Cross-Lingual Generalization in Sentiment Analysis

1.0 Introduction & Objective

This report details an experiment conducted to assess and compare the cross-lingual generalization capabilities of two distinct Transformer-based models: a monolingual model (bert-base-uncased) and a multilingual model (xlm-roberta-base).

The primary objective was to quantify how well a model trained exclusively on English sentiment data performs when tested on other languages (Spanish, Hindi, and French). This baseline performance was then compared against a multilingual model that was trained on a dataset containing all four languages. The task for both models was 3-class sentiment analysis.

2.0 Dataset Details

2.1 Source and Content

The dataset used for this experiment is the **cardiffnlp/tweet_sentiment_multilingual** dataset. This dataset, sourced from Hugging Face, contains tweets annotated for sentiment across multiple languages.

2.2 Languages and Size

For this demonstration, a specific subset of the data was used. The experiment focused on four languages:

- English
- Spanish
- Hindi
- French

A balanced subset of **1,000 samples** was selected for each language, resulting in a total working dataset of 4,000 samples.

2.3 Labels and Splits

The dataset is pre-labeled for 3-class sentiment (e.g., 0, 1, 2, which typically represent negative, neutral, and positive sentiments).

Each language's 1,000-sample subset was split into:

- **Training Set:** 80% (800 samples)
- **Test Set:** 20% (200 samples)

This resulted in four distinct test sets: test-english, test-spanish, test-hindi, and test-french, each containing 200 samples.

3.0 Models & Rationale

Two different models were selected to represent the monolingual and multilingual approaches.

3.1 Model 1: Monolingual Baseline

- **Model:** bert-base-uncased

- **Rationale:** This model was chosen to serve as the monolingual baseline. It is a powerful and widely-used model that was pre-trained almost exclusively on English text. The hypothesis is that while it will perform reasonably on the English test set, it will fail to generalize to other languages it has not been trained on and whose token representations it does not know. It was configured with a classification head for 3 labels.

3.2 Model 2: Multilingual Contender

- **Model:** xlm-roberta-base (XLM-R)
- **Rationale:** This model was chosen for its strong, built-in multilingual capabilities. It was pre-trained on a massive corpus (CommonCrawl) spanning 100 languages. This shared multilingual vocabulary and embedding space is hypothesized to allow the model to "transfer" its understanding of the sentiment task from one language to another. For example, by seeing sentiment-labeled examples in English and Spanish, it should be able to better predict sentiment in French. It was also configured with a classification head for 3 labels.

3.3 Preprocessing and Tokenization

Each model used its corresponding official tokenizer:

- **Model 1 Tokenizer:** AutoTokenizer.from_pretrained('bert-base-uncased')
- **Model 2 Tokenizer:** AutoTokenizer.from_pretrained('xlm-roberta-base')

For both models, all text was truncated or padded to a uniform max_length of **128 tokens**.

4.0 Training Setup

4.1 Training Data Configuration

The core difference in the experimental setup was the data provided to each model during the fine-tuning (training) phase.

- Monolingual (BERT) Training:

This model was trained only on the English training data (train-english), which consisted of 800 samples.

- Multilingual (XLM-R) Training:

This model was trained on a combined multilingual dataset (train_multilingual). This set was created by concatenating the 800-sample training sets from all four languages, resulting in a total training set of 3,200 samples (800 x 4).

4.2 Hyperparameters

To ensure a fair comparison, both models were trained using identical hyperparameters, as specified in the TrainingArguments.

Hyperparameter	Value	Rationale
learning_rate	2e-5	Standard learning rate for fine-tuning Transformers.
per_device_train_batch_size	16	Batch size for training.

per_device_eval_batch_size	16	Batch size for evaluation.
num_train_epochs	2	A short training run for this demonstration.
weight_decay	0.01	Standard regularization.

4.3 Evaluation Metrics

The models were evaluated on the held-out 200-sample test set for *each* of the four languages. The primary metrics used for comparison were:

- **Accuracy**
- **F1-Score (Macro):** This metric was chosen as it is well-suited for multi-class classification. It computes the F1 score for each label independently and then takes the unweighted average, treating all classes as equally important.

5.0 Performance Comparison & Analysis

5.1 Final Results

The cross-lingual evaluation yielded the following performance metrics. The F1 (Macro) score and Accuracy are reported for each model against each language-specific test set.

Model	Test Language	Accuracy	F1 (Macro)
BERT (Trained EN)	**English**	**0.575**	**0.518**
BERT (Trained EN)	Spanish	0.370	0.308
BERT (Trained EN)	Hindi	0.295	0.266
BERT (Trained EN)	French	0.415	0.397
XLM-R (Trained Multi)	**English**	**0.650**	**0.647**
XLM-R (Trained Multi)	Spanish	0.615	0.605
XLM-R (Trained Multi)	Hindi	0.410	0.399
XLM-R (Trained Multi)	French	0.670	0.668

5.2 Performance Analysis

The results, also visualized in the notebook's bar charts, clearly demonstrate the limitations of the monolingual model and the significant advantages of the multilingual approach.

- **BERT (Monolingual):** The English-trained BERT model achieved a 0.575 Accuracy and 0.518 F1-Macro score on the English test set. However, its performance collapsed when evaluated on the other languages. Its accuracy on Spanish (0.370) and Hindi (0.295) was extremely low, indicating a complete failure to generalize the sentiment task.
- **XLM-R (Multilingual):** The XLM-R model, trained on the mixed-language dataset, showed strong cross-lingual generalization. It achieved high scores on Spanish (0.615 Accuracy) and French (0.670 Accuracy).
- **Key Finding:** A striking result is that the multilingual XLM-R model **outperformed the monolingual BERT model even on English** (0.650 Accuracy vs. 0.575). This suggests that

training on a more diverse linguistic dataset made the model's fundamental understanding of "sentiment" more robust, which benefited its performance even on the primary language.

5.3 Key Insights & Conclusion

1. **Monolingual models (like bert-base-uncased) do not inherently generalize** to languages they were not pre-trained on, even if fine-tuned on a related task in a high-resource language.
2. **Multilingual models (like xlm-roberta-base) are specifically designed for this transfer.** By training on a mixed-language dataset, the model learns a shared representation of the task that can be applied across multiple languages.
3. **Training on diversity improves robustness.** The most significant insight is that the multilingual model's superior performance on English suggests that exposure to different linguistic structures can strengthen its core understanding of a task, making it a better model overall.