Taller Aula Invertida: La IA y su papel en la desinformación

MY. (FAC) Mario Francisco Gómez Ortega

Docente: Jaider Ospina Navas

Escuela Superior de Guerra "General Rafael Reyes Prieto"

CEM 2025 – Maestría en Ciberseguridad y Ciberdefensa Nacional

Electiva Habilidades Prácticas en el Ciberespacio

Bogotá D.C., Colombia

2025

Taller Aula Invertida: La IA y su papel en la desinformación

El "aula invertida" (también conocida como Flipped Classroom o aprendizaje invertido) es un modelo pedagógico que cambia el enfoque tradicional de la enseñanza. En esencia, invierte lo que se hace dentro y fuera del aula.

La IA y su papel en la desinformación - Contexto

El documento de estudio profundiza en el creciente problema de la desinformación, resaltando su impacto en áreas cruciales como la salud y la política y distinguiendo entre diferentes tipos de información falsa. El documento explora el doble rol de la Inteligencia Artificial (IA), tanto como facilitadora de la creación y difusión de bulos, especialmente a través de modelos generativos, como herramienta esencial para combatirla. Se presenta FacTeR-Check, una herramienta desarrollada por el grupo AIDA de la Universidad Politécnica de Madrid, que utiliza IA para la verificación y monitoreo de desinformación, destacando su enfoque multilingüe y análisis en redes sociales. Finalmente, se reflexiona sobre el futuro de la IA en esta lucha, considerando avances como la IA explicativa y la importancia de la privacidad.

Cuestionario

1. **¿Cuál es la diferencia fundamental, según el texto, entre "misinformation" y "disinformation"?**

Respuesta:
Según el texto, la diferencia fundamental entre *misinformation* y *disinformation* radica en la intencionalidad con la que se difunde la información falsa:

- Disinformation: Información falsa generada deliberadamente con el fin de causar perjuicio.
- Malinformation: Información que, aun siendo verdadera, se distribuye de una manera sesgada o fuera de contexto con un propósito malicioso.
- Misinformation: Información falsa que se comparte sin intención de causar daño, pues el emisor presume que es verdadera.

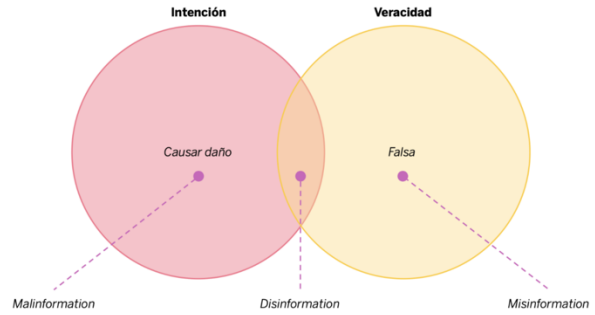


Figura 1. Diferencias entre *misinformation, disinformation* y *malinformation*

Referencias: p. 15, 16.

2. **Según el Reuters Institute Digital News Report 2023, ¿qué tendencia preocupante se observa en España con respecto al interés por las noticias?**

Respuesta:
El Reuters Institute Digital News Report 2023 señala una disminución significativa en el interés por las noticias en España**:** del 85% en 2015 al 51% en 2023**,** es decir, una caída de 34 puntos porcentuales**.** Además, los datos de desconfianza de los lectores en los medios de comunicación llegó a su récord (40%) en los nueve años de encuesta, especialmente entre los menores de 45 años.

Referencia: p. 13.

3. **¿Cómo se comparan, según los experimentos de Vosoughi, Roy y Aral (2018), la velocidad y facilidad de difusión de noticias falsas frente a las verdaderas?**

Respuesta:
Los experimentos muestran que las noticias falsas se difunden más rápido y más fácilmente que las verdaderas. En particular, el 1% de las noticias falsas más difundidas alcanzaron entre 1.000 y 100.000 personas, mientras que el 1% de las noticias verdaderas más difundidas rara vez superó las 1.000 personas**.**

Referencia: p. 21.

