HERRAMIENTAS GRATUITAS Y ESTUDIOS CIENTÍFICOS:

Desde Chile, iniciativas buscan combatir la información falsa

FABIÁN RIVAS

Un sitio que permite comprobar si un dato es verídico, un detector de *bots* en redes sociales y una investigación que creará una funcionalidad para distinguir si un texto fue redactado por una IA son algunos proyectos. ALEXIS IBARRA O.

"¿Es cierto que el incendio del Cerro San Cristóbal fue intencionado?", le preguntamos al sitio CheckNews.ai al día siguiente de producirse el siniestro, cuando recién aparecían noticias con esa hipótesis. "Hay evidencia que sugiere que los incendios en el Cerro San Cristóbal podrían haber sido intencionales", se lee a los pocos segundos, precisando que solo hay un 30% de probabilidades de que la información sea falsa.

Más taxativo es cuando se le pregunta si es verdad que Felipe VI abdicó a la corona española: "La probabilidad de que la noticia de que 'Felipe VI haya abdicado a la corona española' sea falsa es del 100%".

CheckNews.ai es una herramienta creada en Chile, gratuita y que usa la inteligencia artificial para desentrañar información falsa y dar un veredicto de si lo que se está consultando es verdadero o no.

Su creador es Daniel Atik —un español radicado en Chile, gerente de producto de CodeGPT—quien comenzó el sitio como un experimento para ver si un modelo grande de lenguaje (LLM), tal como es ChatGPT, puede validar la calidad de una información basándose en información que hay internet.

Funciona así, dice Atik: "Comienzas con una consulta, por ejemplo, '¿Es cierto que tal persona hizo esto?', y la herramienta convierte esa frase en una consulta optimizada para buscadores en internet".

"El resultado es sorprendente porque busca en la información más actualizada que hay en la web y ubica en el tiempo la pregunta para determinar si son hechos recientes o antiguos", aclara.

El sitio, finalmente, responde verificando si es que en sitios confiables aparece algo sobre esa información y entrega un porcentaje de posibilidades de que la información sea verdadera o falsa.

AMENAZA **IMPORTANTE**

El 6 de noviembre, Naciones Unidas publicó un informe y un plan de acción para combatir la desinformación y la manipulación en redes sociales ya que representan "amenazas importantes para la vida en sociedad y la estabilidad", según dice el texto. El informe —producto de 10 mil contribuciones recogidas en 134 países— da cuenta de la importancia que tiene para esta entidad combatir este creciente fenómeno.

De ahí la importancia de estudiarlo en la academia y de desarrollar herramientas poderosas que ayuden a la población a distinguir información verdadera de la que no lo es.

Marcelo Mendoza, investigador de la U. Católica y del Instituto Milenio Fundamentos de los Datos (IMFD), camina por esa senda. "Trabajamos en una herramienta que podría distinguir si un contenido es generado por un humano o por ChatGPT. Esto es relevante ya que con estas herramientas es posible inundar con desinformación las redes sociales", dice el investigador.

"A nivel internacional ya se ha llegado a la conclusión de que un humano no puede distinguir textos generados por ChatGPT", afirma Mendoza. "Entonces, lo que estamos haciendo es usar IA para determinar si el texto fue generado por humanos o por otra IA. Para ello hemos analizado herramientas existentes, como GPTZero, que hacen este trabajo".

El investigador explica que un texto generado por algoritmos deja huellas: "Por ejemplo, el tipo de palabras que se usa y con qué frecuencia. Los humanos nos expresamos en forma más desordenada desde el punto de vista del léxico".

"Al concluir nuestra investigación esperamos crear un detector de textos generados por IA que probablemente sea mejor que GPTZero, ya que hemos analizado en cuáles dimensiones es posible engañar a un detector como ese", precisa Mendoza.

