Documento de trabajo · Revisión interna | v1.0 | 2026-11-20

¿Cuántos ejemplos de patuá jamaicano reducen la diferencia entre dos modelos de lenguaje?

Análisis secundario de nueve condiciones de entrenamiento en la prueba de inferencia JamPatoisNLI

Adrian Dunkley | info@caribbeanaiassociation.com

Resumen

¿Conserva un modelo entrenado en inglés su ventaja sobre un modelo multilingüe cuando ambos aprenden con ejemplos etiquetados de patuá jamaicano? Reconstruimos la diferencia de exactitud en nueve condiciones de JamPatoisNLI. RoBERTa en inglés aventajó a XLM-RoBERTa en 11,50 puntos porcentuales sin pares locales y en 1,33 puntos con 249 pares. En algunas condiciones intermedias la diferencia aumentó ligeramente. La prueba mide inferencia escrita. No evalúa directamente el reconocimiento de voz ni la calidad de una conversación. [1]

Pregunta de investigación

En las mismas frases de prueba en patuá jamaicano, ¿cómo cambió la diferencia entre RoBERTa en inglés y XLM-RoBERTa multilingüe al aumentar el número de pares locales etiquetados?

Datos y métodos

Armstrong, Hewitt y Manning crearon pares de frases para una prueba con tres respuestas posibles. El modelo lee una premisa y un segundo enunciado. Debe decidir si el segundo se deduce de la premisa, la contradice o queda sin resolver. Los autores recogieron las premisas de textos publicados, sobre todo de redes sociales. Una hablante de patuá escribió los segundos enunciados. El conjunto de datos describe las clases y la división de los ejemplos. [1, 2]

Extrajimos las dos columnas de RoBERTa con parámetros ajustables de la tabla 9. Ambos modelos ya se habían entrenado en inferencia con textos en inglés. Conservamos todos los tamaños de entrenamiento publicados para esos modelos. Cada unidad aporta un ejemplo por clase, por lo que multiplicamos el número indicado por tres. Cada valor es la media de tres ejecuciones sobre los mismos 200 pares reservados para la prueba. [1] La última fila indica 83 tríos, equivalentes a 249 ejemplos según esta regla. La tabla 5 enumera 250 pares disponibles (84, 83 y 83 por clase), y los autores también describen la condición completa como 250. Usamos el total de los tríos equilibrados. La tabla no permite saber si el par adicional entró en una ejecución. [1]

Para cada tamaño, restamos la exactitud del modelo multilingüe a la del modelo en inglés. Identificamos también el primer tamaño ensayado en que la media publicada alcanzó el 70% y el 75%. Estos valores describen la curva observada. No definen requisitos para una aplicación.

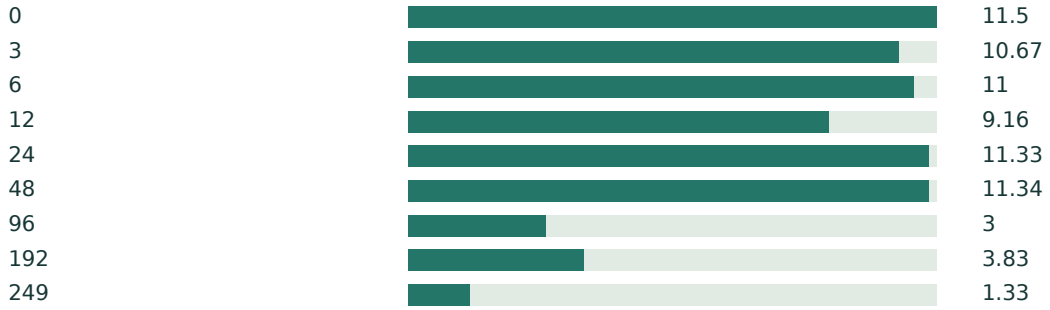
Resultados

Sin pares locales, RoBERTa en inglés obtuvo el 67,50% y XLM-RoBERTa el 56,00%. Con 249 pares alcanzaron el 76,50% y el 75,17%. El primero ganó 9,00 puntos y el segundo 19,17. La diferencia entre ambas ganancias explica la reducción de 10,17 puntos en la ventaja inicial. [1]

La ventaja llegó a 11,34 puntos con 48 pares, bajó a 3,00 con 96, subió a 3,83 con 192 y cayó a 1,33 con 249. El modelo en inglés alcanzó primero el 70% con 48 pares y el 75% con 192. El multilingüe lo hizo con 96 y 249 pares. Un tamaño intermedio que no se ensayó podría haber alcanzado esos valores. [1]

Ventaja del modelo en inglés según el número de pares locales

puntos porcentuales



Las etiquetas indican el número de pares de entrenamiento en patuá jamaicano. Cada barra muestra la exactitud de RoBERTa en inglés menos la de XLM-RoBERTa. Ambos usaron la misma prueba. Fuente: Armstrong y colaboradores, tabla 9 [1].

Discusión

Una selección basada solo en el resultado sin pares locales favorecería con claridad al modelo en inglés. Para un equipo capaz de etiquetar 249 ejemplos, la diferencia observada sería mucho menor. Conviene comparar ambos modelos con distintos presupuestos de etiquetado. El estudio no permite elegir entre ellos para reconocer la voz de una persona que llama.

La mayor caída entre condiciones consecutivas ocurrió al pasar de 48 a 96 pares. Una réplica podría concentrar allí sus tamaños de muestra. Varias selecciones independientes de ejemplos ayudarían a distinguir el efecto del número de etiquetas del efecto de los ejemplos concretos que entran en cada ejecución.

Limitaciones

Ambos modelos respondieron a los mismos elementos de prueba y las medias agrupan ejecuciones repetidas. Para calcular un intervalo de confianza pareado hacen falta las predicciones individuales y los identificadores de ejecución. La ventaja final de 1,33 puntos es descriptiva.

La prueba mide inferencia escrita en una colección pequeña y seleccionada. No representa por muestreo las parroquias de Jamaica, los grupos de edad ni los usos de la lengua. Los modelos conversacionales recientes, el habla espontánea, las variantes ortográficas y la alternancia entre lenguas requieren evaluaciones propias. [1, 2]

Reproducir el análisis

En la tabla 9 de [1], copia las dos últimas columnas de modelos de todas las filas. Conserva el orden.

Multiplica por tres la primera columna. Resta la exactitud de XLM-RoBERTa a la de RoBERTa en inglés y redondea a dos decimales.

En cada columna de exactitud, busca la primera fila que alcance el 70% y la primera que alcance el 75%. Para cada comparación, usa el total de pares locales.

Tabla de datos

Ejemplos por clase	Pares locales totales	Exactitud de RoBERTa en inglés (%)	Exactitud de XLM-RoBERTa (%)	Diferencia (puntos porcentuales)
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

Referencias

[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.

Créditos de los datos

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