4. **¿Qué ventaja clave ofrecen las redes latentes de difusión sobre los modelos epidemiológicos para el estudio de la desinformación?**

Respuesta:
La ventaja clave de las redes latentes de difusión frente a los modelos epidemiológicos es que no solo permiten predecir cómo evolucionará la propagación de piezas de información, sino que también permiten conocer quién propaga la información y cómo lo hace, mientras que los modelos epidemiológicos solo detectan el flujo anómalo sin identificar a los agentes involucrados.

Los modelos epidemiológicos son modelos matemáticos donde la población se divide en diferentes grupos y se definen unas probabilidades para pasar de un grupo a otro. Veamos un ejemplo con el modelo SIR, que no se usa para detectar desinformación per se pero es muy simple y nos permite entender el concepto.

Figura 1.3. Esquema del modelo epidemiológico SIR (susceptible, infectada, recuperada). Los círculos representan las distintas poblaciones y las flechas, la probabilidad de que un individuo de una población pase a otra, es decir, existe una probabilidad β de que un individuo de la población (s)usceptible pase a la población (i)nfectada. Asimismo, existe una probabilidad γ de que un individuo de la población (i)nfectada pase a la población (r)ecuperada

Red latente de difusión; es un modelo generativo de redes sociales que permite modelizar la difusión de información entre los individuos de una red a lo largo del tiempo.

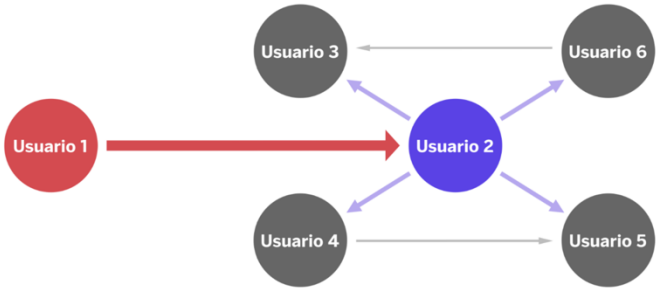


Figura 1.4. Ejemplo de red latente de difusión compuesta por seis usuarios. Las flechas indican la dirección en la que fluye la información, y su grosor es proporcional a la intensidad de la influencia. Así, el Usuario 1 es capaz de influir con más intensidad en el Usuario 2 comparado con la capacidad de influencia que tiene el Usuario 4 sobre el Usuario 5

Referencia: p. 23, 24.

5. ¿Qué son los "grandes modelos de lenguaje" y cuál es su principal riesgo en el contexto de la desinformación?

Respuesta:
Los grandes modelos de lenguaje (Large Language Models - LLM, por sus siglas en inglés) son modelos de IA entrenados con grandes cantidades de texto que pueden generar contenido textual de alta calidad. Su principal riesgo es que pueden generar desinformación de apariencia realista en tiempo récord, facilitando campañas de desinformación a gran escala en malas manos.

Referencia: p. 28.

6. ¿Cómo facilita la accesibilidad de los modelos de IA la generación de desinformación?

Respuesta:

La accesibilidad de los modelos de inteligencia artificial facilita la generación de desinformación porque reduce considerablemente la barrera técnica y económica para su uso. Gracias a los avances de la tecnología actual, ya no se necesitan conocimientos especializados ni infraestructura costosa para utilizar herramientas capaces de generar contenido falso de alta calidad.

En el documento se señala que, en cuestión de semanas, se ha pasado de necesitar "equipos profesionales de alto rendimiento y amplios conocimientos de informática" a poder ejecutar estos modelos "en equipos portátiles domésticos de alta gama con un conocimiento mínimo" (p. 27, párrafo 2). Este fenómeno ha "bajado la barrera de entrada de manera significativa", facilitando así su uso tanto para fines creativos legítimos como para actividades maliciosas.

