****

**Department of Computer Science**

**Master of Science in Computer Science**

**CCS 6204- NATURAL LANGUAGE PROCESSING**

**ENHANCING POS TAGGING FOR LOW-RESOURCE LANGUAGES: A CASE STUDY ON DHULUO**

Table of Contents

[Abstract 3](#_Toc169604764)

[CHAPTER ONE: 3](#_Toc169604765)

[Introduction 3](#_Toc169604766)

[Problem Statement 5](#_Toc169604767)

[Objectives 5](#_Toc169604768)

[Main Objectives 5](#_Toc169604769)

[Specific Objectives 5](#_Toc169604770)

[CHAPTER TWO: LITERATURE REVIEW 6](#_Toc169604771)

[CHAPTER THREE: Methodology 10](#_Toc169604772)

[Dataset 11](#_Toc169604773)

[Random Forest 12](#_Toc169604774)

[Support Vector Machine (SVM) 12](#_Toc169604775)

[Deep Learning Models 12](#_Toc169604776)

[Results and Discussions 12](#_Toc169604777)

[Word Cloud Representation 13](#_Toc169604778)

[Model Evaluation Metrics 14](#_Toc169604779)

[Summary of Results 14](#_Toc169604780)

[RandomForestClassifier 15](#_Toc169604781)

[SVM 15](#_Toc169604782)

[RNN 15](#_Toc169604783)

[LSTM 15](#_Toc169604784)

[Future Improvements: 16](#_Toc169604785)

[Conclusion 16](#_Toc169604786)

[References 17](#_Toc169604787)

# Abstract

This paper addresses the development of a Part-of-Speech (POS) tagger for Dhuluo, a Nilotic language spoken predominantly by the Luo people in Kenya. The scarcity of computational resources for Dhuluo and other under-resourced languages poses significant challenges for natural language processing (NLP) tasks.

This study explores multiple machine learning and deep learning approaches to create an effective POS tagging system for Dhuluo. We experimented with traditional machine learning algorithms, specifically Random Forest and Support Vector Machine (SVM), and advanced deep learning models, including Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) networks. The models were trained and evaluated using a dataset of 54,941 annotated word-tag pairs.

Results indicate that while traditional machine learning methods provide a baseline accuracy, deep learning models, particularly LSTM, significantly improve the performance of the POS tagger by effectively capturing the sequential dependencies in the language. This paper contributes to the field of computational linguistics by offering insights into the development of linguistic tools for under-resourced languages, demonstrating the potential of advanced NLP techniques in enhancing linguistic analysis and processing for such languages.

**Keywords:** Natural language processing, Deep Learning, Random Forest, RNN, SVM, LSTM.

# CHAPTER ONE:

## Introduction

The increasing need for computational linguistic tools for under-resourced languages has driven the development of a Part-of-Speech (POS) tagger for Dhuluo, a Nilotic language spoken by the Luo people in Kenya. As natural language processing (NLP) technologies evolve, the focus has predominantly been on well-resourced languages like English, Spanish, and Mandarin, which have abundant linguistic resources and computational tools. This imbalance leaves languages like Dhuluo with minimal digital representation and fewer tools for linguistic analysis and processing.

Part-of-Speech tagging is a critical task in NLP, involving the annotation of words in a text with their respective parts of speech, such as nouns, verbs, adjectives, etc. This process is foundational for higher-level NLP tasks including syntactic parsing, machine translation, and information retrieval. Effective POS tagging enhances the performance of these applications by providing accurate syntactic and semantic interpretations of text. For Dhuluo, the development of a POS tagger is not just an academic exercise but a step towards linguistic preservation and the integration of the language into modern digital communication systems.

Despite the growing interest in under-resourced languages, several challenges impede the development of linguistic tools for these languages. These challenges include the lack of large, annotated corpora, limited linguistic research, and the absence of computational resources tailored to these languages​​. Dhuluo, with its rich oral tradition and limited written resources, exemplifies these challenges. Therefore, creating a POS tagger for Dhuluo involves not only the application of existing NLP techniques but also significant data preparation and adaptation of models to suit the linguistic characteristics of the language.

Recent advancements in machine learning and deep learning have opened new avenues for developing POS taggers for under-resourced languages. Traditional machine learning algorithms, such as Random Forest and Support Vector Machines (SVM), have been widely used for POS tagging, Karimov, R. (2018). These models, while effective for well-resourced languages, often struggle with the intricacies of languages with fewer annotated examples and complex morphological structures. In contrast, deep learning models, particularly Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) networks, have shown promise in handling sequential data and capturing the contextual relationships between words.

