the original 70 fine-grained tags of KenCorpus into 12 balanced classes. The performance of the present model on an unbalanced test set is compared with that of [2], which reported 85.2% accuracy for Dholuo on a 20-language multilingual evaluation, and without correcting for class imbalance, the accuracy of 79.05% for this present model is a more conservative and arguably more reliable estimate of real-world performance.

2. Materials and Methods

2.1. The Dataset

The main data source was the KenCorpus Dholuo POS dataset [5], which consists of 4,794 sentences and 53,357 tokens from Kenyan media and social media (Twitter). These 70 fine-grained tags were summarized to 12 universal POS categories: NN (Noun), V (Verb), PRON (Pronoun), ADV (Adverb), CONJ (Conjunction), DET (Determiner), ADP (Adposition), ADJ (Adjective), INTER (Interjection), X (Non-Dholuo/other), NUM (Numeral), PUNCT (Punctuation).

2.2. Data Preprocessing

The data was preprocessed using Python in Google Colab, where the following were done:

Tag-Token Separation: Each token and its POS tag were merged together in a single column in the original data set. A Python script was developed to automatically detect the

delimiters and to create separate columns for token and POS-tag, making the dataset suitable for supervised sequence-labeling techniques.

A programmatic mapping script was developed, which could be used to remap and standardize the 70+ original fine-grained tags to the 12 target tags. The transformation of sub-tags of nouns to the general class NN, such as NN+V, NN+Pron, NNS and its sub-tags. Manual inspection and correction were done after automated reassignment to correct cases where there was ambiguity or errors in automated mapping. This was a two-step process to ensure consistency and consistency of annotation.

Whitespace Normalization: to prevent mismatching of the tokenizer in the model training and avoid alignment errors, all the leading, trailing and multiple spaces between tokens and tags were removed.

Data Frame Construction: The data was organized into data frames having 3 columns: sentence ID, word (token) and POS label. The integer index for every POS tag in a string was obtained via a Label Encoder from scikit-learn and used to train the model.

2.3. Handling Class Imbalance

The data set was very imbalanced, with Verb (V) having 4511 examples, Noun (NN) 3808 examples, Adjective (ADJ) 2 examples, and Adposition (ADP) 2 examples. The strategy of resampling was a mixed one: minority classes were randomly duplicated to get 3,808 instances, and majority classes were undersampled to get 3,808 instances to balance the 12 classes to a total of 45,696 tokens. Prior to balancing (pre-balancing) and after balancing (post-balancing), the dataset is visualized in figures 1 and 2, respectively.

2.4. Model Architecture

The model has been modified from DistilBERT-base-cased with 6 layers of transformer encoder and multi-head self-attention. There were additions of: complementary distributional semantic representations using FastText, Word2Vec and word2vec contextual embeddings. Since Dholuo follows the agglutinative language pattern, FastText's character n-gram decomposition is a good feature to utilize as it captures the substructure of the morphology, which remains meaningful even for out-of-vocabulary tokens. Word2Vec brings in global distributional semantics, which may get lost by FastText's subword emphasis. DistilBERT's hidden states provide sentence-level disambiguation that static embedding doesn't provide. This study did not include a formal ablation of the individual contributions of each embedding, which is recognized as a possible avenue for future research. The three embedding types were then amalgamated and fed into a pre-classifier dense layer and a final 12-class SoftMax classification head. The model was end-to-end optimized using the AdamW optimizer, along with cross-entropy loss. Figure 3 below shows the architecture of the proposed model.

2.5. Training and Baselines

The hyperparameters were optimized using random search, and not the grid search, to enable wider exploration of the search space while having lower computational cost. The learning rates of 1×10^{-3}, 1×10^{-4}, 1×10^{-5}, and 1×10^{-6} were tested; higher learning rates led to instability in training, and 1×10^{-5} yielded the most stable convergence. Sizes of 4, 8, 16, and 32 batches were tested – 4 batches required more time to train the model, but yielded better convergence rate, and 16 batches were the best balance between memory and convergence rate. Also tested were embeddings of 128, 256, 512 and 768, which were selected based on the validation performance versus the computational expense. The model was trained for 20 epochs and started to experience a mild level of overfitting from epoch 8 onwards, with epoch 10 being selected as the checkpoint since it did not overfit the model excessively, as it had the lowest validation loss and the highest F1-score.

A linear-chain CRF with hand-engineered features (word identity, context window, affixes, capitalization) and a two-layer BiLSTM with 128 units per layer, dropout 0.3 and Adam optimizer were used as baselines.