El mismo investigador con la colaboración de Sebastián Valenzuela (U. Católica), Marcelo Santos (U. Diego Portales) y Eliana Providel (U. de Valparaíso), creó un detector de *bots*, pensado para identificar en X (antiguo Twitter) las cuentas controladas por programas automatizados o *bots*. Según Valenzuela, BotCheckCL a diferencia de otras herramientas, "es mucho más preciso detectando *bots* en Chile dado que fue entrenado con contenido y el lenguaje tal como se usa en el país", dice.

Aunque aclara que tras la adquisición de Twitter por Elon Musk cada vez se ha restringido el acceso a los investigadores que quieren realizar estudios sobre *bots* o diseminación de noticias falsas.

En el campo de las investigaciones, el mismo Valenzuela dirige el área científica del International Panel on the Information Environment (IPIE), un organismo con sede en Suiza y cuyo objetivo es proporcionar conocimiento científico sobre las amenazas al entorno de la información en el mundo.

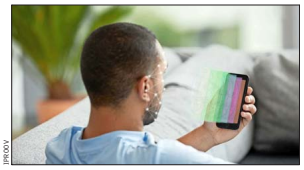
"Estamos recolectando información sobre qué iniciativas funcionan y cuáles no para mitigar el fenómeno de la desinformación", añade Valenzuela. Para ello han estudiado la literatura científica de los últimos 15 años, sobre todo los estudios que analizan empíricamente y con evidencia qué iniciativas dan resultados.

"El tema de fondo no es que circule más o menos información falsa, sino que el fenómeno de la desinformación termina incidiendo en las creencias de las personas, desde sus preferencias políticas, a tomar decisiones importantes concernientes a la salud, como se vio en la pandemia del covid", dice.

En IPIE analizaron 12 soluciones posibles o intervenciones para reducir los efectos de la desinformación en el sistema de creencias de las personas. Entre ellas, las correcciones o *fact checking*, el etiquetado de contenido (por ejemplo, cuando WhatsApp rotula que un mensaje ha sido reenviado muchas veces), redireccionar con información a fuentes confiables, la moderación de contenidos o los programas de alfabetización digital, entre otras.

"De las 12 soluciones posibles encontramos que dos tienen evidencia robusta de que sí funcionan", añade Valenzuela. Estas son las correcciones o *fact checking*, es decir, cuando alguien o un medio corrige la información falsa. Lo otro que da resultado es el etiquetado de información advirtiendo que es un contenido controversial, que se trata de una noticia en proceso o que es generada por un medio estatal, por ejemplo".

"Hay iniciativas, como la alfabetización digital, que uno tiende a pensar de que dan resultados por su mayor cercanía a las personas, pero son poco masivas. Además, las personas no solo comienzan a desconfiar de la información falsa, sino que también de la veraz y emanada por medios reputados", aclara el investigador. "No hay una única solución para mitigar la desinformación. De ahí la importancia de combinar iniciativas dependiendo del contexto", concluye.



IPROOV

El engaño mediante video o *deep fake* incluso puede ser usado para autenticarse en sistemas que usan la biometría para dar acceso a información.

CONTRA VIDEOS ENGAÑOSOS

***Deep fake* (engaño profundo), así se conoce a la técnica con que se crean o modifican videos para que una persona aparezca diciendo o haciendo cosas que no hizo. "Un ejemplo es la aparición de Tom Cruise en un *spot* publicitario que él no grabó ni autorizó la aparición de su imagen. Todo fue realizado con inteligencia artificial", dice Daniel Molina, vicepresidente de iProov para Latinoamérica, empresa con presencia en Chile y que se especializa en biometría.**

Ellos han desarrollado una tecnología que puede detectar videos falsos. "Usamos IA para analizar una serie de características del video, desde sus metadatos hasta las sombras que se reflejan en la cara o pequeños movimientos en los músculos del rostro", aclara.

Esto permite detectar, dice, cuando se "inyecta" un video falso en una videollamada o cuando se quiere usar un video para autenticar a una persona en un sistema de verificación biométrico.