Large Language Models (LLMs) de código abierto para la anotación de contenido político: Evidencia de toxicidad e incivilidad digital

Palabras clave: inteligencia artificial, LLMs, toxicidad, incivilidad

Abstract

Este trabajo evalúa la capacidad de 16 Large Language Models (LLMs) de código abierto para realizar tareas de anotación de contenido político. Los modelos se desplegaron localmente para identificar toxicidad e incivilidad en la esfera digital sobre un novedoso conjunto de datos de eventos de protesta que comprende casi cinco millones de interacciones digitales. Los resultados muestran que Nous-Hermes 2 y sus versiones mejoradas superan las anotaciones con clasificación zero-shot de otros LLMs. Además, Mistral-Openorca, con un número menor de parámetros, es capaz de realizar la tarea con un alto rendimiento, lo que confirma cómo algunos modelos pequeños afinados desplegados localmente podrían ofrecer un buen balance entre rendimiento, costos de implementación y tiempo de procesamiento.

Datos y tarea de clasificación

Ejecutamos una clasificación zero-shot para identificar toxicidad en mensajes publicados en Twitter, renombrado como X, usando diversos LLMs de código abierto desplegados localmente con la temperatura al mínimo para asegurar replicabilidad (Gruber y Weber, 2024; Weber y Reichardt, 2023)[1]. La clasificación fue ejecutada sobre una muestra balanceada de un conjunto de datos novedosos de casi cinco millones de mensajes de tres eventos de protesta en América Latina: (a) protestas contra el coronavirus y las medidas de la reforma judicial en Argentina durante agosto de 2020; (b) protestas contra los recortes presupuestarios en educación en Brasil durante mayo de 2019; y (c) el estallido social en Chile derivado de las protestas contra el alza en el pasaje de metro en octubre de 2019[2].

Primero, ejecutamos una clasificación de toxicidad en todo el conjunto de datos que comprende mensajes publicados en español y portugués, usando el algoritmo de toxicidad de Perspective desarrollado por Jigsaw y Google. Nuestro conjunto de datos incluye mensajes de Argentina

[1] Es importante destacar que, aunque no hay acuerdo sobre el rol de la temperatura en las alucinaciones de los modelos, mantenerla al mínimo constriñe las respuestas creativas y en efecto debiese asegurar replicabilidad en una tarea de clasificación, en particular al utilizar estructuras zero o few-shot.

[2] Descargamos todos los mensajes usando hashtags que alcanzaron más de 50.000 publicaciones durante los meses indicados en cada país. Para esto utilizamos el ya descontinuado acceso de investigación académica de la API de Twitter y descargamos los mensajes en formato JSON.

(*n* = 551.761), Brasil (*n* = 1.272.148) y Chile (*n* = 3.125.254). Las puntuaciones AUC-ROC para toxicidad son 0.94 para español y 0.88 para portugués.

La clasificación con Perspective API implicó 2.411 horas de cómputo y el proceso fue desplegado completamente en una Raspberry Pi 5, un microcomputador con una CPU ARM de bajo consumo eléctrico. Esto nos permitió reducir nuestra huella de carbono en un 96% para las tareas de clasificación. Utilizamos Perspective API como línea base para nuestra evaluación comparativa con LLMs de código abierto. En consecuencia, seleccionamos una muestra aleatoria balanceada de 1.000 mensajes sobre los cuales ejecutamos una clasificación zero-shot utilizando las definiciones centrales de toxicidad del equipo de Jigsaw y Google y la siguiente instrucción como mensaje de sistema: "*Clasifica la categoría del comentario como 'tóxico' o 'no tóxico'. Tóxico: Comentarios groseros, irrespetuosos o poco razonables que pueden hacer que alguien abandone la discusión o deje de compartir su punto de vista. No tóxico: Comentarios civiles o agradables que probablemente no desalienten la conversación*".

Análisis de la tasa de error y distancia de Jaccard

En el trabajo se encuentran disponibles una serie de indicadores de rendimiento y bondad de predicción con validación cruzada para el análisis de las tasas de error de los LLMs. También analizamos el desempeño de los modelos con el tiempo de cómputo y el número de parámetros de entrenamiento. De momento, presentamos este mapa de calor de las distancias de Jaccard entre las anotaciones de cada LLM. La tendencia es que los modelos con peor rendimiento tienden a ser más similares, mientras que, curiosamente, los modelos con mejor rendimiento muestran índices en torno a 0,30, por tanto, no son muy similares a pesar de su buen rendimiento común. Esto implica que es posible realizar anotaciones conjuntas de LLMs o ensamblar distintos clasificadores.