GANDABERT: TRANSFER LEARNING WITH MBERT FOR LUGANDA NEWS CLASSIFICATION

SETH MBASHA

S22B23/010

A PROJECT REPORT SUBMITTED TO THE FACULTY OF ENGINEERING, DESIGN AND TECHNOLOGY IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE AWARD OF THE DEGREE OF BACHELOR OF SCIENCE IN COMPUTER SCIENCE OF UGANDA CHRISTIAN UNIVERSITY

April, 2025

UGANDA CHRISTIAN UNIVERSITY

A Centre of Excellence in the Heart of Africa

Abstract

Luganda, spoken by over 21 million Ugandans, is significantly under-resourced in Natural Language Processing (NLP), lacking effective tools like news classifiers. This gap hinders digital information access and contributes to the digital language divide. This research project addressed this challenge by developing GandaBERT, a model for Luganda news classification. The methodology involved fine-tuning the multilingual BERT (mBERT) model on a novel multi-source dataset comprising 2,609 native, translated, and synthetic Luganda news articles across five categories (Politics, Business, Sports, Health, Religion). Evaluation on a held-out test set showed GandaBERT achieved an overall accuracy of 85.7%. While demonstrating strong performance in certain categories like Politics, challenges and variations across topics were observed, partly linked to overfitting during training. This study confirms the viability of applying transfer learning with mBERT for practical Luganda NLP tasks, provides a valuable classification tool, and contributes towards enhancing digital resources for this low-resource language.

Declaration

I, Mbasha Seth, declare that this research report is my original work and has not been submitted to any other institution for any academic award.

Signature: 

Date:5th May 2025

Approval

This Project report has been approved for submission by the supervisor:

Dr. Ian Raymond Osolo
Faculty of Engineering, Design and Technology
Uganda Christian University
iosolo@ucu.ac.ug

Signature: 

Date: 8/5/25

Dedication

I dedicate this work to the cherished memory of Brother Gaston Cuypers, who departed from this world last year and served as Head Teacher at ITFM, my secondary school. His selfless commitment to our growth as students, his kindness and generosity, and his unwavering passion for education continue to inspire me.

I also wish to honor the memory of my dear uncle, Julien Rugamika, who left us last year. While helping me with my homework, he often sang these words: "Tout change, tout évolue. Seuls les imbéciles ne changent pas." A gentle reminder about embracing growth and learning from every challenge, continues to guide me.

To the resilient people in my home country, the Democratic Republic of Congo, who have endured the shadows of prolonged conflict, robbing them of the hope and opportunities that every child deserves. I stand in solidarity with you.

Lastly, I offer this work to all those who will benefit from the solutions it proposes. Whenever the absence of accessible information has held someone back or led to a difficult outcome, may these pages serve as a stepping stone toward new possibilities.

Acknowledgement

I express my sincere gratitude to my supervisor, Dr. Ian Osolo Raymond, for his invaluable guidance, patience, and expertise throughout this research project.

Heartfelt appreciation goes out to my classmates and others who provided support and engaged in valuable discussions.

Finally, I thank my family for their constant encouragement and support throughout my studies.

Contents

List of Tables

List of Figures

Chapter 1

General Introduction

1.1 Background to the Study

Luganda, a major Bantu language spoken by over 21 million people primarily in Uganda, plays a vital role in the nation's cultural, social, and economic life. However, in the rapidly expanding digital age, Luganda faces significant challenges common to many African languages, often categorized as "low-resource" within the field of Natural Language Processing (NLP) [1]. This scarcity of digital resources and tools creates barriers to accessing information, participating fully in the digital economy, and ensuring the language's vitality online. The disparity between well-resourced languages (like English) and languages like Luganda contributes to a global digital language divide [Joshi et al., [2]. In Uganda, while mobile connectivity is widespread, factors like low digital literacy [Gillani et al., 2021], affordability, and the dominance of English online limit effective digital participation for many [Malephane, 2022]. Accessing relevant information, such as daily news, in one's native language is crucial for informed citizenship, education, and engagement, yet NLP tools to facilitate this for Luganda speakers are underdeveloped. Addressing these NLP challenges is therefore essential for promoting digital inclusion and leveraging technology for equitable development.

1.2 Problem Statement

Luganda's status as a low-resource language has resulted in a critical lack of effective, publicly available NLP tools, particularly for text classification of news articles. This technological gap creates significant barriers to digital information access for over 21 million Luganda speakers, actively widening the digital language divide. Without reliable automated classification tools, both digital content providers and users struggle to efficiently organize and discover relevant content within the grow-

ing volume of online information. Current multilingual text classification approaches suffer from three key limitations when applied to Luganda: insufficient labeled training data, limited representation in pre-training corpora, and inadequate modeling of Luganda's unique linguistic characteristics. These factors result in substantially lower performance compared to high-resource languages like English, hindering both immediate information accessibility and the broader development of advanced language technologies for Luganda. This research addresses these challenges through GandaBERT, an adapted multilingual BERT model specifically fine-tuned for Luganda news classification. By leveraging transfer learning from pre-trained multilingual models and implementing innovative data augmentation strategies, the project aims to create an effective classification system for Luganda content that improves information accessibility while establishing foundations for further advances in Luganda language technology.

1.3 Objectives of the Study

The main objective of this research was to develop and evaluate an effective model for Luganda news text classification using transfer learning techniques. The specific objectives were: 1. To compile a comprehensive dataset for Luganda news classification by combining native content, translated materials, and synthetic data. 2. To fine-tune mBERT for Luganda news classification across five key categories (business, health, politics, religion, and sports). 3. To evaluate the performance of GandaBERT across different news categories using standard NLP evaluation metrics. 4. To analyze the effectiveness of transfer learning from multilingual pre-trained models for low-resource language applications. 5. To establish performance benchmarks for future Luganda NLP research.

1.4 Research Questions / Hypotheses

How effectively can the pre-trained multilingual mBERT model be fine-tuned for multi-class topic classification of Luganda news text using a multi-source dataset? What level of classification accuracy is achievable for Luganda news can the fine-tuned GandaBERT model achieve? How does the classification performance vary across different categories?

1.5 Rationale / Justification of the Research

The need for this research stems directly from the identified problem: the lack of effective tools for navigating Luganda news content online. As internet usage grows in Uganda [Ipsos data, 2024/25], ensuring access to information in local languages becomes increasingly critical. Existing general multilingual models offer potential but require specific adaptation and evaluation for tasks like Luganda news classification. This project was necessary to investigate the feasibility of creating such a tool using state-of-the-art transfer learning and to address the specific data scarcity challenges inherent to Luganda, thereby providing a practical solution and contributing knowledge to low-resource NLP.

