**Luchando contra la desinformación mediante la inteligencia Artificial**

**Alumno**

**My. Diego Alejandro Esmeral Madrid**

**Electiva - Habilidades Prácticas en el Ciberespacio**

**Docente:**

**Jaider Ospina Navas**

**Escuela Superior de Guerra “General Rafael Reyes Prieto”**

**Curso de Estado Mayor 2025**

**01 de julio de 2025**

**Cuestionario**

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

La misinformation es información falsa difundida sin intención de causar daño, mientras que la disinformation se genera y difunde con intención maliciosa de perjudicar.

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

España ha experimentado un descenso notable en el interés por las noticias. En 2015, el 85% de la población afirmaba tener un alto o muy alto interés por las noticias; para 2023, ese porcentaje cayó al 51%. Esta pérdida de 34 puntos porcentuales se combina con un aumento de la desconfianza hacia los medios (40%), especialmente entre los menores de 45 años. Esta situación genera un entorno propicio para la desinformación, ya que una ciudadanía menos informada y más desconfiada es más vulnerable a los bulos.

1. **¿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?**

El estudio muestra que las noticias falsas se difunden más rápido, más lejos y con mayor profundidad que las verdaderas. El 1% de las noticias falsas más compartidas alcanzó entre 1.000 y 100.000 personas, mientras que el 1% de las verdaderas rara vez superó las 1.000. Este comportamiento se debe a que las noticias falsas suelen ser más novedosas, impactantes o emocionales, lo que favorece su viralización. Esta diferencia en los patrones de propagación ha sido aprovechada por investigadores para desarrollar filtros de detección temprana.

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

Mientras que los modelos epidemiológicos permiten modelar la propagación de la información de manera anónima, las redes latentes de difusión permiten identificar quién influye en quién. Esto significa que se puede determinar no solo que hay un flujo anómalo de información, sino también qué usuarios específicos están impulsando la desinformación, cómo se estructura la red de influencia y quiénes son los nodos clave (influencers, receptores, propagadores). Esta capacidad analítica es crucial para intervenir con precisión y detener campañas organizadas de desinformación.

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

Los grandes modelos de lenguaje (Large Language Models, LLM) como GPT o BERT son algoritmos basados en arquitecturas tipo Transformer, entrenados con enormes volúmenes de texto. Su principal riesgo en el contexto de la desinformación es su capacidad para generar texto coherente, persuasivo y falso con apariencia verídica en segundos. A pesar de los esfuerzos por integrarlos con filtros de seguridad, estos modelos pueden ser inducidos a crear narrativas engañosas. Además, no comprenden la veracidad, solo patrones estadísticos, por lo que pueden generar contenido falso sin advertencia ni discernimiento.

1. **¿Cómo facilitar la accesibilidad de los modelos de IA la generación de desinformación?**

El desarrollo de modelos abiertos y la reducción de los requisitos técnicos para ejecutarlos han democratizado el acceso a herramientas de IA. Esto permite que cualquier usuario, sin necesidad de conocimientos especializados ni equipos avanzados, pueda generar contenido falso, manipulaciones visuales o incluso audios sintéticos de manera rápida y barata. Esta accesibilidad ha reducido las barreras de entrada para las campañas de desinformación, incluso a escala masiva mediante bots o plataformas de micro trabajo como Fiverr o Amazon Mechanical Turk.

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

Las "cajas negras" se refieren a modelos de IA complejos, como las redes neuronales profundas, cuyo funcionamiento interno no es comprensible para los usuarios o incluso para los desarrolladores. Aunque estos modelos son potentes y precisos, no explican por qué han tomado una decisión, lo que limita la transparencia y dificulta detectar sesgos o errores. El reto es desarrollar técnicas de IA explicativa (XAI) que permitan descomponer las decisiones del modelo, ofrecer razones comprensibles y, con ello, aumentar la confianza del público y facilitar la supervisión humana.

1. **¿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. Aunque aún es un objetivo futuro, los avances recientes (como GPT-4) apuntan hacia una IA cada vez más generalista y autónoma. Esto puede ser un arma de doble filo: por un lado, una AGI podría detectar y neutralizar desinformación con alta eficacia; por otro, también podría ser utilizada para generar campañas de desinformación altamente sofisticadas, difíciles de detectar y con gran capacidad de persuasión. Además, los desinformadores siempre actúan primero, lo que da ventaja estratégica al uso ofensivo frente al defensivo.

