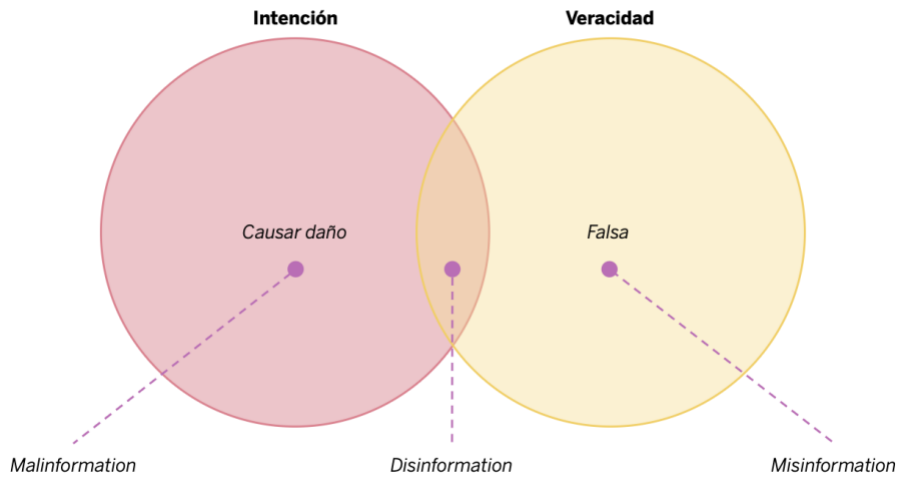
LUCHANDO CONTRA LA DESINFORMACIÓN MEDIANTE LA INTELIGENCIA ARTIFICIAL"

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

Tipos de información falsa: radica la diferencia en el tipo de falsedad y tipo de intencionalidad con la que se distribuye. La desinformación abarca un problema de gran amplitud en un ecosistema mediático hibrido.

- **Misinformation:** se refiere a información falsa difundida sin intención de engañar.
- **disinformation:** es difundida de forma deliberada con la intención de engañar.
- **Malinformation:** información que aun siendo verdadera se distribuye de manera sesgada o fuera de contexto.

Figura 1. *Clasificación de los tipos de información falsa.*



- **bulo:** Contenido intencional falsos y de apariencia verdadera con el fin de generar engaño y difundido por plataforma o medio de comunicación.
- **Infoxicación:** sobreabundancia de información

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

Se ha detectado un descenso preocupante en el interés de la población española por las noticias y en su confianza hacia los medios.

Se ha pasado de un 85% de personas que indicaban tener un interés alto muy alto por las noticias en 2015 al 51% en 2023, es decir, 34 puntos porcentuales menos.

Los datos de confianza de los lectores en los medios de comunicación tampoco ayudan: la desconfianza en las noticias llega a su récord (40%) en estos nueve años de encuesta, sobre todo entre los menores de 45 años.

El papel de la inteligencia artificial en la desinformación, observando los riesgos a los que nos enfrentamos, pero también analizando la necesidad de confiar en la IA para luchar contra las nuevas formas de desinformación.

La década de las fact checkers (creación y crecimiento de las organizaciones de verificación para la lucha de la desinformación y educar a la ciudadanía para no caer en falsedades.

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?**

Agentes maliciosos que abusan del sistema de flujo de información, generando desinformación con el objetivo de influir y manipular la opinión pública.

- **Trolls:**

  Cuentas que buscan crear conflictos

- **Bots**

  Cuentas completamente automatizadas

- **Cyborgs**

  Cuentas parcialmente automatizadas.

- **Cuentas falsas**

- **Chatbot**
  programa informático diseñado para simular una conversación con personas, ya sea a través de texto o voz

Las noticias falsas se difunden más rápido y con mayor alcance que las verdaderas, especialmente en temas **políticos, económicos y de salud**, presentan alta velocidad, ubicuidad y multiplicidad de medios.

Se requiere un enfoque multidisciplinario y multifacético basada en una sólida educación con el desarrollo de nuevas tecnologías disruptivas (IA) que permitan contrarrestar la desinformación por medio de patrones de difusión para desarrollar filtros de detección rápida de noticias falsas (analogía con modelos epidemiológicos de propagación de un

virus) estudiando distintos niveles de complejidad por medio **de secuencias de activación.**

- 1% noticias falsas difundieron entre 1.000 a 100.000 personas
- 1% noticias verdaderas se difundieron rara vez a 1.000 personas

Salud

- Movimientos antivacunas encontraron un caldo de cultivo ideal con la pandemia del covid-19.
- Curas milagrosas
- Negacionismo climático
- Desinformación cambio climático
- Desinformación generando prejuicios y estigmatización en caso LGTBI

Político

- Herramienta de desinformación socavar la confianza de las instituciones democráticas, perfiles falsos y manipulación psicológica.

