Report: Which Model is Smarter? Comparing a Monolingual and Multilingual Model

1. Introduction: What Was This Project About?

The goal of this project was to find out if a multilingual model (one that knows many languages) is better at a language task than a monolingual model (one that only knows English).

I used a "zero-shot" experiment. This means I **trained both models only on English** and then **tested them on 15 different languages** to see if they could handle languages they had never been trained on.

2. What I Used: Models and Data

Dataset: XNLI

I used the **XNLI (Cross-lingual Natural Language Inference)** dataset. This dataset has pairs of sentences, and the model must decide if they:

Entailment: (The first sentence proves the second one is true)
Neutral: (The sentences are not related)
Contradiction: (One Sentence contradict second one)

This dataset is perfect because it has the same text in 15 different languages (like English, French, Spanish, Arabic, and Chinese).

The Models

I compared two different models:

bert-base-uncased (The Monolingual Model):

This is a "monolingual" model.

It was pre-trained **only on English text**. It does not know any other language.

xlm-roberta-base (The Multilingual Model):
This is a "multilingual" model.
It was pre-trained on **text from 100 different languages**. It has a massive vocabulary and understands the grammar of many languages.

3. How I Did It: The Experiment Setup

To make it a fair test, I trained both models in the exact same way:

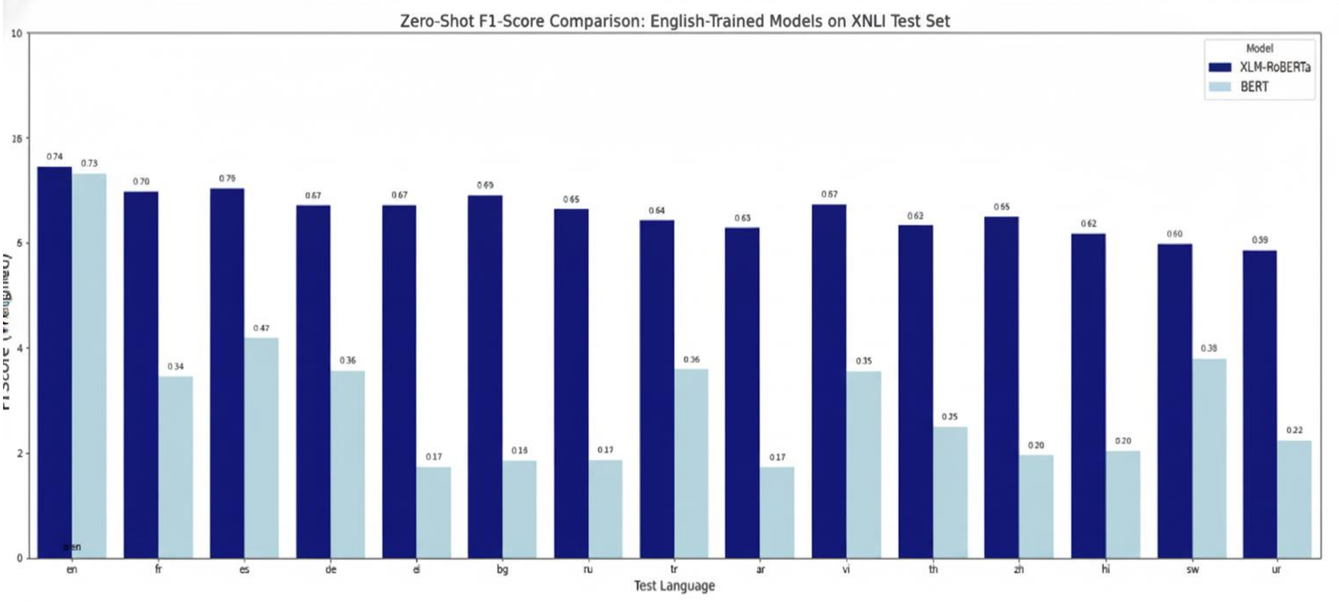
Training: I fed both models a training set of **13,000 training samples**, which were **all in English**.
Testing: I then tested them on the test set for all 15 languages to see how they would perform.

My settings (Hyperparameters):

Learning Rate: 2e-5
Batch Size: 64
Number of Epochs: 5
Optimizer: AdamW

4. What I Found: The Results

This is the most important part of the experiment. The chart below shows the F1-Score (a measure of accuracy) for both models across all 15 languages.