This paper explores the development of a POS tagger for Dhuluo using a combination of traditional machine learning and deep learning approaches. We begin with a baseline model using Random Forest, which provides insights into the feature importance and the complexity of the tagging task. This is followed by an SVM model that leverages a bag-of-words representation to enhance tagging accuracy. The limitations of these traditional models, particularly in capturing sequential dependencies, necessitate the use of more advanced deep learning techniques.

Deep learning models like RNN and LSTM are particularly suited for languages like Dhuluo due to their ability to capture long-range dependencies and context within text sequences. RNNs process sequences of words, maintaining a hidden state that captures information about previous words, which is crucial for POS tagging. However, RNNs can struggle with long-term dependencies due to vanishing gradient problems. LSTM networks address this issue by incorporating memory cells that can retain information over long sequences, making them highly effective for sequential data tasks like POS tagging.

## Problem Statement

The problem statement for this paper focuses on the need to bridge the gap in linguistic analysis and processing for under-resourced languages like Dhuluo by exploring the effectiveness of both traditional machine learning algorithms and advanced deep learning models in developing a robust POS tagger system. The study aims to address the challenges faced in adapting existing Natural Language Processing (NLP) techniques to suit the linguistic characteristics of under-resourced languages and to enhance the performance of POS tagging systems for languages with limited digital resources

## Objectives

### Main Objectives

To develop a Part-of-Speech (POS) tagger specifically tailored for the Dhuluo language.

### Specific Objectives

1. To explore and compare the performance of traditional machine learning algorithms, including Random Forest and Support Vector Machine (SVM), in POS tagging for Dhuluo.
2. To investigate the effectiveness of advanced deep learning models, particularly Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) networks.
3. To Train and evaluate the machine learning and deep learning models using a dataset of 54,941 annotated word-tag pairs

# CHAPTER TWO: LITERATURE REVIEW

Part of speech (POS) tagging is a fundamental task in natural language processing (NLP) that involves assigning a grammatical category to each word in a sentence. With the increasing demand for accurate and efficient NLP systems, researchers have turned to artificial intelligence (AI) techniques such as deep learning (DL) and machine learning (ML) to enhance POS tagging performance. Gupta et al. conducted a study on aspect-based sentiment analysis of mobile reviews, highlighting the importance of understanding user sentiments in text data. In a similar vein, recent research articles between 2017 and 2021 have focused on AI-oriented POS tagging to provide updated information on the design of POS taggers. This systematic literature review aims to analyze the methodologies, techniques, and evaluation metrics used in DL- and ML-based POS tagger systems to provide insights for new researchers entering this domain.

The study by Chiche and Yitagesu, emphasizes the importance of systematic article selection processes to obtain relevant research articles on POS tagging implementation using AI methods. By selecting recent journal articles published between 2017 and February 2021, the authors were able to review and discuss various parameters such as proposed methods, weaknesses, strengths, and evaluation metrics. This three-phase approach allowed for a comprehensive analysis of recent trends in POS tagging using AI methods, highlighting challenges in DL/ML-based POS tagging and providing future research directions in this domain. By focusing on new ML and DL methods written in English and excluding papers with keywords like survey, review, and analysis, the study ensured a targeted analysis of cutting-edge approaches in AI-oriented POS tagging.

In their methodology, Chiche and Yitagesu outlined a two-stage process for conducting the systematic literature review. Stage-1 involved identifying information resources and keywords related to POS tagging to obtain an initial list of articles, while Stage-2 applied specific criteria to select the most relevant articles for further analysis. This approach allowed the researchers to identify the state-of-the-art in AI-oriented POS tagging design and address gaps in the field to effectively assign parts of speech within sentences. By examining the strengths and weaknesses of the proposed methodologies and evaluating performance metrics, the study provides a comprehensive overview of the advancements and recent trends in DL- and ML-based solutions for POS tagger systems.

The MasakhaPOS project, spearheaded by Alabi et al. (2022), is a groundbreaking initiative focusing on part-of-speech tagging for a diverse set of African languages, shedding light on the intricate challenges encountered during the annotation process. The research underscores the complexities and variations in agreement levels among language teams, emphasizing the need for meticulous data quality control measures and the strategic division of annotated sentences into training, development, and test sets to ensure the robustness and reliability of the POS corpus developed. By adhering to established standards in the field, the MasakhaPOS project not only contributes significantly to the advancement of NLP technologies but also serves as a pivotal resource for researchers and practitioners interested in developing NLP tools for underrepresented languages. Furthermore, the project's alignment with the broader efforts in NLP research, particularly in adapting pre-trained language models to diverse linguistic contexts through multilingual adaptive fine-tuning, highlights the critical role of linguistic diversity in shaping inclusive language technology development and advancing computational linguistics in linguistically diverse regions, as emphasized by Alabi et al. (2022).