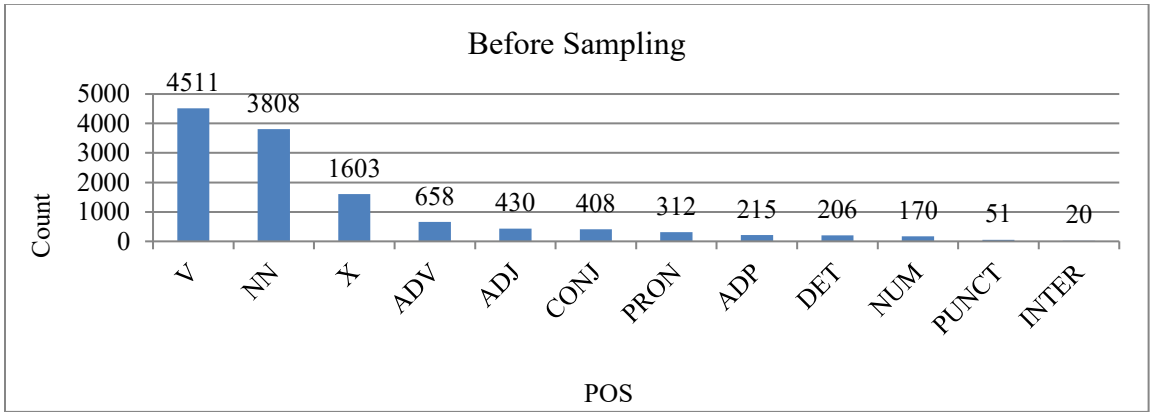

Fig. 1 Visualization of the dataset before balancing

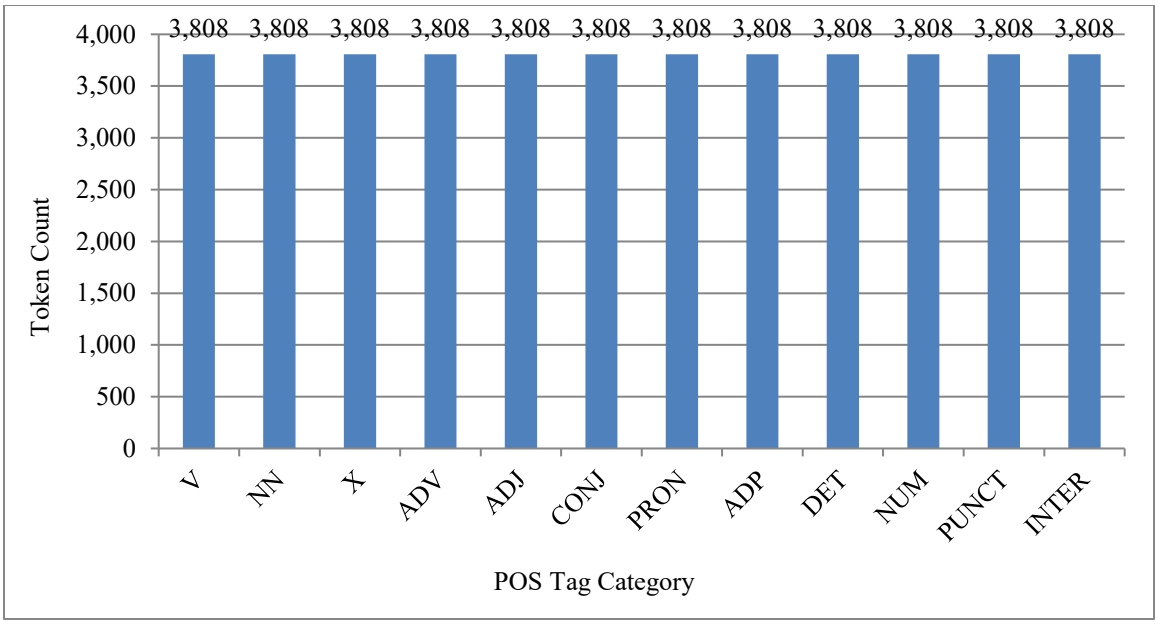

Fig. 2 Visualization of the dataset after balancing

Fig. 3 DistilBERT model Architecture

3. Results and Discussion

3.1. Results

3.1.1. The Dataset Preprocessing Outcomes

After the preprocessing, the data set consisted of 4,794 sentences and 53,357 tokens in a pre-processed, clean and structured format with 12 categories of POS. Before resampling, the distribution was highly skewed (V: 4,511; NN: 3808; ADJ:430; ADP:215). Following the mixed resampling, the 12 classes were approximately equal in size, with 3,808 instances per class, totaling 45,696 tokens.

