Aula Invertida: Luchando Contra la Desinformación Mediante la Inteligencia Artificial

MY. Carlos Augusto Uribe Vergara

Escuela Superior de Guerra “General Rafael Reyes Prieto”

Curso de Estado Mayor - CEM 2025

Electiva - Habilidades Prácticas en el Ciberespacio

Mag. Jaider Ospina Navas

01 de julio de 2025

Luchando Contra la Desinformación Mediante la Inteligencia Artificial

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.

En concordancia con el contexto anterior se van a realizar las siguientes preguntas:

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

   La diferencia radica principalmente en la intencionalidad con la que se difunde la información falsa. Misinformation se refiere a información incorrecta difundida sin intención de causar daño; el emisor la comparte creyendo que es verdadera. En cambio, disinformation implica una intención maliciosa: se crea y distribuye información falsa con el objetivo de engañar, manipular o perjudicar a individuos o grupos. Esta distinción es esencial para comprender las estrategias de combate, ya que el enfoque ante la desinformación intencionada debe ser más riguroso. **Referencia: p. 15**

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

   El informe destaca una caída significativa del interés por las noticias en España. En 2015, un 85% de los ciudadanos manifestaban tener un alto o muy alto interés por la información, pero esta cifra cayó al 51% en 2023. Esta disminución de 34 puntos porcentuales refleja una preocupante desconexión entre la población y los medios informativos. Esta tendencia se agrava con el aumento de la desconfianza, que alcanza

un 40% de escepticismo hacia los medios, especialmente entre 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?**
   Las noticias falsas se propagan considerablemente más rápido y con mayor alcance que las verdaderas. El 1% de las noticias falsas más virales alcanzaron entre 1.000 y 100.000 personas, mientras que el 1% de las noticias verdaderas llegó a superar los 1.000 receptores. Esta diferencia se atribuye a la naturaleza sensacionalista y emocionalmente provocadora de las noticias falsas, que aumenta la probabilidad de ser compartidas. Adicional estos patrones tan distintos de difusión, han llevado a desarrollar filtros de detección rápida de noticias falsas. **Referencia: p. 21**

4. **¿Qué ventaja clave ofrecen las redes latentes de difusión sobre los modelos epidemiológicos para el estudio de la desinformación?**
   La ventaja fundamental es que las redes latentes permiten identificar específicamente qué usuarios propagan información y cómo se interrelacionan entre sí. A diferencia de los modelos epidemiológicos, que analizan la propagación de forma anónima y estadística, basado en el modelo SIR (Susceptibles, infectadas y recuperadas), las redes latentes asignan valores de influencia entre nodos, revelando a los actores clave, "influencers" y receptores pasivos. Esto posibilita intervenciones estratégicas en redes sociales para mitigar la desinformación desde su origen. **Referencia: pp. 24-25**

5. ¿**Qué son los "grandes modelos de lenguaje" y cuál es su principal riesgo en el contexto de la desinformación?**
   Los grandes modelos de lenguaje, como GPT, son sistemas de IA entrenados con enormes cantidades de texto para generar respuestas coherentes y fluidas. En el contexto de la desinformación, su principal riesgo es la capacidad de producir rápidamente textos convincentes con coherencia humana, aunque factualmente falsos, sin necesidad de conocimientos técnicos por parte del usuario. Esto permite escalar campañas de

desinformación con mínima inversión de recursos y gran efectividad. **Referencia: pp. 28-29**

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

   La accesibilidad se refiere a la facilidad con la que usuarios comunes pueden utilizar estos modelos. Gracias a plataformas abiertas, software libre y hardware comercial asequible, cualquier persona puede crear contenido falso (textos, imágenes, audios) sin conocimiento experto. Esto reduce las barreras de entrada para generar desinformación de calidad profesional, lo que multiplica los actores potenciales de amenaza**. Referencia: pp. 27-28**

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

   Las "cajas negras" son modelos de IA (como redes neuronales profundas) cuyo funcionamiento interno es opaco, lo que impide entender por qué toman ciertas decisiones. El desafío es desarrollar modelos igualmente precisos pero que puedan explicar de forma comprensible sus procesos de inferencia, lo cual es esencial para su aceptación social y regulación ética. **Referencia: p. 49**

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

   La AGI se refiere a una IA con capacidades cognitivas similares a las humanas, capaz de resolver tareas generales. Si bien podría ser una herramienta poderosa para detectar y combatir desinformación, también podría generar contenido falso extremadamente realista y adaptativo. La AGI podría automatizar la desinformación a gran escala con poco esfuerzo, reforzando la necesidad de vigilancia y regulación, ya que el problema es que el desinformador siempre tiene la ventaja, debido a que es el primero en actuar. **Referencia: p. 49**

