|  |
| --- |
| **Detecting Machine Generated text in African Languages** |
|  |
| **COS760 Project Report**  **Shiza Butt, Vutomi Mohube, Lesedi Ntsele**  **June 2025** |
|  |
|  |
|  |
|  |

1. Abstract

Large Language Models (LLMs) have revolutionized text generation, profoundly influencing everyday content creation. However, their widespread use raises critical authenticity concerns, particularly in sensitive areas like news reporting and education. While considerable efforts exist for detecting machine-generated texts in high-resource languages, low-resource African languages remain largely neglected. Addressing this gap, our project evaluates modern multilingual models for detecting machine-generated content in four low-resource African languages: Northern Sotho, Xhosa, Venda, and Tsonga. Utilizing a balanced dataset containing human-written and AI-generated texts, this study trains classifiers employing multilingual embeddings (XLM-RoBERTa, AfroLM) and interpretable machine learning methods. Through rigorous evaluation, including 5-fold cross-validation, accuracy, precision, recall, and F1-score assessments, our research identifies linguistic patterns that differentiate human-authored texts from machine-generated ones. This work significantly advances natural language processing for African languages, promotes fairness in AI technologies, and contributes to building trustworthy digital communications system.

1. **Problem Statement**

This project investigates the detection of machine-generated text in low-resource African languages, focusing on the following research questions:

How accurately can multilingual models detect machine-generated text in low-resource African languages?

What linguistic features distinguish machine-generated and human-generated text in these languages?

Can pre-trained models like AfriBERTa and XLM-RoBERTa be effectively fine-tuned for this task?

While large language models are capable of generating fluent text, reliable detection methods for identifying AI-generated content remain largely underdeveloped for African languages due to limited data availability and linguistic diversity. This project aims to evaluate modern multilingual models on this detection task, while also analysing the specific linguistic patterns exploited by these models and assessing the challenges posed by synthetic data generation.

1. Introduction

The advent and proliferation of LLMs have profoundly reshaped content creation, enabling sophisticated machine-generated texts that closely resemble human-written communication. Despite their benefits, these advancements pose significant challenges in contexts where authenticity and trustworthiness are paramount, including news, education, and public communication. The rapid integration of machine-generated texts has necessitated reliable methods to distinguish human-generated content from AI-generated content, sparking considerable research efforts primarily concentrated on high-resource languages like English, Chinese, and Spanish Ippolito et al., 2020

However, low-resource African languages continue to face major data and resource limitations, limiting the effectiveness of AI technologies across diverse linguistic contexts and deepening existing digital divides (Nekoto et al., 2020). Consequently, there is a critical need to develop specialized methodologies for detecting machine-generated texts tailored explicitly to African languages such as Northern Sotho, Xhosa, Venda, and Tsonga, which this study seeks to address. This project aims to bridge the existing gap by assessing the capabilities of contemporary multilingual models in distinguishing AI-generated from human-written texts in these languages, promoting fairness and linguistic inclusivity in NLP research.

1. **Background/Literature survey**

A substantial body of literature explores methods to detect AI-generated content.

Ippolito et al. (2020) underscore the complexity involved in reliably detecting machine-generated texts, highlighting that AI-generated content becomes harder to distinguish as models improve and generate texts increasingly indistinguishable from human-authored materials. They demonstrate that effective detection depends heavily on subtle linguistic and stylistic patterns, which sophisticated language models can effectively mimic. Moreover, they argue for the need to continuously update detection methodologies as generative models evolve, underscoring the importance of linguistic diversity in developing generalized detection systems.

Nekoto et al. (2020) emphasize the significant gaps in NLP resources for African languages, highlighting the challenges posed by limited datasets, insufficient computational resources, and inadequate attention from the broader NLP community. They advocate for participatory and collaborative methods that integrate local linguistic expertise, an approach particularly relevant when dealing with low-resource languages such as Northern Sotho, Xhosa, Venda, and Tsonga.

