Clustering Analysis on Sentiment Tagged Parallel Corpus for Luganda and Swahili

Arno Jooste
University of Pretoria
21A Bernard Road, Poortview
Gauteng South Africa
u21457451@tuks.co.za

DECLARATION OF ORIGINALITY

I certify that this assignment represents my own work. I have not used any unauthorised or unacknowledged assistance or sources in completing it including free or commercial systems or services offered on the internet or text generating systems embedded into software.

ABSTRACT

The development of NLP resources for low-resource African languages, such as Luganda and Swahili, is difficult due to a lack of linguistic resources, financing, and qualified staff. This study leverages clustering and topic modelling approaches on a sentiment-tagged parallel corpus to investigate and improve the potential of these languages in NLP applications. Using K-means clustering and Latent Dirichlet Allocation (LDA), we investigate the effect of feature extraction approaches such as TF-IDF on clustering performance. Our findings show distinct patterns in sentiment and topic correlations for Luganda and Swahili, providing insight into sentiment structures within low-resource language corpora. This paper makes a methodological contribution to NLP research on African languages, with the goal of closing the linguistic resource gap and setting up new avenues for more advanced sentiment applications.

1. INTRODUCTION

Africa is a continent rich in linguistic and cultural diversity [6], with approximately 1,250 to 3,000 spoken languages across various ethnic groups. In the digital age, African languages face a crucial need for technological development to support sectors such as the economy, politics, education, and healthcare. These technological advancements rely on linguistic resources that enable applications in machine translation, sentiment analysis, data analysis, and beyond [8; 9]. However, a lack of funding, documentation, and trained personnel remains a significant barrier to advancing natural language processing (NLP) for African languages [9; 1; 3].
This study focuses on NLP techniques such as clustering and topic modelling to enhance research on low-resource languages, specifically Luganda and Swahili. To address the gap, we explore clustering algorithms applied to sentiment-tagged corpora in these languages, particularly with K-Means, examining unique challenges that limited language resources present. This research also investigates the impact of feature extraction methods, especially TF-IDF, and topic modelling with LDA on clustering performance. Finally, we explore clustering's potential to uncover underlying sentiment structures and identify patterns within these corpora, examining correlations between sentiment labels and topic themes.

2. PROBLEM STATEMENT

The development of NLP resources for low-resource African languages remains limited due to a shortage of funding, documentation, and trained personnel [9; 1; 3]. This study focuses on enhancing research efforts for Luganda and Swahili by applying clustering and topic modeling techniques to sentiment-tagged corpora. Given the constraints of limited resources, this research explores the use of clustering algorithms, particularly K-Means, to reveal meaningful insights in sentiment data. Additionally, it investigates the role of feature extraction methods, such as TF-IDF, and topic modeling with LDA in influencing clustering outcomes. By examining how these methods impact clustering performance, this study aims to uncover underlying sentiment structures and identify patterns and correlations between sentiment labels and topic themes.

3. LITERATURE SURVEY

Recent research has focused on creating a benchmark for sentiment analysis across multiple African languages [7]. The study by Muhammad et al. introduced AfriSenti, a significant benchmark dataset for sentiment analysis in 14 African languages. This dataset addresses the scarcity of NLP resources for African languages, enabling researchers to develop and evaluate models for sentiment analysis. Their paper highlights the data collection methodology, annotation guidelines, and challenges encountered during curation. It also presents baseline experiments demonstrating the dataset's utility and potential impact on future research in African language NLP [7].
Another study by Aryal et al. made use of the AfriSenti benchmark and evaluated the performance of state-of-the-art transformer models on a sentiment analysis task for 12 African languages. The authors investigated the impact of language-specific models versus multilingual models, as well as the role of data quantity in model performance. Their findings reveal that language-specific models generally out-

perform multilingual models, emphasising the importance of language-specific training data. Furthermore, more training data leads to better model performance, even for language-specific models. However, the optimal model choice varies across languages, with multilingual models potentially performing better for languages with limited data. Aryal et al.'s research highlights the need for increased language-specific data and models to enhance sentiment analysis capabilities for African languages [2].