The BiLSTM-LAN introduces a hierarchically-refined label attention network that explicitly leverages label embeddings and captures potential long-term label dependencies through hierarchical attention mechanisms (Zhang et al., 2019). By incrementally refining label distributions for each word in the sequence, the model enhances the representation of label sequences, leading to improved tagging accuracy across tasks such as POS tagging, NER, and CCG super tagging.

In comparative studies with top-performing methods reported in the literature, the BiLSTM-LAN has demonstrated its efficacy and superiority. Huang et al. (2015) utilized BiLSTM-CRF, while Ma and Hovy (2016), Liu et al. (2017), and Yang et al. (2018) explored character-level representations in conjunction with BiLSTM-CRF. Additionally, Zhang et al. (2018c) employed S-LSTM-CRF, a graph recurrent network encoder, while Yasunaga et al. (2018) showcased the benefits of adversarial training in improving tagging accuracy. Xin et al. (2018) proposed a compositional character-to-word model combined with LSTM-CRF. In this landscape, the BiLSTM-LAN model has emerged as a competitive and effective solution, outperforming baseline models such as BiLSTM-softmax and BiLSTM-CRF across multiple languages and achieving statistically significant improvements in accuracy (Zhang et al., 2019).

The architecture of the BiLSTM-LAN model is designed to consist of stacked attentive BiLSTM layers, each processing a sequence of vectors to generate hidden state vectors and label distributions. By performing attention over label embeddings, the model derives marginal label distributions that are used to calculate a weighted sum of label embeddings. This process results in a packed label vector that, when combined with input word vectors, forms the hidden state vector for the current layer. The model's emphasis on label attention and hierarchical encoding distinguishes it as a multi-layered BiLSTM-softmax sequence labeler, with each layer contributing to a deeper representation of both input and output sequences (Zhang et al., 2019). Through this approach, the BiLSTM-LAN achieves significantly better accuracies and higher efficiencies compared to BiLSTM-CRF and BiLSTM-softmax models, even without external training data. Moreover, the model offers enhanced interpretability due to its visualizable label embeddings and distributions, providing insights into the tagging process and model decisions.

The paper authored by Yong Jiang, Kewei Tu, and Xinyu Wang, discusses the use of contextual string embeddings for sequence labeling tasks such as Named Entity Recognition (NER), Chunking, Aspect Extraction (AE), Dependency Parsing (DP), and Semantic Dependency Parsing (SDP). The authors evaluate their models using various metrics such as F1 score, accuracy, unlabeled attachment score (UAS), labeled attachment score (LAS), and labeled F1 score. They employ different optimization techniques and hyperparameters based on the task structure, such as using SGD optimizer for sequence-structured tasks and Adam optimizer for graph-structured tasks. The study also includes a detailed configuration evaluation and mentions the benefits of ensemble methods for combining predictions efficiently. The paper references other works in the field of computational linguistics and machine learning, providing a comprehensive overview of the research landscape in this area.

The paper, Yoann Dupont (2017) delves into the realm of named entity recognition (NER) in the French language using machine learning techniques. Named entity recognition is a crucial task in natural language processing that involves identifying and classifying named entities such as persons, organizations, locations, and more in text data. Yoann Dupont's research focuses on enhancing the performance of NER systems specifically for French text by exploring different features that can improve the accuracy and efficiency of entity recognition.

In the study, Dupont conducts experiments to evaluate the effectiveness of various feature sets in the context of French named entity recognition. By analyzing different types of features and their impact on the NER task, the author aims to identify the most relevant and informative features for accurately recognizing named entities in French text. The research contributes to the broader field of natural language processing by providing insights into the specific challenges and nuances of NER in the French language.

