Escuela Superior de Guerra “General Rafael Reyes Prieto”

Luchando contra la desinformación mediante la inteligencia Artificial

My. Alex Danny Guerrero Cortés

Electiva - Habilidades Prácticas en el Ciberespacio

Jaider Ospina Navas

Curso de Estado Mayor 2025

01 de julio de 2025

Cuestionario

1. Diferencia fundamental entre "misinformation" y "disinformation":

La diferencia fundamental radica en la intencionalidad de quien la difunde. La "misinformation" (información errónea) se refiere a contenido falso que es compartido sin la intención deliberada de causar daño; el emisor cree en su veracidad. Por otro lado, la "disinformation" (desinformación) es información falsa creada y difundida con la intención explícita de engañar y causar perjuicio. El documento también introduce el concepto de "malinformation," que es información verdadera, pero sacada de contexto o sesgada con un propósito malicioso. Desde una perspectiva de ciberseguridad, la "disinformation" representa una amenaza más directa y activa, a menudo ligada a campañas de influencia orquestadas, mientras que la "misinformation" puede propagarse orgánicamente, pero sigue siendo un vector de riesgo.

2. Tendencia preocupante en España según el *Reuters Institute Digital News Report* 2023:

Según el *Reuters Institute Digital News Report* 2023, una tendencia preocupante en España es el descenso significativo del interés por las noticias. El porcentaje de personas con un interés alto o muy alto por las noticias cayó del 85% en 2015 al 51% en 2023, lo que representa una disminución de 34 puntos porcentuales. Paralelamente, la desconfianza de los lectores en los medios de comunicación alcanzó un récord del 40% en nueve años de encuesta, siendo más pronunciada entre los menores de 45 años. Este panorama crea un caldo de cultivo ideal para la desinformación, ya que una audiencia menos interesada y más desconfiada es más susceptible a narratives falsas, un desafío crítico para la estabilidad informacional y la ciber resiliencia social.

3. Comparación de la difusión de noticias falsas vs. verdaderas según Vosoughi, Roy y Aral (2018):

El documento proporcionado no detalla los experimentos o las conclusiones específicas de Vosoughi, Roy y Aral (2018) sobre la velocidad y facilidad de difusión de

noticias falsas frente a las verdaderas. Aunque se menciona a estos autores en la bibliografía, la información concreta sobre sus hallazgos en este ámbito no se encuentra en el texto. Para un análisis completo, sería necesario consultar la fuente original de dicho estudio.

4. Ventaja clave de las redes latentes de difusión sobre los modelos epidemiológicos para el estudio de la desinformación:

El texto proporcionado no describe explícitamente la ventaja clave que ofrecen las redes latentes de difusión sobre los modelos epidemiológicos para el estudio de la desinformación. Si bien el informe aborda la propagación de la desinformación y la necesidad de nuevas tecnologías para combatirla, una comparación detallada entre estos dos tipos de modelos y sus ventajas específicas no se encuentra en las secciones accesibles del documento.

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" (LLMs) son sistemas de inteligencia artificial avanzados, como GPT-3 (ChatGPT), capaces de generar textos altamente convincentes y personalizados. Su principal riesgo en el contexto de la desinformación radica en su capacidad para producir argumentos muy persuasivos, incluso sin base científica, haciendo extremadamente difícil diferenciar el contenido generado por IA del escrito por un humano. Esta facilidad de generación permite a actores malintencionados desplegar agentes automáticos (bots) que simulan comportamiento humano para difundir texto malicioso a gran escala, por ejemplo, en campañas de propaganda política. Desde la ciberseguridad, esto representa una proliferación de amenazas de ingeniería social a niveles sin precedentes.

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

La accesibilidad de los modelos de IA facilita la generación de desinformación al reducir drásticamente la barrera de entrada para la creación de contenido falso y sofisticado. El desarrollo de modelos de código abierto, como *Stable Diffusion* para la generación de imágenes, ha democratizado el acceso a técnicas muy potentes. Esto permite que incluso individuos sin conocimientos técnicos avanzados puedan generar texto, imágenes y audio realistas y engañosos. Esta facilidad de uso y disponibilidad masifica la capacidad de producir y diseminar desinformación, transformando la amenaza de algo reservado a actores con recursos a un riesgo generalizado.