Analysis of the Results

bert-base-uncased (The English-only model):

On the English test set, BERT did very well, with a score of 0.73.
On *every other language* (like 'fr', 'es', 'ar'), the model failed completely. Its score was around ~0.25, which is the same as randomly guessing.
This happened because BERT's vocabulary is only English. It sees French or Chinese words as unknown [UNK] tokens.

xlm-roberta-base (The Multilingual model):

On the English test set, XLM-R performed similarly to BERT, with a score of 0.74
On all other 14 languages, XLM-R performed **much, much better** than BERT.
This proves that even though it was only fine-tuned on English, its "background knowledge" of other languages allowed it to transfer what it learned.

5. A Closer Look: Good and Bad Examples

Looking at specific examples helps to understand *why* the model failed or succeeded.

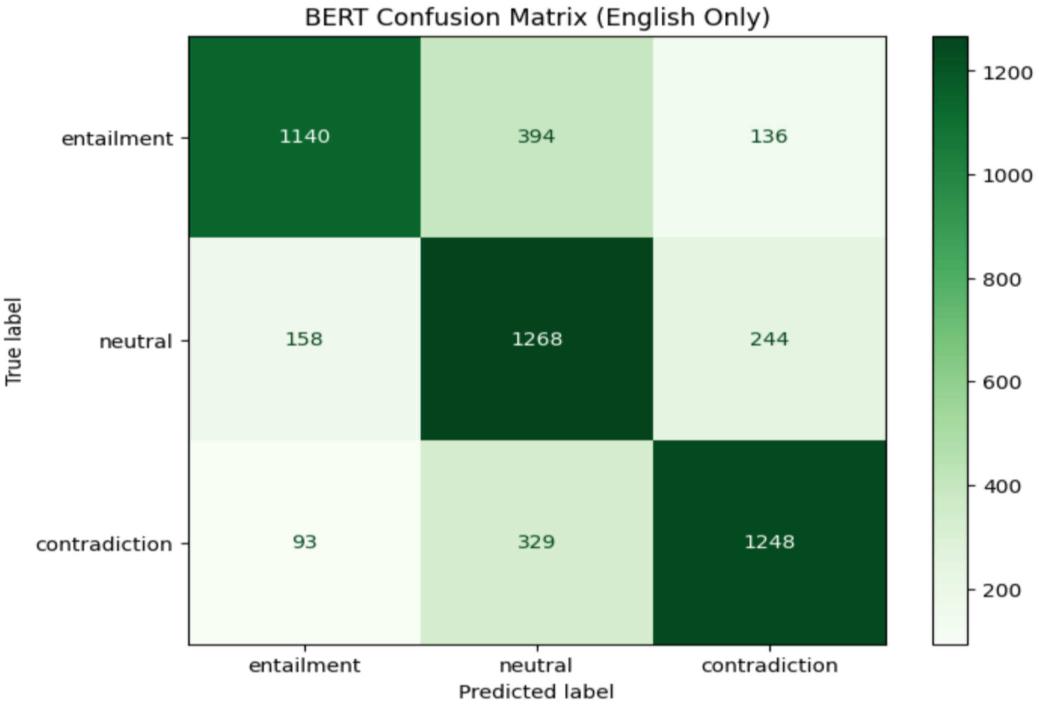
Example of a Success (XLM-R on a non-English language):