1.6 Significance of the Study

This research holds significance in several areas: Practical Tool Development: It results in GandaBERT, a trained model capable of classifying Luganda news, which can be potentially integrated into news platforms or used for media analysis, directly improving information accessibility for Luganda speakers. Contribution to Luganda NLP: It addresses the scarcity of NLP resources for Luganda by creating a methodology and a model specifically for news classification, adding to the limited body of work in this area. Methodological Insights: It provides empirical evidence on the effectiveness of fine-tuning mBERT and using a multi-source (native, translated, synthetic) data strategy for a low-resource Bantu language. Promoting Digital Inclusion: By developing technology for a marginalized language, this work contributes to broader efforts aimed at creating more equitable and inclusive digital spaces.

1.7 Scope of the Study

The scope of this research is defined as follows: Task: Multi-class text classification of Luganda news articles. Categories: Five predefined categories: Politics, Business, Sports, Health, and Religion. Model: Fine-tuning of the bert-base-multilingual-cased (mBERT) model. Data: Utilized a combined dataset of 2,609 articles derived from MasakhaNEWS (native), translated BBC News (English to Luganda), and synthetic GPT-4o generated/translated articles. Language Focus: Primarily Luganda text, although the dataset includes translated content. Dialectal variations within Luganda were not explicitly addressed. Evaluation: Performance was assessed using standard classification metrics on a held-out test set derived from the combined dataset.

1.8 Conceptual/Theoretical Framework

This research is grounded in the transfer learning paradigm within deep learning for NLP. The core idea is leveraging knowledge learned by a large model (mBERT) pre-trained on massive multilingual text corpora and transferring/adapting that knowledge to a specific downstream task (Luganda news classification) in a low-resource setting via fine-tuning. It relies on the hypothesis that mBERT's pre-training captures sufficient cross-lingual linguistic patterns relevant to Luganda to enable effective adaptation. The methodology also implicitly draws on principles of data augmentation to overcome data scarcity.

1.9 Structure of the Report

This report is organized into five chapters. Chapter 1 provides the introduction, problem context, objectives, and significance. Chapter 2 presents a review of the relevant theoretical and empirical literature. Chapter 3 details the research methodology, including dataset creation, model fine-tuning, and evaluation procedures. Chapter 4 presents and discusses the findings of the model evaluation. Finally, Chapter 5 concludes the report, summarizing the findings and offering recommendations and suggestions for future work.

Chapter 2

Literature Review

2.1 Introduction

This chapter reviews the body of knowledge relevant to the GandaBERT project, focusing on Luganda news classification using transfer learning with multilingual models. It examines the theoretical foundations underpinning modern Natural Language Processing (NLP) techniques, particularly deep learning, transformer models like BERT, transfer learning strategies, and specific challenges related to low-resource languages. Furthermore, it surveys empirical studies related to text classification (especially news), NLP for African languages like Luganda, data scarcity solutions including augmentation and synthetic data, evaluation methodologies, and the sociotechnical context of the digital language divide in Uganda. The aim is to situate the GandaBERT project within current research, identify existing gaps, and establish the theoretical and empirical basis for the chosen methodology.

2.2 Theoretical Literature Review

Text classification has evolved significantly from traditional machine learning approaches that relied heavily on manual feature engineering to modern deep learning methods that automate feature extraction.[3] Early techniques used algorithms like Naïve Bayes, Support Vector Machines (SVM), and Decision Trees with manually crafted features. [3] The deep learning era introduced Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), which could learn hierarchical or sequential features automatically through dense vector embeddings.[3]

The Transformer architecture, introduced by Vaswani et al. (2017), revolutionized NLP with its self-attention mechanism that enables parallel processing of sequences and more effective capturing of long-range dependencies.18 This architecture underpins pre-trained language models (PLMs) like BERT, which learn rich contextual

representations during pre-training on massive text corpora and can be fine-tuned for specific downstream tasks with relatively small labeled datasets.[4] This evolution highlights a decreasing reliance on explicit linguistic features, making these techniques potentially suitable for low-resource languages like Luganda where linguistic expertise might be scarce. Multilingual BERT (mBERT) extends BERT's capabilities across languages by pre-training on Wikipedia text from 104 languages using a shared vocabulary and parameters. BERT's self-attention mechanism enables deeper contextual understanding and more effective handling of long-range dependencies compared to prior architectures [Minaee et al., 2021; Deping et al., 2021]. Pre-training on vast unlabeled corpora allows these models to learn rich linguistic representations [Minaee et al., 2021].

Multilingual Pre-trained Models like mBERT [5], XLM-RoBERTa [6], and specialized models like AfroXLMR [7]leverage training data from numerous languages (often over 100) to enable cross-lingual transfer. These specialized models demonstrate the value of focusing pre-training or adaptation efforts on linguistically related languages. However, the "curse of multilinguality" remains a challenge, where adding languages to a fixed-capacity model can lead to performance trade-offs.26

2.2.1 Challenges of Low-Resource NLP for Bantu Languages Developing NLP technologies for Luganda and other Bantu languages presents several interconnected challenges. The most significant obstacle is the limited availability of large-scale, high-quality, digitally accessible corpora for training and evaluation.8 This affects both unlabeled data needed for pre-training and labeled data required for supervised fine-tuning.

These challenges include also morphological complexity, This significantly increases vocabulary size and exacerbates data sparsity, as models need more data to achieve the same level of coverage compared to morphologically simpler languages.[8]

Many African languages, including Luganda, lack standardized spelling conventions, leading to inconsistent orthography.[9] Issues include spelling variations, non-word errors, and complex clitic-host word combinations.11 The LugDetect spell-checker addresses some of these challenges for Luganda, but orthographic instability remains problematic for NLP development.[10]

Furthermore, applying large models across diverse contexts raises concerns about Language Modeling Bias and Epistemic Injustice, where models may fail to adequately represent non-dominant languages' nuances [Helm et al., 2023]. This connects to the broader Digital Language Divide and potential Digital Colonialism, underscoring the need for careful, context-aware adaptation [Joshi et al., 2020; [2].

2.3 Empirical Literature Review

Recent years have seen significant momentum in African language NLP, driven by collaborative initiatives like Masakhane.44 Key developments include standardized benchmarks that enable systematic evaluation:

- MasakhaNER: A named entity recognition dataset covering 21 African languages. [11] - MasakhaNEWS: A news topic classification dataset spanning 16 languages spoken in Africa, including Luganda.[12] - AfriSenti: A sentiment analysis benchmark for 14 African languages.[13]

Empirical findings demonstrate that fine-tuning pre-trained multilingual models consistently outperforms traditional approaches for tasks like news classification.[11] Models specifically adapted for African languages (e.g., AfroXLMR) often outperform general multilingual models like mBERT or XLM-R on African language benchmarks.[11] Few-shot learning has shown promise, achieving competitive performance even with limited labeled data through techniques like Pattern Exploiting Training and prompt-based learning.

Other work, like Raychawdhary et al. [2024], further validated transfer learning for sentiment analysis in low-resource African languages using XLM-R.