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

En el contexto de la Inteligencia Artificial Explicativa (XAI), las "cajas negras" se refieren a los sistemas de IA cuyo funcionamiento interno es opaco, lo que dificulta comprender cómo llegan a decisiones o clasificaciones específicas. El desafío asociado es la ausencia de transparencia, lo que menoscaba la confianza en estos sistemas, especialmente cuando se utilizan para clasificar contenido como desinformación. Para generar confianza, es imperativo que estos sistemas puedan ofrecer razones y argumentos concretos que justifiquen sus decisiones, permitiendo a los usuarios comprender y confiar en la tecnología. En ciberseguridad, la falta de aplicabilidad en sistemas de detección de amenazas puede impedir la auditoría y la mejora continua, así como generar escepticismo entre los analistas.

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

El concepto de Inteligencia Artificial General (AGI) implica que, a medida que los grandes modelos de lenguaje (y potencialmente otros modelos de visión o audio) continúen evolucionando, adquirirán capacidades superiores tanto para generar como para apoyar la desinformación. Aunque una AGI también podría verificar información de manera más fiable, el informe señala que el desinformador mantendrá una ventaja inicial al actuar primero antes de ser verificado. Esto sugiere una carrera armamentística cibernética

continua donde la capacidad ofensiva de la IA para crear desinformación avanzada podría crecer a la par, o incluso superar temporalmente, la capacidad defensiva de la IA para detectarla.

9. Normativas europeas importantes mencionadas en relación con la IA y la privacidad:

El informe menciona las siguientes normativas europeas clave en relación con la IA y la privacidad:

- **Reglamento General de Protección de Datos (RGPD):** Es la normativa fundamental que rige el tratamiento de datos personales, estableciendo los derechos de los interesados y las obligaciones de los responsables del tratamiento.
- **Libro Blanco de la IA:** Este documento, enmarcado en la Estrategia Digital de la Unión Europea, sienta las bases para el desarrollo de la IA en Europa, con un énfasis particular en las implicaciones relativas a los datos personales.
- **Ley de Inteligencia Artificial:** Se menciona como un avance normativo próximo que establecerá un marco legal más específico para la IA.

Estas normativas reflejan el compromiso de la UE con un desarrollo de la IA ético y centrado en el ser humano, crucial para la ciberseguridad y la protección de los derechos individuales en la era digital.

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

FacTeR-Check garantiza el cumplimiento de la normativa de protección de datos mediante un diseño intrínseco que integra los principios de privacidad y protección de datos desde su concepción. Al analizar datos de redes sociales, como los tuits, la herramienta verifica que provengan de perfiles públicos o, en caso de ser privados, que los usuarios

hayan otorgado su consentimiento explícito para su uso, en conformidad con la normativa de protección de datos. Además, se asegura de que se consideren principios de la IA como la aplicabilidad, la seguridad y la supervisión humana, especialmente cuando las decisiones basadas en el tratamiento de datos puedan afectar los derechos de las personas. Utiliza una base de datos previamente verificada.

Preguntas de Formato Ensayo

11. Análisis de las diferentes formas en que la Inteligencia Artificial puede ser utilizada tanto para generar como para combatir la desinformación:

La Inteligencia Artificial se ha erigido como una espada de doble filo en el ecosistema de la información. Desde la perspectiva de la ciberseguridad, entender su dualidad es fundamental.

La IA como generadora de desinformación (vector de ataque):

- **Generación de Texto y Bots:** Los Grandes Modelos de Lenguaje (LLMs) como ChatGPT han perfeccionado la creación de textos coherentes y persuasivos, haciendo indistinguible el contenido generado por IA del humano. Esto permite la proliferación de bots sofisticados que simulan comportamientos humanos para difundir propaganda política o narratives maliciosas a escala masiva. La capacidad de personalización del mensaje magnifica su potencial de manipulación.
- **Manipulación de Medios Visuales:** Herramientas como DALL-E y Stable Diffusion, mediante la generación de imágenes a partir de texto, permiten crear ilustraciones hiperrealistas de falsedades, confiriendo mayor credibilidad a las noticias falsas. Los *"deepfakes"* son el ejemplo más palpable de esta amenaza, alterando la percepción de la realidad visual.
- **Manipulación de Audio:** La síntesis y clonación de voz basada en IA puede generar audios que imitan a figuras públicas o crean conversaciones fabricadas. Estas capacidades se combinan con la generación de texto, permitiendo producir discursos completos y automatizados con la voz de una persona deseada en tiempo real.

