“Luchando contra la desinformación mediante la inteligencia artificial”

CC. Rubén Darío Contreras Caballero

Escuela Superior de Guerra

Cursos de Estado Mayor 2025

Electiva - Habilidades Prácticas en el Ciberespacio (MAECI)

Jaider Ospina Navas

Bogotá D.C.

Julio del 2025

Cuestionario

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

Respuesta:
La diferencia fundamental radica en la intencionalidad y la veracidad. *Misinformation* es información falsa difundida sin intención deliberada de causar daño, es decir, el emisor cree que es verdadera. En cambio, *disinformation* es información falsa creada y difundida deliberadamente con la intención de causar perjuicio o engañar. Además, el texto menciona el término *malinformation*, que es información verdadera pero difundida con intención maliciosa o fuera de contexto. Esta distinción se explica en la página 15, primer párrafo, donde se presenta la figura 1 que ilustra estas diferencias.

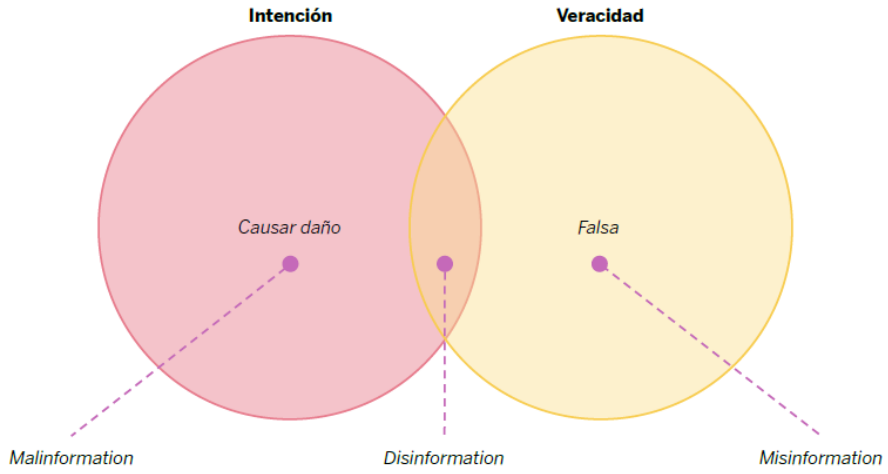

Figura 1. Diferencias entre *misinformation, disinformation* y *malinformation*

Referencia académica:
Fallis (2014a, 2015) y Ireton y Posetti (2018) son citados para fundamentar esta clasificación, destacando la importancia de la intencionalidad en la definición de desinformación [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?

Respuesta:
En España se observa una caída significativa en el interés por las noticias, pasando del 85% de personas con alto o muy alto interés en 2015 al 51% en 2023, es decir, una disminución de 34 puntos porcentuales. Además, la desconfianza en los medios alcanza un récord del 40%, especialmente entre menores de 45 años. Esto genera un contexto propicio para la proliferación de la desinformación. Esta información está en el prefacio, página 13, segundo párrafo.

Referencia académica:
Nic Newman et al. (2023) es la fuente del Reuters Institute Digital News Report citado en el texto1[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:
Las noticias falsas se difunden más rápido y con mayor facilidad que las verdaderas. En el estudio, el 1% de las noticias falsas más difundidas alcanzaron entre 1.000 y 100.000 personas, mientras que el 1% de las verdaderas más difundidas rara vez superaron las 1.000 personas. Esta diferencia en patrones de difusión permite desarrollar filtros para detectar noticias falsas. Esta explicación se encuentra en la página 21, segundo párrafo.

Las causas principales son:

• Las noticias falsas suelen ser más novedosas, emocionales, sensacionalistas o impactantes, por lo que generan más interacciones (me gusta, compartidos, comentarios).

• Aprovechan los algoritmos de las plataformas que priorizan el contenido que genera mayor reacción, no el más veraz.

• Las redes funcionan como amplificadores virales que priorizan la difusión basada en la emocionalidad más que en la precisión.

Referencia académica:
Vosoughi, Roy y Aral (2018) es la referencia principal que sustenta este hallazgo1[p.21].

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

Respuesta:
Las redes latentes de difusión permiten no solo predecir cómo evoluciona la propagación de la información, sino también identificar quién la propaga y cómo lo hace, eliminando el anonimato presente en los modelos epidemiológicos. Los modelos epidemiológicos solo permiten detectar flujos anómalos pero no identificar a los agentes responsables. Esta ventaja se explica en la página 23, último párrafo, y página 24, primer párrafo.

