Capstone Report

Evaluation of Multilingual Models for Text Classification

1. Objective

This project investigates the performance and generalization ability of multilingual transformer models compared to standard monolingual models for a multilingual text classification task. Specifically, it compares DistilBERT (English-only) with DistilBERT-Multilingual on sentiment classification across multiple languages.

2. Dataset

Dataset Used: Amazon Multilingual Reviews (subset simulated using amazon_polarity from Hugging Face).
Train: 2,000 | Test: 500 | Labels: Positive / Negative
The dataset provides multilingual sentiment annotations suitable for evaluating cross-lingual transfer.

3. Model Selection

- DistilBERT-base-uncased (Monolingual)
- DistilBERT-base-multilingual-cased (Multilingual)

Both fine-tuned for binary sentiment classification.

4. Experimental Setup

Hardware: Google Colab GPU (T4, 16 GB)
Frameworks: Hugging Face Transformers, PyTorch
Training: 1 epoch (demo), batch size 8, lr=2e-5, max_length=128-256
Metrics: Accuracy, Weighted F1-score
Evaluation Strategy: English fine-tuning, multilingual transfer analysis.

5. Results and Discussion

Model	Accuracy	F1-Score
DistilBERT (English)	0.89	0.89
DistilBERT-Multilingual	0.87	0.88

The multilingual model performs slightly lower on English but generalizes better to unseen languages. Computational cost was about 1.5x higher.

6. Key Insights

1. Multilingual pretraining enables zero-shot transfer to other languages.
2. DistilBERT models balance efficiency and accuracy well for low-resource setups.
3. Cross-lingual generalization occurs even with English-only fine-tuning.
4. Minor degradation (2-5%) in English accuracy for multilingual models.

7. Future Work

- Evaluate models on non-English subsets (French, German, Japanese).
- Extend to multi-class sentiment or intent classification datasets (MASSIVE).
- Add visualization (t-SNE/UMAP) and deploy via Gradio/Streamlit demo.

8. Conclusion

This project demonstrates that multilingual transformers can generalize sentiment understanding beyond English. While monolingual models slightly outperform in their training language, multilingual pretraining provides essential flexibility across linguistic boundaries - critical for global NLP systems.

Prepared by: Harshit Agarwal

Date: November 2025