3.1.2. DistilBERT Model Performance

Table I shows a comparison of the performance of the DistilBERT model against baseline models across various performance metrics.

Table 1. Comparison of model performance on Dholuo POS tagging test set

Metric	DistilBERT	CRF	BiLSTM
Accuracy	0.7905	0.5130	0.5755
Precision	0.8063	0.4990	0.5444
Recall	0.7905	0.5130	0.5755
F1-Score	0.7924	0.4930	0.5529

3.1.3. Comparative Performance

The DistilBERT model outperforms the CRF baseline with an accuracy score of 83.00% and the F1 score of 78.86%, which is 27.75% and 29.94% improvement, respectively. It boosted the accuracy of the BiLSTM by 21.5 percentage points and the F1-score of the BiLSTM by 23.95 percentage points, respectively. The transformer-based approach clearly has margins on the margins.

The datasets and the granularity of the tags are different from previous Dholuo POS tagging projects, making direct comparisons difficult. Without any annotated training data, De Pauw et al. [11] achieved the accuracy of 48.9%–51.3% and the present model achieved 79.05% – a significant increase due to training directly on annotated Dholuo data. Dholuo reported a higher raw accuracy of 85.2% when evaluated using CRF in their evaluation of MasakhaPOS [2] and 89.4% average accuracy with AfroXLMR-large across twenty languages. The figures were not derived from a test set that is balanced across the classes, and the Dholuo AfrolMR figure was not taken out of the 20-language average, so comparisons are only indicative and not exact. The F1 score of 0.7924 achieved here is an average across the 12 tag classes; many minority classes (which could be buried by a raw accuracy figure) are represented here, and this is arguably a more conservative and more diagnostically useful measure of tagging reliability in the field than raw accuracy.

Care must be taken in comparing different languages because the accuracy depends on the size of the data set, the granularity of the tag set and the morphological typology. The results presented by Kamba [8] and Amazigh [9] are also higher accuracy, though both are based on a smaller tag inventory and annotated corpus than the 4794-sentence subset of KenCorpus that was used here.

However, the Igbo HMM result [6] obtained on a similar, although smaller, data set, did not yield as high an accuracy as the model presented here, even though it was based on Dholuo's more complex morphology (agglutination).

This accuracy is 27–30 percentage points better than the accuracy achieved by [11] in their projections of the Dholuo accuracy onto the other languages. The F1 score of 0.7924 is a much more rigorous and informative score than the previous studies.

A two-proportion significance test was conducted to determine if the reported gains in accuracy by DistilBERT over the baselines were due to chance on the test set (n = 20) at the token level. Both DistilBERT's improvement over CRF ($z = 42.55$, $p < 0.001$) and its improvement over BiLSTM ($z = 33.75$, $p < 0.001$) were statistically significant, indicating that the improvements in accuracy were not due to sampling error.

Table 2. Statistical comparison of the model's performance

Model Comparison	Accuracy Difference	z-statistic	p-value
DistilBERT vs. CRF	+27.75 pp	42.55	$p < 0.001$
DistilBERT vs. BiLSTM	+21.50 pp	33.75	$p < 0.001$

3.2. Discussions

3.2.1. Why DistilBERT Excelled

The morphology ambiguity of the Dholuo language is context sensitive, like the word “Otieno” can be a proper name or the word ‘night’, while assigning a tag to a specific token, DistilBERT can capture all the tokens of a sentence at once by using its self-attention mechanism. The orthographic case distinction was maintained in the cased form, which is relevant to the Dholuo disambiguation of case. Agglutinated forms, such as “Anindo”, were also helped by FastText's ability to split words into sub-words.

3.2.2. Limitation of Baseline Models

The CRF's F1 of 0.493 is due to the difficulties of hand-engineered features with a morphologically rich language, and the inability of the CRF to deal with minority classes and agglutinated forms. The BiLSTM was not as good as CRF, but it was not capable of modelling long-range dependencies, and it did not have pre-trained Dholuo embeddings that were trained on a small dataset.

