DETECTING MACHINE GENERATED TEXT IN LR AFRICAN LANGUAGES

Group 23

Shiza Butt

Vutomi Mohube

Lesedi Ntsele

INTRODUCTION / PROBLEM STATEMENT

- The rapid rise of large language models make it increasingly hard to tell human-written from machine-generated text.
- This is especially challenging in low-resource African languages where detection tools are scarce.
- ***Can we develop robust models to accurately detect machine-generated texts in African languages, and what linguistic patterns differentiate them from human texts?***
- African languages remain underrepresented in NLP research and datasets
- Fake content threatens trust in civic domains (e.g., news, education).
- Most existing tools are built for high-resource languages like English.

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-R** be effectively fine-tuned for this task?

DATASET & LANGUAGES

- We chose the **Vukuzenzele dataset** for its authentic civic content and multilingual representation of low-resource South African languages. Since it only contained human-written text, we generated matching machine samples using GPT-based models **(Falcon and Chat GPT)** to enable binary classification.
- To ensure quality and balance, we **cleaned** the data, removed noise, and applied length-based filtering to keep the human and machine text pairs comparable.
- We prompt-engineered GPT models to produce **civic-themed, coherent,** and **sentiment-aligned** texts that matched human samples in tone and length.

DATASET & LANGUAGES (cont)

Language	Human Source	Machine source	Sample size
Northern Sotho (nso)	Vukuzenzele	Chat GPT + Falcon	1,000 (500 + 500)
Tsonga (tso)	Vukuzenzele	Chat GPT + Falcon	1,000 (500 + 500)
Venda (ven)	Vukuzenzele	Chat GPT + Falcon	1,000 (500 + 500)
Xhosa (xho)	Vukuzenzele	Chat GPT + Falcon	1,000 (500 + 500)

Total sample size : 4000 and specific languages were chosen to balance **representativity** with limited **computational capacity**.

MODELLING APPROACH

- We chose **AfriBERTa** for its specialisation in African languages, making it ideal for capturing region-specific linguistic nuances.
- In contrast, **XLM-RoBERTa** is a high-performing multilingual model with strong generalisation across low-resource languages.
- This complementary pairing allowed us to leverage both local relevance and cross-lingual robustness
- To improve reliability, we combined their predictions using a **simple ensemble approach** (averaging softmax probabilities), to enhance consistency across inputs.
- We also applied **LIME** to interpret model predictions and highlight which words influenced decisions.
- Finally, we built a user-friendly **Streamlit demo** that loads both models locally, applies ensemble logic, and visually presents LIME explanations.

LIVE DEMO

RESULTS & EVALUATION STRATEGY

Computed in evaluate_model.py

Overall Performance

Accuracy: 99.95%, F1 Score: 0.9995, Precision: 0.9990, Recall: 1.0000

Language Robustness

- Tso, Xho, Ven: 100% accuracy and F1
- Nso: Slight dip to 99.8% — may reflect subtle linguistic noise

Text-Length Robustness

- Short (≤15 words): 99.77%
- Medium (16–30 words): 100%
- Long (>30 words): 100%

These results show that our model was **overfitting** to the dataset.

MITIGATION & CHALLENGES

- Overfitting was evident when tested on truly unseen civic-domain samples
- We conducted extensive prompt engineering to match sentiment, coherence, and length with human-written data. GPT-based models often struggled with **prompt leakage**, **unnatural phrasing,** and English token **contamination**.Making the distinction between the human and machine significant.
- To address this, we switched to **manual generation** with Chat-GPT in batches of 200, curating each for tone, length, and fluency
- We ensured balance across languages and classes, even categorising samples by word length.Every strategy was repeated multiple times across languages, but overfitting remained persistent
- **Limitations** in dataset size and computational resources restricted deeper augmentation or pre-training
- Ultimately, our efforts reflect the **difficulty** of producing natural machine text in low-resource settings despite strong modelling techniques

MITIGATION & CHALLENGES (cont)

Approach tried	Observed outcome
DistilGPT-2	Frequently reused human text, generated short or repetitive sentences
Mistral-7B	Repeated identical phrases across multiple samples with minimal **coherence**
Gemini & LLaMA	Weak handling of African languages; often **ignored** the prompt or **hallucinated** content
Falcon	Mirrored basic human patterns but lacked **sentiment** alignment and topic control
Manual GPT batch generation	Best performance in tone and sentiment matching, but lacked **semantic** flow between sentences

Overfitting persisted across both the AfriSenti and Vuk'uzenzele datasets, despite multiple rounds of refinement,generation adjustments and after exploring back translation too. The problem was consistent across all models and setups.

INSIGHTS AND LESSONS LEARNT

- **Text generation is not trivial**, especially in low-resource African languages. Most LLMs failed to produce fluent, semantically consistent, and culturally grounded civic text.
- **Overfitting is not just a modelling issue**- it reflects flaws in **data quality, diversity, and realism**. When the machine-generated text is too easy to distinguish, models learn shortcuts.
- **Multilingual models can't replace language-specific nuance**, especially when data is synthetic.
- **Ensembling and explainability** improved transparency but couldn't solve the core generalisation problem.

POSSIBLE WORKAROUNDS / FUTURE DIRECTIONS

- **Human-in-the-loop generation**
  → Collaborate with fluent/native speakers to refine or generate machine text with better tone, coherence, and sentence flow.
- **Prompt templating with hard constraints**
  → Use stricter, structured prompts that enforce length, sentiment, and multi-sentence continuity across generations.
- **Domain-restricted generation**
  → Limit generation to one civic theme at a time (e.g., only health or only education) to improve internal consistency within samples.
- **Mixed-generator training**
  → Label which generator created each machine sample and train models to recognise shared vs. model-specific generation patterns.

Q & A