Building Text and Speech Benchmark Datasets and Models for Low-Resourced East African Languages: Experiences and Lessons

1 Introduction

Africa is home to over 2000 languages, yet most remain absent from the global NLP ecosystem. The lack of text and speech datasets has hindered the development of machine translation, sentiment analysis, and speech technologies for African contexts. In Uganda alone, more than 41 indigenous languages are spoken, but only a few have any digital representation. Addressing this gap is critical for inclusive technological development and language preservation. This work focuses on building open benchmark datasets and baseline models for five East African languages: Luganda, Runyankore-Rukiga, Lumasaba, Acholi, and Swahili through participatory, community-driven approaches.

2 Methodology

2.1 Data collection

We adopted a participatory approach, mobilizing linguists, translators, and communities to contribute text and speech data. Tools such as Mozilla Pontoon, Common Voice, and the custom Yogera application facilitated translation, crowdsourced voice contributions, and validation. Each translation underwent linguistic validation by experts. Sentences followed guidelines ensuring readability, semantic accuracy, and cultural neutrality. Speech data was reviewed for clarity, speaker diversity, and dialectal representation.

We developed monolingual corpora consisting of approximately 400,000 sentences in Luganda, 200,000 in Swahili, and about 40,000 sentences each in Acholi, Runyankore-Rukiga, and Lumasaba. In addition, a parallel corpus of 40,000 English sentences was translated into each of the five East African languages, producing aligned bilingual datasets, alongside sentiment-tagged parallel corpora in Luganda and Swahili. For speech data, the collection yielded 582 hours of Luganda audio (438 hours validated) and 1,100 hours of Swahili audio (393 hours validated), contributed by more than 1,700 speakers across a broad age range (18–80 years) and multiple dialects, thereby ensuring diversity and representativeness.

3 Baseline Models

The baseline experiments underscore the utility of the developed datasets for a range of NLP tasks. For **machine translation, a Marian MT model** achieved **BLEU scores of 26.0 (English→Luganda) and 24.6 (Luganda→English)**. Topic modeling with NMF uncovered coherent clusters in 20,000 Luganda sentences, while **topic classification** showed strong performance, with **SVM (F1 score of 0.967)** and **BERT (F1 score of 0.983)** as the best models. **Sentiment classification** on the Luganda corpus also produced high performance with **Stacking Classifier (F1 score of 0.90)** and **MultinomialNB (F1 score of 0.88)**. In speech, **Coqui STT** trained on 300 hours of Luganda data achieved a **WER of 23%**, with further gains from Wav2vec2, underscoring the benefit of transfer learning.

4 Conclusion

This work shows that open, community-driven datasets can enable robust NLP baselines for African languages despite challenges of scarcity and diversity. Community participation and transfer learning were key to ensuring quality and improving performance. The released benchmark datasets and baseline models for five East African languages provide a foundation for inclusive NLP research and future cross-lingual applications.