Adelani et al. introduced MasakhaNEWS, a new benchmark dataset for news topic classification in 16 African languages. The dataset aims to address the lack of resources for NLP tasks in African languages. The authors evaluate various models, including classical machine learning models and state-of-the-art language models. To address the challenge of limited data, they explore zero-shot and few-shot learning techniques, such as prompting language models and parameter-efficient fine-tuning. Their results demonstrate the potential of these techniques, especially prompting, for achieving reasonable performance in low-resource settings.

In summary, recent studies have made significant strides in developing NLP resources for African languages, exemplified by datasets such as AfriSenti and MasakhaNEWS. However, challenges remain, particularly in securing language-specific models and comprehensive resources for low-resource languages. This literature highlights the necessity for continued research in this area. By focusing on clustering and topic modeling for Luganda and Swahili, this study aims to contribute valuable insights and methodologies to enhance sentiment analysis capabilities for these languages.

4. METHODOLOGY

This section outlines the methodology and includes a discussion on the data that was used during this study.

4.1 Data

The dataset used for this study is created by the Harvard Dataverse group [4]. The dataset, named Sentiment Tagged Parallel Corpus for Luganda and Swahili, comprises of 10,000 sentence pairs in English, Luganda, and Swahili. The Luganda and Swahili translations were produced by language experts and professional translators in collaboration with Makerere University. The sentiment labels were assigned to the English sentences, and corresponding translations were assigned the same sentiment label. Our study also leveraged other data for tasks such as parts-of-speech tagging. Dione et al. introduced MasakhaPOS, a large-scale part-of-speech (POS) dataset for 20 African languages. The dataset addresses the lack of POS resources for African languages and facilitates NLP research [5]. The paper discusses the challenges of applying universal dependency guidelines to African languages. It then presents baseline experiments using various models, including conditional random fields and multilingual language models. The study also explores cross-lingual transfer learning techniques, demonstrating that transferring knowledge from related languages can significantly improve POS tagging performance, especially when using parameter-efficient fine-tuning methods. The findings highlight the importance of language family and morphosyntactic properties in selecting appropriate source languages for transfer learning. For the purposes of this study,

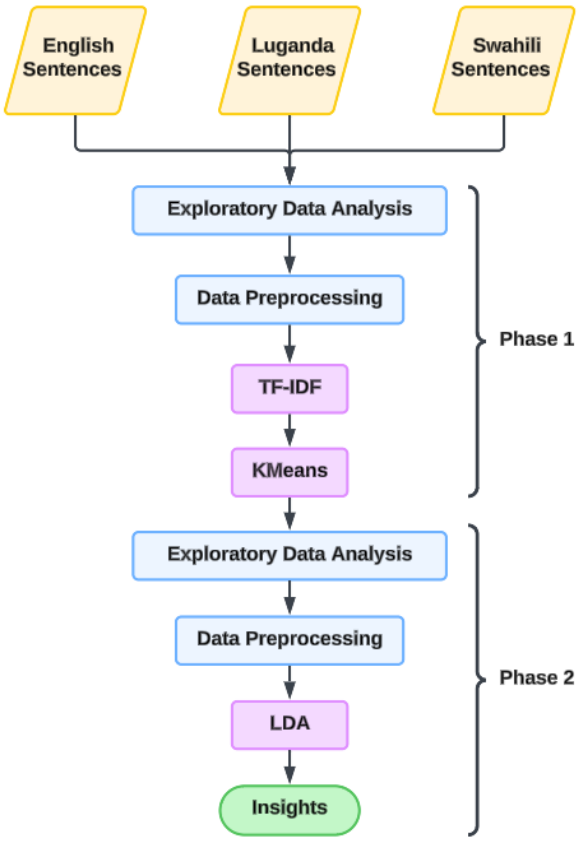

Figure 1: Methodology Pipeline

we only made use of the Luganda and Swahili POS tags such that we could tag the data from [4] with the appropriate tags. The English texts were tagged using Python libraries. Other data used for this study was only used for validating Luganda and Swahili words' contexts. This data included [1]Sunbird/salt-dataset from Hugging Face, the [2]Luganda-English parallel corpus by R. Kimera, and the public English-Swahili corpus availabe on the [3]Translators Without Borders website.

