

Facebook entrena inteligencia artificial para detectar memes de odio

Facebook dio a conocer una iniciativa el martes para hacer frente los «memes de odio» usando inteligencia artificial respaldada por la colaboración externa (crowdsourcing) para identificar ese tipo de publicaciones.

La red social explicó que ya creó una base de datos de 10.000 memes -imágenes a veces con texto para transmitir un mensaje específico que se presume humorístico- como parte de la intensificación de sus acciones contra los discursos de odio.

Facebook dijo que está dando acceso a investigadores a esa base de datos como parte de un «desafío de memes de odio» para desarrollar algoritmos mejorados que permitan detectar mensajes visuales con contenido odioso, con un premio de 100.000 dólares.

«Estos esfuerzos estimularán a la comunidad de investigadores de IA en general a probar nuevos métodos, comparar su trabajo y cotejar sus resultados para acelerar el trabajo en la detección del discurso de odio multimodal», dijo Facebook en una publicación de blog.

La red se está apoyando fuertemente en la inteligencia artificial para filtrar contenido cuestionable durante la pandemia de coronavirus, que ha reducido su capacidad de moderación humana como resultado de los confinamientos.

El informe de transparencia trimestral de la compañía detalla que Facebook eliminó unos 9,6 millones de publicaciones por violar las políticas de «discurso de odio» en los primeros 3 meses de este año, incluyendo 4,7 millones de contenidos «vinculados al odio organizado».

Guy Rosen, vicepresidente de integridad de Facebook, dijo que con inteligencia artificial «podemos encontrar más contenido y ahora podemos detectar casi el 90% del contenido que eliminamos

antes de que alguien nos lo informe».

Facebook se comprometió hace un año a «interrumpir» la conducta de odio organizada tras los letales ataques en una mezquita de Nueva Zelanda, que provocaron un «llamado a la acción» por parte de los gobiernos para frenar la propagación del extremismo en línea.

Los sistemas automatizados y la inteligencia artificial pueden ser útiles, dijo Facebook, para detectar contenido extremista en varios idiomas y analizar texto incrustado en imágenes y videos para comprender su contexto completo.

Mike Schroepfer, director de tecnología de Facebook, dijo a los periodistas en una conferencia telefónica que una de las técnicas que ayudaron a este esfuerzo fue un sistema para identificar imágenes «casi idénticas», con el que es posible vigilar la publicación de imágenes y videos maliciosos con pequeños cambios para evadir la detección.

«Esta tecnología puede detectar coincidencias casi perfectas», dijo Schroepfer.

Heather Woods, profesora de la Universidad del Estado de Kansas que estudia memes y contenido extremista, elogió la iniciativa de Facebook y la inclusión de investigadores externos.

«Los memes son claramente complejos, no solo porque son multimodales, al incorporar imágenes y texto, como señala Facebook, sino porque son contextuales», dijo Woods.

«Me imagino que los matices y la especificidad contextual de los memes seguirán siendo un desafío para Facebook y otras plataformas que buscan eliminar el discurso de odio», añadió.

Con información de El Nacional