Cuestionario Lectura

“Luchando contra la Desinformación mediante la inteligencia artificial

Fundación BBVA”

2024

Asignatura: Habilidades prácticas en el ciberespacio

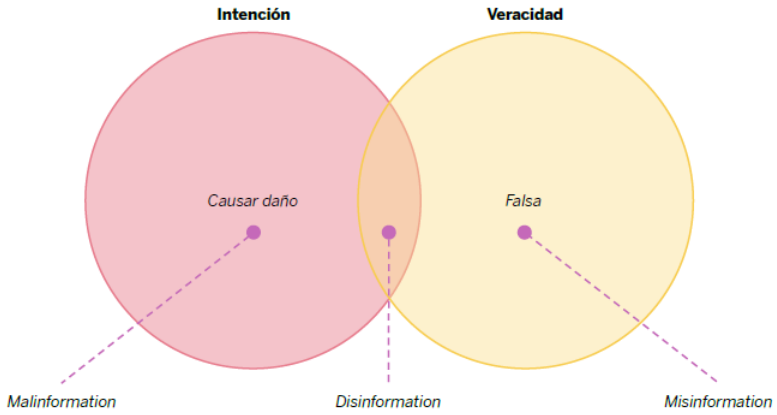
Estudiante: MY. LÓPEZ SALGUERO VÍCTOR ALFONSO

1. ¿Cuál es la diferencia fundamental entre *misinformation* y *disinformation*?

La diferencia esencial entre estos dos conceptos radica en la **intención del emisor**. La *misinformation* (información errónea) se refiere a contenido falso o inexacto que se difunde sin intención deliberada de causar daño. Un ejemplo común puede ser una persona que comparte un remedio casero ineficaz para una enfermedad creyendo que es útil. En contraste, la *disinformation* es creada y propagada de forma **intencional** con el propósito de manipular, engañar o provocar perjuicio a individuos, grupos o instituciones, y suele estar asociada a campañas de propaganda, manipulación política o desestabilización social.

El documento también introduce un tercer concepto: *malinformation*, que se refiere a información verdadera difundida **fuera de contexto o con sesgo**, con una clara intención de causar daño. Este tipo puede incluir la publicación de datos personales con el fin de amenazar, extorsionar o desacreditar a alguien.

Estos tres conceptos configuran el espectro de la desinformación digital y resultan fundamentales para diseñar estrategias diferenciales de detección, respuesta y prevención en plataformas digitales y medios de comunicación.



2. ¿Qué tendencia preocupante se observa en España según el Reuters Institute Digital News Report 2023?

El informe del Reuters Institute resalta una **disminución alarmante del interés y la confianza del público español en las noticias**. Entre 2015 y 2023, el porcentaje de personas con un alto interés en las noticias bajó del 85 % al 51 %, una caída de 34 puntos porcentuales. Este descenso coloca a España entre los países con mayor pérdida de interés en la información periodística a nivel global.

Además, la **desconfianza en los medios** también alcanzó niveles récord, situándose en un 40 % en 2023. Este fenómeno es especialmente pronunciado entre los menores de 45 años, lo cual indica un deterioro en la percepción de credibilidad de los medios tradicionales entre las generaciones jóvenes.

La combinación de desinterés y desconfianza contribuye a un entorno donde los ciudadanos recurren con mayor frecuencia a redes sociales y otras fuentes no verificadas para informarse, lo que amplifica la **vulnerabilidad colectiva ante la desinformación** y el sesgo informativo.

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

El estudio de Vosoughi, Roy y Aral (2018) demostró empíricamente que las noticias falsas tienen una **velocidad de propagación significativamente superior** a las noticias verdaderas en redes sociales, en particular en Twitter. Según los resultados, las noticias falsas tienen un **70% más de**

probabilidad de ser retuiteadas que las verdaderas, y el 1 % más viral de estas noticias falsas puede alcanzar entre 1.000 y 100.000 personas, mientras que el mismo porcentaje de las verdaderas rara vez supera los 1.000 usuarios.