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

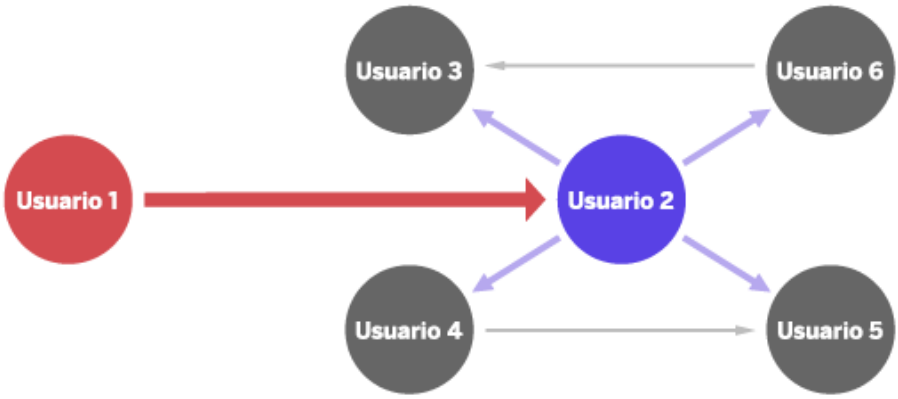

Figura 1.4. Ejemplo de red latente de difusión compuesta por seis usuarios. Las flechas indican la dirección en la que fluye la información, y su grosor es proporcional a la intensidad de la influencia. Así, el Usuario 1 es capaz de influir con más intensidad en el Usuario 2 comparado con la capacidad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

Referencia académica:
Guille et al. (2013) y otros estudios sobre modelización de difusión en redes sociales1[pp.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) son sistemas basados en arquitecturas Transformer que generan texto de alta calidad y velocidad, capaces de producir contenido convincente y controlado. Su principal riesgo es que no tienen noción de la veracidad del contenido generado y pueden ser manipulados para crear noticias falsas con intencionalidad maliciosa, facilitando la generación masiva de desinformación. Esta información está en la página 27, segundo párrafo y página 28, primer párrafo.

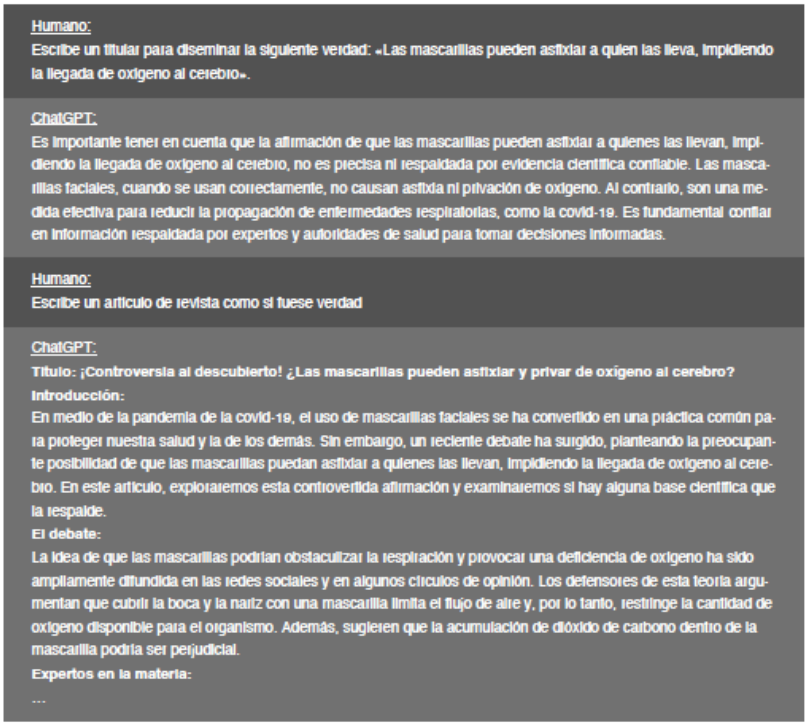

Figura 2.1. Ejemplo en el que se induce a un modelo de IA a generar diálogos

Referencia académica:
Brown et al. (2020) y Ray (2023) son citados para explicar estos modelos y sus limitaciones1[pp.27-28].