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

   Se mencionan el Reglamento General de Protección de Datos (RGPD), que garantiza los derechos sobre el uso de datos personales, y el Libro Blanco de la Inteligencia Artificial, que establece principios como la explicabilidad, la transparencia y la supervisión humana para el desarrollo de IA. Estas normas buscan equilibrar el uso de IA con la protección de los derechos fundamentales. **Referencia: p. 49**

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

    FacTeR-Check analiza únicamente datos públicos o aquellos con consentimiento explícito, evitando vulnerar el RGPD. No perfila usuarios ni retiene datos personales innecesarios. Además, incorpora principios de explicabilidad y control humano en sus procesos de verificación y monitorización en redes, alineándose con los estándares europeos de privacidad y ética. **Referencia: p. 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.**

    La Inteligencia Artificial (IA) desempeña un papel ambivalente en el contexto de la desinformación. Por un lado, ha facilitado la creación y difusión de contenido engañoso mediante herramientas cada vez más accesibles y sofisticadas. Por ejemplo, los grandes modelos de lenguaje basados en arquitectura Transformer permiten la generación de textos persuasivos y coherentes a partir de afirmaciones falsas. La accesibilidad de estos modelos, su capacidad para redactar artículos y titulares falsos, y su utilidad en redes sociales hacen que puedan emplearse con fines desinformativos a gran escala **(p. 28)**.

    La IA también ha posibilitado la manipulación de imágenes, audios y videos mediante técnicas como el inpainting, el clonado de voz y los deepfakes. Estas tecnologías permiten, por ejemplo, alterar rostros, cambiar escenarios o generar voces sintéticas que simulan a figuras públicas, incrementando la credibilidad de contenidos falsos **(pp. 31-**

36). Además, su integración con plataformas de microencargos hace posible coordinar campañas de astroturfing y desinformación en redes sociales **(p. 25)**.

Por otro lado, la IA ofrece herramientas poderosas para combatir la desinformación. El proyecto CIVIC desarrolló la herramienta FacTeR-Check, que combina similitud semántica, inferencia del lenguaje natural y bases de datos de hechos verificados para detectar y etiquetar afirmaciones como verdaderas, falsas o neutrales **(pp. 37-41)**. Además, permite la monitorización automatizada de redes sociales como Twitter mediante técnicas de extracción de palabras clave, detección de idioma, análisis semántico e inferencia, lo que facilita el seguimiento y análisis de la diseminación de bulos **(pp. 42-44)**.

En conclusión, el informe demuestra que la IA es una tecnología de doble filo: puede escalar la creación de desinformación, pero también fortalecer las capacidades de verificación y análisis si se aplica con responsabilidad, transparencia y control humano.

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?**
    La Inteligencia Artificial Explicativa (XAI) se presenta en el informe como una dimensión crítica en el desarrollo de sistemas confiables para combatir la desinformación. Su papel principal consiste en hacer comprensibles las decisiones que toman los modelos de IA, permitiendo que tanto expertos como usuarios generales puedan entender por qué un sistema clasificó una afirmación como verdadera o falsa **(p. 48)**.

    En términos de confianza pública, la XAI contribuye a legitimar el uso de estas tecnologías, ya que la transparencia en los procesos decisionales es clave para que el público acepte los resultados generados por los modelos. Además, tiene un valor educativo: al explicar por qué un contenido es considerado desinformación, los usuarios pueden aprender a reconocer patrones y estrategias engañosas por sí mismos, fortaleciendo su alfabetización mediática.

No obstante, el desarrollo de la XAI enfrenta importantes obstáculos. El más significativo es que muchos de los modelos actuales, especialmente los más eficaces, funcionan como "cajas negras", es decir, su funcionamiento interno es opaco y difícil de interpretar. Estos modelos se componen de millones de parámetros numéricos que interactúan de forma no lineal, lo que complica su traducción a explicaciones humanas comprensibles **(p. 49)**.

Otro desafío es que avanzar hacia modelos explicables podría implicar comprometer su rendimiento. Encontrar un equilibrio entre explicabilidad y precisión es uno de los retos centrales que aún deben resolverse para que la XAI sea adoptada de forma generalizada. A pesar de estos retos, el texto remarca que la IA explicativa no solo es deseable, sino indispensable para garantizar la transparencia, corregir sesgos y reforzar la utilidad ética y social de los sistemas de verificación automatizada.

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?**

El documento distingue entre dos tipos de modelos para estudiar la propagación de la desinformación en redes sociales: los modelos epidemiológicos y las redes latentes de difusión. Los modelos epidemiológicos, como el modelo SIR, se inspiran en la dinámica de propagación de enfermedades y clasifican a la población en grupos como susceptibles, infectados y recuperados **(p. 23)**. Aunque permiten predecir la evolución global del fenómeno, su principal limitación es que no identifican a los individuos involucrados ni las relaciones entre ellos. Se trata de modelos anónimos, útiles para detectar flujos anómalos, pero no para actuar sobre actores específicos.