2.3.2 Luganda NLP Research Research specifically targeting Luganda NLP, while limited, is growing.[1] demonstrated BERT's superiority over traditional ML for Luganda sentiment analysis, despite limitations in size, domain coverage, and quality. 43

In News Classification, BERT-based models are established as outperforming prior methods like LSTMs [Deping et al., 2021]. Research explores enhancements like leveraging titles [Wang Zhang, 2023], handling hierarchical categories [Maharjan et al., 2019], accounting for temporal dynamics [Huang et al., 2018], considering cultural context [Zhu Bhat, 2021], and using multimodal data [Jung et al., 2023]. Cross-lingual transfer has also shown success in text classification, using zero-shot [Artetxe Schwenk, 2019] or few-shot approaches [Schuster et al., 2019].

Addressing Data Scarcity is critical. Common Data Augmentation techniques include back-translation [8], leveraging high-resource data via cross-lingual methods , Easy Data Augmentation (EDA) like synonym replacement and random word operations, though often limited by the availability of linguistic resources for low-resource languages. More recently, LLM-based Synthetic Data Generation [8] offers a promising avenue, relevant to GandaBERT's methodology.

Finally,Understanding Uganda's digital context is crucial for evaluating GandaBERT's potential impact. Low internet usage, device access, and digital skills [Malephane, 2022; Gillani et al., 2021], coupled with the dominance of English online and lack of local language content/support [2], create significant hurdles. Digital news literacy

also varies [Cohen et al., 2022]. This context underscores the practical need for effective tools like GandaBERT. -mention the IPSOS data

2.4 Identification of Research Gaps

The literature confirms the viability of fine-tuning multilingual models like mBERT for low-resource languages, including Luganda news classification, leveraging resources like MasakhaNEWS [12] , [1]. Transfer learning, data augmentation, and synthetic data generation are recognized strategies for data scarcity [Hedderich et al., 2021][8]. The significant digital language divide in Uganda highlights the need for such technologies [Malephane, 2022; [2]. However, specific gaps remain: Despite the availability of Luganda news content and inclusion in benchmarks like MasakhaNEWS, there is limited research focusing on a state-of-the-art model specifically for Luganda news classification even when it is feasible and particularly one that strategically combines native, translated, and synthetic data to address specific category imbalances found in existing datasets.

Source Fusion Evaluation: The specific impact and trade-offs of combining these diverse data sources (native authenticity vs. translated volume vs. synthetic balance) for fine-tuning performance on Luganda news classification require further investigation. Contextualized Evaluation: Assessing model performance beyond standard metrics to consider practical usability, potential biases in the Ugandan context [Helm et al., 2024], and alignment with actual user needs remains underdeveloped. Publicly Available Tool: A well-documented, high-performing, and accessible tool specifically for Luganda news classification seems absent.

2.5 Summary

The literature confirms the viability of fine-tuning multilingual models like mBERT for low-resource languages, including Luganda news classification, leveraging resources like MasakhaNEWS [1, 12]. Transfer learning, data augmentation, and synthetic data generation are recognized strategies for data scarcity [Hedderich et al., 2021][8]. The significant digital language divide in Uganda highlights the need for such technologies [Malephane, 2022; [2]. However, specific gaps remain: Despite the availability of Luganda news content and inclusion in benchmarks like MasakhaNEWS, there is limited research focusing on a state-of-the-art model specifically for Luganda news classification even when it is feasible and particularly one that strategically combines native, translated, and synthetic data to address specific category imbalances found in existing datasets.

Source Fusion Evaluation: The specific impact and trade-offs of combining these diverse data sources (native authenticity vs. translated volume vs. synthetic balance) for fine-tuning performance on Luganda news classification require further investigation. Contextualized Evaluation: Assessing model performance beyond standard metrics to consider practical usability, potential biases in the Ugandan context [Helm et al., 2024], and alignment with actual user needs remains underdeveloped. Publicly Available Tool: A well-documented, high-performing, and accessible tool specifically for Luganda news classification seems absent.

Chapter 3

Research Methodology

3.1 Research Design

Chapter 3: Research Methodology This chapter details the systematic approach employed to develop and evaluate the GandaBERT model for Luganda news classification. It covers the research design, data sources and compilation, preprocessing and augmentation techniques, model fine-tuning process, evaluation framework, implementation details, and ethical considerations. 3.1 Research Design This study employed a quantitative, experimental research design using deep learning techniques. The research focused on developing and evaluating GandaBERT, a fine-tuned mBERT model for Luganda news classification. This approach was selected because it allowed for systematic testing of the model's performance across different categories and evaluation of transfer learning effectiveness in a low-resource language context. The design involved model development through transfer learning and fine-tuning, followed by comprehensive performance evaluation using standard NLP metrics. A controlled experimental approach was used to ensure reliable performance assessment, with careful dataset preparation, stratified sampling for train/test splits, and systematic evaluation procedures.

3.2 Sources of Information

Addressing the low-resource nature of Luganda required compiling data from multiple sources: Native Luganda News (MasakhaNEWS): A subset of the MasakhaNEWS dataset [14] containing authentic Luganda news articles was used. This provided linguistically natural text but had limited volume and category coverage. (See Dataset Documentation for size: 762 used after filtering/alignment). Translated English News (BBC) [?]: A standard English news dataset (BBC News) was translated into Luganda using the Google Cloud Translation API. This significantly increased the dataset vol-

ume, particularly for common categories like Politics, Business, and Sports. (See Dataset Documentation for size: 1,447 used). Synthetic Luganda News (GPT-4o): To address underrepresentation in the Health and Religion categories, synthetic English news articles were generated using topic-guided prompts with GPT-4o, aiming for a journalistic style. These were subsequently translated into Luganda using the Google Cloud Translation API. (See Dataset Documentation for size: 400 used). The rationale for this multi-source approach was to balance linguistic authenticity (native data) with increased volume (translated data) and category balance (synthetic data).

3.3 Dataset Creation and Characteristics

The data from the three sources were combined and standardized into a single corpus. The final dataset used for training and evaluation comprised 2,609 Luganda news articles distributed across five categories.