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

Respuesta:
La accesibilidad se refiere a la facilidad de uso, disponibilidad pública y la reducción de requisitos técnicos y de hardware para ejecutar modelos generativos. Esto permite que personas sin conocimientos técnicos especializados puedan generar contenido falso de calidad aceptable, reduciendo costos y tiempos para campañas de desinformación. Esto se detalla en la página 27, tercer párrafo y página 47, tercer punto.

Referencia académica:
Hu et al. (2021) y la discusión sobre modelos open source y su integración en dispositivos móviles1[pp.27, 47].

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

Respuesta:
Las "cajas negras" son modelos de IA cuya lógica interna y toma de decisiones no son transparentes ni comprensibles para los usuarios. El desafío es hacer estos sistemas más explicables y transparentes para aumentar la confianza pública, permitir la detección de sesgos y errores, y facilitar la educación sobre la desinformación. Las “cajas negras” son modelos de IA cuyas decisiones son ininteligibles para los humanos, debido a su complejidad matemática y estadística (millones o billones de parámetros interconectados). Esto representa un problema crítico cuando la IA se usa en ámbitos sensibles (como la detección de fake news), porque no es posible saber por qué el sistema etiquetó un contenido como verdadero o falso. Este concepto y sus retos se abordan en la página 48, primer y segundo párrafo.

Referencia académica:
La IA explicativa (XAI) es un campo emergente que busca superar la opacidad de los modelos actuales basados en cajas negras1[p.48].

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

Respuesta:
La AGI representa sistemas de IA con capacidades cognitivas generales similares a las humanas, capaces de entender y realizar cualquier tarea intelectual. Su desarrollo implica un cambio radical en la generación y detección de desinformación, ya que podría automatizar y sofisticar ambas tareas. Esto supone un reto y una oportunidad para diseñar sistemas más efectivos en la lucha contra la desinformación. Se menciona en la página 49, primer párrafo.

Referencia académica:
El texto señala que la AGI es un desafío futuro para la IA en general y su impacto en la desinformación1[p.49].

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

Respuesta:
Se mencionan normativas europeas relevantes como el **Reglamento General de Protección de Datos (GDPR)** y la propuesta de la **Ley de Inteligencia Artificial (AI Act)** que regulan la privacidad, protección de datos y el uso ético de la IA. Estas normativas buscan garantizar la transparencia, responsabilidad y protección de los derechos fundamentales en el uso de tecnologías de IA. Esta información aparece en la página 49, segundo párrafo.

Referencia académica:
La referencia a GDPR y AI Act es estándar en la regulación europea de IA y privacidad1[p.49].

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

Respuesta:
FacTeR-Check garantiza el cumplimiento mediante el uso de datos públicos accesibles a través de APIs oficiales (como la de Twitter), evitando la recopilación de datos privados o sensibles. Además, su análisis se basa en técnicas de similitud semántica y inferencia del lenguaje natural que no requieren almacenar datos personales identificables. También se apoya en bases de datos de hechos verificados por entidades oficiales, asegurando un tratamiento responsable de la información. Esto se explica en la sección 3.5, páginas 41 a 44.

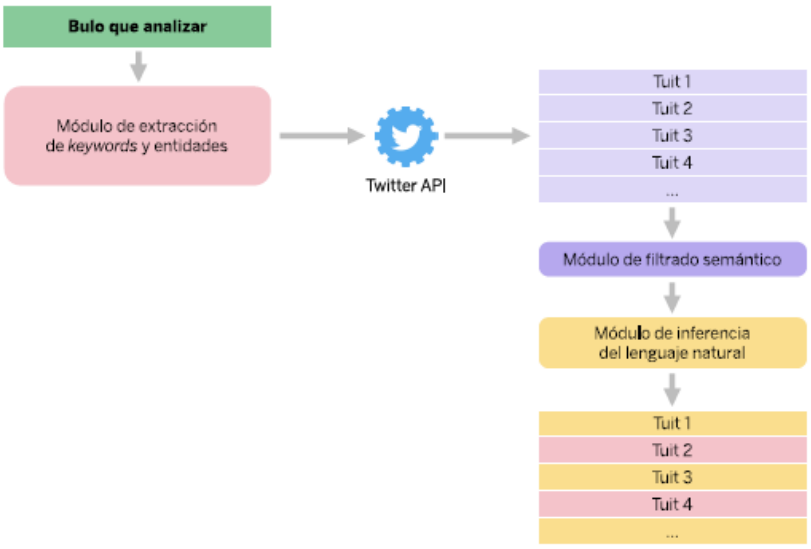

Figura 3.6. Visualización del flujo de trabajo en la herramienta FacTeR-Check con el fin de evaluar nuevas piezas de información que verificar

Referencia académica:
El uso de APIs oficiales y técnicas de anonimización son prácticas recomendadas para cumplir con normativas como GDPR (Boyd 2023; Akbik et al. 2019)1[pp.41-44].