- **Sinergias Avanzadas:** La combinación de estas técnicas (e.g., texto a voz, audio a vídeo) multiplica la sofisticación de la desinformación, haciendo que los ataques sean más difíciles de detectar y contrarrestar.
-

La IA como herramienta para combatir la desinformación (vector de defensa):

- **Fact-Checking Automatizado y Semiautomatizado:** Ante el volumen inabarcable de información, la IA se vuelve indispensable para la detección eficiente y precisa de la desinformación. Herramientas como *FacTeR-Check* utilizan análisis de similitud semántica y de inferencia del lenguaje natural para verificar la veracidad del contenido.
- **Monitorización de Redes Sociales:** La IA permite el monitoreo a gran escala de redes sociales, identificando patrones de propagación de bulos basados tanto en el contenido como en el comportamiento de los actores que los difunden. Esto posibilita la visualización de la evolución de las falsedades y la optimización de las respuestas de las organizaciones de verificación.
- **Detección de Bots Sofisticados:** Aunque los bots son cada vez más avanzados, las técnicas de IA también se desarrollan para identificarlos y mitigar su impacto en la diseminación de desinformación.
- **Identificación de Marcadores Forenses:** Las técnicas de *deep learning* pueden detectar "marcas" o artefactos sutiles en imágenes y audios manipulados por IA, revelando su naturaleza sintética y ayudando en la verificación forense digital.

En resumen, la IA acelera la "carrera armamentística" entre atacantes y defensores en el dominio de la información. Como futuros expertos en ciberseguridad, nuestro rol es comprender profundamente estas dinámicas para desarrollar y desplegar contramedidas inteligentes que superen las tácticas adversarias.

12. Discusión sobre el papel de la Inteligencia Artificial Explicativa (XAI) en la mejora de la confianza pública y en la educación de los usuarios, y sus obstáculos:

La Inteligencia Artificial Explicativa (XAI) es una rama crítica de la IA, especialmente relevante en ciberseguridad y la lucha contra la desinformación, donde la confianza es paramétrica.

Papel de XAI en la mejora de la confianza pública:

La XAI busca transformar los modelos de "caja negra" en sistemas transparentes y comprensibles. Al proporcionar razones y argumentos concretos detrás de las decisiones de la IA —por ejemplo, por qué un contenido es clasificado como desinformación— se permite a los usuarios comprender el razonamiento del sistema. Esta transparencia es vital para generar confianza pública en las tecnologías de detección de desinformación. En un entorno donde la confianza en las noticias está en declive, un sistema explicable puede mitigar el escepticismo, permitiendo a los usuarios verificar la lógica subyacente y, potencialmente, descubrir y corregir sesgos o errores en el modelo.

Papel de XAI en la educación de los usuarios:

Más allá de la confianza, la XAI tiene un papel educativo crucial. Al exponer cómo la IA identifica patrones y anomalías en la desinformación, los usuarios pueden desarrollar una mayor "alfabetización mediática". Comprender las heurísticas de detección (e.g., lenguaje sesgado, inconsistencias visuales, patrones de difusión) empodera a los individuos para evaluar críticamente la información por sí mismos, incluso sin el apoyo directo de la IA. La XAI no solo detecta, sino que enseña a detectar.

Principales obstáculos para su desarrollo:

- **Complejidad Intrínseca de los Modelos:** Muchos modelos avanzados de IA logran su alta precisión a través de arquitecturas complejas que, por naturaleza, son difíciles de interpretar. El desafío principal es hacerlos explicables sin comprometer significativamente su rendimiento o precisión (*"black box" problem*).

- **Trade-off entre Explicabilidad y Rendimiento:** Frecuentemente, existe una tensión entre la complejidad del modelo (que puede llevar a un mayor rendimiento) y su capacidad de ser explicable. Modelos más simples son más fáciles de explicar, pero pueden ser menos precisos para detectar la desinformación más sofisticada.
- **Escalabilidad y Rendimiento en Tiempo Real:** Implementar XAI en sistemas de detección de desinformación a gran escala y en tiempo real presenta desafíos computacionales significativos. Generar explicaciones detalladas para cada decisión puede ser costoso en términos de recursos y tiempo de procesamiento.
- **Subjetividad y Nuances de la Desinformación:** La desinformación a menudo opera en zonas grises de interpretación. Explicar por qué un mensaje es desinformativo puede requerir un razonamiento que es inherentemente matizado y difícil de articular de manera objetiva y comprensible para una IA.
- **Riesgo de Explotación Adversaria:** Una explicación demasiado detallada de cómo funciona un sistema de detección de IA podría, paradójicamente, ofrecer a los actores maliciosos una "hoja de ruta" para eludir las defensas. Deben diseñarse sistemas que revelen suficiente información para generar confianza sin exponer vulnerabilidades críticas.

Desde el punto de vista de la ciberseguridad, los obstáculos de la XAI son retos de ingeniería de seguridad que deben abordarse mediante investigación en robustez, interpretabilidad segura y privacidad por diseño.