Economía

- Difusión de rumores tiene un impacto en los mercados financieros

Violencia de genero

- Discriminación y violencia de genero.
- Desinformación puede alimentar la intolerancia, odio, discriminación en minorías religiosas.

Migración

- Divisiones sociales y violación de derechos humanos a los migrantes.
- Promover estereotipos sobre las posibles amenazas de seguridad que los migrantes

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

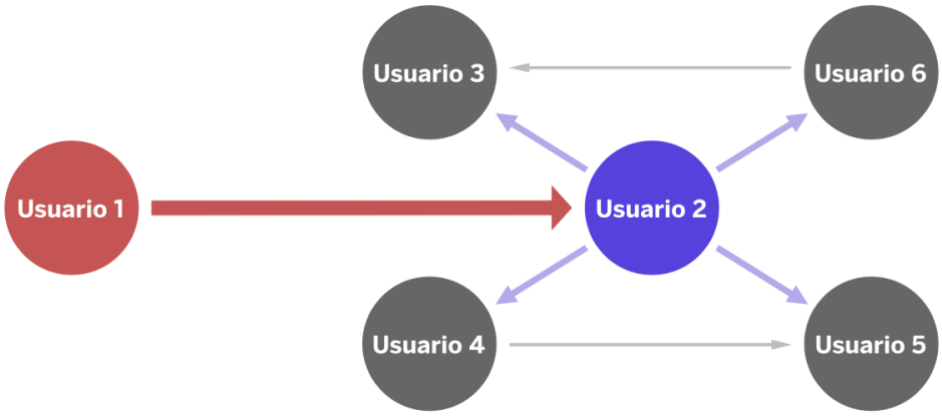
Modelos de redes latentes de difusión

los Modelos generativos de redes latentes de difusión permiten identificar la difusión de información entre usuarios clave de una red y dinámicas en la propagación de la desinformación a lo largo del tiempo, superando la generalización de los modelos epidemiológicos.

Permite predecir los siguientes elementos:

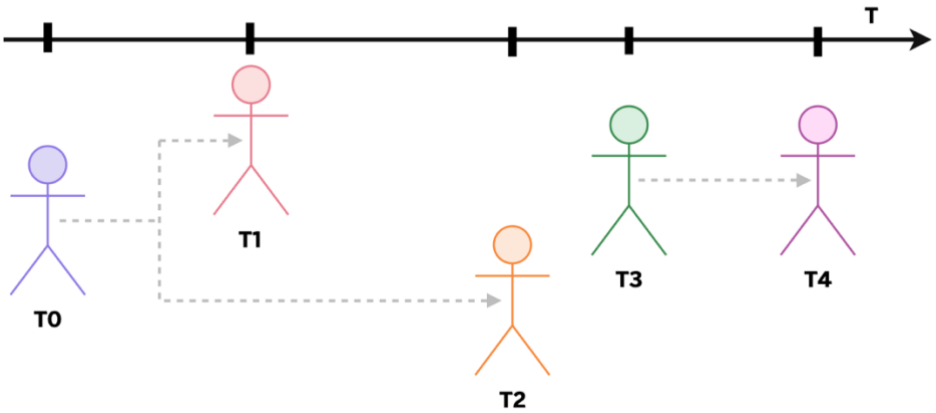
- la propagación de la información
- Quien la propaga (actores y comunidades influyentes)
- Como lo hace.
- Identifica la estructura de las relaciones dentro de la red
- Densidad de conexiones
- Crear a partir de cascadas de difusión como los usuarios influyen mutuamente.

Figura 2. *Modelo de red latente de difusión compuesta por seis usuarios.*



Las fechas indican la dirección como fluye la información y el espesor de la línea es proporcional a la intensidad de la influencia.

Figura 3. *Modelo de red latente de difusión compuesta por cinco usuarios.*



El individuo azul habla de un temo en un momento (TO) e interactúa con rojo y naranja, que también han hablado sobre el mismo en un momento (T1) y (T2)

Modelo epidemiológico SIR

Figura 4. *Modelo epidemiológico SIR (susceptible, infectada, recuperada)*

S: personas susceptibles de contagiarse

I: Personas infectadas

R: Personas recuperadas

β: Probabilidad que un individuo susceptible se contagie

ɣ: Probabilidad que un individuo contagiado se recupere.