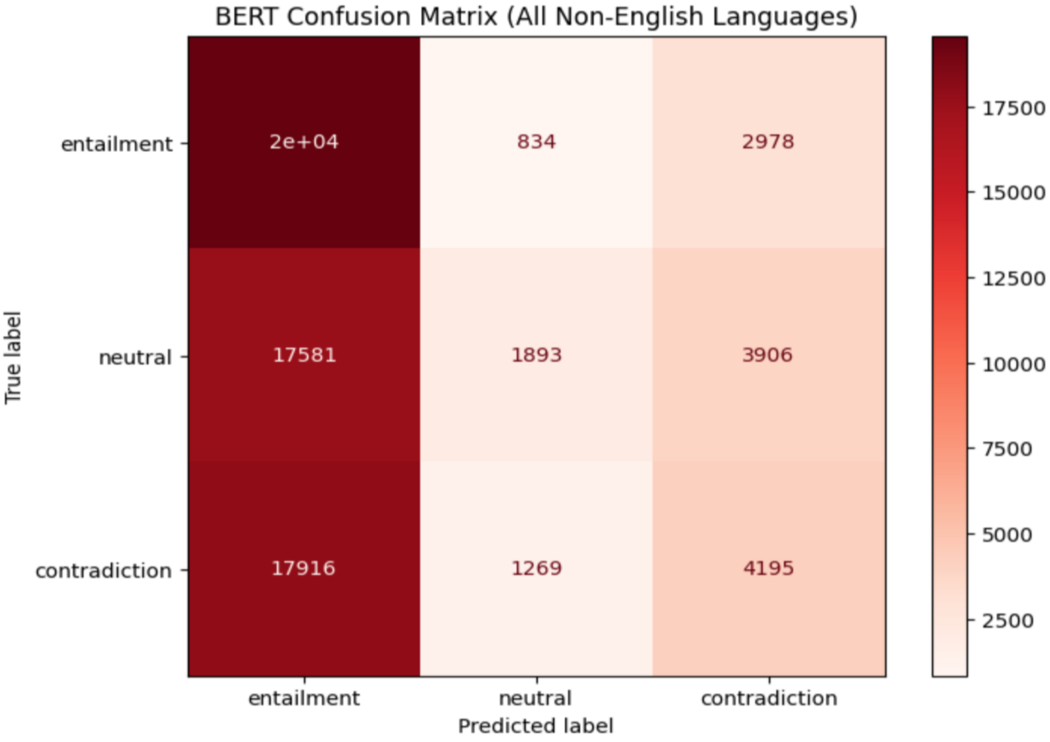
Language: Spanish (es)
Premise: "Un hombre está tocando una guitarra en la calle." *(Translation: "A man is playing a guitar on the street.")*
Hypothesis: "Una persona está haciendo música." *(Translation: "A person is making music.")*
Model Prediction: Entailment
Correct Answer: Entailment
Analysis: This is a strong example of "transfer learning." Even though the model was only fine-tuned on English, xlm-roberta-base understands that "hombre" (man) is a type of "persona" (person) and that "tocando una guitarra" (playing a guitar) is a form of "haciendo música" (making music). The monolingual bert-base-uncased would have seen these as [UNK] tokens and failed.

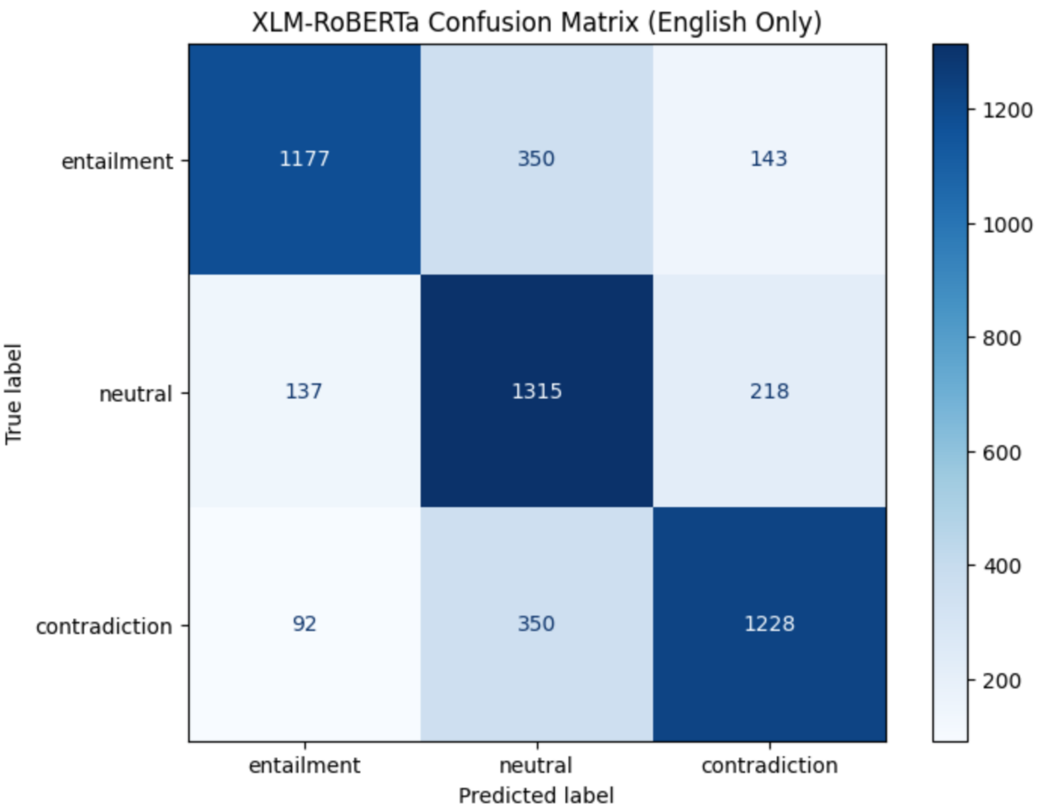
Example of a Failure (XLM-R on a non-English language):