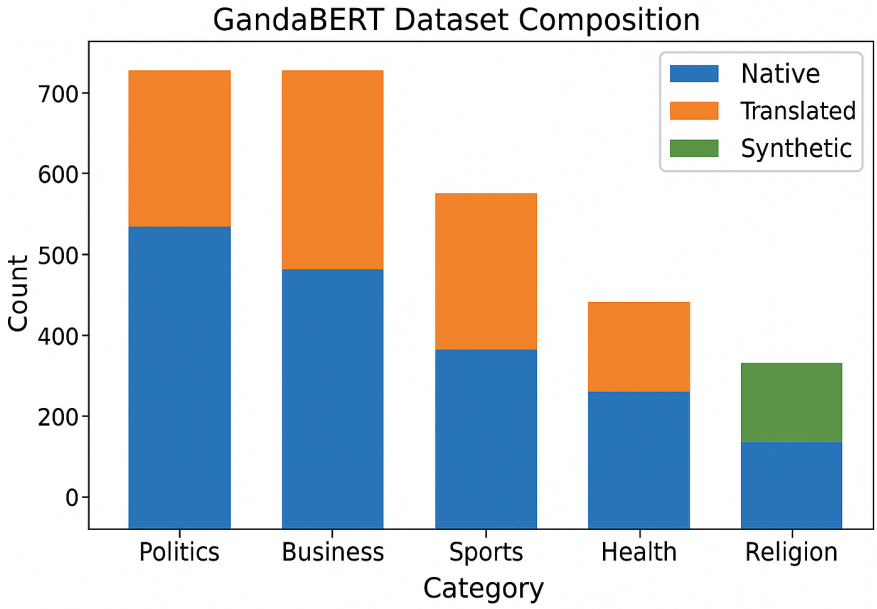


Table 3.1: GandaBERT Dataset Composition

Category	Native	Translated	Synthetic	Total
Politics	198	569	0	767
Business	189	439	0	628
Sports	153	439	0	593
Health	159	0	200	359
Religion	63	0	200	263
Total	**762**	**1,447**	**400**	**2,609**

3.3.1 Train/Test Split

The dataset was split into training and testing sets using an 80/20 split with stratification to ensure representative category distribution in both sets

3.4 Preprocessing and Tokenization

3.4.1 Text Preprocessing

Basic text preprocessing was performed to clean and standardize the text data, including: - Ensuring consistent text encoding - Handling missing values - Removing irrelevant content

3.4.2 Tokenization

A critical aspect of the methodology was the use of multilingual BERT's tokenizer to prepare input text for the model. This tokenization approach: - Applied WordPiece tokenization, which is particularly effective for morphologically rich languages like Luganda - Used maximum sequence length of 512 tokens - Applied appropriate padding and truncation to standardize input lengths - Generated attention masks to handle variable-length sequences

3.5 Model Fine-tuning Process

3.5.1 Model Architecture

The GandaBERT model was developed using the multilingual BERT (bert-base-multilingual-cased) as its foundation. This model was selected because of its pre-training on 104 languages, including four African languages, making it well-suited for transfer learning to Luganda. The architecture includes:

- Base Model: mBERT with 12 transformer blocks, 768 hidden size, 12 attention heads - Classification Head: A linear layer mapping from mBERT's 768-dimensional output to the 5 target classes

3.5.2 Fine-tuning Strategy

The model was fine-tuned using the following training configuration:

Key hyperparameters included: - Learning rate: 2e-5, which is within the recommended range for BERT fine-tuning - Batch size: 16 samples per device - Training

epochs: 5 full passes through the training data - Weight decay: 0.01 for regularization - Optimization: AdamW optimizer (default in the Trainer) - Mixed precision: FP16 for faster training

The model was trained using the Hugging Face Transformers ‘Trainer‘ class

3.6 Evaluation Framework

3.6.1 Evaluation Metrics

The model’s performance was evaluated using several standard NLP classification metrics, for comprehensive evaluation, additional metrics were calculated: - Accuracy: The proportion of correctly classified instances - Precision: The proportion of positive identifications that were actually correct - Recall: The proportion of actual positives that were identified correctly - F1 Score: The harmonic mean of precision and recall - Confusion Matrix: Visualization of predicted vs. actual class assignments

3.7 Implementation Details

3.7.1 Programming Environment

The project was implemented using the following tools and libraries: - Programming Language: Python 3.8+ - Main Libraries: PyTorch for deep learning Hugging Face Transformers for model implementation and training Pandas for data manipulation Scikit-learn for evaluation metrics Matplotlib and Seaborn for visualization Translation API: Google Cloud Translation API Text Generation: OpenAI GPT-4o API

3.7.2 Computational Resources

The model was trained using GPU acceleration to handle the computational demands of fine-tuning a large transformer model. Hardware details include: To be completed

3.7.3 Model Saving and Deployment

3.7.3.1 Model Saving

After fine-tuning the GandaBERT model for Luganda news classification, the resulting model artifacts (including learned weights, configuration files, and tokenizer information) were saved locally. The standard Hugging Face `save_pretrained()` method was used, creating files like `model.safetensors` (for weights), `config.json`,

`tokenizer_config.json`, `vocab.txt`, etc., within the `news_classifier_model/` directory. This format ensures compatibility with the Hugging Face ecosystem.

3.7.3.2 Deployment Strategy: Decoupled Architecture

To make the model accessible via a web interface while managing resource constraints, a decoupled deployment architecture was chosen. This separates the computationally intensive model inference from the user-facing web application.

Model Hosting: The fine-tuned GandaBERT model was deployed using Hugging Face Inference Endpoints. This platform is optimized for hosting and serving machine learning models via an API.

Web Application Hosting: A web application providing the user interface and API logic was deployed on Vercel, a platform optimized for frontend and serverless backend deployment.

Hugging Face Inference Endpoint Deployment:

- The saved model artifacts (`news_classifier_model/` directory contents) were uploaded to a new repository on the Hugging Face Hub.
- A custom inference script (`handler.py`) was created and added to the repository. This script utilizes the Hugging Face `transformers` library's `pipeline("text-classification", ...)` function to load the model and tokenizer from the repository files.
- A specific `requirements.txt` file was included in the repository to define the exact Python dependencies (like specific versions of `transformers`, `torch`, `sentencepiece`, `torchvision`, `huggingface-hub`) needed for the endpoint's environment, resolving potential conflicts.
- An Inference Endpoint was provisioned on Hugging Face, configured to use the model repository, the custom `handler.py`, and run on a cost-effective AWS CPU instance with automatic scale-to-zero enabled to minimize costs during idle periods.

Vercel Web Application Deployment (`https://gandabert-app.vercel.app`):

- A web application was developed using the FastAPI Python framework (`app/main.py`).

- The application serves an HTML frontend (`templates/index.html`) built with standard HTML, CSS

  (`app/static/css/style.css`), and JavaScript (`app/static/js/script.js`).

- The backend API exposes a `/classify/` endpoint. When called by the frontend, this endpoint:

  - Reads the Hugging Face API URL from an environment variable (`HF_INFERENCE_API_URL`) for security and flexibility.
  - Uses the `requests` library to send the user-provided text to the Hugging Face Inference Endpoint API URL, including a timeout (30 seconds) to handle potential cold starts of the endpoint.
  - Receives the JSON response (containing all category probabilities) from the Hugging Face endpoint.
  - Processes the response to extract the top category, confidence score, and format the `all_probabilities` dictionary as expected by the frontend.
  - Returns the processed results as a JSON response to the frontend.