Para el caso de desinformación se usa secuencias de activación ajustando el modelo SIR.

- Calcular tamaño de grupo susceptibles (S) así como (β y ɣ) que mejor se adapten a los datos observados para predecir cuantas personas habrá por grupos y hacer simulaciones.
- Este modelo permite detectar el flujo de información, pero no quien lo está causando.

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

   Son sistemas de (IA) generativa capaces de generar texto, video e imágenes convincentes. Su principal riesgo es la generación de contenido falso difícil de distinguir de la verdad, especialmente si se usan de forma malintencionada.

Los modelos y análisis de la propagación de la información en redes sociales con (IA) generativas en texto, video, audio e imágenes (CHAPGPT 4), los cuales se relacionan las siguientes aplicaciones así: secuencias y manipulación de texto (DALL E 2 , AUG X LABS, SLIDESAI y TRANFORMER), generación y manipulación de imágenes (Stable diffusión), economía de micro encargos (gig economy) y redes neuronales profundas para edición y manipulación de imágenes (DEEPART.IO, RUNWAYML, NVIDIA STYLEGAN y DEEP FAKES) convirtiéndolo en una herramienta muy peligrosa porque pueden realizar:

- Suplantación de personas.

- Aprender patrones y características visuales
- Generación de imágenes realistas (utiliza redes generativas adversarias (GAN) y redes neuronales convencionales (CNN).
- Aprendizaje basado en datos
- Variedad en la generación
- Control creativo
- Transferencia de estilo
- Manipular fotogramas para pasar de imágenes a videos
- Manipulación de audio es una modalidad en avance con capacidades de desinformación, la cual intenta imitar la voz humana con los siguientes softwares (clonado (FEW SHOT), manipulación y sintetizacion).
- Las técnicas de texto a imagen son utilizadas con los siguientes softwares (OPENAI JUKEDECK, AMPER MUSIC, MAGENTA NSYNTH, JUKIN MEDIA EKO, AIVA, DEEPDUB.AI)

Se debe tener mucho cuidado porque estos modelos no entienden la veracidad de la información generando riesgos, puede alucinar generando información falsa y tampoco tiene mecanismo para frenar dicha generación de desinformación.

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

Al estar disponibles públicamente, incluso personas sin conocimientos técnicos pueden usarlos para generar y difundir desinformación.

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

Se refiere a la falta de transparencia en las decisiones que toman los modelos de (IA). El desafío es comprender y justificar cómo se llegó a una conclusión, especialmente en contextos sensibles como la lucha contra la desinformación.

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

La AGI implicaría máquinas con capacidades cognitivas similares a los humanos. Aunque promete soluciones más inteligentes, también conlleva riesgos si no se controla adecuadamente, pues podría amplificar la desinformación en lugar de reducirla.

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

Reglamento General de Protección de Datos (RGPD) de la comisión europea 2016 y el libro blanco de la (IA) de la comisión europea 2020, (toda persona tiene derecho a la

protección de los datos de carácter personal que le conciernan) así mismo, la Ley de Servicios Digitales (DSA) como marcos legales clave.

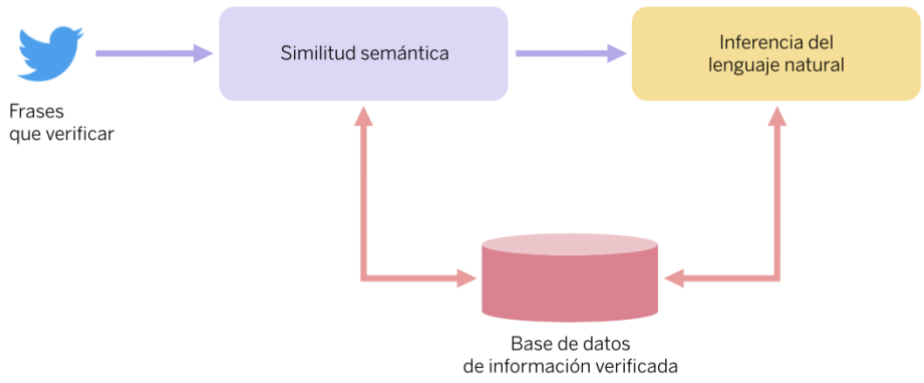
10. **¿Cómo garantiza FacTeR-Check el cumplimiento de la normativa de protección de datos al analizar redes sociales?**

La arquitectura FacTeR-Check es una herramienta implementada en el proyecto (CIVIC) la cual permite detectar y contrarrestar la desinformación mediante el uso de distintas herramientas de (IA), analizando el proceso de verificación, monitorización y lucha contra la desinformación en redes sociales mediante las API (interfaz de programación de aplicaciones).