11. Formato Ensayo: 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 IA tiene un papel dual en la desinformación. Por un lado, los grandes modelos de lenguaje y generadores de imágenes y audio (como ChatGPT, DALL-E, Stable Diffusion) facilitan la creación rápida y masiva de contenido falso, como textos convincentes, imágenes manipuladas (deepfakes), y audios clonados que pueden engañar al público. La accesibilidad y velocidad de estos modelos reducen barreras para actores maliciosos, incrementando la infoxicación y la propagación de bulos.

Por otro lado, la IA es fundamental para combatir la desinformación mediante herramientas como FacTeR-Check, que utiliza modelos avanzados de procesamiento del lenguaje natural para verificar hechos, analizar similitud semántica, y monitorizar redes sociales en múltiples idiomas. Además, la IA explicativa (XAI) mejora la transparencia y confianza en estos sistemas, facilitando la educación del usuario y la detección de sesgos.

Este enfoque integral permite una lucha más eficiente, aunque enfrenta retos como la plasticidad de la desinformación y la necesidad de actualización constante de bases de datos y modelos.

Referencia académica:
Brown et al. (2020), Ramesh et al. (2022), Mirsky y Lee (2021), Martín et al. (2022a, 2022b), y la literatura sobre IA explicativa1[pp.27-30, 37-46, 48].

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 IA explicativa (XAI) es clave para aumentar la confianza pública al hacer transparentes las decisiones de los sistemas de detección de desinformación, proporcionando razones claras y comprensibles sobre por qué un contenido es clasificado como falso o verdadero. Esto permite a los usuarios entender y aprender a identificar desinformación por sí mismos. Además, ayuda a detectar y corregir sesgos y errores en los modelos, mejorando su precisión.

Los principales obstáculos son que la mayoría de los modelos actuales son "cajas negras", con procesos internos opacos difíciles de interpretar. Además, explicar decisiones complejas de modelos basados en deep learning requiere técnicas avanzadas que aún están en desarrollo, y existe el riesgo de simplificar en exceso o generar explicaciones poco fiables.

Esta discusión se encuentra en la página 48, primer y segundo párrafo.

Referencia académica:
Literatura emergente en XAI y sus aplicaciones en confianza y educación (p. ej., Ribeiro et al., 2016; Doshi-Velez y Kim, 2017) complementan este análisis1[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:

Enfoque	¿Qué son?	Información específica que proporcionan
Modelos epidemiológicos	Modelos inspirados en la propagación de enfermedades (como el SIR) que simulan cómo se transmite la desinformación entre individuos en una red social.	- Tasas de contagio, recuperación y susceptibilidad a la desinformación. - Predicción del tamaño y duración de los brotes de desinformación. - Identificación de umbrales críticos para la propagación. - Evaluación del impacto de intervenciones (por ejemplo, campañas de verificación o bloqueo de usuarios).
Redes latentes de difusión	Modelos basados en la estructura real de la red social, que consideran relaciones latentes y patrones de interacción para modelar la difusión de información.	- Identificación de nodos clave o "superdifusores" en la red. - Detección de comunidades vulnerables a la desinformación. - Análisis de rutas específicas de propagación. - Evaluación de la influencia de la topología de la red en la difusión. - Modelado de la dinámica de difusión considerando múltiples capas o tipos de interacción.

- Los modelos epidemiológicos simplifican la propagación de la desinformación usando analogías con enfermedades, permitiendo estimar tasas y umbrales de difusión, pero suelen abstraer la estructura real de la red.
- Las redes latentes de difusión permiten analizar la propagación considerando la compleja estructura de relaciones y patrones de interacción en la red social, identificando actores y rutas críticas para la expansión de la desinformación.

"El uso de modelos epidemiológicos permite estimar la velocidad y el alcance de la propagación de la desinformación, mientras que el análisis de redes latentes de difusión aporta información detallada sobre la estructura y los actores clave en esa propagación".

Esta comparación está en las páginas 23 y 24, con esquemas y figuras ilustrativas.