Por otro lado, las redes latentes de difusión ofrecen una visión mucho más detallada y personalizada. Estas redes permiten identificar quién influye a quién y con qué intensidad, representando la propagación de desinformación como una red social dinámica **(p. 24)**. Son capaces de clasificar a los usuarios en roles como propagadores (influencers), receptores pasivos y controladores del discurso. Estas redes requieren gran cantidad de datos y procesamiento, pero permiten intervenciones más precisas.

En síntesis, los modelos epidemiológicos ofrecen una visión general y macro de la propagación, mientras que las redes latentes permiten un análisis micro, identificando actores clave y rutas específicas de transmisión de la desinformación.

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?

El informe señala que la accesibilidad de las herramientas de inteligencia artificial generativa es uno de los principales factores que pueden potenciar la desinformación. Hoy en día, muchos de estos modelos, como los generadores de texto (GPT) o los generadores de imágenes (StyleGAN, DALL-E, entre otros), están disponibles de forma abierta o gratuita, lo cual reduce las barreras técnicas y económicas para su uso **(pp. 27-28)**. Esto permite que cualquier persona, incluso sin conocimientos técnicos, pueda generar contenido falso de forma rápida, convincente y masiva.

La facilidad de acceso, combinada con la calidad cada vez más alta de los resultados generados por estas IA, incrementa el riesgo de que se usen con fines maliciosos. Por ejemplo, se pueden generar noticias falsas, imágenes manipuladas, audios con voces clonadas y videos falsos realistas **(pp. 31-36)**. Esto crea un entorno donde la desinformación puede escalar rápidamente y alcanzar a audiencias amplias antes de que sea verificada o desmentida.

El texto también advierte que la capacidad para crear campañas automatizadas de desinformación puede combinarse con técnicas como la personalización de mensajes y el uso de redes de cuentas falsas, lo cual amplifica aún más el alcance e impacto de estos contenidos falsos **(p. 25)**.

Para mitigar estos riesgos, el documento sugiere varias estrategias. Primero, mejorar las capacidades de verificación automatizada mediante herramientas como FacTeR-Check, que permite comparar afirmaciones con hechos previamente verificados utilizando IA

(pp. 37-41). Segundo, implementar sistemas de monitorización en redes sociales para detectar rápidamente patrones de propagación sospechosos **(pp. 42-44)**. Finalmente, se destaca la necesidad de fomentar la alfabetización digital y mediática de la población, de manera que los usuarios puedan reconocer y resistir mejor los intentos de manipulación **(p. 18)**.

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 identificando si existen normativas en nuestro país similares.**

    El documento resalta que el uso de inteligencia artificial en la lucha contra la desinformación debe realizarse con responsabilidad ética y respetando los derechos fundamentales, especialmente la privacidad de los usuarios. Para ello, se mencionan como marco de referencia dos normativas clave en el contexto europeo: el Reglamento General de Protección de Datos (RGPD) y el Libro Blanco sobre la Inteligencia Artificial **(p. 49)**.

    El RGPD establece obligaciones para el tratamiento de datos personales, como el consentimiento informado, el derecho a la transparencia, la minimización del uso de datos y la protección frente a decisiones automatizadas. Esto implica que cualquier herramienta basada en IA que analice o clasifique información generada por usuarios, como las publicaciones en redes sociales, debe cumplir con estos principios para garantizar que no se vulneren los derechos de privacidad.

    El Libro Blanco de la IA, por su parte, promueve un uso ético, transparente y supervisado de los sistemas de inteligencia artificial. Propone principios como la explicabilidad de las decisiones, la trazabilidad de los datos utilizados y la intervención humana en los procesos automáticos, especialmente en contextos sensibles como la moderación de contenido en redes.

En línea con estos principios, la herramienta FacTeR-Check desarrollada en el proyecto CIVIC se diseñó para cumplir con la legislación vigente. Solo accede a información pública o con consentimiento, evita el perfilado de usuarios y prioriza la supervisión humana en su funcionamiento **(p. 50)**.

Respecto a la existencia de normativas similares en otros países, el texto no hace mención específica a regulaciones en América Latina o Colombia. Por tanto, no se puede afirmar desde el documento si existen normativas equivalentes. Lo que sí se infiere es que se requiere una regulación armonizada a nivel internacional para enfrentar los desafíos éticos que plantea la IA, en particular cuando se usa en la esfera pública para intervenir en el discurso social y político.

Fuera del texto, se tiene en el caso de Colombia, una normativa robusta en materia de protección de datos personales, encabezada por la Ley 1581 de 2012. Además, Colombia cuenta con una autoridad de control, la Superintendencia de Industria y Comercio (SIC), que vigila el cumplimiento de estas normas.

Bibliografía

Martín García, A., Panizo Lledot, Á., D'Antonio Maceiras, S. A., Huertas Tato, J., Villar Rodríguez, G., Anguera de Sojo Hernández, Á., & Camacho Fernández, D. (2024). *Luchando contra la desinformación mediante la inteligencia artificial*. Fundación BBVA. https://www.fbbva.es