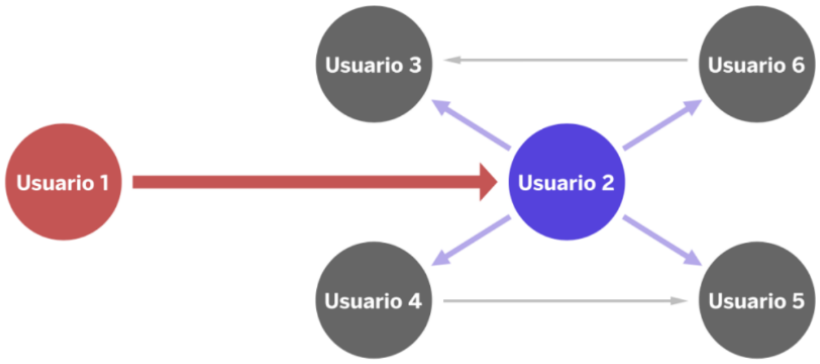
El motivo detrás de esta propagación acelerada no se debe a la acción de *bots*, sino al comportamiento humano: las noticias falsas tienden a evocar emociones más intensas (sorpresa, indignación, miedo), lo que incentiva su diseminación. Este hallazgo es fundamental para entender que **el combate a la desinformación debe considerar factores psicológicos y sociales**, además de tecnológicos

4. ¿Qué ventaja clave ofrecen las redes latentes de difusión sobre los modelos epidemiológicos?

Los modelos epidemiológicos aplicados a la información (como el modelo SIR: Susceptible-Infectado-Recuperado) tratan la propagación de contenido como si fuera una enfermedad, observando tasas de contagio y expansión sin identificar actores individuales. Aunque útiles, **estos modelos no permiten entender quién influye en quién** ni rastrear dinámicas interactivas concretas.

Las redes latentes de difusión, en cambio, permiten **inferir las relaciones individuales de influencia entre usuarios**, identificar nodos clave (controladores, difusores o receptores), segmentar comunidades según su comportamiento informativo y comprender las dinámicas temporales de la propagación.

Este enfoque permite una intervención más dirigida: por ejemplo, se pueden desactivar cuentas altamente influyentes o intervenir narrativas en puntos de máxima propagación, logrando un **control más quirúrgico y eficiente de las campañas de desinformación.**



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* (Large Language Models o LLMs) son sistemas de IA entrenados con billones de palabras que utilizan arquitecturas como Transformer (e.g., GPT, BERT) para generar lenguaje natural con un alto grado de coherencia semántica. Son capaces de escribir artículos, responder preguntas o generar conversaciones, incluso simulando estilos y tonos específicos.

En el contexto de la desinformación, el principal riesgo radica en que **permiten generar contenido falso, persuasivo y convincente a gran velocidad**, reduciendo los costos y habilidades necesarias para crear narrativas engañosas. Estos modelos no comprenden la verdad ni verifican la información; simplemente optimizan la coherencia gramatical y contextual, lo que los hace especialmente peligrosos si se usan para manipulación deliberada.

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

El fácil acceso a modelos como GPT o Stable Diffusion ha reducido considerablemente la **barrera técnica** para crear contenido falso sofisticado. Anteriormente, generar desinformación a escala requería equipos profesionales; hoy, **una persona con acceso a internet puede crear narrativas falsas en múltiples idiomas, incluyendo imágenes y audio manipulados**, desde un ordenador doméstico.

Esto implica una **democratización del potencial destructivo**, pues ahora actores individuales, sin necesidad de recursos estatales o tecnológicos avanzados, pueden llevar a cabo campañas de desinformación. Además, los modelos están migrando hacia versiones más ligeras y ejecutables en dispositivos móviles, lo que amplifica su potencial de uso malicioso.

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

Una *caja negra* en IA se refiere a un modelo complejo (como redes neuronales profundas) cuyas decisiones no pueden ser comprendidas fácilmente ni siquiera por sus desarrolladores. Aunque

estos modelos tienen gran poder predictivo, su **falta de transparencia** genera desconfianza, especialmente cuando se usan para clasificar información como falsa o verdadera.