In response to these challenges, multilingual models such as XLM-RoBERTa (Conneau et al., 2020) and AfroLM (Dossou & Emezue, 2021) have emerged as promising solutions. XLM-RoBERTa, trained on extensive multilingual corpora, provides robust cross-lingual transfer capabilities that facilitate effective performance even when training data for a specific language is limited. AfroLM specifically targets African languages, leveraging targeted pretraining strategies to enhance performance in low-resource scenarios. Both models have shown promise for NLP tasks, including text classification, named entity recognition, and sentiment analysis, thus forming a foundational basis for detecting machine-generated content.

Additionally, the issue of interpretability in detecting machine-generated texts has gained prominence. Explainability methods such as SHAP (Lundberg & Lee, 2017) and LIME (Ribeiro et al., 2016) provide crucial insights into how models differentiate human-generated from AI-generated text, improving trust and transparency in AI systems. Such explainability is particularly critical when deploying technologies in linguistically and culturally diverse contexts, such as those found across the African continent.

Despite these advancements, a critical gap remains in explicitly applying and evaluating these methods on African languages. This project's exploration aims to address precisely this gap, providing an essential contribution to the existing literature by tailoring detection methodologies explicitly to low-resource African languages.

1. Methodology

This section describes the methods and steps applied to the project in the detecting of machine generated texts on African languages.

* 1. Preprocessing

We utilized the Vukuzenzele dataset, a publicly available civic-oriented resource developed by the South African government, covering multiple low-resource South African languages. As the dataset initially included only human-written texts, synthetic machine-generated texts were created to enable binary classification (human vs. machine).

Machine-generated texts were produced using GPT-based models, specifically ChatGPT and Falcon. Structured prompt engineering was applied to ensure generated texts closely aligned with the original dataset in terms of topic, tone, sentiment, coherence, and length. To maintain data balance and comparability, the final dataset consisted of 500 human-written and 500 machine-generated samples for each language — Northern Sotho, Xhosa, Venda, and Tsonga — totaling 4,000 samples.

A comparison of blue and pink bars

AI-generated content may be incorrect.

**Figure 1** Above shows the distribution of languages and type of text after preprocessing

To ensure high-quality data, we applied the following preprocessing steps:

* **Data Cleaning:** Removal of noise, irrelevant text, and emojis.
* **Length Filtering:** Ensuring comparable length distributions between human and machine samples.
* **Sentiment and Tone Alignment:** Maintaining thematic and emotional consistency across both categories

Synthetic Data Generation

Synthetic data generation followed a controlled and iterative process:

* **Prompt Engineering:** Carefully crafted prompts guided GPT models to generate civic-themed texts consistent with human-written counterparts.
* **Batch Generation and Manual Curation:** Synthetic texts were generated in batches of approximately 200 samples. Each batch underwent manual inspection to eliminate unnatural phrasing, prompt leakage, thematic drift, and contamination from English tokens. This ensured fluency and high-quality generation across all languages.
  1. AfriBERTa on Vukuzenzele

We fine-tuned AfriBERTa, specifically trained on African languages, providing nuanced linguistic sensitivity. The training setup was:

* Batch Size: 32 samples per batch.
* Learning Rate: 2e-5.
* Epochs: 5 epochs, chosen to balance accuracy against potential overfitting.
* Sequence Length: Texts tokenized to a maximum sequence length of 256 tokens.

AfriBERTa’s specialized regional knowledge significantly enhanced its ability to capture subtle linguistic patterns unique to the dataset.

* 1. XLM- RoBERTa on Vukuzenzele

We fine-tuned the multilingual XLM-RoBERTa model, chosen for its robust cross-lingual generalization. Training parameters were:

* Batch Size: ~16 samples per batch (reduced to accommodate computational limitations due to the model's size).
* Learning Rate: 3e-5.
* Epochs: 4 epochs to reduce overfitting risk.
* Sequence Length: Consistently maintained at 256 tokens per sample.