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

La AGI se refiere a una IA con capacidades cognitivas similares a las humanas. Aunque aún es un objetivo futuro, los avances recientes (como GPT-4) apuntan hacia una IA cada vez más generalista y autónoma. Esto puede ser un arma de doble filo: por un lado, una AGI podría detectar y neutralizar desinformación con alta eficacia; por otro, también podría ser utilizada para generar campañas de desinformación altamente sofisticadas, difíciles de detectar y con gran capacidad de persuasión. Además, los desinformadores siempre actúan primero, lo que da ventaja estratégica al uso ofensivo frente al defensivo.

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

FacTeR-Check asegura la protección de datos mediante varios mecanismos:

* Utiliza solo datos públicos o con consentimiento explícito.
* No realiza perfilados intrusivos ni toma decisiones automatizadas sin supervisión humana.
* Aplica principios de la IA explicativa, seguridad y transparencia.
* Los datos analizados provienen de fuentes verificadas por entidades de fact-checking.  
  Todo el diseño de la herramienta se alinea con los principios del RGPD y del Libro Blanco de la IA, asegurando que el proceso de verificación y monitoreo no vulnere los derechos fundamentales de los usuarios

1. **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 actúa como un arma de doble filo en el ecosistema informativo contemporáneo. Por un lado, se ha convertido en una herramienta poderosa para generar desinformación. A través de los grandes modelos de lenguaje (LLM) como GPT y de modelos generativos de imagen (GANs, Diffusion, StyleGAN), cualquier usuario —incluso sin conocimientos técnicos— puede crear noticias falsas, deepfakes, audios clonados y microtextos automatizados en múltiples idiomas y con apariencia de veracidad. Además, estos modelos son altamente accesibles, lo que reduce los costes y aumenta la escala y velocidad de las campañas de desinformación.

Por otro lado, la IA también ha demostrado ser un recurso valioso para combatir este fenómeno. La herramienta FacTeR-Check, desarrollada en el Proyecto CIVIC, utiliza análisis de similitud semántica e inferencia de lenguaje natural para verificar afirmaciones en tiempo real, comparándolas con bases de datos de hechos verificados. También permite monitorizar redes sociales, identificar influencers de desinformación, detectar patrones de propagación y categorizar mensajes como falsos, neutros o verificados. Este enfoque semiautomático permite una verificación rápida, multilingüe y precisa, con respeto a los principios de privacidad y legalidad.

1. **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 IA explicativa (XAI, por sus siglas en inglés) tiene un rol crucial en la confianza pública hacia los sistemas de verificación automatizada. Esta disciplina busca dotar de transparencia a modelos complejos, permitiendo a los usuarios entender por qué una afirmación fue clasificada como verdadera, falsa o dudosa. Esta comprensión favorece la aceptación de los sistemas automatizados, facilita la supervisión humana, permite detectar sesgos y errores, y, sobre todo, educa al usuario para que pueda identificar desinformación por sí mismo.

Sin embargo, el desarrollo de la XAI enfrenta obstáculos significativos. El principal es la opacidad de las llamadas "cajas negras": modelos como los Transformers, que operan con millones de parámetros, no son comprensibles ni siquiera para sus creadores. Explicar sus decisiones requiere desarrollar nuevas técnicas que aún están en fases tempranas. Otro obstáculo es lograr un equilibrio entre rendimiento y explicabilidad: los modelos más precisos suelen ser los menos transparentes. Además, existe el reto de adaptar los niveles de explicación al perfil del usuario (experto, periodista, ciudadano común), lo cual añade una capa de complejidad adicional al desarrollo de estas herramientas.

1. **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?**

Los modelos epidemiológicos y las redes latentes de difusión ofrecen dos enfoques distintos para analizar la diseminación de la desinformación. Los primeros, como el modelo SIR, se inspiran en la propagación de enfermedades: dividen a la población en grupos (susceptibles, infectados y recuperados) y utilizan parámetros como tasa de infección y recuperación. Permiten estimar el volumen global de personas expuestas y simular escenarios futuros. Sin embargo, son anónimos: no permiten identificar quién difunde ni cómo lo hace.