El desafío es desarrollar mecanismos de *IA explicativa* (Explainable AI, XAI) que permitan comprender, auditar y justificar las decisiones automatizadas. Esto es esencial para validar la equidad del sistema, detectar sesgos y **garantizar la responsabilidad algorítmica**, sobre todo en entornos sensibles como la información pública, la justicia o la seguridad.

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

La AGI es una forma de IA con capacidades cognitivas comparables a las humanas, capaz de razonar, aprender y adaptarse a contextos diversos. En el ámbito de la desinformación, esta tecnología podría representar un arma de doble filo.

En positivo, permitiría **detectar, interpretar y desactivar campañas de desinformación** con gran precisión y rapidez. Pero también podría ser usada para **crear desinformación hiperpersonalizada**, diseñada para resonar emocional y culturalmente con objetivos específicos, optimizando los tiempos, formatos y canales de difusión. La automatización y escalabilidad que ofrece la AGI pone en riesgo la capacidad de respuesta de los sistemas humanos o tradicionales de verificación.

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

Las principales normativas citadas son:

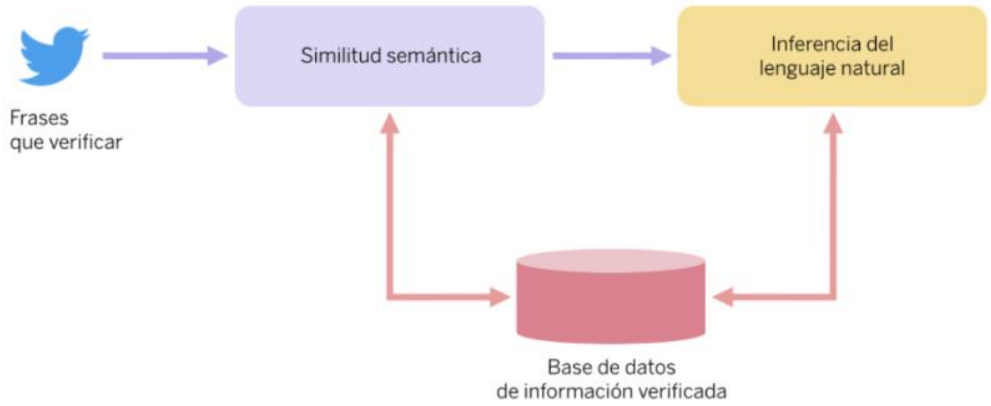
- **Reglamento General de Protección de Datos (RGPD)**: regula el tratamiento lícito, transparente y justo de datos personales. Establece principios como la minimización de datos, consentimiento informado y derecho al olvido.
- **Libro Blanco sobre la Inteligencia Artificial**: propone un enfoque centrado en el ser humano para el desarrollo de la IA, con principios éticos y jurídicos.

- **Ley de Inteligencia Artificial (AI Act)**: busca regular el uso de IA en función de su riesgo (bajo, medio, alto), protegiendo derechos fundamentales como la privacidad, libertad de expresión y no discriminación

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

FacTeR-Check implementa un enfoque de *privacy by design*, lo que significa que **la privacidad está integrada desde la concepción del sistema**, no como una capa añadida posterior. Solo analiza perfiles públicos, y si necesita acceder a información privada, lo hace con consentimiento explícito.

Además, incorpora principios de IA responsable: explicabilidad, seguridad, transparencia y supervisión humana. Estos elementos aseguran que el sistema sea **técnicamente eficaz pero también jurídicamente compatible y éticamente sólido.**



PREGUNTAS FORMATO ENSAYO

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) representa un paradigma dual en el ecosistema de la desinformación, actuando tanto como facilitadora de su generación como herramienta para combatirla. En su faceta más problemática, los avances en modelos generativos —particularmente los grandes modelos de

lenguaje (LLMs), los generadores de imágenes (como DALL·E o Stable Diffusion) y los sintetizadores de voz— han facilitado la creación de *deepfakes*, textos persuasivos y narrativas falsas con una apariencia de legitimidad alarmante. Estos contenidos no solo circulan rápidamente en redes sociales, sino que también pueden ser adaptados a contextos culturales y lingüísticos diversos gracias a la capacidad multilingüe de los modelos actuales.

