Detectar propaganda en redes sociales con Inteligencia Artificial

ESCUELA DE INGENIERÍA
FACULTAD DE INGENIERÍA

Gracias a herramientas de análisis de texto, los investigadores buscan identificar campañas de propaganda en redes sociales y reconocer su alcance y efecto en los receptores.

1 Departamento de Ciencia de la Computación. Escuela de Ingeniería, Pontificia Universidad Católica de Chile.
2 Facultad de Comunicación y Letras, Universidad Diego Portales.
3 Doctorado en Ciencias de la Ingeniería, Pontificia Universidad Católica de Chile.
4 Doctorado en Ingeniería y Ciencias con la Industria. Pontificia Universidad Católica de Chile.

Investigadores principales

Marcelo Mendoza (1)
Marcelo Santos (2)
Miguel Fernández (3)
Carlos Muñoz (4)

Propaganda y desinformación en internet

Las redes sociales han expandido los daños que pueden provocar las campañas de propaganda encubiertas:

Amplifican la visibilidad de los contenidos por contagio emocional, alcanzando a más individuos.

Fomentan la polarización y radicalización de las opiniones: al propagarse por contagio emocional, disminuyen las barreras analíticas de la información.

¿Qué es la propaganda encubierta?
Son mensajes que buscan manipular la opinión o el comportamiento del receptor. Utilizan principalmente el género noticioso.

La irrupción de la inteligencia artificial

Las herramientas de lenguaje basadas en Inteligencia Artificial, como Chat GPT, han acrecentado aún más el problema. No solo pueden ser un riesgo para el "ecosistema de información", sino para toda la democracia:

Los textos que producen no se distinguen de los generados por humanos.

La producción de mensajes aumenta exponencialmente.

¿Cómo combatirlas?

Hoy existen numerosas iniciativas de "fact checking", pero no dan abasto:

La cantidad de información que circula es inabarcable.

No toda la propaganda se basa en información incorrecta. A veces el enfoque o el lenguaje utilizado puede convertir información verdadera en propaganda encubierta.

Herramientas con modelos IA

Los investigadores están creando herramientas con modelos de lenguaje basados en IA que permitan caracterizar y reconocer técnicas lingüísticas de propaganda. Además, buscan identificar los efectos que las campañas coordinadas tienen en las redes sociales.

El objetivo es combatir estas campañas exponiéndolas y entregando mecanismos a las audiencias para entender la naturaleza de la información que lee.

Herramienta 1

IDENTIFICADOR DE CONTENIDOS

Reconoce y caracteriza el uso de técnicas de propaganda en textos.

El problema de la falta de datos

Para que una herramienta de IA pueda identificar contenidos específicos –como la propaganda encubierta– debe ser "entrenada": se le entregan millones de textos como los que debe detectar, para que aprenda a reconocerlos.

El primer desafío de los investigadores fue generar estos textos o "base de datos".

Fuente de data inicial

Los investigadores partieron de una base de datos de noticias y comentarios publicados en Instagram y Facebook, en la cual estaban demarcadas 18 técnicas de propaganda (como desacreditar al oponente o apelar al miedo).

Pero presentaba dos desafíos:

1

Desbalance
Algunas técnicas estaban sobrerrepresentadas, lo que afecta el aprendizaje del modelo.

2

ENG

Idioma
Los textos estaban en inglés.

Aumentar la data con IA generativa

GPT: Le pidieron que parafraseara los textos, para crear nuevos.

Humanos redactan textos de propaganda y revisan la calidad de los generados con IA.

GPT: Pidieron una traducción al español, para entrenar la herramienta en ambos idiomas.

Crear esta instrucción para GPT fue clave en la investigación. Usaron la estrategia "sintonización del prompt": afinaron la instrucción de manera de alcanzar un resultado conocido que habían definido previamente.

Herramienta en funcionamiento

Una vez entrenado el algoritmo, la herramienta es capaz de destacar los fragmentos que contienen propaganda e identificar la estrategia usada.

Apelación al miedo

"Debemos detener a esos refugiados, porque son terroristas"

Sobresimplificación

"El Presidente Trump ha estado un mes en el gobierno y el precio del gas se ha disparado"

Herramienta 2

IDENTIFICADOR DE EFECTOS EN EL RECEPTOR

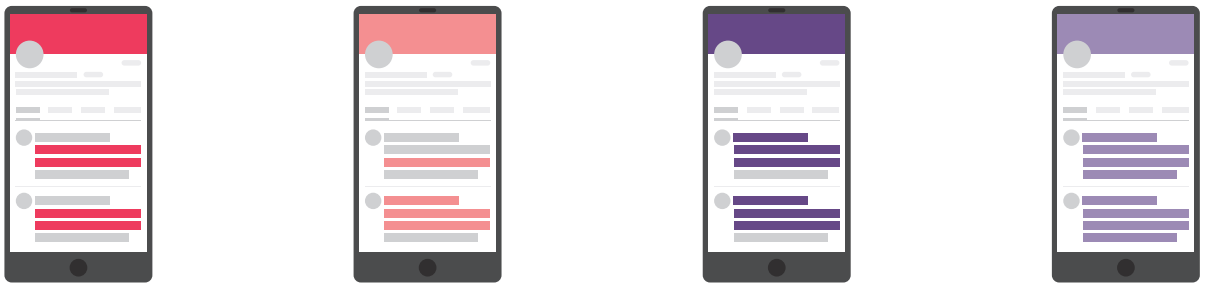
Caracteriza los efectos de la propaganda en el comportamiento en redes sociales del receptor.

Los investigadores se preguntaron si las estrategias de propaganda se replican en los usuarios, es decir, **¿existe el "contagio propagandístico"?**

Y crearon un modelo para identificar distintos perfiles de usuarios, lo que permite reconocer campañas coordinadas:

Se categoriza cada comentario, y luego se pondera el historial del usuario

Las categorías no son excluyentes



Haters

Suelen utilizar un lenguaje grosero, que denigra y ofende a otras personas.

Hiperpartisanos

Se identifican con causas, se articulan en movimientos o partidos políticos.

Usuarios neutrales

Usan un lenguaje respetuoso con las demás personas.

Bots sociales

Cuentas operadas por algoritmos, programadas para un fin específico.

Si la mayoría de los comentarios que replican una estrategia de propaganda corresponden a hiperpartisanos o bots, esto sugiere coordinación.

Caso de estudio

Los investigadores aplicarán ambas herramientas en las elecciones presidenciales de Chile en 2025.

Utilizarán noticias y comentarios publicados en las redes de META y en medios de prensa nacional.

Identificadores abiertos y gratuitos

Las herramientas generadas serán de uso público: estarán disponibles para consultas online de manera gratuita.

Pueden ser especialmente útiles para periodistas, ONGs dedicadas a velar por la transparencia en redes sociales y organizaciones de fact checking.

Los modelos serán de código abierto.

Ojo con:

La información que recibimos siempre contendrá la perspectiva de su autor o, a veces, de quien la comparte.