4.2 Methodology Pipeline

The overview of the methodology followed can be seen in Figure 1.

Initially, the English, Luganda, and Swahili sentences were extracted and subjected to basic exploratory data analysis (EDA). This EDA revealed slight misspellings in the sentiment codes (e.g. 'Pos', 'pos', 'Neg', 'neg') but confirmed an even distribution across positive and negative sentiments. Subsequent preprocessing steps included removing stop words. For Luganda and Swahili, stopword lists from the ([4]mukairnlpv1) library and a GitHub repository by Liam Doherty ([5]more-stoplists) were utilised. The final

[1]SALT-dataset
[2]English-Luganda-Parallel-corpus
[3]gamayun-5k-english-swahili
[4]mukairnlpv1
[5]more-stoplists

preprocessing steps involved encoding sentiment labels, tokenization, and lemmatization. The phase concluded with TF-IDF and K-Means clustering.

The second phase delved deeper into the data through parts-of-speech (POS) tagging. English text was tagged using NLTK's *pos_tag* library, while Luganda and Swahili data leveraged pre-tagged words from the [6]masakhane-pos [5] dataset to map and tag our data. After filtering for adjectives and nouns, the data underwent further preprocessing, including bigram creation using Word2Vec, and filtering out the most common and uncommon words. Finally, Latent Dirichlet Allocation (LDA) was applied for topic modelling, and insights were extracted from the identified topics and previous phase's clusters.

5. RESULTS

This section discusses the results of this study.

5.1 Clustering

The MiniBatchKMeans library was used to train and make predictions. The Elbow-method was followed to select the optimal number of clusters to be used for each language's data. We concluded that the optimal number of clusters for the English data was *k=100*, for the Luganda data it was *k=60*, and for the Swahili data it was *k=80*. With these optimal number of clusters we continued to make predictions. The predictions were used and the clustering performances were evaluated by visualising the clusters using PCA, and t-SNE scatter plots. The initial plots of applying PCA and t-SNE showcased poor performance due to overlapping of clusters. Example of the PCA and t-SNE plots for the English data are illustrated in Figures 2 and 3. We then experimented with a pipeline of applying PCA and t-SNE in sequence and the resulting visualisations illustrated much better clustering performance. The result for the English data is illustrated in Figure 4.

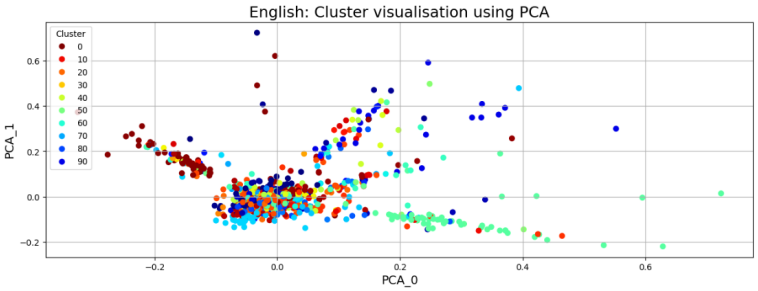

Figure 2: PCA plot for English data

5.2 Topic Modelling

The [7]LdaMulticore library provided by Gensim was used to perform the topic modelling across all three sets of texts. The durations (in seconds) of time taken to perform the topic modelling for each language are listed in Table 1.