|  |  |  |  |
| --- | --- | --- | --- |
| Paper Title | Author(s) | Year | Key Points |
| Aspect-Based Sentiment Analysis of Mobile Reviews | Gupta et al. | N/A | Highlights the importance of understanding user sentiments in text data. |
| Systematic Literature Review on AI-Oriented POS Tagging | Chiche and Yitagesu | 2022 | Emphasizes systematic article selection for relevant research on POS tagging with AI methods. Reviews recent trends, methodologies, and evaluation metrics in DL/ML-based POS tagging. |
| MasakhaPOS: Part-of-Speech Tagging for African Languages | Alabi et al. | 2022 | Focuses on POS tagging for African languages. Highlights challenges in annotation and emphasizes the need for data quality control. Contributes to NLP for underrepresented languages and aligns with broader efforts in adapting pre-trained models to diverse linguistic contexts. |
| Hierarchically-Refined Label Attention Network for Sequence Labeling | Zhang et al. | 2019 | Introduces a hierarchically-refined label attention network for POS tagging. Demonstrates superior performance over BiLSTM-CRF and BiLSTM-softmax models. Emphasizes the model’s interpretability and efficiency. |
| Bidirectional LSTM-CRF for Sequence Tagging | Huang et al. | 2015 | Utilizes BiLSTM-CRF for sequence tagging. Establishes a baseline for comparing advanced POS tagging models. |
| Character-Level Representations for Sequence Labeling | Ma and Hovy | 2016 | Explores character-level representations in conjunction with BiLSTM-CRF. Enhances the performance of sequence labeling tasks. |
| Integrating Graph Recurrent Networks for POS Tagging | Zhang et al. | 2018 | Employs S-LSTM-CRF, a graph recurrent network encoder. Highlights the benefits of integrating graphical models with sequence labeling tasks. |
| Adversarial Training in Improving Tagging Accuracy | Yasunaga et al. | 2018 | Demonstrates the use of adversarial training to enhance tagging accuracy. |
| Compositional Character-to-Word Model with LSTM-CRF | Xin et al. | 2018 | Proposes a compositional character-to-word model combined with LSTM-CRF. Highlights improvements in tagging performance through this approach. |

# CHAPTER THREE: Methodology

Three primary approaches were employed in developing the Dhuluo POS tagger: Random Forest, Support Vector Machine (SVM), and deep learning models (RNN and LSTM).