La IA permite automatizar la creación de campañas masivas de desinformación que antes requerían recursos humanos y financieros significativos. El empleo de bots conversacionales alimentados por IA, capaces de sostener diálogos coherentes y emocionalmente efectivos, ha transformado las redes sociales en un entorno donde la frontera entre lo humano y lo artificial se desdibuja peligrosamente.

En contraposición, el mismo potencial computacional está siendo aprovechado para el desarrollo de herramientas como **FacTeR-Check**, plataforma creada por el grupo AIDA de la Universidad Politécnica de Madrid. Esta solución se basa en múltiples técnicas: análisis de similitud semántica, inferencia de lenguaje natural y modelos multilingües tipo *Transformer*, para identificar afirmaciones potencialmente falsas y verificar su contenido frente a bases de datos de hechos comprobados. FacTeR-Check permite también monitorizar redes sociales en tiempo real, evaluando patrones de diseminación y perfiles generadores de desinformación.

Esta dualidad evidencia que la IA no es buena ni mala en sí misma; su impacto depende de cómo, quién y con qué propósito se utilice. La clave reside en diseñar y regular entornos tecnológicos donde los principios de ética, transparencia y responsabilidad guíen la implementación de sistemas de IA, especialmente en entornos informativos que afectan directamente la democracia y la confianza pública.

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 IA Explicativa o *Explainable Artificial Intelligence* (XAI) surge como respuesta a uno de los desafíos más apremiantes del siglo XXI: la opacidad algorítmica. En el contexto de la

desinformación, donde la clasificación de información como "verdadera" o "falsa" puede afectar la percepción pública, la libertad de expresión y la reputación de personas o instituciones, es fundamental que los sistemas automatizados puedan **justificar y clarificar sus decisiones**.

Actualmente, muchos modelos de IA, especialmente los basados en redes neuronales profundas, funcionan como *cajas negras*: generan resultados eficaces, pero sin explicar con claridad cómo han llegado a ellos. Esta falta de transparencia debilita la confianza de los usuarios, impide la auditoría externa, y puede ocultar sesgos algorítmicos que perpetúen injusticias o desinformación.

La IA Explicativa busca "abrir la caja negra" mediante métodos que permitan visualizar los factores más influyentes en la decisión del modelo, destacar qué variables activan determinadas respuestas o permitir simulaciones para comprobar la coherencia de los resultados. En contextos de verificación informativa, esto se traduce en la necesidad de explicar por qué un contenido ha sido clasificado como sospechoso o falso, y sobre qué base de hechos verificados se ha tomado dicha decisión.

No obstante, los desafíos son considerables. Por un lado, existe una tensión entre explicabilidad y precisión: muchos modelos altamente precisos son también más complejos y, por ende, menos interpretables. Por otro lado, no existe aún una estandarización internacional sobre qué se considera una "explicación suficiente", ni sobre cómo debe presentarse al usuario final para que sea comprensible.

La XAI es, por tanto, una pieza central para consolidar la **legitimidad democrática de los sistemas de verificación automatizada**, reforzando su aceptación social y minimizando el riesgo de abuso o error.

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?

Tanto los modelos epidemiológicos como las redes latentes de difusión han sido utilizados para analizar la propagación de la desinformación, aunque su enfoque y alcance difieren notablemente. Los modelos epidemiológicos, como los de tipo SIR (Susceptibles, Infectados,

Recuperados), se inspiraron en la propagación de enfermedades infecciosas y permiten estudiar el comportamiento general de cómo una pieza de información se difunde dentro de una población. Ofrecen métricas como tasa de propagación, picos de infección (difusión) y fases de estabilización. Sin embargo, **su gran limitación es que operan de forma anónima y agregada**, es decir, no identifican quién difunde, ni cómo ni por qué.