Además, al estar disponibles modelos de código abierto como Stable Diffusion o GPT-j, y plataformas de trabajo colaborativo como Hugging Face, cualquier persona con acceso a internet puede clonar, adaptar y usar estos modelos para diseminar falsedades.

El riesgo se acentúa cuando se combinan estos modelos con plataformas de la gig economy (como Fiverr o Mechanical Turk), que permiten contratar tareas de generación y difusión de contenido por muy bajo costo y de forma casi anónima. Esto ha facilitado la proliferación de campañas de desinformación organizadas sin requerir grandes inversiones (p. 25, párrafo 2).

Referencias: p.25 y p.27

7. ¿Qué son las "cajas negras" en el contexto de la IA explicativa y cuál es el desafío asociado?

Respuesta:

En el contexto de la Inteligencia Artificial Explicativa (XAI, por sus siglas en inglés), las "cajas negras" hacen referencia a modelos de IA, especialmente redes neuronales profundas y modelos Transformer, cuyos procesos internos son tan complejos (por manejar millones de parámetros) que no permiten entender cómo se llega a una determinada decisión o resultado**.**

Según el texto, las "cajas negras" en IA se refieren a modelos complejos (como los basados en deep learning) cuyos procesos internos no son comprensibles o transparentes para los humanos.

El desafío principal es que, al no entender cómo se toman las decisiones, es difícil confiar, validar o corregir los resultados, lo que limita la explicabilidad y la adopción segura de estos sistemas, es decir; la falta de transparencia: si no se puede explicar cómo y por qué el modelo toma una decisión, entonces es muy difícil confiar en sus resultados, auditar su comportamiento, corregir sesgos o hacer pedagogía con los usuarios**.** En la lucha contra la desinformación, esto es especialmente problemático, pues se requiere confianza pública y trazabilidad.

Referencias: p.48

8. ¿Qué implicaciones tiene el concepto de "Inteligencia Artificial General (AGI)" para la lucha contra la desinformación?

Respuesta:
La AGI (Inteligencia Artificial General) es una IA capaz de pensar y aprender de forma similar a un humano**,** con competencia en múltiples tareas. En el contexto de la desinformación, su desarrollo implica una doble implicación:

1. Mayor capacidad para generar desinformación más engañosa.
2. Mayor capacidad para detectarla y verificarla automáticamente.

Sin embargo, en el texto también se menciona que el desinformador siempre llevará una ventaja, pues es quien actúa en este intercambio de información antes de ser verificado.

Referencia: p.49, párrafo 4.

9. ¿Qué normativas europeas importantes se mencionan en relación con la IA y la privacidad?

Respuesta:
Se mencionan dos normativas europeas clave:

1. Reglamento General de Protección de Datos (RGPD)**:** Aprobado el 27 de abril de 2016, establece el derecho a la protección de los datos personales y regula los principios, derechos y obligaciones en su tratamiento (Comisión Europea 2016).
2. Libro Blanco de la Inteligencia Artificial**:** Se enmarca en la Estrategia Digital de la Unión Europea. Esta describe el desarrollo ético y seguro de la IA en Europa, con énfasis en la privacidad, explicabilidad, seguridad y supervisión humana (Comisión Europea 2020).

Referencias: p.49 y 50.

10. ¿Cómo garantiza FacTeR-Check el cumplimiento de la normativa de protección de datos al analizar redes sociales? Preguntas de Formato Ensayo.

Respuesta:
FacTeR-Check garantiza el cumplimiento del RGPD y demás normativas de protección de datos mediante:

- El uso exclusivo de perfiles públicos en redes sociales.
- El análisis de tuits representativos**,** sin incluir datos personales sensibles.
- El diseño desde el inicio bajo los principios de privacidad por defecto y por diseño, garantizando transparencia, explicabilidad y respeto a los derechos fundamentales.

