NLP Track Capstone Project:
Evaluation of Multilingual Models

Objective

Fine-tune and evaluate transformer-based language models on a multilingual text classification task.
You will compare a monolingual model (BERT) with a multilingual model (e.g., mBERT or XLM-RoBERTa) to analyze performance differences across languages.

This project focuses on transfer learning in multilingual contexts — how models generalize across different linguistic structures, scripts, and semantics.

Core Tasks

1. **Dataset Selection**
    - Choose a multilingual or cross-lingual dataset suitable for text classification or sentiment analysis.
      Example datasets:
        - **Amazon Multilingual Reviews (English, German, French, Japanese)**
        - **XNLI (Cross-lingual Natural Language Inference)**
        - **MASSIVE or MTOP (Intent classification in multiple languages)**
        - **Twitter Sentiment Multilingual Dataset**
        - **Hate Speech / Offensive Language in Multiple Languages (from Hugging Face)**
    - You may also create a multilingual subset from multiple monolingual datasets (e.g., combine English IMDB reviews with translated versions).

2. **Model Selection and Comparison**
    - Choose **two models**:
        - A **standard (monolingual)** model such as `bert-base-uncased`, `distilbert-base-uncased`, or `roberta-base`.
        - A **multilingual model** such as `bert-base-multilingual-cased`, `xlm-roberta-base`, or `distilbert-multilingual-nli-stsb`.
    - Fine-tune both models on your dataset (you can use a subset if resources are limited).
    - Compare their performance on **multiple languages** (for example, train on English and evaluate on French, Spanish, etc.).

3. **Evaluation and Analysis**

  - ○ Use standard metrics: Accuracy, F1-score, Confusion Matrix.
  - ○ Report language-wise performance, how accuracy or F1 changes per language.
  - ○ Discuss:
    - ■ How multilingual pretraining affects transfer to unseen languages.
    - ■ Any degradation in non-English performance.
    - ■ Computational trade-offs between multilingual and monolingual models.

4. **Visualization & Reporting**
   - ○ Visualize results across languages (bar charts, confusion matrices, etc.).
   - ○ Include short qualitative examples (where the model succeeded or failed in non-English text).
   - ○ Conclude with observations on multilingual transfer learning effectiveness.
   - ○ Set up a demo on Gradio/Streamlit for others to try out the model

Deliverables

1. **Code notebook** (Colab / Jupyter):
   - ○ Dataset loading and preprocessing
   - ○ Model fine-tuning and evaluation for both models
   - ○ Result visualization and brief interpretation
   - ○ Set up a demo on Gradio/Streamlit for others to try out the model
2. **Short report / presentation (2–4 pages or slides)** covering:
   - ○ Dataset details (languages, size, labels)
   - ○ Models used and rationale
   - ○ Training setup and hyperparameters
   - ○ Performance comparison and analysis
   - ○ Key insights on multilingual generalization

Optional Extensions

- Add zero-shot evaluation, train on English, test on a new language.
- Explore translation-based augmentation (Google Translate API or open datasets).
- Experiment with language-specific fine-tuning vs. joint multilingual fine-tuning.