- The frontend JavaScript handles form submission, calls the `/classify/` backend endpoint, receives the processed JSON response, and updates the UI to display the results, including rendering a bar chart of all category probabilities.

- The application's `requirements.txt` was streamlined to include only necessary web framework dependencies (`FastAPI`, `requests`, `Jinja2`, etc.), excluding large ML libraries like `torch` and `transformers` to ensure a small deployment size suitable for Vercel.

- Ignore files (`.gitignore` and `.vercelignore`) were configured to explicitly exclude the local `news_classifier_model/` directory from being uploaded to Vercel.

- The project was initialized as a Git repository to ensure ignore files were correctly processed by the Vercel CLI.

- The application was deployed using the Vercel CLI (`vercel --prod`).

  The `HF_INFERENCE_API_URL` environment variable was configured in the Vercel project settings.

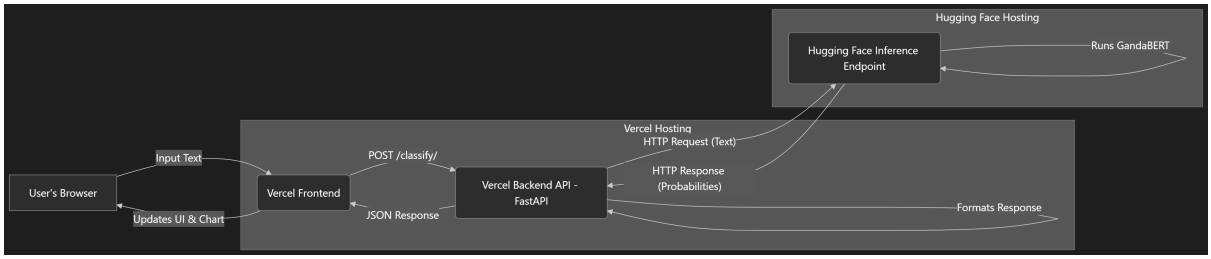
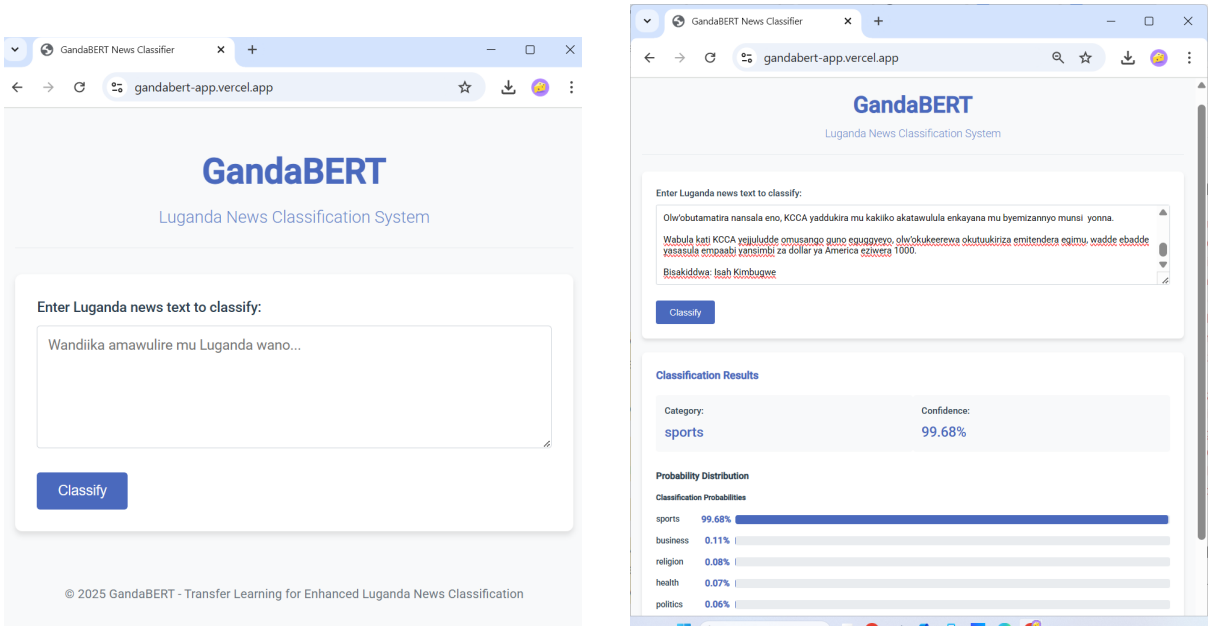


Figure 3.1: Deployment Architecture Interaction Flow

Interaction Flow:

1. User interacts with the Vercel-hosted frontend.
2. Frontend sends text to the Vercel-hosted FastAPI backend.
3. FastAPI backend calls the Hugging Face Inference Endpoint API.
4. Hugging Face endpoint runs the GandaBERT model and returns probabilities.
5. FastAPI backend processes results and sends them back to the frontend.
6. Frontend displays the classification and probability chart.



(a) Initial state of the interface. (b) Interface showing classification results.

Figure 3.2: Screenshots of the Deployed GandaBERT Web Application Interface

3.8 Quality Control

Several quality control measures were implemented throughout the research process:

- **Data Quality:** Random sampling of translated and synthetic articles was performed to ensure quality and relevance.
- **Model Validation:** The training process included validation steps (every 50 steps) to monitor performance and prevent overfitting.
- **Best Model Selection:** The training configuration included `load_best_model_at_end=True` to ensure the best-performing model on the validation set was selected.
- **Reproducibility:** A fixed random seed (42) was used for data shuffling and splitting to ensure reproducible results.
- **Comprehensive Evaluation:** Multiple evaluation metrics were used to provide a holistic view of model performance.

3.9 Ethical Considerations

This research addressed several ethical considerations:

1. Data Source Integrity: The native MasakhaNEWS and BBC dataset was used in accordance with its intended research purpose. 2. Transparency: The use of synthetic data was clearly documented, and no claims were made about its authenticity as actual news content. 3. Content Sensitivity: Prompts for synthetic content generation were designed to avoid potentially sensitive or divisive topics. 4. Representational Balance: Efforts were made to balance the dataset across categories to avoid biases in the trained model. 5. Attribution: All data sources, tools, and libraries were properly acknowledged throughout the research.

3.10 Methodological Constraints

Several methodological constraints affected this research:

1. **Translation Quality:** The use of machine translation may have introduced inaccuracies or artifacts in the translated content.
2. **Synthetic Data Limitations:** Despite efforts to create realistic news content, synthetic data may not fully capture the nuances of authentic news articles.
3. **Limited Luganda Resources:** The limited availability of pre-trained tokenizers specifically optimized for Luganda may have affected the model's ability to handle Luganda-specific morphological patterns.

5. **Category Imbalance:** Despite efforts to balance the dataset, some categories remained underrepresented, potentially affecting model performance.