[6]masakhane-pos
[7]Gensim: LdaMulticore

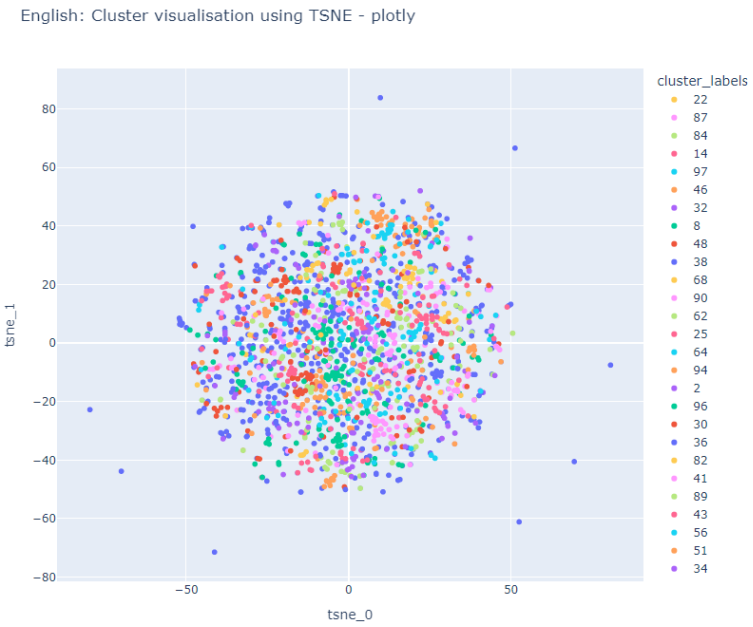

Figure 3: t-SNE plot for English data

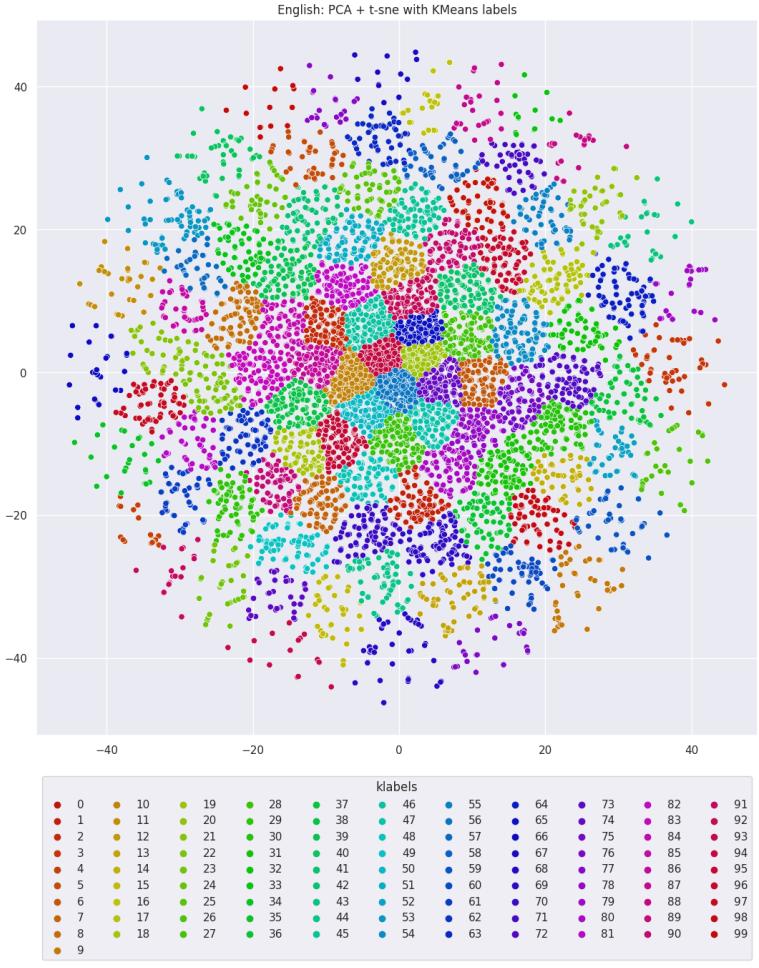

Figure 4: PCA + t-SNE plot for English data

There is a clear difference between the durations of performing topic modelling across the different sets of text data. Our preprocessing (during phase 2) supports this large difference due to the English data containing much more adjectives and nouns than the other two languages' data. For each language, dictionaries and corpora were built using adjectives, nouns, and bigrams from the documents. Each dictionary was filtered to exclude bigrams that appeared in fewer than 20 instances, ensuring only commonly occurring terms were retained. Then, each document was transformed into a bag-of-words format, with the dictionaries capturing unique tokens and the corpora representing document structures. This approach provided a tailored vocabulary and document representation for subsequent analysis steps. the number of unique tokens and number of documents for each language are reported in Table 2. These results clearly support the differences in durations of performing topic modelling.

Language	Duration (seconds)
English	196.15
Luganda	147.66
Swahili	26.57

Table 1: Comparison of duration to perform topic modelling across different languages' text data

Language	No. of unique tokens	No. of documents
English	305	9997
Luganda	148	7147
Swahili	7	754

Table 2: Overview of number of unique tokens and documents for each language

5.3 Insights

This section discusses the insights obtained from the clustering analysis and topic modelling. To keep it concise, we will only be discussing the results of the first clusters of each language.