Referencias: p.49 y 50.

11. Analice las diferentes formas en que la Inteligencia Artificial puede ser utilizada tanto para generar como para combatir la desinformación, basándose en los ejemplos y conceptos presentados en el texto.

Respuesta:

La inteligencia artificial (IA) desempeña un rol doble en el ecosistema de la desinformación: puede ser una herramienta poderosa tanto para su generación como para su detección y combate. Por un lado, el desarrollo de modelos generativos como los grandes modelos de lenguaje (LLM), generadores de imágenes y video, y los modelos de síntesis de voz permite a actores maliciosos crear contenido falso altamente convincente con muy pocos recursos. Por ejemplo, la síntesis de texto falso**,** los deepfakes visuales y auditivos, y la combinación de múltiples modelos para crear entrevistas ficticias o discursos manipulados son técnicas ampliamente utilizadas para propagar falsedades de forma creíble y masiva.

Por otro lado, herramientas como FacTeR-Check utilizan la misma tecnología IA (especialmente los modelos Transformer) para combatir la desinformación a través de análisis de similitud semántica, inferencia del lenguaje natural y verificación contra bases de datos de hechos verificados. Esta herramienta permite automatizar parte del trabajo de las organizaciones de fact-checking, aumentando la eficacia y escalabilidad en la lucha contra los bulos.

En conclusión, la IA representa un campo de batalla tecnológico tanto como generadores así como verificadores, los cuales se perfeccionan mutuamente. Mientras que unos producen contenido falso más sofisticado, otros desarrollan sistemas más precisos para detectarlo y refutarlo.

12. Discuta el papel de la Inteligencia Artificial Explicativa (XAI) en la mejora de la confianza pública en los sistemas de detección de desinformación y en la educación de los usuarios. ¿Cuáles son los principales obstáculos para su desarrollo?

Respuesta:

La Inteligencia Artificial Explicativa (XAI) es esencial para mejorar la confianza pública en los sistemas de detección automática de desinformación. Su función principal es

proporcionar transparencia, es decir, explicar por qué un sistema toma una decisión concreta, como clasificar un contenido como desinformación. Esta capacidad explicativa permite a los usuarios entender no solo el resultado, sino también el razonamiento detrás de él, lo cual es crucial para evitar la percepción de arbitrariedad (p. 48).

Además, la XAI tiene un potencial pedagógico de educar a los usuarios, ayudándolos a identificar patrones de desinformación por sí mismos, lo que refuerza la alfabetización mediática. También permite mejorar la toma de decisiones de los científicos de datos y diseñadores de sistemas (p. 48).

Sin embargo, los principales obstáculos para el desarrollo de la XAI radican en la complejidad técnica de los modelos actuales, conocidos como "cajas negras", que operan con millones de parámetros y son inherentemente difíciles de interpretar. Por lo tanto, avanzar hacia una IA más explicable requiere superar barreras tecnológicas significativas sin sacrificar la precisión y el control humano, lo cual sigue siendo un reto en la investigación contemporánea.

Referencias: p.48

13. Compare los modelos epidemiológicos y las redes latentes de difusión como enfoques para estudiar la propagación de la desinformación en las redes sociales. ¿Qué información específica puede obtenerse de cada tipo de modelo?

Respuesta:

Los modelos epidemiológicos y las redes latentes de difusión representan dos enfoques complementarios para el estudio de la propagación de la desinformación en redes sociales.