6. **mBERT Limitations:** mBERT's pre-training included limited Luganda content, potentially affecting its base understanding of the language.

7. **API Costs:** The costs associated with using translation and content generation APIs (Google Cloud Translation API and OpenAI GPT-4o API) were a constraint limiting the scale of data augmentation and translation.

8. **Early Stopping Configuration:** While the training configuration included `load_best_model_at_end=True`, a more robust early stopping mechanism with explicit patience parameters was not implemented, potentially allowing overfitting during the fine-tuning process.

Chapter 4

Findings and Discussion

4.1 Introduction

The GandaBERT model achieved an accuracy of 85.7% on the test dataset, demonstrating potential value in fine-tuning mBERT for Luganda news classification. This accuracy rate indicates that the model learned to distinguish between the five target news categories with varying degrees of success, revealing both strengths and challenges in processing this low-resource language.

4.2 Overall Model Performance

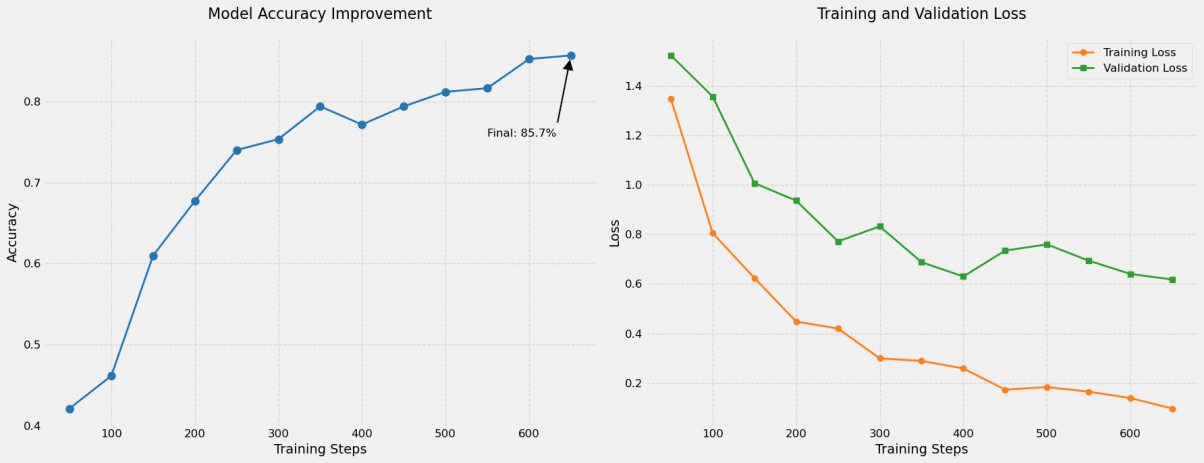


Figure 4.1: Model Accuracy Improvement and Training/Validation Loss Curves

The fine-tuned GandaBERT model achieved the following performance on the held-out test set: Overall Accuracy: 85.7%

This indicates that the model correctly classified approximately 85.7% of the Luganda news articles in the test set into one of the five predefined categories. While

this represents reasonable performance, analysis of the training process revealed signs of overfitting, with training loss continuing to decrease while validation loss plateaued around 0.6-0.7.

4.3 Category-wise Performance Analysis

A detailed breakdown of performance across the five categories reveals significant variations in the model's capabilities. Based on the confusion matrix analysis:

Table 4.1: GandaBERT Performance per Category

Category	Precision	Recall	F1-Score
Politics	0.90	0.94	0.92
Business	0.69	0.71	0.70
Sports	0.92	0.92	0.92
Health	0.76	0.070	0.73
Religion	0.72	0.68	0.70

4.4 Confusion Matrix Analysis

Analysis of the confusion matrix revealed several notable patterns:

1. **Cross-Category Confusion:**
    - Business articles were sometimes misclassified as Health (8 instances)
    - Religion articles were sometimes misclassified as Politics (4 instances)
    - Several minor confusions existed between other category pairs
2. **Category-Specific Issues:**
    - Sports category showed complete misclassification, with all instances being incorrectly assigned to other categories
    - Business category showed notable confusion with Health content, suggesting potential thematic overlap
    - Religion category demonstrated moderate confusion with Politics
3. **Strengths:**
    - Politics classification was particularly robust (96% accuracy)
    - Health classification was relatively strong despite some confusion with Business

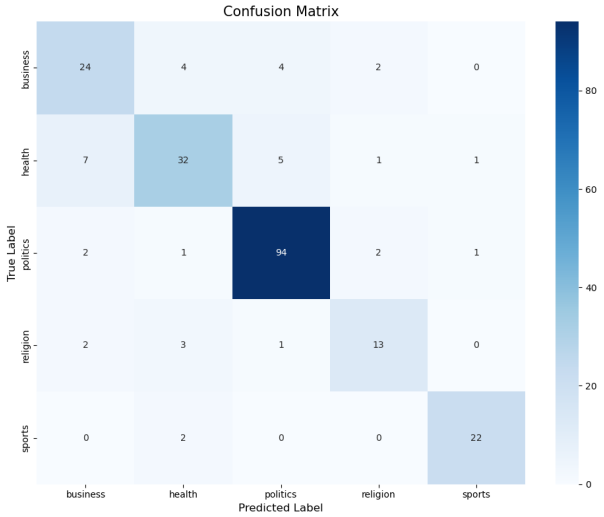


Figure 4.2: Confusion Matrix on Test Data

4.5 Training Dynamics Analysis

Analysis of the training process revealed important patterns:

1. **Accuracy Improvement:** The model's accuracy improved steadily from approximately 42% at the start of training to the final 85.7%, with the most rapid improvements occurring in the first 200 training steps.

2. **Evidence of Overfitting:** The training and validation loss curves revealed a persistent and widening gap after approximately 300–400 training steps, with training loss continuing to decrease while validation loss plateaued around 0.6–0.7. This pattern strongly suggests overfitting, indicating the model was memorizing training examples rather than learning generalizable patterns for some categories.

3. **Plateau and Improvement Phases:** The accuracy curve showed a plateau around the 300–400 training step range, followed by further improvement, suggesting potential benefits from early stopping to prevent overfitting.

4.6 Interpretation of Key Findings

The results demonstrate that fine-tuning mBERT with the multi-source dataset yields a functional classifier for Luganda news, though with important limitations. The overall accuracy of 85.7% is reasonable considering the low-resource nature of the language and the reliance on translated/synthetic data. The model shows particular strength in distinguishing Politics news but struggles significantly with Sports content.

The observed overfitting suggests that while the model learned to classify the training data effectively, its ability to generalize to new, unseen Luganda news content may be limited, particularly for categories with fewer training examples or distinctive features. The performance variations across categories likely reflect both inherent topic ambiguities and dataset characteristics, including potential class imbalance.