En contraste, las **redes latentes de difusión** permiten reconstruir la topología completa de la interacción social: identifican nodos, vínculos, direccionalidad de la influencia y temporización de los intercambios. Esto permite modelar no solo el volumen, sino el recorrido exacto de la información, determinando quién inició una narrativa falsa, quién la amplificó (por ejemplo, *influencers*), y quiénes fueron meros receptores. Este nivel de granularidad es indispensable para tomar decisiones estratégicas orientadas a detener la propagación o desacreditar fuentes clave.

Además, las redes latentes son adaptables y pueden incorporar atributos como ideología, localización geográfica o idioma, facilitando análisis más contextuales. Esto convierte a las redes latentes en una herramienta poderosa para el estudio no solo de cómo se propaga la desinformación, sino también de **quién la produce, con qué intención y en qué condiciones sociales o tecnológicas**.

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?

Uno de los aspectos más disruptivos de la evolución reciente de la IA es su creciente accesibilidad. El acceso a modelos como GPT-4, DALL·E o Stable Diffusion, que antes requería infraestructura computacional especializada, hoy es posible desde interfaces web o incluso desde aplicaciones móviles. Esta democratización tecnológica conlleva beneficios indiscutibles en términos de innovación, pero también **riesgos considerables** en cuanto a la facilidad para producir desinformación.

La posibilidad de generar imágenes hiperrealistas de eventos falsos, crear discursos políticos atribuidos falsamente a líderes o clonar voces con alta fidelidad está al alcance de cualquier persona con conexión a internet. Esto ha multiplicado las campañas de manipulación, desde fraudes financieros hasta propaganda política, y ha desbordado la capacidad de respuesta de los sistemas tradicionales de verificación.

Frente a este panorama, el documento sugiere varias estrategias de mitigación:

- **Regulación del acceso y uso** de modelos generativos mediante licencias, trazabilidad de uso y control ético del desarrollo.
- **Implementación de marcas de agua digitales**, que permitan rastrear contenidos generados por IA.
- **Educación en alfabetización digital y mediática**, que forme a la ciudadanía para detectar patrones típicos de manipulación visual, textual o sonora.

En definitiva, el reto consiste en **encontrar un equilibrio entre apertura tecnológica e integridad informativa**, sin obstaculizar la innovación, pero previniendo su uso malicioso.

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 pais similares.

El uso de IA para combatir la desinformación plantea una paradoja fundamental: al tiempo que busca proteger a la sociedad de los efectos negativos de la información manipulada, puede amenazar **derechos fundamentales** como la privacidad, la libertad de expresión o el derecho a la información si no se implementa con criterios éticos rigurosos.

El Reglamento General de Protección de Datos (RGPD) de la Unión Europea establece los principios esenciales para el tratamiento lícito de datos personales. Esto incluye el consentimiento informado, la minimización de datos, la transparencia y el respeto a los derechos del interesado. Toda herramienta de IA utilizada para monitorizar redes sociales —

como FacTeR-Check, debe garantizar que analiza únicamente perfiles públicos o aquellos que han dado consentimiento explícito.

FacTeR-Check ejemplifica el modelo de *privacy by design*, integrando la privacidad desde la concepción del sistema. Además, incorpora principios de *IA responsable*: explicabilidad, supervisión humana y capacidad de corrección.

En Colombia, la **Ley 1581 de 2012** establece un marco para la protección de datos personales, y si bien comparte principios similares al RGPD, aún carece de una normativa específica para IA. Esto supone un vacío legal que debe ser abordado en el corto plazo, especialmente en sectores críticos como seguridad nacional, ciberdefensa o gobernanza digital.

En resumen, **el desarrollo tecnológico sin guía ética puede reproducir las mismas dinámicas de opresión, exclusión y manipulación** que pretende combatir. Por ello, la ética y la privacidad deben ser pilares estructurales en el diseño y despliegue de IA en contextos informativos.

Referencias

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/publicaciones/luchando-contra-la-desinformacion-mediante-la-inteligencia-artificial/