This multilingual modelling approach effectively captured broader linguistic features applicable across low-resource languages, complementing AfriBERTa’s language-specific strengths.

Ensemble Approach

To leverage the complementary strengths of AfriBERTa and XLM-RoBERTa, we implemented an ensemble prediction strategy by averaging their SoftMax probabilities. This method consistently yielded improved classification robustness and reliability.

* 1. Model Evaluation

Model evaluation was comprehensive, employing multiple classification metrics:

* Accuracy, Precision, Recall, and F1-score to measure overall model performance.
* Cross-validation: Stratified 5-fold cross-validation ensured robust performance estimates.
* Length-based performance assessment: Evaluated separately for short (≤15 words), medium (16–30 words), and long (>30 words) samples to identify performance variability by text length.

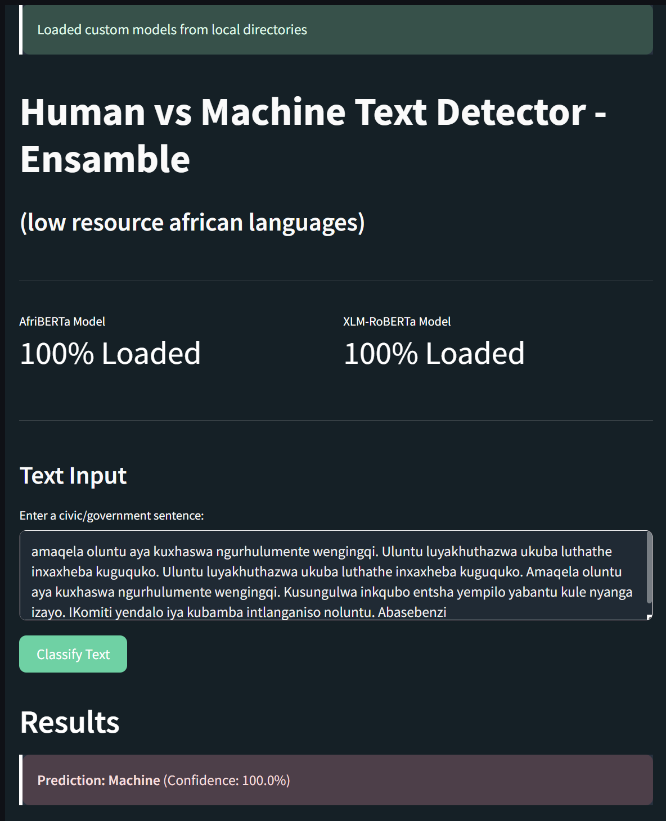
Initial evaluation results indicated excellent performance (Accuracy: 99.95%, F1-score: 0.9995 (Table 1)), though deeper analysis revealed overfitting. Consequently, extensive manual adjustments and additional prompt engineering were performed to enhance the generalizability of the synthetic texts and minimize dataset-specific learning biases.

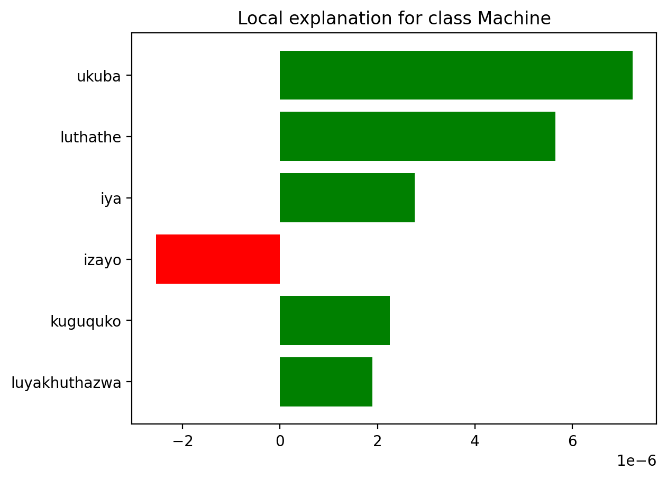
* 1. Model Deployment

To practically utilize the trained models, we developed an accessible deployment solution:

* **Streamlit Application**: A user-friendly web interface enabled real-time inference using both AfriBERTa and XLM-RoBERTa models.
* **Interactive Ensemble Predictions**: Integrated ensemble logic provided clear, reliable classifications.
* **Interpretability with LIME**: Immediate visualization of key features influencing model decisions enhanced transparency and user trust.