Proposed methodology

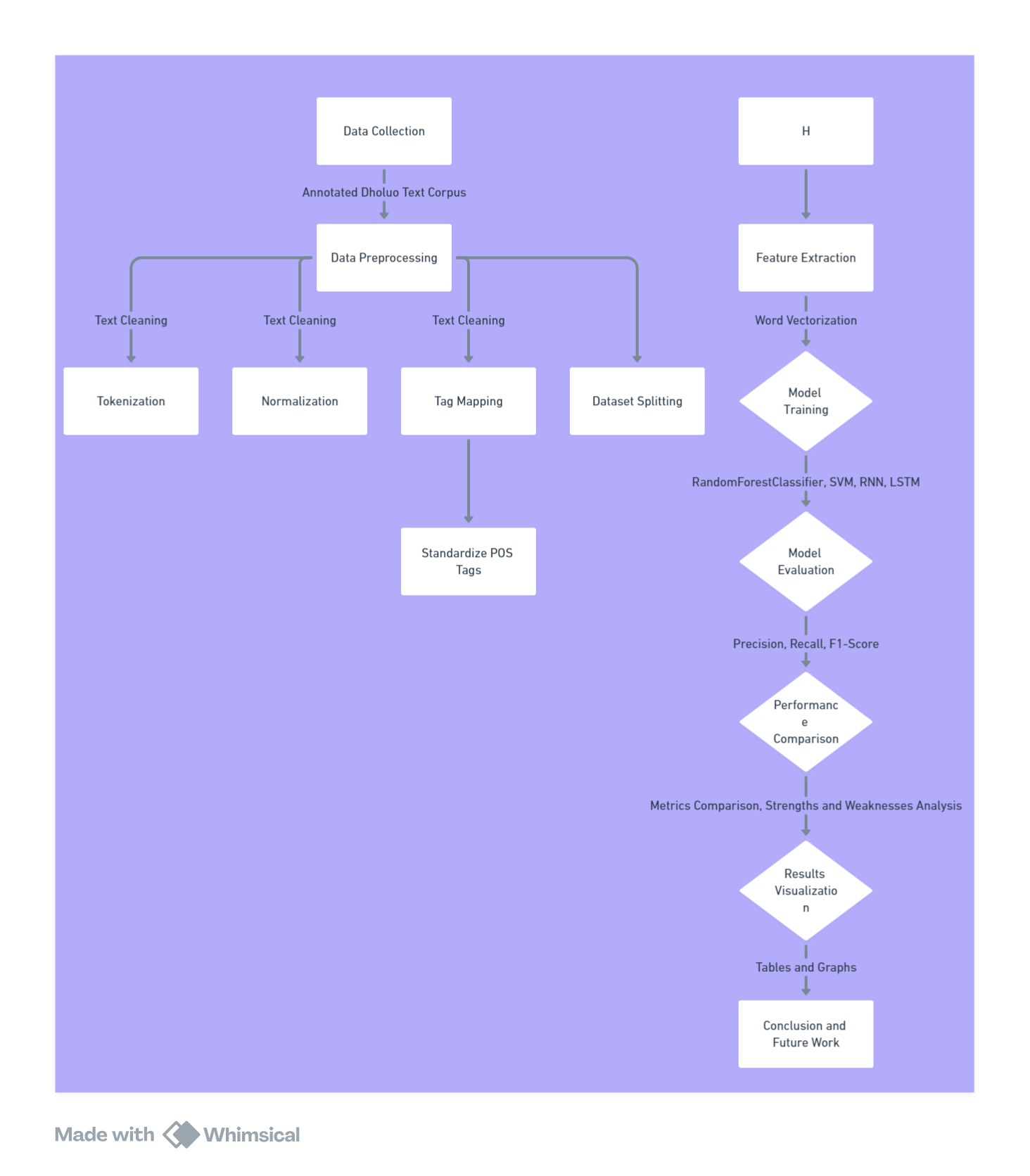


Figure 1: Proposed methodology

## Dataset

The dataset used for this study comprises a collection of Dhuluo words annotated with their respective POS tags kencorpus (2022). The dataset includes 54,941 word-tag pairs distributed across several POS categories such as nouns (NN), verbs (V), adjectives (Adj), and more complex tags like Conj+NN and NN+Det. Data preprocessing involved handling missing values mapping the similar tags classes together and ensuring the dataset was suitable for training and testing​​.

## Random Forest

The Random Forest classifier was used as a baseline model. It achieved an accuracy of approximately 61.64%. Despite being a non-sequential model, it provided insights into the feature importance and the complexity of the tagging task.

## Support Vector Machine (SVM)

The SVM model was trained using a bag-of-words representation of the dataset. It slightly improved the tagging accuracy to about 61.89%. SVM's effectiveness in handling high-dimensional data was beneficial, though it struggled with the nuances of sequential dependencies in language.

## Deep Learning Models

Deep learning approaches, specifically RNN and LSTM, were employed to capture the sequential nature of language more effectively.

* **RNN:** The RNN model demonstrated a progressive learning curve, achieving an accuracy of 63.98% over five epochs. Despite its limitations with long-range dependencies, it provided a substantial improvement over traditional machine learning model.
* **LSTM:** The LSTM model, designed to handle long-term dependencies, achieved the highest accuracy of 64.83%. Its architecture enabled better context preservation, leading to more accurate tagging of complex sentences.

# Results and Discussions

In this section, we present the evaluation results for different machine learning models used for POS tagging: RandomForestClassifier, Support Vector Machine (SVM), Recurrent Neural Network (RNN), and Long Short-Term Memory (LSTM). We compare these models based on their precision, recall, and F1-score to determine their effectiveness and highlight their respective strengths and weaknesses.

## Word Cloud Representation

A word cloud can provide a visual representation of the most frequent tags in the dataset, giving insight into the data distribution and the importance of various POS tags.