5.3.1 Insights on English data

Cluster 0 from the English data revealed several insights into the clustering and topic modeling analysis. The cluster leaned slightly toward negative sentiment (shown in Figure 5, with more negative sentences (≈62) than positive (≈48), suggesting a focus on complaints or emotionally charged discussions. Top words from this cluster such as "faulty," "hate," and "disappointed" suggested connections to personal relationships or emotional situations, potentially indicating dissatisfaction. Figure 6 illustrates that the terms largely aligned with topic 0, which included "angry," "sad," and "miserable," reinforcing a theme of complex social or emotional interactions.

5.3.2 Insights on Luganda data

Cluster 0 displayed a balanced sentiment distribution (shown in Figure 7, slightly leaning towards negative sentiment (≈90 negative versus ≈85 positive mentions), hinting at discussions that may have involved both challenges and positive

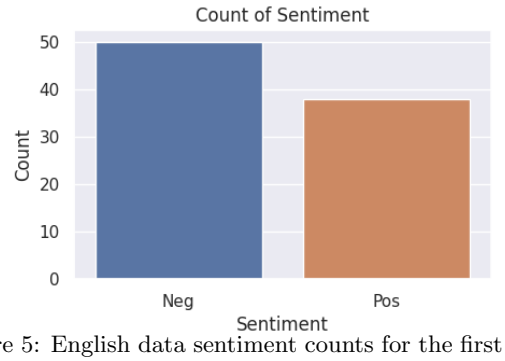

Figure 5: English data sentiment counts for the first cluster

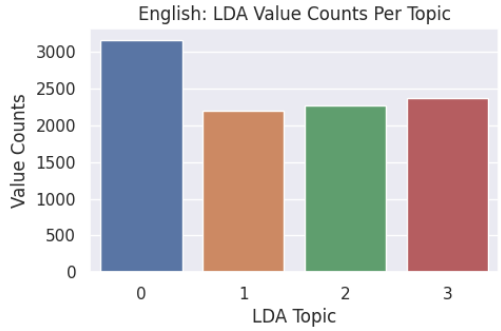

Figure 6: English data topic counts for the first cluster

moments in personal or educational contexts. The presence of words such as "omwana" (child context), "eziyigiriza" (educational context), and "omulungi" (beautiful context) suggested family and social themes, touching on education and admiration. Figure 8 illustrates that the terms largely aligned with topic 0, indicating a focus on family dynamics. While terms such as "embi" (bad) hinted at struggles, the inclusion of "mmwagala" (love or similar contexts) and "omulungi" highlighted supportive, positive interactions.