**Figure 2** is the UI for the user friendly application





**Figure 3** shows the model’s explainability to the decision of the model. The app can be found here: [GitHub-repo](https://github.com/shizaamir1615/human-vs-machine-text-detector.git)

This deployment framework enabled both practical detection capabilities and pedagogical demonstrations of model interpretability and performance.

* 1. Data description and analysis

5.6.1 Dataset Attributes

Each text sample in the dataset included structured attributes for consistency and detailed analysis:

* Text: The main textual content (sentences or paragraphs).
* Source: Original data source (Vukuzenzele or GPT models).
* Label: Binary label indicating the text origin (Human or Machine-generated).
* Language: Language code (Nso, Tso, Ven, Xho).
* Length Category: Short (≤15 words), Medium (16–30 words), Long (>30 words).

Timestamp**:** Publication or generation date.

* + 1. Text length Distribution

Exploratory analysis of text length revealed:

* The majority of samples fell within the medium-length category (16–30 words), supporting model comparability.
* Long texts (>30 words) provided richer context for learning nuanced linguistic patterns.

Short texts (≤15 words) presented challenges for the models, requiring sensitivity to subtle linguistic cues.

* + 1. Linguistic Diversity and Quality

Qualitative analysis indicated that:

* Human-written texts exhibited natural flow, culturally relevant idiomatic expressions, and coherent semantic transitions.
* Machine-generated texts occasionally showed subtle linguistic imperfections such as unnatural phrasing, slight thematic drift, or contamination from English tokens.
* Despite rigorous prompt engineering, consistent differences emerged between human and synthetic texts, aiding binary classification but posing a challenge for generalization.
  + 1. Sentiment and Thematic Alignment

Sentiment analysis revealed that generated texts generally matched human texts in thematic alignment (e.g., civic themes like health, education, community welfare). Nonetheless, machine-generated texts occasionally lacked emotional subtlety, providing identifiable linguistic markers for the detection task.

1. Results and Analysis

To evaluate the effectiveness of the ensemble model in detecting machine-generated text across languages, we assessed the model using standard classification metrics: precision, recall, F1-score, and accuracy. The analysis was conducted separately for both human-written and machine-generated classes.

* 1. Language-wise Model Performance

Across low-resource African languages **Xhosa**, **Tsonga**, and **Venda** the model achieved **perfect scores** in all evaluation metrics (Precision, Recall, F1-Score = 1.00), indicating an exceptional ability to differentiate between human and machine-generated content in these languages. This consistency suggests that the model effectively captures linguistic patterns unique to both human and machine-generated texts within these language structures.

**Table 1** Model Metrics for each language.

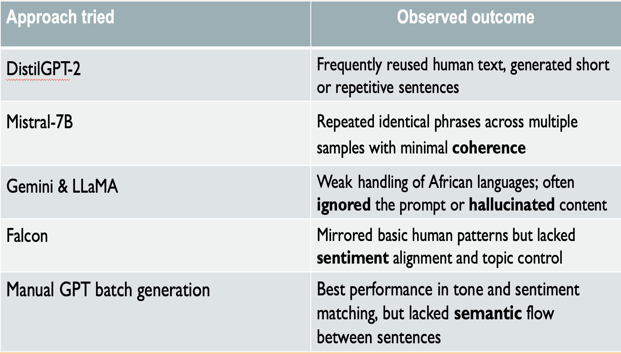
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| LANGUAGE | CLASS | PRECISION | RECALL | F1-SCORE | SUPPORT |
| Tsonga (Tso) | Human | 1.00 | 1.00 | 1.00 | 179 |
|  | Machine | 1.00 | 1.00 | 1.00 | 153 |
| Venda (Ven) | Human | 1.00 | 1.00 | 1.00 | 144 |
|  | Machine | 1.00 | 1.00 | 1.00 | 149 |
| Xhosa (Xho) | Human | 1.00 | 1.00 | 1.00 | 207 |
|  | Machine | 1.00 | 1.00 | 1.00 | 79 |
| Nothern Sotho (Nso) | Human | 0.98 | 0.99 | 0.99 | 164 |