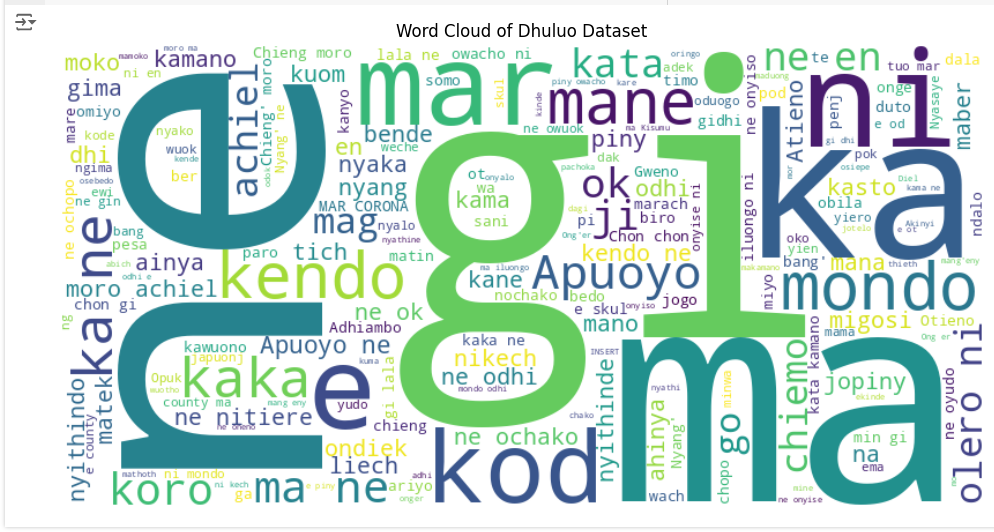


Figure 2: Word cloud word representation

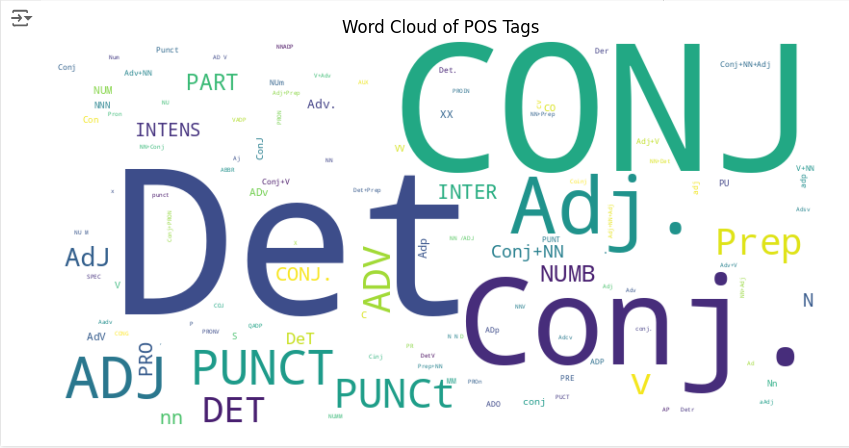


Figure 3: WordCloud Tag representations

## Model Evaluation Metrics

The following metrics were used to evaluate the models:

* **Precision**: Measures the accuracy of positive predictions.
* **Recall**: Measures the ability of the model to identify all relevant instances.
* **F1-Score**: The harmonic mean of precision and recall, providing a balanced measure.

## Summary of Results

The table below summarizes the performance metrics for each model:

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Precision | Recall | F1-Score |
| RandomForestClassifier | 0.736 | 0.671 | 0.675 |
| SVM | 0.733 | 0.673 | 0.675 |
| RNN | 0.726 | 0.717 | 0.706 |
| LSTM | 0.713 | 0.704 | 0.696 |

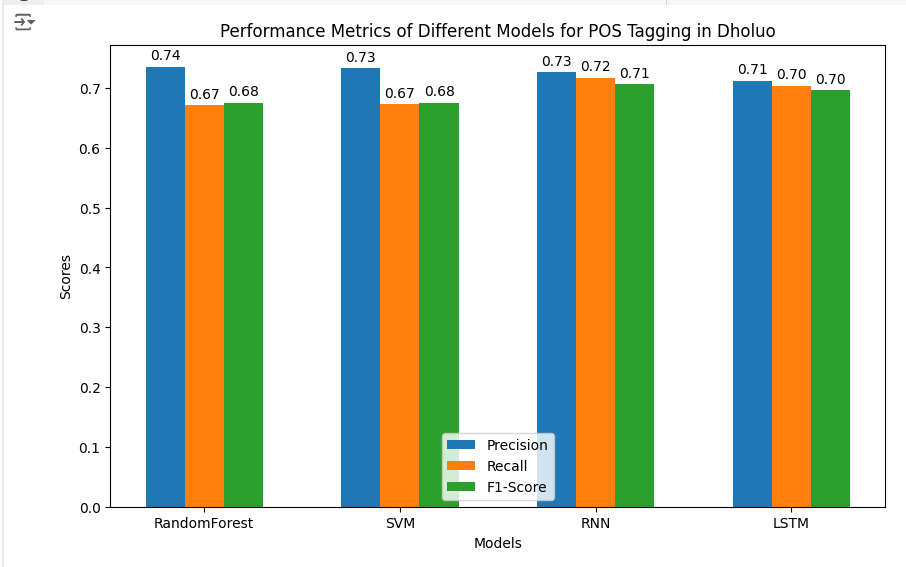
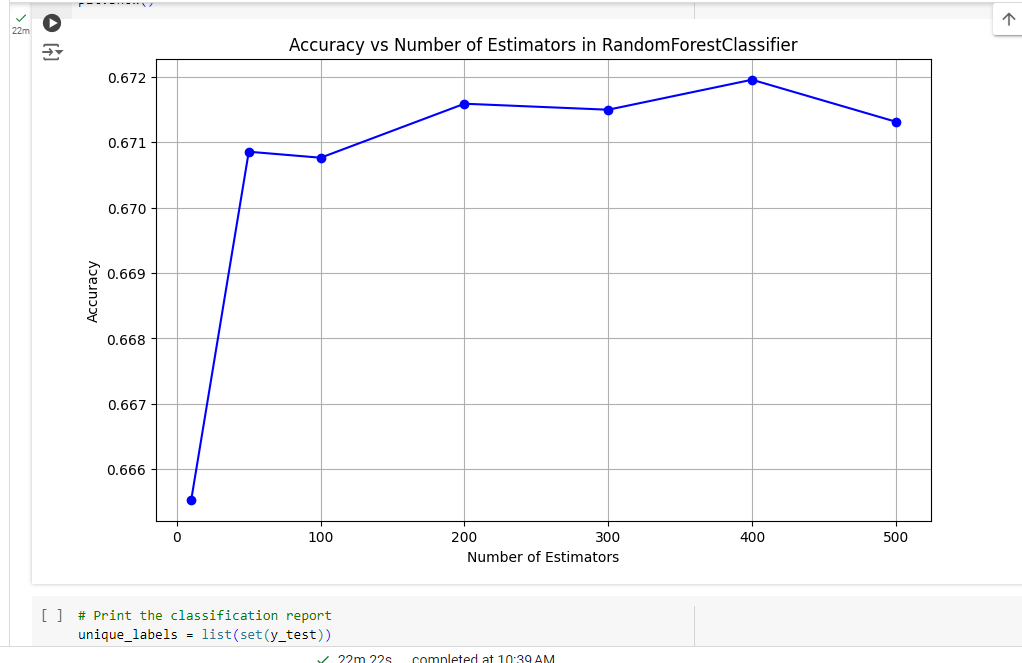


