**English–Sepedi Machine Translation Evaluation**

**Project Title**

**Evaluation of NLLB and M2M100 Models on English–Sepedi Machine Translation using Corrected FLORES Dataset**

**Overview**

This project evaluates the translation performance of two state-of-the-art multilingual machine translation models:

* facebook/nllb-200-distilled-600M
* facebook/m2m100\_418M

The target language is Sepedi (Northern Sotho). We used the Corrected FLORES dev set to benchmark translation quality. The evaluation includes both quantitative metrics and qualitative error analysis.

**Objectives**

1. Load a high-quality English–Sepedi evaluation dataset.
2. Generate translations using NLLB and fine-tuned M2M100 models.
3. Evaluate outputs using BLEU, chrF and similarity scoring.
4. Identify translation challenges specific to Sepedi using manual error analysis.

**Dataset**

* **Source**: Corrected FLORES
* **Columns**:
  + English: Source sentences
  + Sepedi: Reference (ground truth) translations
* **Sample Size**: 5 English–Sepedi sentence pairs used for evaluation

**🧪 Models Used**

| **Model** | **Description** |
| --- | --- |
| facebook/nllb-200-distilled-600M | A distilled version of Meta’s No Language Left Behind model, supporting 200 languages |
| facebook/m2m100\_418M | is a 418-million-parameter multilingual transformer model that enables direct many-to-many translation across 100 languages without relying on English as an intermediate language. |

**🛠️ Setup and Installation**

**Ensure you have the following dependencies installed:**

* !pip install transformers torch
* !pip install -U datasets huggingface\_hub fsspec
* !pip install datasets
* !pip install transformers datasets evaluate sacrebleu
* from transformers import AutoTokenizer, AutoModelForSeq2SeqLM
* import evaluate
* from sacrebleu import corpus\_bleu, corpus\_chrf

**How to Run**

1. Launch the notebook:
2. Steps followed in the notebook:(Run all cells to get generated .txt files

* Load the Corrected FLORES dataset.
* Select and tokenize source texts.
* Translate using each model:
  + Set source language to "en" (M2M100) or "eng\_Latn" (NLLB).
  + Set target language to "nso" (M2M100) or "nso\_Latn" (NLLB) using forced\_bos\_token\_id.
* Generate translations and decode the outputs.
* Compute BLEU and chrF scores using sacrebleu.
* Use difflib.SequenceMatcher to calculate sentence-level similarity.
* Save predictions and evaluation results to a .txt file.

**📈 Evaluation Metrics**

* **BLEU**: Measures n-gram overlap between predicted and reference translations.
* **chrF**: Measures character-level precision and recall, effective for morphologically rich languages.
* **Similarity Score**: Uses difflib to compare translations on a scale from 0.0 to 1.0.
* **Error Analysis**: Logs worst examples (low similarity)

**📋 Output**

* 2 Detailed log file including:
  + Source, reference, and predicted sentences
  + Similarity scores
  + BLEU and chrF scores
  + worst-case translations for manual inspection

**🧩 Challenges Faced**

* Language token mismatches (nso vs nso\_Latn) during model generation.
* Copying behavior in translations (predicted == source).
* BLEU and chrF may not fully capture fluency or adequacy in Sepedi.
* Limited high-resource training data for low-resource languages like Sepedi.

**📚 References**

* Hugging Face Transformers