3.2.3. Error Analysis

Three common error patterns were identified from a qualitative look at the misclassified tokens on the test set. The first kind of error is due to agglutination: In the case of forms like Anindo and Odwaro, the model paid more attention to the part that looks like a pronoun (A-, O-) than the rest of the form, which often turns them into PRON rather than V. Second, context-dependent homograph errors: words like ne and gi that serve multiple semantic functions as determiners, pronouns or conjunctions in different positions in a sentence were sometimes wrongly tagged in less common syntactic constructions. Third, under-detection in the minority class (INTER class: only 20 instances with an original size of 20 before resampling) was defaulted to ADV or X, reflecting the limitations of duplication-based oversampling for generating true diversity in the training set. Check for tonal ambiguity was not directly examined because KenCorpus is comprised of written, prosodically annotated social-media text, which is noted as a limitation rather than a finding.

3.2.4. Role of Data Preprocessing and Resampling.

The model was trained using systematic preprocessing and mixing of resamples, and this systematic process was essential to obtain an appropriate precision(80.63%) and recall(79.05%). However, the classes with duplicated samples were oversampled, resulting in the potential for generalization error for them. In addition to mixed resampling, AdamW's weight decay (0.01) was used to regularize the model weights and validation loss-based checkpoint selection was also used to avoid overfitting by keeping the epoch with the lowest validation loss as the final model instead of the epoch with the lowest training loss. This research should be continued in the direction of back-translation, synonym substitution/crosslingual transfer for augmentation of minority classes.

3.2.5. Broader Implications

The study presents a viable approach to under-resourced language POS tagging by pre-processing and normalizing existing corpora, resampling with multiple resampling techniques, fine-tuning a small transformer (DistilBERT), incorporating subword embeddings into the model, and thorough testing the model. The methodology can be used for other Kenyan and African languages like Luhya, Kalenjin, Mijikenda, Somali, etc.

4. Future work

- KenCorpus is a collection of Kenyan media and Twitter text data, which makes the Kenyan model trained and evaluated on informal, social-media-register Dholuo. Other formal registers (such as administrative documents, literature or broadcast transcripts) were not assessed and may be different because of the distributional shift in vocabulary and sentence structure. Recommendations are offered for further research on this work and alternative approaches for expanding it to other types of text, including domain adaptation methods proposed for fine-tuning on a small sample of formal register text.
- Since there is currently no formal morphological analyzer for Dholuo, this was done implicitly by using the subword decomposition feature in FastText, instead of explicit morpheme segmentation. Future models would benefit from collaborating with Dholuo linguists to build a rule-based or statistical morphological analyzer, to be able to split the agglutinated forms into their morphemes before they are tagged, and directly tackle the misclassification pattern for the Dholuo pronouns.
- That is, along with the duplication-based oversampling of ADJ and ADP, two techniques are presented to ensure that the examples generated for the minority classes are genuinely novel: the back-translation technique and the Dholuo-lexicon-based synonym substitution technique. Finally, it is suggested that morphologically similar Nilotic languages could provide cross-lingual transfer as another way to increase the size of the training dataset without any new annotation.

5. Conclusion

This study introduced a DistilBERT deep learning framework for Part-of-Speech (POS) tagging of the Dholuo language in Kenya with limited resources in the field of NLP. To address the challenges of the small size of annotated data, morphological complexity and extreme class imbalance, the study developed a pipeline for end-to-end processing, which involves data pre-processing, POS tag standardization, class balancing with a hybrid resampling strategy, and transformer-based sequence classification with the introduction of FastText and Word2Vec embeddings. The proposed model was able to effectively utilize the contextual transformer representations, semantic and subword embeddings to capture the contextual and morphological information that is required for the correct tagging of POS in Dholuo.

The experimental results showed that the proposed model had an accuracy of 79.05%, a precision of 80.63%, a recall of 79.05%, and an F1-score of 79.24%, which was significantly better than the CRF and BiLSTM baseline models. The results show that the lightweight transformer models can achieve competitive performance for low-resource languages with lower computational resources. The paper also highlights the importance of carefully preprocessing the data sets, standardizing the labels and using balanced sets for improving the classification accuracy of complex language systems.

Beyond the reported performance improvements, this study offers a method that is practical and replicable, and

could be used by other under-resourced Kenyan and African languages that suffer similar linguistic and resource challenges. The common POS tag set and the preprocessing scheme provide a robust foundation for further language technology development and are key to a more linguistically inclusive AI.

