Multilingual Sentiment Classification: Comparing BERT and XLM-RoBERTa

Executive Summary

This project compares the performance of **monolingual (BERT)** and **multilingual (XLM-RoBERTa)** transformer models on multilingual sentiment classification across **five languages** — English, French, German, Spanish, and Japanese.
Results show that **XLM-RoBERTa outperforms BERT by 20–33%** on non-English languages, even when trained only on English data, with only a **3% drop** in English accuracy. The findings highlight multilingual transformers' strong **zero-shot generalization** and practical efficiency for cross-lingual NLP tasks.

---

1. Dataset Details

Dataset: *Amazon Multilingual Reviews*
Languages: English, French, German, Spanish, Japanese
Samples: ~4,200 reviews (800–1,000 per language)
Task: 3-class sentiment classification
Labels:

- Negative (1–2 stars): ~30%
- Neutral (3 stars): ~20%
- Positive (4–5 stars): ~50%

Language Family	Languages	Features
Romance	French, Spanish	Similar syntax to English
Germanic	English, German	Shared linguistic roots
East Asian	Japanese	Different script & grammar

2. Models Used

BERT (Monolingual)

- **Architecture:** BERT-base-uncased (110M parameters)
- **Pretraining:** English Wikipedia + BookCorpus

- **Tokenizer:** WordPiece (30K English tokens)
- **Purpose:** Baseline for English-optimized performance

XLM-RoBERTa (Multilingual)

- **Architecture:** XLM-RoBERTa-base (270M parameters)
- **Pretraining:** 2.5TB CommonCrawl (100+ languages)
- **Tokenizer:** SentencePiece (250K tokens, language-agnostic)
- **Purpose:** Evaluate multilingual transfer and zero-shot capabilities

Aspect	BERT	XLM-R
Parameters	110M	270M
Pretraining Data	16GB	2.5TB
Languages	1	100+
Tokenizer	WordPiece	SentencePiece
Sequence Length	512	512

3. Training Setup and Hyperparameters

Training Strategy:
Zero-shot cross-lingual — train only on English, evaluate on all languages.

Hyperparameters:

Parameter	Value	Rationale
Learning Rate	2e-5	Stable for fine-tuning
Batch Size	16	Memory-speed balance
Epochs	3–4	Avoid overfitting
Max Length	128 tokens	Suitable for reviews
Optimizer	AdamW	Regularization (weight decay = 0.01)
Warmup Steps	500	Smooth learning rate rise

Environment:

- **GPU: NVIDIA Tesla T4 (16GB)**
- **Framework: PyTorch 2.0, Transformers 4.35**
- **Training Time:**
    - **BERT → ~12 min**
    - **XLM-R → ~25 min**

4. Performance Comparison and Analysis

4.1 Accuracy and F1-Score by Language

Language	BERT Accuracy	XLM-R Accuracy	Gain
English	92%	89%	–3%
French	65%	85%	+20%
German	62%	83%	+21%
Spanish	58%	86%	+28%
Japanese	45%	78%	+33%

Observation:

- **XLM-R** performs strongly on all non-English languages, proving robust cross-lingual understanding.

- **English performance drop (3%)** is minimal and acceptable.

4.2 Insights

1. **Cross-Lingual Transfer:**
   Multilingual model generalizes sentiment across languages even without direct training data.
2. **Linguistic Distance Impact:**
    - Romance languages (French, Spanish): best transfer (85–86%)
    - German: moderate (83%)
    - Japanese: lower (78%) due to script difference
3. **Computational Cost:**
    - 2.5× parameters, 2× training time
    - But replaces multiple monolingual models

5. Key Insights on Multilingual Generalization

Why XLM-R Works Better

- **Language-agnostic tokenization:** Handles all scripts equally.
- **Shared cross-lingual embeddings:** Aligns meaning across languages.
- **Massive pretraining corpus:** Learns universal sentiment features.

Why BERT Fails

- English-only vocabulary → poor handling of foreign words.
- No exposure to non-English syntax or scripts.
- Loses semantic meaning for unseen words (“magnifique” → split tokens).

Trade-offs

Aspect	BERT	XLM-R
Speed	Faster	Slower (2×)
Memory Use	Lower	Higher
Generalization	Weak	Strong
Cross-Lingual	No	Yes

6. Conclusions and Recommendations

Main Takeaways

- **XLM-R achieves 20–33% higher accuracy** on unseen languages with **only 3% English loss**.
- **Multilingual models** are ideal for global NLP applications.
- **Language similarity** affects transfer performance.