Los modelos epidemiológicos, como el modelo SIR, tratan la desinformación como una enfermedad que se propaga de un usuario a otro. Permiten predecir la evolución global del fenómeno, es decir, cuántos usuarios estarán expuestos a una pieza de desinformación en determinado momento. Sin embargo, tienen una limitación crítica: son anónimos y no identifican quiénes propagan la desinformación ni cómo lo hacen (p. 23)

Figura 1.3. Esquema del modelo epidemiológico SIR (susceptible, infectada, recuperada). Los círculos representan las distintas poblaciones y las flechas, la probabilidad de que un individuo de una población pase a otra, es decir, existe una probabilidad β de que un individuo de la población (s)usceptible pase a la población (i)nfectada. Asimismo, existe una probabilidad γ de que un individuo de la población (i)nfectada pase a la población (r)ecuperada

En cambio, las redes latentes de difusión permiten mapear las relaciones específicas entre usuarios, identificando quién influye en quién y con qué intensidad. Esta capacidad permite

señalar a los usuarios más influyentes, distinguir entre iniciadores e intermediarios de los bulos y analizar la estructura de la red de propagación (p. 24).

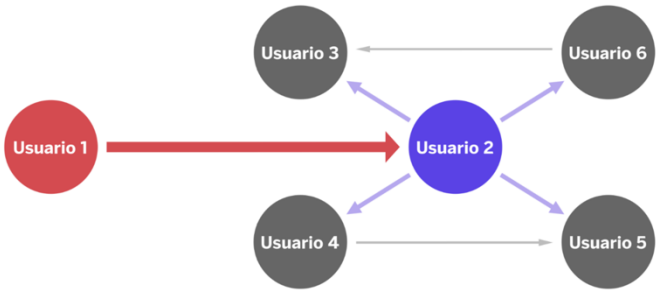


Figura 1.4. Ejemplo de red latente de difusión compuesta por seis usuarios. Las flechas indican la dirección en la que fluye la información, y su grosor es proporcional a la intensidad de la influencia. Así, el Usuario 1 es capaz de influir con más intensidad en el Usuario 2 comparado con la capacidad de influencia que tiene el Usuario 4 sobre el Usuario 5

En resumen, los modelos epidemiológicos ofrecen una visión agregada y predictiva del fenómeno, mientras que las redes latentes permiten una comprensión detallada y operativa sobre las dinámicas individuales de propagación.

14. Examine la relación entre la accesibilidad de las herramientas de IA generativa y el aumento potencial de la desinformación. ¿Qué estrategias se sugieren para mitigar este riesgo?

Respuesta:

La creciente accesibilidad de las herramientas de IA generativa ha reducido los obstáculos técnicos y económicos para crear contenido falso. Hoy en día, cualquier usuario con un computador portátil puede generar textos, imágenes, audios o vídeos falsos usando modelos abiertos como Stable Diffusion o interfaces accesibles como **ChatGPT**.

Esta democratización del acceso ha contribuido al auge de la desinformación, ya que permite a actores no especializados crear contenidos persuasivos a gran escala y en múltiples idiomas. Asimismo, la economía de microencargos (gig economy) facilita contratar personas para generar o propagar desinformación de forma anónima (p. 25).

Para mitigar este riesgo, según el texto, se podría proponer:

- Desarrollar métodos ágiles y escalables de detección (p. 27).
- Apostar por herramientas como FacTeR-Check que combinan análisis semántico, inferencia lingüística y bases de datos verificadas (pp. 37 - 41).
- Adoptar medidas de regulación tecnológica y fomentar la colaboración entre instituciones, academia y desarrolladores de IA (p. 51, párrafo 3).

15. **Analice las consideraciones éticas y de privacidad asociadas con el uso de la Inteligencia Artificial para combatir la desinformación, haciendo referencia a las normativas europeas mencionadas e identificado si existen normativas en nuestro país similares.**

Respuesta:

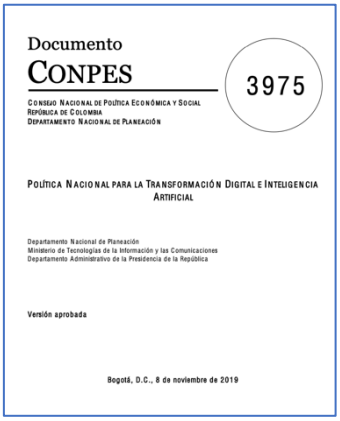
El uso de IA para combatir la desinformación plantea retos éticos y de privacidad significativos. El análisis de contenido generado por usuarios en redes sociales puede implicar la recolección y procesamiento de datos personales, lo cual debe hacerse bajo estrictos principios de legalidad, transparencia y proporcionalidad.