While these perfect scores suggest highly effective classification, they simultaneously raise serious concerns regarding overfitting. Such absolute performance is rarely achievable in real-world natural language tasks, especially on low-resource languages, and likely reflects that the model has memorised systematic artefacts or idiosyncrasies in the synthetic data rather than learning true linguistic generalisations.

The model achieved perfect scores (precision, recall, and F1-score of 1.00) for Tsonga, Venda, and Xhosa across both text classes. This indicates that the classifier was able to distinguish with complete accuracy between human and machine-generated text in these languages. In contrast, performance in Northern Sotho was slightly lower, with the human class achieving a precision of 0.98, recall of 0.99, and F1-score of 0.99. While still excellent, the drop suggests a few misclassifications most likely false positives. These errors may be due to subtle stylistic overlaps between human and machine texts in Northern Sotho, or a more complex linguistic structure that posed challenges for the model.

During dataset construction and model training, we observed that many synthetic machine-generated texts contained recurrent patterns, repetitive phrasing, prompt leakage, and occasional contamination from English tokens. Despite careful prompt engineering and manual review, these artefacts remained detectable, leading the model to memorize superficial statistical patterns rather than generalizing to true semantic or syntactic differences.

To address these challenges, we experimented with multiple mitigation strategies, each introducing different language models and generation techniques. Table 2 (reproduced below) summarizes the various approaches and their observed outcomes:

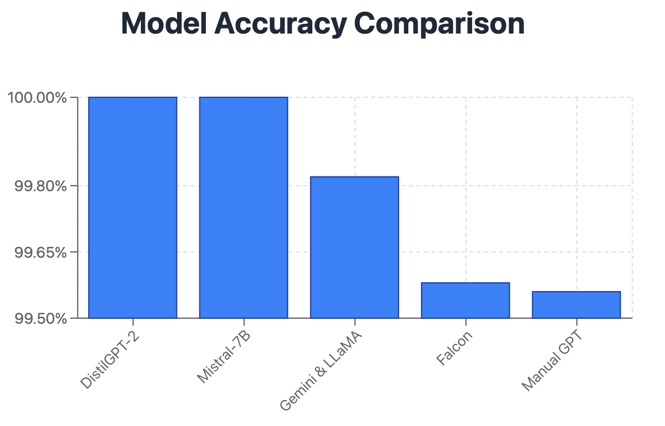


**Table 2** Shows different models used to generate the synthetic

Despite these diverse attempts, overfitting persisted across both AfriSenti and Vukuzenzele datasets. Even after refining prompts, generating additional variation, and experimenting with back-translation techniques, models continued to pick up superficial signals within the synthetic data rather than deeper language patterns.

This challenge was reflected not only in qualitative observations, but also quantitatively in the evaluation metrics. As illustrated in Figure 4, despite the varied data generation approaches, accuracy across all models remained extremely high — clustered very close to 100% — further reinforcing concerns about model overfitting:

**Figure 4** Performance of the different models attempt to generate machine text.



**Reflections and Limitations**

* Perfect scores indicate the model heavily relied on dataset-specific quirks rather than fully generalising.
* Multiple generation models consistently struggled to produce sufficiently diverse and fluent synthetic texts in African languages.
* Even the best performing manual generation approach failed to provide natural sentence-to-sentence semantic flow.
* These findings highlight the urgent need for better high-quality human-annotated datasets for African language evaluation.