Utiliza técnicas de enfoque de verificación semiautomática (*natural lenguaje undertanding) y automática* con el uso de técnicas de machine learning de comprensión del lenguaje humano utilizado el contexto del actor que crea la desinformación y la semántica del lenguaje en diferentes idiomas.

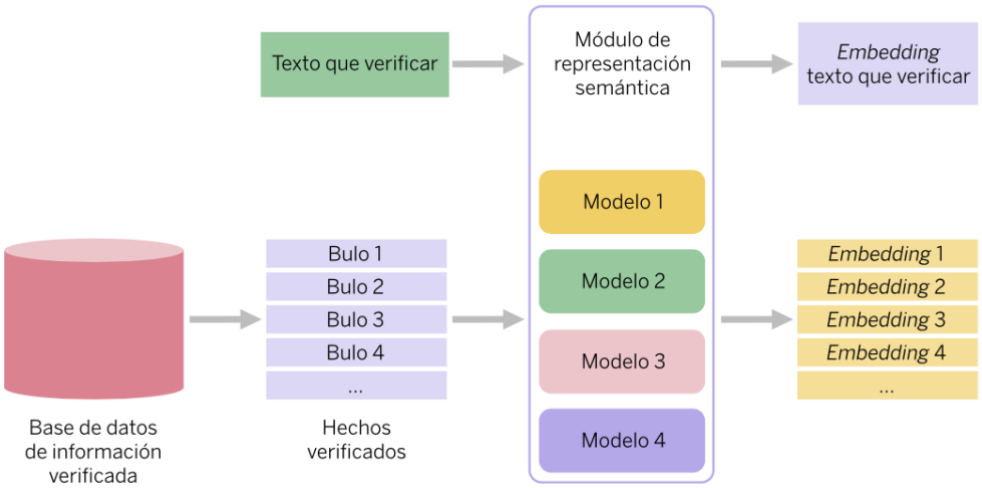
El modelo entrenado verifica el grado de alineamiento entre dos textos, no verifica directamente piezas de información sino contrasta con una base de datos de hechos verificados con entidades de *fact checking* generando fiabilidad en sus respuestas por medio de tres tareas la detección, verificación de frases y recuperación de evidencias.

Figura 5. *Modelo de la herramienta FacTeR-Check.*



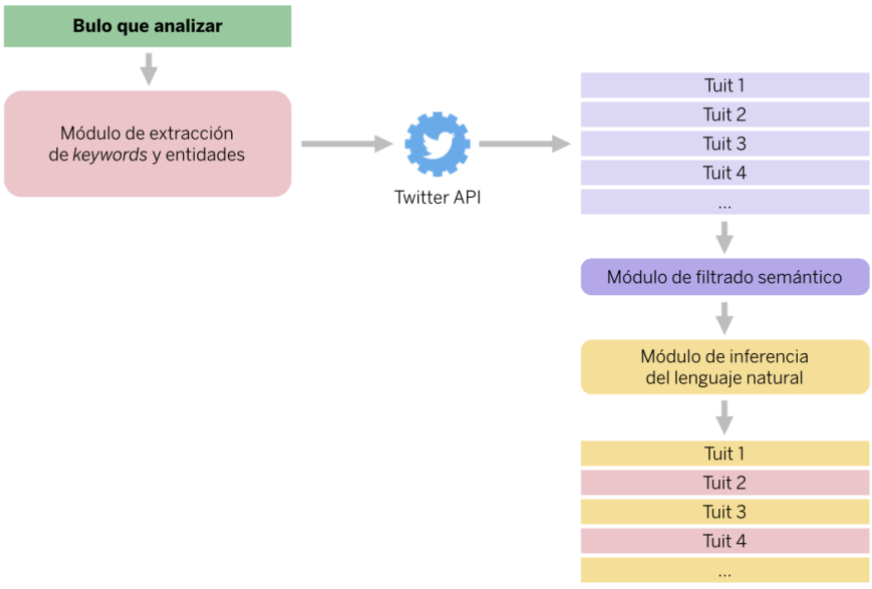
Ofrece una herramienta de verificación de una frase o afirmación si un texto es veraz para el público y organismos como (*fact checking)* con el objetivo de combatir y monitorear el fenómeno de la desinformación, para contener fenómenos como la polarización, resonancia mediática, influencia y presión social.