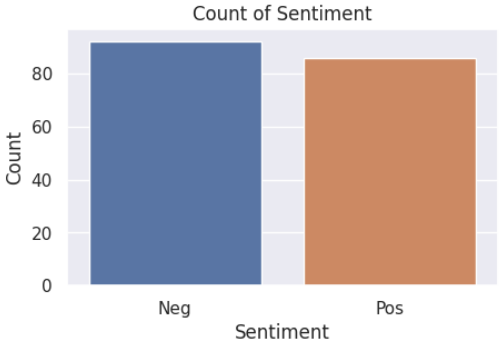

Figure 7: Luganda data sentiment counts for the first cluster

5.3.3 Insights on Swahili data

Cluster 0 showed a slight bias toward negative sentiment (illustrated in Figure 10, with approximately 15 negative sentiments compared to approximately 13 positive, hinting at a subtle portrayal of both challenges and resilience in personal or physical contexts. Top words from the cluster such as "pain" and "severe" suggested themes of endurance and struggle. Words such "kuendana" (to match) and "kudumisha" (to maintain) implied sustained efforts,

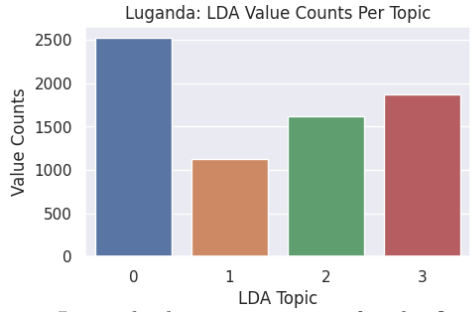

Figure 8: Luganda data topic counts for the first cluster

while common connectors ("ya," "za") mostly linked this cluster to topic 0 (illustrated in Figure 10 which contained words about everyday topics such as health or well-being, with sentiment balancing between adversity and perseverance in social or work contexts.

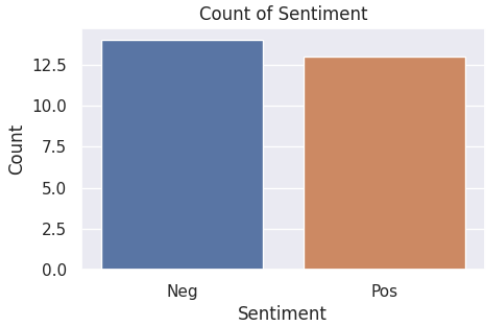

Figure 9: Swahili data sentiment counts for the first cluster

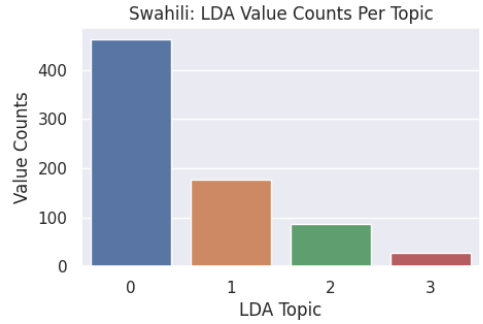

Figure 10: Swahili data topic counts for the first cluster

6. CONCLUSIONS

This study has demonstrated that clustering and topic modelling techniques can effectively reveal underlying sentiment structures within Luganda and Swahili parallel corpora. By employing methods such as K-Means clustering and LDA, and leveraging feature extraction with TF-IDF, we were able to identify patterns and correlations between sentiment labels and thematic topics across both languages. The use of PCA and t-SNE for dimensionality reduction improved visualisation, making distinct clusters more identifiable and providing insights into language-specific data structures. Additionally, the comparison of processing times and token distributions highlights the significant variance between English and African language datasets, stressing the importance of customised approaches for low-resource languages. These findings contribute to the growing field of NLP for African languages, emphasising the potential of clustering techniques in facilitating sentiment analysis and enriching language technology applications for under-resourced languages. Future work should explore more sophisticated algorithms and cross-linguistic transfer learning to further enhance clustering efficacy and NLP model performance for low-resource languages.