Although promising results are obtained, the study is restricted to the small size of the available Dholuo corpus and the oversampling of the extremely underrepresented POS categories, which might impact generalization. It is suggested that the corpus needs to be expanded further with more varied sources of text; the use of larger and more complex data augmentation and cross-lingual transfer learning methods should be tested; large-scale pre-trained language models specifically tailored to Dholuo or Africa should be developed; and morphological analysis could be integrated to improve the accuracy of tagging. The proposed framework also paves the way for the application of transformer-based approaches for other downstream NLP tasks like dependency parsing, named entity recognition, and machine translation for under-resourced African languages.

The proposed framework not only enhances the robustness of POS tagging but also provides a viable platform for future Dholuo language applications such as machine translation, speech recognition, grammar error detection and correction, information retrieval, and educational language tools. The methodology could also be modified to work with other under-resourced Kenyan languages with similar morphological properties.

Conflicts of Interest

The authors do not have any competing financial or personal interests to declare.

There are no relationships that might have seemed to have affected the work covered in this paper.

Funding Statement

This research was supported by the SG-NAPI award, supported by the German ministry of education and research, BMBF, through UNESCO-TWAS.

Data Availability

The standardized dataset used is publicly available on the KenCorpus huggingface repository at https://huggingface.co/datasets/Kencorpus/Dholuo_POS

Acknowledgments

This publication could not have been possible without the support of the SG-NAPI award, supported by the German ministry of education and research, BMBF, through UNESCO-TWAS. The authors express their gratitude to the resources and environment provided by Maseno University, which facilitated the research. Thanks to the annotators working on the preparation of the data set.

References

[1] Mary Dibits, and Pius A. Owolawi, and Sunday O. Ojo, "An Hybrid Part of Speech Tagger for Setswana Language using Voting Method," *International Conference on Intelligent and Innovative Computing Applications*, pp. 245-253, 2022. [CrossRef] [Google Scholar]

[2] Cheikh M. Bamba Dione et al., "MasakhaPOS: Part-of-Speech Tagging for Typologically Diverse African Languages," *arXiv Preprint*, pp. 1-18, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[3] Ikechukwu I. Ayogu et al., "A Comparative Study of Hidden Markov Model and Conditional Random Fields on a Yorùbá part-of-Speech Tagging Task," *2017 International Conference on Computing Networking and Informatics (ICCNI)*, Lagos, Nigeria, pp. 1-6, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[4] Tusarkanta Dalai, Tapas Kumar Mishra, and Pankaj K. Sa, "Part-of-Speech Tagging of Odia Language using Statistical and Deep Learning based Approaches," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 22, no. 6, pp. 1-24, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[5] Barack Wanjawa et al., "KenCorpus: A Kenyan Language Corpus of Swahili, Dholuo and Luhya for Natural Language Processing Tasks," *Journal for Language Technology and Computational Linguistics*, vol. 36, no. 2, pp. 1-27, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[6] Iheanetu Olamma, Michael Kingsley, and Sunday Ojo, "Hidden Markov-based Part-of-Speech Tagger for Igbo Resource-Scarce African Language," *Proceedings of the First International Workshop on NLP Solutions for Under Resourced Languages*, Association for Computational Linguistics, Trento, Italy, pp. 118-123, 2019. [Google Scholar] [Publisher Link]

[7] Alebachew Chiche, Hiwot Kadi, and Tibebu Bekele, "A Hidden Markov Model-based Part of Speech Tagger for Shekki'noono Language," *International Journal of Computing*, vol. 20, no. 4, pp. 587-595, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[8] Benson N. Kituku, Musumba George, and Peter Wagacha, "Kamba Part of Speech Tagger using Memory-based Approach," *International Journal on Natural Language Computing*, vol. 4, no. 2, pp. 43-53, 2015. [CrossRef] [Google Scholar]

[9] Rkia Bani et al., "Toward Accurate Amazigh Part-of-Speech Tagging," *IAES International Journal of Artificial Intelligence*, vol. 13, no. 1, pp. 572-580, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[10] Mohamed Amine Cheragui, Abdelhalim Hafedh Dahou, and Amin Abdedaiem, "Exploring BERT Models for Part-of-Speech Tagging in the Algerian Dialect: A Comprehensive Study," *Proceedings of the 6th International Conference on Natural Language and Speech Processing (ICNLSP)*, Association for Computational Linguistics, pp. 140-150, 2023. [Google Scholar] [Publisher Link]

[11] Guy De Pauw, Naomi Maajabu, and Peter Waiganjo Wagacha, "A Knowledge-light Approach to Luo Machine Translation and Part-of-Speech Tagging," *Proceedings of the Second Workshop on African Language Technology (AfLaT)*, Valletta, Malta, pp. 15-20, 2010. [Google Scholar]