***Figura 6**. Flujo de trabajo de la herramienta FacTeR-Check por medio de generación de embeddings (representación de datos en vectores) de los bulos (ensamble de varios modelos)*



Estos modelos se basan en la arquitectura (Transformer XLM-R y MiniLM y el conjunto de datos mSTSb para determinar los embeddings por la herramienta distancia coseno para calcular la cercanía en el hiperespacio en el grado de similitud semántica y luego pasa a el módulo de naturaleza de lenguaje para analizar el nivel de implicación entre el texto y los hechos verificados.

***Figura 7**. Visualizacion del flujo de trabajo de la herramienta FacTeR-Check*



Así mismo, anonimiza los datos personales y trabaja con metadatos y análisis agregado para respetar el (RGPD) y evitar la identificación directa de usuarios.

Actualmente el sistema realiza la veracidad y aproximación del uso del contenido y sus fuentes de la información, pero se debe seguir explorando los campos de estilo y el contexto de las desinformaciones.

11. **Analice las diferentes formas en que la Inteligencia Artificial puede ser utilizada tanto para generar como para combatir la desinformación.**

La (IA) puede facilitar la creación de bulos mediante modelos generativos como los grandes modelos de lenguaje, que producen textos o imágenes falsas difíciles de detectar. Al mismo tiempo, la IA se emplea para detectar patrones de desinformación, analizar redes sociales, identificar cuentas sospechosas y verificar hechos (fact-checking). FacTeR-Check es un ejemplo concreto: usa IA para monitorizar y clasificar información potencialmente falsa en varios idiomas y plataformas, aplicando procesamiento de lenguaje natural y análisis de contexto. Así, la IA se convierte en un arma de doble filo: peligrosa si se usa sin control, pero poderosa si se orienta al bien común.

12. **Discuta el papel de la Inteligencia Artificial Explicativa (XAI) en la mejora de la confianza pública en los sistemas de detección de desinformación.**

La (IA) explicativa busca hacer comprensibles y transparentes las decisiones de los modelos, lo que es crucial para que usuarios y entidades confíen en los sistemas que detectan bulos permitiendo tener una mejor compresión de cómo y por qué se clasifica el contenido como desinformación generando una mayor confianza en esta tecnología.

La (IA) explicativa permite descomponerla desinformación en forma efectiva determinando piezas de texto contrastables y ayudando a los usuarios a entender porque ciertas afirmaciones son incorrectas personalizando el grado de detalle y tipo de explicación en función del usuario aprendiendo a detectar la desinformación.

Sin embargo, los modelos actuales suelen ser "cajas negras", dificultando la interpretación. Esto limita su aceptación en ámbitos sensibles como política o salud. Obstáculos principales son la complejidad técnica, la escasa estandarización y la falta de soluciones universalmente aplicables. Aun así, el desarrollo de (XAI) es esencial para lograr una (IA) ética y confiable.

13. **Compare los modelos epidemiológicos y las redes latentes de difusión como enfoques para estudiar la propagación de la desinformación en redes sociales.**

Los modelos epidemiológicos tratan la desinformación como un virus, con tasas de contagio y recuperación, lo que permite estimar su propagación. Son útiles para analizar tendencias generales. En cambio, las redes latentes identifican relaciones ocultas y actores clave, permitiendo una intervención más precisa y personalizada. Mientras los primeros

son más abstractos y globales, los segundos ofrecen un análisis detallado de la estructura y dinámica real de las redes sociales.

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?**

La disponibilidad de IA generativa en plataformas abiertas ha permitido que individuos sin formación técnica creen fácilmente contenido falso. Esto multiplica la amenaza de desinformación viral. El documento sugiere fomentar la transparencia, crear filtros de verificación y desarrollar contramedidas como IA para detección automática de bulos, además de campañas de alfabetización mediática para el público.

15. **Analice las consideraciones éticas y de privacidad asociadas con el uso de la IA para combatir la desinformación.**

Aunque la (IA) es útil para analizar y combatir desinformación, su uso plantea riesgos como la vigilancia masiva, el sesgo algorítmico o la falta de consentimiento informado. Por eso, herramientas como FacTeR-Check deben ajustarse al RGPD y otras normativas como la DSA. En Colombia, aún no existe una legislación específica para IA, aunque se han planteado lineamientos desde MinTIC y CONPES relacionados con la ética y privacidad en tecnología. La armonización con estándares internacionales será clave para un uso responsable y legalmente sólido.