Figure 4: Performances metrics of the different models for POS tagging

### RandomForestClassifier

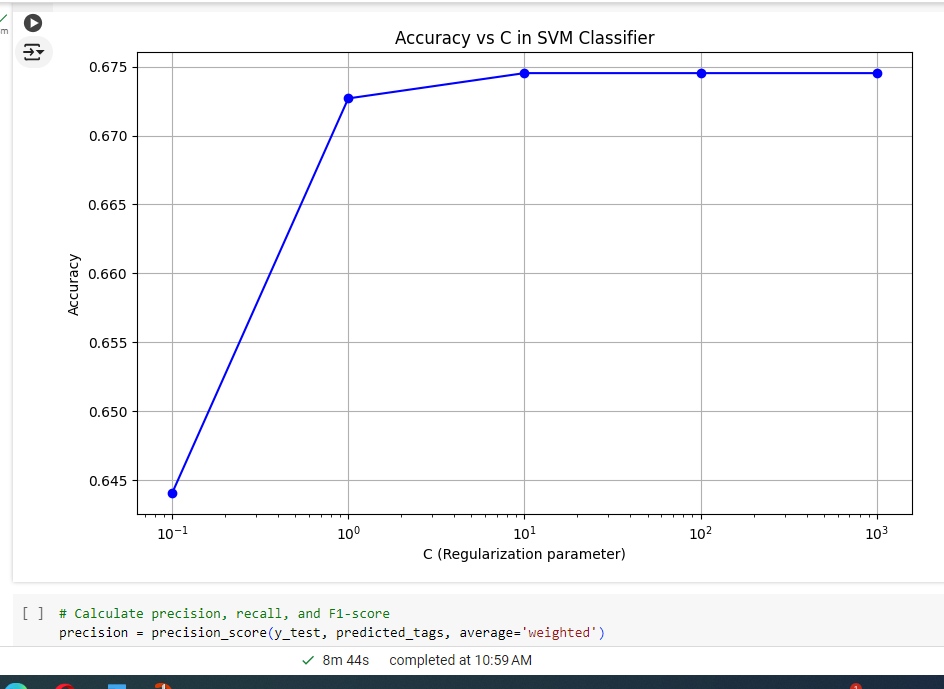
The RandomForestClassifier achieved a precision of 0.736, recall of 0.671, and an F1-score of 0.675. This model demonstrates a good balance between precision and recall, making it suitable for POS tagging where general accuracy is important.

The diagram below shows visualization of accuracy against number of estimators



### SVM

The Support Vector Machine model achieved a precision of 0.733, recall of 0.673, and an F1-score of 0.675. The performance metrics are very close to those of the RandomForestClassifier, indicating that SVM is also a strong candidate for POS tagging tasks. However, the computational cost of training SVMs can be higher compared to Random Forest.

The diagram below shows the SVM accuracy  


### RNN

The Recurrent Neural Network model achieved a precision of 0.726, recall of 0.717, and an F1-score of 0.706. This model shows improved recall compared to Random Forest and SVM, suggesting that RNN is better at identifying relevant instances. However, its slightly lower precision indicates that it may produce more false positives.

### LSTM

The Long Short-Term Memory model achieved a precision of 0.713, recall of 0.704, and an F1-score of 0.696. While LSTM's performance is slightly lower than that of the RNN, it still demonstrates strong capabilities in sequence prediction tasks. The results indicate that LSTM, despite being complex, might need further hyperparameter tuning or more training data to outperform simpler models like Random Forest and SVM in this specific task.

Comprehensive Discussion

Performance Analysis:

The RandomForestClassifier and SVM performed comparably, with very close values for precision, recall, and F1-score. This suggests that both models are robust for POS tagging tasks and can be chosen based on other factors like training time and computational efficiency.