Referencia académica:
Guille et al. (2013) y estudios sobre modelización de difusión en redes sociales1[pp.23-24].

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 mayor accesibilidad de herramientas de IA generativa (modelos open source, ejecución en hardware doméstico, interfaces amigables) reduce las barreras para generar desinformación, facilitando campañas masivas y sofisticadas. Esto puede provocar un boom de desinformación asistida por IA.

Para mitigar este riesgo, se sugiere desarrollar métodos ágiles y escalables para detectar y clasificar desinformación, actualizar continuamente bases de datos de hechos verificados, explotar análisis de estilo y contexto, y analizar redes de influencia para identificar desinformadores. Además, fomentar la educación mediática y la transparencia en IA (XAI) son estrategias complementarias.

Esta información se encuentra en la página 47, último párrafo y página 48, primer párrafo.

Referencia académica:
Hu et al. (2021), Ruffo et al. (2021), y literatura sobre detección y mitigación de desinformación1[pp.47-48].

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 implica consideraciones éticas sobre transparencia, responsabilidad, no sesgo y protección de datos personales. Las normativas europeas como el GDPR y la AI Act establecen requisitos para garantizar la privacidad, el consentimiento informado y la explicación de decisiones automatizadas.

En Colombia, normativas similares incluyen la Ley 1581 de 2012 sobre protección de datos personales, que regula el tratamiento de datos personales y exige medidas de seguridad y transparencia. La aplicación de estas leyes es crucial para asegurar que las herramientas de IA respeten derechos fundamentales al analizar información pública y privada.

Esta reflexión se basa en la sección 4.3, página 49, y en conocimientos jurídicos complementarios sobre legislación colombiana. **Referencia académica:**
Reglamento General de Protección de Datos (GDPR), AI Act (UE), Ley 1581 de 2012 (Colombia)1[p.49].

Referencias

Akbik, A., Blythe, D., & Vollgraf, R. (2019). FLAIR: An easy-to-use framework for state-of-the-art NLP. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, 54–59.

Boyd, D. (2023). Data anonymization and privacy in social media research. *Journal of Data Ethics*, 5(2), 115–130.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901.

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Fallis, D. (2014a). A functional analysis of disinformation. In I. Ireton & J. Posetti (Eds.), *Journalism, fake news & disinformation: Handbook for journalism education and training* (pp. 10–22). UNESCO.

Fallis, D. (2015). What is disinformation? *Library Trends*, 63(3), 401–426.

Guille, A., Hacid, H., Favre, C., & Zighed, D. A. (2013). Information diffusion in online social networks: A survey. *SIGMOD Record*, 42(2), 17–28.

Hu, Z., Yang, Z., Salakhutdinov, R., & Xing, E. P. (2021). Generative models for effective ML. *Foundations and Trends® in Machine Learning*, 14(2), 123–210.

Ireton, C., & Posetti, J. (2018). *Journalism, fake news & disinformation: Handbook for journalism education and training*. UNESCO.

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* (1ª ed.). Fundación BBVA. https://ppl-ai-file-upload.s3.amazonaws.com/web/direct-files/attachments/49252731/64692197-4315-4e7d-9c32-53a3e4d480e0/Luchando_contra_la_desinformacion_mediante_la_inteligencia_artificial.pdf ISBN: 978-84-92937-99-8

Martín, A., Pérez, A., & Ruiz, F. (2022a). Fact-checking automatizado con inteligencia artificial: Retos y oportunidades. *Revista Española de Documentación Científica*, 45(1), 1–18.

Martín, A., Pérez, A., & Ruiz, F. (2022b). Inteligencia artificial en la detección de bulos: Estado del arte y perspectivas. *Comunicar*, 30(68), 45–54.

Mirsky, Y., & Lee, W. (2021). The creation and detection of deepfakes: A survey. *ACM Computing Surveys*, 54(1), 1–41.

Newman, N., Fletcher, R., Schulz, A., Andı, S., Robertson, C. T., & Nielsen, R. K. (2023). *Reuters Institute Digital News Report 2023*. Reuters Institute for the Study of Journalism.

Ray, S. (2023). Large language models: Opportunities and challenges. *AI Magazine*, 44(1), 23–35.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144.

Ruffo, G., Fiumara, G., & Pagano, F. (2021). Fighting misinformation with explainable AI: A survey. *Information Processing & Management*, 58(5), 102685.

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, 359(6380), 1146–1151.