El texto señala como marco normativo fundamental el Reglamento General de Protección de Datos (RGPD) de la Unión Europea, que establece que toda persona tiene derecho a la protección de sus datos personales (p. 49). Además, el Libro Blanco sobre la IA (2020) fija principios como la explicabilidad, la supervisión humana y el respeto a los derechos fundamentales (p. 50).

El FacTeR-Check se alinea con estos principios al operar solo sobre datos públicos y asegurarse de que los procesos de inferencia y perfilado cumplan con los estándares de transparencia y protección de derechos (p. 50).

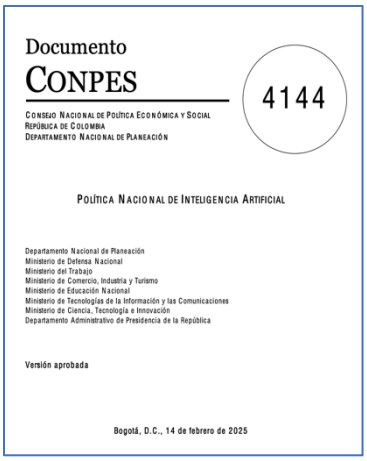
En cuanto a Colombia, si bien no se menciona en el texto, el país cuenta con una normativa vigente, como la Ley 1581 de 2012 sobre protección de datos personales, y su reglamentación en el Decreto 1377 de 2013, que se alinean en parte con el RGPD europeo, aunque con menor especificidad respecto al uso de IA.

Igualmente, se cuenta con:


Documento
CONPES
3975
Consejo Nacional de Política Económica y Social
República de Colombia
Departamento Nacional de Planeación
Política Nacional para la Transformación Digital e Inteligencia Artificial
Departamento Nacional de Planeación
Ministerio de Tecnologías de la Información y las Comunicaciones
Departamento Administrativo de la Presidencia de la República
Versión aprobada
Bogotá, D.C., 8 de noviembre de 2019

- CONPES 3975 de 2019 Política Nacional para la Transformación Digital e Inteligencia Artificial: Este documento establece lineamientos para el desarrollo responsable de la IA en Colombia. Promueve el uso ético de la tecnología, destacando principios como la transparencia, la privacidad, la responsabilidad, la equidad y la explicabilidad**.**

- Estrategia Nacional de IA "Colombia hacia una sociedad e inteligencia artificial" (2020): Inspirada en la OECD y la UNESCO, este plan orienta el desarrollo de IA centrada en el ser humano. Reconoce la necesidad de garantizar la explicabilidad de los algoritmos, evitar sesgos, y respetar los derechos fundamentales, incluyendo la libertad de expresión.
- Hoja de Ruta Adopción Ética y Sostenible de Inteligencia Artificial Colombia (Febrero de 2024). La hoja de ruta tiene objetivos ambiciosos que reflejan la visión de una Colombia líder en la adopción ética y sostenible de la inteligencia artificial.



- CONPES 4144 de 2025: marca un avance decisivo hacia una gobernanza ética y responsable de la IA en Colombia. Su implementación refuerza el marco normativo para el desarrollo de sistemas de detección de desinformación, alineándolos con estándares internacionales de protección de datos, transparencia, equidad y derechos humanos. Cualquier iniciativa tecnológica en este campo deberá cumplir no solo con la Ley 1581 y el RGPD europeo (como se exige en el informe), sino también con los nuevos estándares nacionales de evaluación de impacto ético, supervisión humana y protección a poblaciones vulnerables.

Muchas gracias.