1. **Discussion and Conclusion**

This study addressed three core research questions on detecting machine-generated text in African languages:

**RQ1:** Multilingual models like AfriBERTa and XLM-RoBERTa achieved near-perfect classification performance. However, these results largely reflect overfitting to artefacts present in the synthetic data rather than robust linguistic understanding.

**RQ2:** Human-written texts exhibited natural coherence, semantic flow, and culturally relevant phrasing, while machine-generated texts displayed repetition, weak sentence transitions, and occasional hallucinations. These shallow artefacts enabled easy classification but limit model generalisability.

**RQ3:** While both models fine-tuned successfully, their strong performance depended heavily on dataset artefacts. Without more diverse and natural synthetic data, fine-tuning remains limited in real-world applicability.

Despite experimenting with multiple generation models (DistilGPT-2, Mistral-7B, Gemini, Falcon, and manual GPT), none produced fully fluent and diverse synthetic texts for African languages. The main bottleneck remains the difficulty of generating high-quality, linguistically coherent synthetic data for low-resource languages.

**Future Work**

To overcome these challenges, several directions are recommended:

* **Human-in-the-loop generation:** Incorporate native speakers to refine machine-generated text, improving tone, coherence, and cultural appropriateness.
* **Structured prompt design:** Develop strict multi-sentence prompts that enforce continuity, sentiment alignment, and topic consistency across generated texts.
* **Domain-controlled generation:** Restrict generation to narrow civic themes (e.g., health, education) to improve semantic consistency within samples.
* **Mixed-generator supervision:** Track which generator produced each synthetic sample, allowing the classifier to model across different generation styles rather than overfitting to a single generator's quirks.
* **Synthetic data evaluation frameworks:** Establish formal metrics or automated quality checks to assess coherence, fluency, and semantic flow in generated datasets before using them for training.

**Conclusion**

This work demonstrates the technical feasibility of detecting machine-generated text in African languages but also exposes critical weaknesses in current synthetic data generation methods. Ultimately, progress in this field will depend less on model architecture and more on the availability of rich, diverse, and linguistically authentic training data. Building such datasets will be essential for developing reliable, generalisable machine-generated text detectors for underrepresented languages.

References

Our GitHub Repository : [GitHub-repo](https://github.com/shizaamir1615/human-vs-machine-text-detector.git)

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., ... & Stoyanov, V. (2020). *Unsupervised cross-lingual representation learning at scale*. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 8440–8451). Association for Computational Linguistics. https://aclanthology.org/2020.acl-main.747/

Dossou, B. F., & Emezue, C. C. (2021). *AfroXLMR: A self-active learning model for low-resource African language understanding*. arXiv. <https://arxiv.org/abs/2104.08821>

Ippolito, D., Duckworth, D., Callison-Burch, C., & Eck, D. (2020). *Automatic detection of generated text is easiest when humans are fooled*. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 1808–1822). Association for Computational Linguistics. https://aclanthology.org/2020.acl-main.486/

Lundberg, S. M., & Lee, S.-I. (2017). *A unified approach to interpreting model predictions*. In *Advances in Neural Information Processing Systems*, 30. https://proceedings.neurips.cc/paper\_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

Nekoto, W., Marivate, V., Matsila, T., Fasubaa, T., Fagbohungbe, T., Ogundepo, A., ... & Adelani, D. I. (2020). *Participatory research for low-resourced machine translation: A case study in African languages*. In *Findings of the Association for Computational Linguistics: EMNLP 2020* (pp. 2144–2160). https://aclanthology.org/2020.findings-emnlp.327/

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). *"Why should I trust you?": Explaining the predictions of any classifier*. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). https://doi.org/10.1145/2939672.2939778

South African Government. (n.d.). *Vukuzenzele dataset* [Data set]. GitHub. <https://github.com/masakhane-io/masakhane-news>