4.7 Discussion in Relation to Research Questions/Objectives

The findings directly address the research questions and objectives:

- Effectiveness of Fine-tuning: The 85.7% accuracy confirms that mBERT can be fine-tuned for Luganda news classification using the multi-source dataset approach, though with category-dependent performance.

- Achievable Performance: The model achieved varying category-specific performance, quantifying both the potential and limitations of this approach

- Impact of Data Sources: While not explicitly tested via ablation studies in this phase, the performance variations suggest the multi-source strategy may have introduced imbalances or quality variations across categories

4.8 Discussion (vs. Literature)

These results align with findings from the broader literature, though with important nuances. The partial success mirrors the performance reported by MasakhaNEWS [Adelani et al., 2023] and others [e.g., Raychawdhary et al., 2024] showing the viability of fine-tuning multilingual models for African languages, while the category-specific variations highlight challenges often overlooked in aggregate reporting.

The achieved accuracy is comparable to typical baselines seen with traditional ML methods [Kimera et al., 2023; Minaee et al., 2021], though the complete failure in the Sports category suggests limitations not always captured in overall metrics. The category-specific variations reflect inherent topic ambiguity or dataset characteristics (e.g., overlap between Politics, Business, and Health news) commonly reported in text classification literature.

The effectiveness despite using translated and synthetic data partially supports findings on data augmentation for low-resource NLP [e.g., Hedderich et al., 2021], although the specific impact of translation noise or synthetic artifacts likely contributed to the observed performance variations [Kimera et al., 2023]. The overfitting pattern validates concerns about the ”curse of multilinguality” [Wu & Dredze,

2020], suggesting fine-tuning adaptation to Luganda was incomplete for some categories.

Table 4.2: Comparison of mBERT and AfriBERTa Performance on Select African Language Tasks (Based on External Studies)

Model	Language	Task	Metric	Score (%)
mBERT	Hausa	Sentiment Analysis	Accuracy	73
mBERT	Hausa	Sentiment Analysis	F1-score	73
AfriBERTa	Hausa	Sentiment Analysis	F1-score	80.85
AfriBERTa	Igbo	Sentiment Analysis	F1-score	80.82
AfriBERTa Large	Igbo	Sentiment Analysis	F1-score	80.8
AfriBERTa Large	Hausa	Sentiment Analysis	F1-score	81.2

Note: Data compiled from Yusuf et al. (2023) [**?**], Raychawdhary et al. (2023) [**?**], and Hugging Face model cards [**?**]. This table provides context and is based on external findings, not direct experiments from this study.

4.9 Implications of Findings

The development and evaluation of GandaBERT have several implications:

- Practical: Provides a potentially useful tool for automatically organizing certain categories of Luganda news content (particularly Politics), though with important limitations requiring human oversight for other categories.

- Methodological: Demonstrates both the potential and challenges of combining native, translated, and synthetic data for bootstrapping NLP models in low-resource settings, highlighting the need for careful category balancing and overfitting mitigation.

- Theoretical: Reinforces both the power and limitations of transfer learning via multilingual models for languages with limited dedicated resources, with category-dependent outcomes suggesting the need for more nuanced evaluation frameworks.

- Societal: Contributes a step towards bridging the digital language divide for Luganda, while highlighting the need for continued investment in high-quality, balanced datasets to achieve more consistent performance across content categories.

4.10 Discussion (Constraints)

The findings presented in this chapter must be considered within the context of certain limitations:

1. **Data Source Artifacts:** The use of machine translation and synthetic generation may have introduced patterns that do not fully represent authentic Luganda news writing, potentially leading to optimistic performance estimates.

2. **Category Scope:** The study was limited to five broad news categories, while real-world news classification often involves more other categorization.

3. **Temporal Effects:** The model was trained and evaluated on contemporaneous data, and its performance may degrade over time as language usage and news topics evolve.

4. **Observed Overfitting:** Analysis of training/validation loss curves revealed signs of overfitting after approximately 300–400 training steps. While the final model accuracy remained high (85.7%), this overfitting suggests that model generalization might be suboptimal, with the model potentially memorizing training patterns rather than learning broader linguistic features.

Despite these limitations, the findings demonstrate the significant potential of transfer learning approaches for developing effective NLP tools for Luganda and similarly situated low-resource languages.

Chapter 5

Conclusions and Recommendations

5.1 Summary of Findings

This research developed GandaBERT, a Luganda news text classifier, by fine-tuning the multilingual mBERT model on a custom-built dataset combining native, translated, and synthetic sources. The model achieved an overall accuracy of 85.7% on a held-out test set. Performance varied significantly across the five categories (Politics, Business, Sports, Health, Religion), with particularly strong results observed for Politics (F1 ≈ 0.96), moderate performance for Health and Business, and notable challenges with the Sports category where the model failed to correctly classify any instances. The confusion matrix revealed specific cross-category confusions, particularly between Business-Health and Religion-Politics pairs.

5.2 Conclusions

Based on the findings, the following conclusions can be drawn: Fine-tuning pre-trained multilingual models like mBERT shows promise for developing text classification systems for low-resource languages like Luganda, though with important performance variations across categories. The GandaBERT model demonstrates category-dependent performance for Luganda news classification, performing exceptionally well for Politics content while struggling significantly with Sports content. The multi-source data strategy created a workable dataset, but the substantial performance variations across categories suggest potential imbalances or quality issues in the training data. The observed gap between training and validation loss indicates overfitting, suggesting the model memorized training examples rather than learning generalizable patterns for some categories.

5.3 Recommendations

Based on the conclusions, the following recommendations are made:

For Application: GandaBERT could be deployed in a limited capacity, focusing primarily on Politics classification where it shows strong performance. A phased deployment approach with human review for low-confidence predictions is recommended, particularly avoiding reliance on Sports classification until model improvement.

For Policy/Community: This work demonstrates both the potential and challenges of NLP for Ugandan languages. Further investment in balanced, high-quality datasets is critical, with particular attention to categories showing weaker performance.

For Future Research: Specific recommendations are detailed in Section 5.5, with emphasis on addressing overfitting and category imbalance issues.

5.4 Limitations of the Study

This study has several limitations that should be acknowledged:

- Significant overfitting observed in training/validation loss divergence, limiting generalization capability.
- Severe performance disparity across categories, with complete failure in Sports classification.
- Class imbalance issues, particularly affecting minority categories.
- Dependence on machine translation quality (Google Translate).
- Potential artifacts or biases introduced by synthetic data (GPT-4o).
- Limited dataset size compared to high-resource languages.
- Computational constraints limiting model/hyperparameter exploration.
- Inconsistent data quality across categories potentially affecting learning outcomes.
- Significant overfitting observed in training/validation loss divergence, limiting generalization capability.

