Working paper · Internal review | v1.0 | 2026-11-20

How many Jamaican Patois examples close the gap between two language models?

A secondary analysis of nine training conditions in the JamPatoisNLI inference benchmark

Adrian Dunkley | info@caribbeanaiassociation.com

Abstract

We examined whether an English-trained model retained its initial lead over a multilingual model as both received more labelled Jamaican Patois examples. We reconstructed the difference in test accuracy across nine conditions from JamPatoisNLI. English RoBERTa led XLM-RoBERTa by 11.50 percentage points without local training pairs and by 1.33 points with 249 pairs. The gap widened slightly at some intermediate settings. These results concern written sentence inference and provide no direct measure of speech recognition or dialogue quality. [1]

Research question

On the same held-out Jamaican Patois sentences, how did the difference between English RoBERTa and multilingual XLM-RoBERTa change with the number of locally labelled training pairs?

Data and methods

Armstrong, Hewitt and Manning created sentence pairs for a three-way test. The model reads a premise and a second statement. It assigns the second statement to one of three classes: follows from the premise, conflicts with it, or remains undecided. The authors collected premises from published writing, mostly social media. A fluent Patois speaker wrote the second statements. The released dataset documents the classes and the split. [1, 2]

We extracted the two unfrozen RoBERTa columns from Table 9. Both models had already trained for inference on English text. We retained every listed training size for these two models. Each training unit in the table supplies one example per class, so we multiplied the reported count by three. Each accuracy is the published mean across three runs on the same 200 held-out sentence pairs. [1] The final row reports 83 class triples, equivalent to 249 examples by this rule. Table 5 lists 250 available training pairs (84, 83 and 83 by class), and the paper also describes the full condition as 250. Our totals report the balanced triple count. The table does not establish whether the extra pair entered an individual run. [1]

For each training size, we subtracted multilingual accuracy from English-model accuracy. We also identified the first tested size where each published mean reached 70% and 75%. These reference levels describe the observed curve. They carry no claim about the accuracy an application requires.

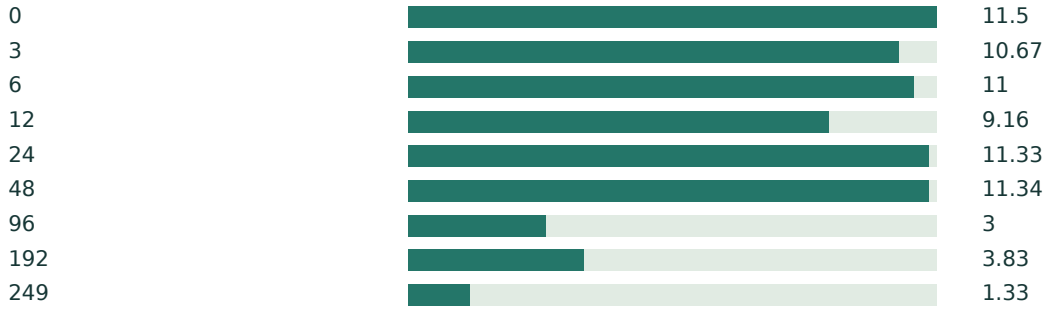
Results

With zero local pairs, English RoBERTa scored 67.50% and XLM-RoBERTa scored 56.00%. With 249 pairs, the scores were 76.50% and 75.17%. The English model gained 9.00 points. The multilingual model gained 19.17. The difference in gains accounts for the 10.17-point reduction in the original gap. [1]

The gap reached 11.34 points at 48 pairs, fell to 3.00 at 96, rose to 3.83 at 192, and fell to 1.33 at 249. The English model first reached 70% at 48 pairs and 75% at 192. The multilingual model first reached those reference levels at 96 and 249 pairs. These are observed training sizes. An untested intermediate size could have reached either level. [1]

English model’s lead at each local training size

percentage points



Labels give the number of Jamaican Patois training pairs. Each bar equals English RoBERTa accuracy minus XLM-RoBERTa accuracy. Both models used the same test set. Source: Armstrong et al., Table 9 [1].

Discussion

A selection based only on the zero-pair result would favour the English model by a wide margin. A team able to label 249 examples would face a much smaller observed difference. The data support testing several training budgets before choosing a model for this task. They do not tell a team which model is better at a different task, such as recognising a caller’s words.

The largest drop in the gap between two adjacent conditions occurred between 48 and 96 pairs. Replication could focus its sample sizes in that interval. Multiple independent draws of training examples would help distinguish the effect of the number of labels from the effect of which examples happened to enter a training run.

Limitations

Both models answered the same test items, and the published means combine repeated runs. The aggregate values cannot support a paired confidence interval for their difference. Raw predictions and training-run identifiers would be needed. The 1.33-point final lead therefore remains a descriptive result.

The test measures written inference in a small, selected corpus. It does not sample Jamaican speakers by parish, age or language use. Newer conversational models, spontaneous speech, spelling changes and code-switching require separate evaluation. [1, 2]

Reproduce the analysis

Read Table 9 of source [1]. Extract the last two model columns for every row. Preserve their order.

Multiply the first column by three. Subtract XLM-RoBERTa accuracy from English RoBERTa accuracy and round to two decimals.

Read down each accuracy column. Identify the first row at or above 70% and the first at or above 75%. Use the total number of local pairs for both comparisons.

Data table

Examples per class	Total local pairs	English RoBERTa accuracy (%)	XLM-RoBERTa accuracy (%)	Difference (percentage points)
0	0	67.5	56.0	11.5
1	3	68.17	57.5	10.67
2	6	69.17	58.17	11.0
4	12	66.83	57.67	9.16
8	24	68.83	57.5	11.33
16	48	70.17	58.83	11.34
32	96	73.0	70.0	3.0
64	192	76.33	72.5	3.83
83	249	76.5	75.17	1.33

References

[1] Armstrong, R.-A., Hewitt, J. and Manning, C. (2022). JamPatoisNLI: A Jamaican Patois Natural Language Inference Dataset. Findings of EMNLP, Tables 5 and 9.

[2] Armstrong, R.-A. JamPatoisNLI dataset and documentation.

Data credits

Armstrong, Hewitt & Manning (2022), Table 9. CAIA calculations. CC BY 4.0