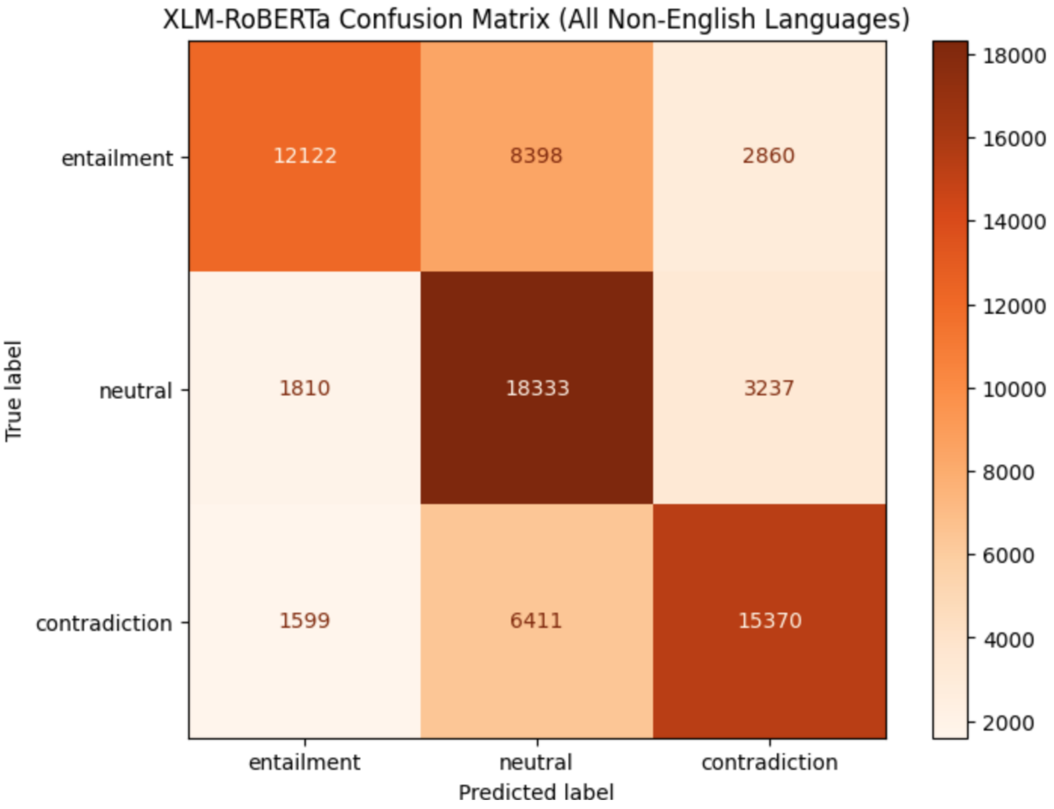
Language: (e.g., Arabic)
Premise: "Le garçon a couru sur la route, mais il n'a pas regardé." *(Translation: "The boy ran onto the road, but he did not look.")*
Hypothesis: "Le garçon est en sécurité." *(Translation : "The boy is safe.")*
Model Prediction: Neutral
Correct Answer: Contradiction
Why it might have failed: This example requires a level of real-world inference. The model correctly understood all the words, but it may have failed to connect the *action* ("ran onto the road without looking") to the *consequence* ("is not safe"). It saw the two sentences as unrelated (Neutral) instead of directly conflicting (Contradiction). This shows that while the model understands language, it can still struggle with complex reasoning.

6. Confusion Matrices


BERT Confusion Matrix (English Only)
True label
entailment
neutral
contradiction
1140
394
136
158
1268
244
93
329
1248
entailment
neutral
contradiction
Predicted label
1200
1000
800
600
400
200


BERT Confusion Matrix (All Non-English Languages)
True label
entailment
neutral
contradiction
2e+04
834
2978
17581
1893
3906
17916
1269
4195
entailment
neutral
contradiction
Predicted label
17500
15000
12500
10000
7500
5000
2500


XLM-RoBERTa Confusion Matrix (English Only)
True label
entailment
neutral
contradiction
1177
350
143
137
1315
218
92
350
1228
entailment
neutral
contradiction
Predicted label
1200
1000
800
600
400
200


XLM-RoBERTa Confusion Matrix (All Non-English Languages)
True label
entailment
neutral
contradiction
12122
8398
2860
1810
18333
3237
1599
6411
15370
entailment
neutral
contradiction
Predicted label
18000
16000
14000
12000
10000
8000
6000
4000
2000

7. What I Learned: Conclusion

This project clearly showed that **multilingual pre-training is essential for cross-lingual tasks.**

A monolingual model like bert-base-uncased is excellent at its one language, but it cannot generalize to others. A multilingual model like xlm-roberta-base can be fine-tuned on one language (like English) and still successfully perform that task in many other languages. This shows the power of its shared, multilingual understanding of language.