Model Comparison & Selection Report – Task 4

Prepared by: Mesfin Mulugeta

Overview

This report presents the comparison and evaluation of different pre-trained transformer models for the Amharic Entity Extraction task. The models were assessed based on their accuracy (F1-score), precision, recall, training time, and feasibility on Colab in terms of RAM usage. The goal is to select the most suitable model for production deployment.

Model Comparison Table

Model	F1 Score	Precision	Recall	Training Time	RAM Usage	Notes
XLM-Roberta-base	0.96	0.97	0.96	3 hr 55 min	High	Best accuracy, reliable, robust
rasyosef/bert-tiny-amharic	0.002	0.01	0.001	3 mins	Very Low	Failed to learn, underfitting
masakhane/afroxlmr-large-ner-masakhaner	-	-	-	Not trained	Exceeded Colab RAM	Not trainable on Colab without cloud GPU

Conclusion & Recommendation

Based on the evaluation metrics, XLM-Roberta-base is the best performing model, achieving an F1 score of 0.96 with high precision (0.97) and recall (0.96). It demonstrates excellent accuracy and robustness, suitable for production environments that have adequate computing resources.
The bert-tiny-amharic model severely underfits, showing that it lacks the capacity for this NER task. It is useful only for debugging pipelines or quick tests.
Training masakhane/afroxlmr-large-ner-masakhaner was infeasible on Colab due to RAM limitations. It would require enterprise-level GPUs (such as A100, V100, or multi-GPU setups).

Therefore, XLM-Roberta-base is selected for production deployment based on its high performance and reliability.