7. ACKNOWLEDGEMENTS

I want to give credit to [8]Harvard Dataverse for providing the public dataset which formed the foundation of this study. I also want to acknowledge the use of Google's Gemini for possible verification of context for Luganda and Swahili words.

8. REFERENCES

[1] G. Adda, S. Stüker, M. Adda-Decker, O. Ambouroue, L. Besacier, D. Blachon, H. Bonneau-Maynard, P. Godard, F. Hamlaoui, D. Idiatov, G.-N. Kouarata, L. Lamel, E.-M. Makasso, A. Rialland, M. Van de Velde, F. Yvon, and S. Zerbian. Breaking the unwritten language barrier: The bulb project. *Procedia Computer Science*, 81:8–14, 2016. SLTU-2016 5th Workshop on Spoken Language Technologies for Under-resourced languages 09-12 May 2016 Yogyakarta, Indonesia.

[2] S. K. Aryal, H. Prioleau, and S. Aryal. Sentiment analysis across multiple african languages: A current benchmark, 2023.

[3] I. I. Ayogu and O. Abu. Automatic diacritic recovery with focus on the quality of the training corpus for resource-scarce languages. In *2020 IEEE 2nd International Conference on Cyberspac (CYBER NIGERIA)*, pages 98–103, 2021.

[4] C. Babirye, J. Tusubira, J. Mukiibi, J. Nakatumba-Nabende, and A. Katumba. Sentiment Tagged Parallel Corpus for Luganda and Swahili, 2023.

[5] C. M. B. Dione, D. I. Adelani, P. Nabende, J. Alabi, T. Sindane, H. Buzaaba, S. H. Muhammad, C. C. Emezue, P. Ogayo, A. Aremu, C. Gitau, D. Mbaye, J. Mukiibi, B. Sibanda, B. F. P. Dossou, A. Bukula, R. Mabuya, A. A. Tapo, E. Munkoh-Buabeng, V. Memdjokam Koagne, F. Ouoba Kabore, A. Taylor, G. Kalipe, T. Macucwa, V. Marivate, T. Gwadabe, M. T. Elvis, I. Onyenwe, G. Atindogbe, T. Adelani, I. Akinade, O. Samuel, M. Nahimana, T. Musabeyezu, E. Niyomutabazi, E. Chimhenga, K. Gotosa, P. Mizha, A. Agbolo, S. Traore, C. Uchechukwu, A. Yusuf, M. Abdullahi, and D. Klakow. MasakhaPOS: Part-of-speech tagging for typologically diverse African languages. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10883–10900, Toronto, Canada, July 2023. Association for Computational Linguistics.

[6] S. T. Hazel, N. K. Macdonald, and M. M. Edwin. The role of women in promoting cultural diversity and tolerance in south africa communities: A theoretical discourse. *Gender and Behaviour*, 22(1):22188–22200, 2024.

[7] S. H. Muhammad, I. Abdulmumin, A. A. Ayele, N. Ousidhoum, D. I. Adelani, S. M. Yimam, I. S. Ahmad, M. Beloucif, S. M. Mohammad, S. Ruder, O. Hourrane, P. Brazdil, F. D. M. A. Ali, D. David, S. Osei, B. S. Bello, F. Ibrahim, T. Gwadabe, S. Rutunda, T. Belay, W. B. Messelle, H. B. Balcha, S. A. Chala, H. T. Gebremichael, B. Opoku, and S. Arthur. Afrisenti: A twitter sentiment analysis benchmark for african languages, 2023.

[8] K. Siminyu, S. Freshia, J. Abbott, and V. Marivate. Ai4d – african language dataset challenge, 2020.

[9] K. Siminyu, G. Kalipe, D. Orlic, J. Abbott, V. Marivate, S. Freshia, P. Sibal, B. Neupane, D. I. Adelani, A. Taylor, J. T. ALI, K. Degila, M. Balogoun, T. I. DIOP, D. David, C. Fourati, H. Haddad, and M. Naski. Ai4d – african language program, 2021.

[8] https://dataverse.harvard.edu/