RNN and LSTM, while typically more powerful due to their ability to capture sequential dependencies, did not significantly outperform the RandomForestClassifier and SVM in this case. This could be attributed to the dataset size, training parameters, or the nature of the POS tagging task itself.

Model Selection Considerations:

RandomForestClassifier: Offers a good balance of precision and recall, is easier to train, and is less computationally intensive compared to deep learning models.

SVM: Performs similarly to Random Forest but with potentially higher computational cost.

RNN and LSTM: Show the potential for higher recall but require more data and computational resources. They might be preferred in scenarios where capturing sequential patterns is crucial.

Error Analysis:

Further analysis using confusion matrices for each model can help identify specific classes where the models struggle. This information can be used to fine-tune models or to apply more targeted preprocessing steps.

Investigating misclassifications and refining feature engineering can also help improve model performance.

## Future Improvements:

Data Augmentation: Increasing the size of the training dataset could help improve the performance of RNN and LSTM models.

Hyperparameter Tuning: Applying grid search or random search to find optimal hyperparameters for each model.

Advanced Techniques: Experimenting with transformer-based models like BERT, which have shown state-of-the-art performance in various NLP tasks.

# Conclusion

The project aimed to evaluate various machine learning models for Part-of-Speech (POS) tagging, specifically RandomForestClassifier, SVM, RNN, and LSTM. The RandomForestClassifier and SVM exhibited comparable performance, both achieving high precision and recall with an F1-score of 0.675, demonstrating their robustness and efficiency for POS tagging tasks with lower computational costs. Despite the generally higher potential of RNN and LSTM to capture sequential dependencies, their performance was only marginally better in recall, suggesting their advantages may require larger datasets and extensive hyperparameter tuning to be fully realized. RandomForestClassifier and SVM are recommended for their balance of performance and efficiency, while RNN and LSTM may be better suited for applications where the sequential nature of data is paramount, albeit with higher computational demands. This project highlights the importance of selecting models based on specific task requirements and computational constraints, pointing to future improvements through data augmentation and advanced techniques like transformer-based models.

# References

*POS tagged data* (2022) *Kencorpus*. Available at: https://kencorpus.maseno.ac.ke/pos-tagged-data/ (Accessed: 12 June 2024).

Cui, L., & Zhang, Y. (2019). Hierarchically-refined label attention network for sequence labeling. *arXiv preprint arXiv:1908.08676*.

Wang, X., Jiang, Y., Bach, N., Wang, T., Huang, Z., Huang, F., & Tu, K. (2020). Automated concatenation of embeddings for structured prediction. *arXiv preprint arXiv:2010.05006*.

Chiche, A., & Yitagesu, B. (2022). Part of speech tagging: a systematic review of deep learning and machine learning approaches. *Journal of Big Data*, *9*(1), 10.

Dione, C. M. B., Adelani, D., Nabende, P., Alabi, J., Sindane, T., Buzaaba, H., ... & Klakow, D. (2023). MasakhaPOS: part-of-speech tagging for typologically diverse African languages. *arXiv preprint arXiv:2305.13989*.

Tambat, I., & Narayankar, M. (2022, September). A Comparative Survey on Parts of Speech Taggers for the Marathi Language. In *2022 4th International Conference on Inventive Research in Computing Applications (ICIRCA)* (pp. 1740-1747). IEEE.

Martin, L., Muller, B., Suárez, P. J. O., Dupont, Y., Romary, L., de La Clergerie, É. V., ... & Sagot, B. (2019). CamemBERT: a tasty French language model. *arXiv preprint arXiv:1911.03894*.

Warjri, S., Pakray, P., Lyngdoh, S. A., & Maji, A. K. (2021). Part-of-speech (pos) tagging using deep learning-based approaches on the designed khasi pos corpus. *Transactions on Asian and Low-Resource Language Information Processing*, *21*(3), 1-24.

Alam Firoj, Chowdhury Shammur Absar, and Noori Sheak Rashed Haider. 2016. Bidirectional LSTMs—CRFs networks for Bangla POS tagging. In *Proceedings of the 2016 19th International Conference on Computer and Information Technology*. IEEE, 377–382.

Martinez, A. R. (2012). Part‐of‐speech tagging. *Wiley Interdisciplinary Reviews: Computational Statistics*, *4*(1), 107-113.

Kanakaraddi, S. G., & Nandyal, S. S. (2018, March). Survey on parts of speech tagger techniques. In *2018 International Conference on Current Trends towards Converging Technologies (ICCTCT)* (pp. 1-6). IEEE.

Karimov, R. (2018). Combined Machine-Learning Approach to PoS-Tagging of Middle English Corpora. *Crossroads. A Journal of English Studies*, (02 (21)), 42-52.