En cambio, las redes latentes de difusión permiten modelar las interacciones específicas entre usuarios. A partir de cascadas de difusión, se puede estimar la influencia de un nodo sobre otro, identificar estructuras jerárquicas (influencers, intermediarios, receptores), detectar núcleos de propagación y analizar la intensidad y dirección del flujo informativo. Esto resulta clave para entender campañas organizadas de desinformación y diseñar intervenciones más precisas. La principal desventaja de estas redes es que requieren grandes volúmenes de datos no siempre disponibles y son sensibles a cambios en la plataforma o en los perfiles de usuario.

1. **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?**

La creciente accesibilidad de herramientas de IA generativa —por ejemplo, modelos disponibles en HuggingFace o servicios como ChatGPT y Stable Diffusion— ha facilitado que cualquier persona pueda generar textos, imágenes, audios y vídeos falsos con una apariencia convincente. Esto ha reducido el costo y el umbral técnico necesario para realizar campañas de desinformación, especialmente en redes sociales. Plataformas como Fiverr o Mechanical Turk permiten contratar tareas específicas para amplificar campañas, lo que refuerza el fenómeno del astroturfing.

Para mitigar este riesgo, el informe propone desarrollar métodos de detección rápidos, escalables y actualizados, como FacTeR-Check. También se sugiere invertir en IA explicativa, establecer controles éticos en el desarrollo de modelos, limitar el acceso irrestricto a modelos potentes, y promover la colaboración con entidades de fact-checking que validen el contenido antes de su propagación. La alfabetización mediática también se destaca como una barrera de contención frente a la difusión acrítica de contenidos generados por IA.

1. **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 identificadas si existen normativas en nuestro país similares.**

El uso de inteligencia artificial (IA) para combatir la desinformación plantea serios desafíos éticos y de privacidad. Desde una perspectiva ética, cualquier sistema automatizado de verificación debe garantizar la **no discriminación**, la **transparencia**, la **supervisión humana** y el **respeto por la libertad de expresión**. Uno de los mayores riesgos radica en la posibilidad de que estos sistemas etiqueten como falsos contenidos legítimos o generen perfiles injustificados de los usuarios, lo cual afectaría sus derechos fundamentales.

En cuanto a la privacidad, el principal foco de atención es el tratamiento de datos personales. Las herramientas como *FacTeR-Check* deben asegurarse de analizar únicamente contenido de **fuentes públicas** o de usuarios que hayan dado su **consentimiento informado**. Además, deben respetar principios como **minimización de datos**, **finalidad específica**, **limitación temporal**, y **explicabilidad de las decisiones automatizadas**.

En el ámbito europeo, el **Reglamento General de Protección de Datos (RGPD)** establece el marco normativo para proteger los datos personales, y el **Libro Blanco sobre la Inteligencia Artificial** (Comisión Europea, 2020) define los principios para un desarrollo ético y confiable de la IA. Estos documentos exigen que cualquier sistema que trate información personal incorpore salvaguardas técnicas y organizativas adecuadas desde su diseño (“**privacy by design**”).

En Colombia, la normativa análoga es la **Ley 1581 de 2012** sobre protección de datos personales, reglamentada por el **Decreto 1377 de 2013** y supervisada por la **Superintendencia de Industria y Comercio (SIC)**. Esta ley establece principios similares a los del RGPD: legalidad, finalidad, libertad, veracidad, transparencia, acceso y circulación restringida, seguridad y confidencialidad. Sin embargo, el enfoque en IA es aún incipiente, por lo que sería necesario avanzar en una regulación específica que incorpore los riesgos éticos de la automatización en contextos como el combate a la desinformación.

En conclusión, si bien la IA puede ser un aliado poderoso contra la desinformación, su uso debe estar estrictamente regulado para evitar violaciones a la privacidad, abusos de poder o daños a derechos fundamentales. La confianza en estas tecnologías dependerá de su capacidad para operar dentro de un marco ético y jurídico robusto.