13. Comparación de modelos epidemiológicos y redes latentes de difusión para el estudio de la propagación de la desinformación, y la información obtenible de cada uno:

El documento proporcionado discute la importancia de comprender la dinámica de propagación de la desinformación y la necesidad de nuevas tecnologías para combatirla. Sin embargo, no ofrece una comparación explícita ni detalla las ventajas clave o la información específica que se puede obtener de los "modelos epidemiológicos" frente a las "redes latentes de difusión" en el contexto de la desinformación. Aunque se mencionan indirectamente en el índice y la bibliografía (por ejemplo, para algoritmos basados en

grafos y para modelado epidemiológico), el texto no profundiza en sus diferencias metodológicas o en el tipo de *insights* únicos que cada uno proporciona. Por lo tanto, con la información suministrada, no es posible realizar esta comparación detallada.

14. Relación entre la accesibilidad de las herramientas de IA generativa y el aumento potencial de la desinformación, y estrategias de mitigación:

La relación entre la creciente accesibilidad de las herramientas de IA generativa y el aumento potencial de la desinformación es directamente proporcional y alarmante desde una perspectiva de ciberseguridad. La democratización de tecnologías como los LLMs (e.g., ChatGPT) para la generación de texto, y modelos como DALL-E 2 o Stable Diffusion para la manipulación visual y de audio, ha reducido significativamente la barrera de entrada para la creación de contenido sintético. Anteriormente, la producción de desinformación altamente convincente requería habilidades técnicas especializadas y recursos considerables. Ahora, cualquier individuo con acceso a estas herramientas puede generar volúmenes masivos de texto, imágenes o audios realistas y engañosos. Esto empodera a actores maliciosos y multiplica la superficie de ataque informacional, facilitando campañas de desinformación más frecuentes, sofisticadas y de mayor alcance.

Estrategias sugeridas para mitigar este riesgo:

El informe plantea varias estrategias esenciales, que pueden ser categorizadas desde una óptica de ciberdefensa:

- **Desarrollo de Tecnologías Robustas de Detección (Defensa Activa):** Es imperativo invertir y desarrollar nuevas tecnologías "punteras y disruptivas" de IA, como FacTeR-Check, que permitan detectar, contrarrestar y combatir eficazmente el fenómeno de la desinformación. Esto incluye modelos capaces de identificar patrones sutiles y artefactos forenses en contenido sintético, así como sistemas que puedan adaptarse rápidamente a las nuevas tácticas de generación.
- **Fomento de la Alfabetización Mediática (Defensa Humanitaria/Concientización):** Una estrategia clave es la educación sólida en

"alfabetización mediática". Capacitar a los ciudadanos para evaluar críticamente la información, reconocer sesgos y fuentes no fiables, y comprender cómo funciona la desinformación, fortalece la "resiliencia cibernética" a nivel individual.

- **Regulación Efectiva de Plataformas (Gobernanza y Aplicación):** Se sugiere la necesidad de una "regulación efectiva de las plataformas de redes sociales" para combatir la difusión de información falsa. Esto implica un marco legal que incentive la responsabilidad de las plataformas en la moderación de contenido y la aplicación de políticas contra la desinformación, actuando como una "ciber-higiene" a nivel de infraestructura.
- **Transparencia y Explicabilidad en la IA (Confianza y Auditoría):** Aunque no es una mitigación directa de la *accesibilidad*, la promoción de la IA explicativa (XAI) mejora la confianza en los sistemas de detección. Si los usuarios y analistas entienden cómo y por qué se detecta la desinformación, se fortalece la aceptación de las contramedidas de IA, facilitando la identificación y corrección de sesgos en los algoritmos de defensa.
- **Diseño con Principios de Privacidad y Ética (Seguridad por Diseño):** Las herramientas de IA para combatir la desinformación deben diseñarse desde el inicio respetando los principios de protección de datos y privacidad. Esto incluye asegurar el consentimiento para el uso de datos (en particular de perfiles privados), y la incorporación de principios como la seguridad y la supervisión humana. Esto garantiza que las soluciones no comprometan los derechos fundamentales mientras combaten la amenaza.
- **Colaboración Multidisciplinar (Respuesta Colectiva):** El informe subraya la necesidad de un enfoque multifacético y multidisciplinar. La colaboración entre *fact-checkers*, instituciones, investigadores y expertos en ciberseguridad es vital para desarrollar estrategias holísticas y adaptativas contra la desinformación.

En conclusión, la accesibilidad de la IA generativa amplifica la amenaza de desinformación, exigiendo una respuesta multifacética que combine avances tecnológicos de detección, educación pública, marcos regulatorios y un enfoque ético en el desarrollo de soluciones de ciberseguridad.