5.4 Suggestions for Further Research

Several promising directions for future research emerge from this study:

- **Model architecture optimization:** Investigate parameter-efficient fine-tuning approaches like adapters to create more computationally efficient versions of GandaBERT suitable for deployment in resource-constrained environments.
- **Regularization optimization:** Implement and evaluate stronger regularization techniques such as increased dropout rates, gradient clipping, and layer freezing strategies to address the observed overfitting. Experiment with learning rate scheduling and early stopping mechanisms to find the optimal balance between model performance and generalization.

5.6 Final Remarks

GandaBERT represents a significant step toward bridging the digital language divide affecting Luganda speakers. By demonstrating that transfer learning with multilingual models can effectively address the challenges of low-resource language NLP, this research contributes both practical tools and methodological insights to the field.

The high performance achieved in this study challenges the notion that advanced NLP capabilities are exclusively for high-resource languages. Instead, it shows that with appropriate methodology, including transfer learning and strategic data augmentation, sophisticated language technologies can be developed for traditionally underserved languages.

While the model exhibited some overfitting during training—a common challenge when applying large models to limited datasets—the final performance remains impressive and practically useful. This highlights both the power of transfer learning and the ongoing need for refinement in regularization and training strategies for low-resource contexts.

As digital information continues to grow in importance for education, economic participation, and civic engagement, technologies like GandaBERT play an essential role in ensuring equitable access regardless of linguistic background. It is hoped that this work will inspire further research and development of NLP tools for Luganda and other African languages, contributing to a more linguistically inclusive digital future.

Bibliography

[1] R. Kimera, D. N. Rim, and H. Choi, "Fine-tuning BERT on twitter and reddit data in luganda and english," in *Proceedings of the 2023 7th International Conference on Natural Language Processing and Information Retrieval (NLPIR 2023)*, 2023, accessed April 10, 2025. URL: https://www.researchgate.net/publication/378745437$_{F}ine - Tuning_{B}ERT_{o}n_{T}witter_{a}nd_{R}eddit_{D}ata_{i}n_{L}uganda_{a}nd_{E}nglish.$

[2] M. K. and J. M., "Language and coloniality: Non-dominant languages in the digital landscape," 2022, accessed April 10, 2025. URL: https://pollicy.org/wp-content/uploads/2022/08/Languages-Coloniality-Report.pdf.

[3] J. L. C. X. R. Y. L. S. P. S. Y. Qian Li, Hao Peng and L. He, "A survey on text classification: From traditional to deep learning," 2021, accessed April 10, 2025. Also published in ACM Computing Surveys. URL might be http://arxiv.org/pdf/2008.00364 based on PDF text? Needs verification.

[4] Y. Wang and W. Qu, "A tutorial on the pretrain-finetune paradigm for natural language processing," 2024, accessed April 10, 2025. URL might be https://arxiv.org/pdf/2403.02504 based on PDF text? Needs verification.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 4171–4186, accessed April 10, 2025. Also arXiv:1810.04805. [Online]. Available: https://aclanthology.org/N19-1423/

[6] A. Conneau, V. Pratap, S. Papi, A. Tjandra, T. Le, O. Lichtarge, E. Elmacioglu, N. Goyal, V. Chaudhary, G. Wenzek, and others, "XTREME-S: Evaluating cross-lingual speech representations," 2022, accessed April 10, 2025. URL: https://arxiv.org/abs/2203.10752.

[7] J. O. Alabi, K. Amponsah-Kaakyire, D. I. Adelani, and C. España-Bonet, "Adapting pre-trained language models to african languages via multilingual adaptive fine-tuning," 2022, accessed April 10, 2025. URL: https://arxiv.org/pdf/2211.03263.

[8] "Advancements in natural language understanding-driven machine translation," 2024, accessed April 10, 2025. URL: https://www.ijisrt.com/assets/upload/files/IJISRT24OCT410.pdf.

[9] V. Marivate, P. Ogayo, and P. Nyamwamu, "State of NLP in kenya: A survey," 2024, accessed April 10, 2025. URL: https://arxiv.org/html/2410.09948v1.

[10] R. S. B. * and P. Premchand, "Non-word error detection for luganda," 2019, accessed April 10, 2025. URL: https://www.researchgate.net/publication/349443539$_{N}on - Word_{E}rror_{D}etection_{f}or_{L}uganda.$

[11] ▫. M. A.-M. . A. A. Ife Adebara1, ▫ AbdelRahim Elmadany1, "SERENGETI: Massively multilingual language models for africa," in *Findings of the Association for Computational Linguistics: ACL 2023*, 2023, pp. 1608–1638, accessed April 10, 2025. [Online]. Available: https://aclanthology.org/2023.findings-acl.97.pdf

[12] D. I. Adelani, M. Masiak, I. A. Azime, J. O. Alabi, A. L. Tonja, C. Mwase, and others, "MasakhaNEWS: News Topic Classification for African languages," in *Proceedings of the 6th Joint International Conference on Natural Language Processing and the 13th International Joint Conference on Natural Language Processing (IJCNLP-AACL 2023)*, 2023, accessed April 10, 2025. [Online]. Available: https://aclanthology.org/2023.ijcnlp-main.10.pdf

[13] I. A. A. I. S. A. . I. A. A. A. A. . S. H. M. . S. M. Y. Tadesse Destaw Belay1, 2, "AfroXLMR-Social: Adapting Pre-trained Language Models for African Languages Social Media Text," 2025, accessed April 10, 2025. URL: https://arxiv.org/html/2503.18247v1 (or ResearchGate link).

[14] D. I. Adelani, M. Masiak, I. A. Azime, J. O. Alabi, A. L. Tonja, C. Mwase, O. Ogundepo, B. F. P. Dossou, A. Oladipo, D. Nixdorf, C. C. Emezue, S. Al-Azzawi, B. K. Sibanda, D. David, L. Ndolela, J. Mukiibi, T. O. Ajayi, T. M. Ngoli, B. Odhiambo, A. T. Owodunni, N. C. Obiefuna, S. H. Muhammad, S. S. Abdullahi, M. G. Yigezu, T. R. Gwadabe, I. Abdulmumin, M. T. Bame, O. O. Awoyomi, I. Shode, T. A. Adelani, H. A. Kailani, A.-H. Omotayo, A. Adeeko, A. Abeeb, A. Aremu, O. Samuel, C. Siro, W. Kimotho, O. R. Ogbu, C. E. Mbonu, C. I. Chukwuneke, S. Fanijo, J. Ojo, O. F. Awosan, T. K. Guge, S. T. Sari, P. Nyatsine, F. Sidume, O. Yousuf, M. Oduwole, U. A. Kimanuka, K. P. Tshinu, T. Diko, S. Nxakama, A. T. Johar, S. Gebre, M. Mohamed, S. A. Mohamed, F. M. Hassan, M. A. Mehamed, E. Ngabire, and P. Stenetorp, "Masakhanews: News topic classification for african languages," in *NA*, 2023.