15. Análisis de las consideraciones éticas y de privacidad asociadas con el uso de la Inteligencia Artificial para combatir la desinformación, con referencia a normativas europeas:

El uso de la Inteligencia Artificial en la lucha contra la desinformación, si bien esencial, plantea consideraciones éticas y de privacidad significativas que deben ser abordadas con rigor para evitar que las soluciones generen nuevos riesgos o vulneren derechos fundamentales.

Consideraciones Éticas y de Privacidad:

- **Protección de Datos Personales:** La principal preocupación es el tratamiento de datos personales, especialmente cuando las herramientas de IA analizan contenido de redes sociales. Es fundamental que cualquier sistema respete los principios de protección de datos y privacidad. Esto implica que la recolección, el procesamiento y el almacenamiento de datos deben ser lícitos, limitados a su finalidad y seguros.
- **Consentimiento y Fuentes de Datos:** Para cumplir con la normativa de protección de datos, las herramientas deben garantizar que los datos analizados (e.g., tuits) provengan de perfiles públicos o, en el caso de perfiles privados, que se haya obtenido el consentimiento expreso del usuario. Esto es crucial para evitar la vigilancia no autorizada o el uso indebido de información personal.
- **Respeto a Derechos Fundamentales:** La intervención de la IA no debe comprometer otros derechos fundamentales, como la libertad de expresión y el derecho a la información. Existe un delicado equilibrio entre combatir la desinformación y evitar la censura o la restricción ilegítima del discurso.
- **Sesgos Algorítmicos y Discriminación:** Aunque no se detalla extensamente en el fragmento como un obstáculo específico en la *lucha* contra la desinformación, un riesgo ético inherente a la IA es la posibilidad de que los algoritmos incorporen sesgos presentes en los datos de entrenamiento. Esto podría llevar a la discriminación en la identificación de desinformación, afectando

desproporcionadamente a ciertos grupos o narratives. La explicabilidad de la IA (XAI) se menciona como una vía para descubrir y corregir posibles sesgos.

- **Transparencia y Responsabilidad (Caja Negra):** La naturaleza de "caja negra" de muchos modelos de IA dificulta comprender cómo se toman las decisiones, lo que plantea cuestiones éticas sobre la responsabilidad y la rendición de cuentas. La XAI es vital para la auditoría y la justificación de las decisiones de la IA, especialmente cuando estas afectan los derechos de las personas.
- **Supervisión Humana:** El documento enfatiza la necesidad de incorporar la supervisión humana en el diseño de los sistemas de IA. Esto asegura que, en última instancia, las decisiones críticas que puedan impactar a los individuos no recaigan únicamente en un algoritmo, sino que haya una intervención y juicio humano.

Normativas Europeas Mencionadas:

- **Reglamento General de Protección de Datos (RGPD):** Es la piedra angular de la protección de datos en la Unión Europea. Establece un marco legal robusto para el procesamiento de datos personales, aplicable a cualquier entidad que trate datos de ciudadanos de la UE, incluyendo las herramientas de IA. Define derechos como el acceso, la rectificación, la supresión y la portabilidad de los datos, y exige principios como la minimización de datos y la privacidad desde el diseño. *FacTeR-Check* busca adherirse a estos principios.
- **Libro Blanco de la IA:** Este documento es parte de la estrategia digital de la UE y busca establecer un marco regulatorio para la IA, prestando especial atención a sus implicaciones en la privacidad y los datos personales. Aboga por una IA "fiable" que respete los valores y derechos europeos.
- **Ley de Inteligencia Artificial (AI Act):** Es una propuesta legislativa ambiciosa que busca regular la IA en función de su riesgo. Establece requisitos estrictos para los sistemas de IA de "alto riesgo", lo que probablemente incluirá aquellos utilizados en la lucha contra la desinformación, dadas sus posibles implicaciones en los derechos fundamentales.

Normativas Similares en Colombia (nuestro país):

El documento proporcionado no ofrece información sobre la existencia de normativas similares en Colombia. Para una respuesta completa sobre las regulaciones colombianas en materia de protección de datos y su aplicación a la IA, sería necesario consultar fuentes externas a este informe. No obstante, en Colombia, la Ley 1581 de 2012 de Protección de Datos Personales es el marco legal principal, que establece principios y derechos similares a los del RGPD en cuanto a la recolección, uso y circulación de datos personales.

En síntesis, la implementación de la IA en la ciberdefensa contra la desinformación debe ser meticulosamente calibrada para asegurar que los beneficios en la protección de la información no se logren a expensas de la privacidad y los derechos fundamentales de los ciudadanos.