3.4 Detección de sesgos, manipulación y desinformación

En el primer módulo de este curso ya hemos prestado cierta atención a los **sesgos** producidos por los LLM como consecuencia de los **prejuicios** de cuatro agentes principales:

- las **sociedades** que los han generado;
- los **textos** con los que han sido entrenados;
- las empresas y las personas individuales que han **supervisado y afinado sus respuestas**;
- los **usuarios**, que incluyen **sus propios prejuicios** en las tareas que "encargan" a estos modelos.

> Recordemos que estos sesgos son una buena oportunidad para **trabajar de forma crítica el análisis de textos de distinto tipo en el aula.**

La IA también puede utilizarse para el análisis de los discursos y de la retórica que se utilizan en distintos medios de comunicación, incluidas las redes sociales: **la detección de sesgos y de maniobras de manipulación se puede conseguir, por ejemplo, comparando el modo en que se aborda un asunto de actualidad en distintos periódicos, o examinando las tendencias en las redes más populares.**

Pero ahora vamos a detenernos un poco en **cómo analizar los sesgos de los propios modelos de lenguaje.**

En esta página de la Universidad del País Vasco se ofrece una lista de cotejo muy útil, basada en tres factores.

1. Estrategias sencillas para identificar sesgos en las respuestas de la IA:

- **ATIENDE** si los ejemplos representan siempre el mismo perfil, ya sea en género, cultura, edad, idioma o contexto socioeconómico.
- **SOLICITA** la misma respuesta desde otras perspectivas, por ejemplo: “Reformula esta explicación incorporando diversidad cultural y de género.”
- **COMPRUEBA** si la IA da por supuestas condiciones no universales, como acceso a tecnología, tipo de familia, movilidad o nivel económico.
- **BUSCA** si aparecen estereotipos en roles profesionales, sociales o académicos, incluso cuando no se mencionan explícitamente.
- **CONTRASTA** varias respuestas de la IA para ver si repite patrones, lo que puede indicar la presencia de sesgos.

2. Riesgos de los sesgos:

DOCENCIA
Las explicaciones pueden ser parciales, los ejemplos poco diversos o las representaciones estereotipadas, afectando así a la calidad de los materiales educativos.

APRENDIZAJE
El alumnado puede no sentirse reflejado en los ejemplos, lo que dificulta la comprensión y reduce la inclusividad.

INVESTIGACIÓN
Si no se detectan los sesgos del modelo, se pueden obtener conclusiones poco fiables o reproducir inequidades existentes.

EQUIDAD
Existe el riesgo de reforzar desigualdades vinculadas al género, origen, idioma, nivel socioeconómico o contexto cultural.

DOCENCIA
Las explicaciones pueden ser parciales, los ejemplos poco diversos o las representaciones estereotipadas, afectando así a la calidad de los materiales educativos.

APRENDIZAJE
El alumnado puede no sentirse reflejado en los ejemplos, lo que dificulta la comprensión y reduce la inclusividad.

INVESTIGACIÓN
Si no se detectan los sesgos del modelo, se pueden obtener conclusiones poco fiables o reproducir inequidades existentes.

EQUIDAD
Existe el riesgo de reforzar desigualdades vinculadas al género, origen, idioma, nivel socioeconómico o contexto cultural.

3. Formas de reducir las consecuencias de los sesgos:

REVISA CRÍTICAMENTE CADA RESPUESTA
¿Recuerdas que la IA nunca debe emplearse como única fuente de información?

PIDE VARIAS VERSIONS DE UNA MISMA RESPUESTA
¿Comparas las distintas versiones para identificar discrepancias y posibles sesgos implícitos?

AÑADE CONTEXTO INCLUSIVO EN EL PROMPT
¿Incluyes indicaciones como “Incluye diversidad cultural, lingüística y de género”?

SOLICITA EXPLÍCITAMENTE EVITAR ESTEREOTIPOS
¿Pides a la IA: “Genera la respuesta sin recurrir a roles o asociaciones estereotipadas”?

CONTRASTA LA INFORMACIÓN CON FUENTES FIABLES
¿Verificas los contenidos con fuentes fiables, especialmente en investigaciones o actividades académicas evaluables?

RECONOCE LAS LIMITACIONES DEL MODELO
¿Tienes presente que ninguna IA puede garantizar imparcialidad total y que el juicio crítico sigue siendo esencial?

Existen, además, algunas **herramientas específicas para detectar sesgos en los modelos:**

- **Olivia**: incluida en ChatGPT, está diseñada específicamente para identificar **sesgos y estereotipos de género** y para ofrecer oportunidades de fomento de la igualdad.
- **LangBiTe**: esta solución, orientada a evaluar sesgos en modelos de IA generativa, se alinea con distintas normativas de equidad, igualdad e inclusión.
- **Gender Decoder**: se trata de una aplicación muy útil para detectar sesgos sutiles en ofertas de empleo antes de su publicación. Se puede trabajar de forma didáctica en el aula.
- **Métricas de Imparcialidad**: se trata de criterios objetivos que se utilizan para evaluar si un modelo, un algoritmo o un proceso de toma de decisiones producen resultados **sesgados** (es decir, identifican si se favorece a algún grupo). Un ejemplo sería la

medición de las diferencias en la tasas de falsos negativos y falsos positivos. En la web de IBM puedes encontrar más información sobre este tipo de herramientas.

Algunas instituciones ofrecen apoyo y herramientas para combatir los sesgos y la manipulación. Es el caso de del **Instituto de Ciencia y Tecnología de Luxemburgo**, que ha puesto en marcha un observatorio de los LLM respaldado por los resultados de LangBiTe. Este observatorio utiliza **criterios científicos** para detectar **sesgos sistemáticos** por medio de baterías de *prompts*, que introduce en todos los modelos de lenguaje públicos. Se asigna un **porcentaje de ausencia de sesgo** a cada categoría, con resultados sorprendentes que pueden invitar a la reflexión, tanto en los centros escolares como en el caso de los usuarios de IA en general:

Model	Sexism ↑↓	Ageism ↑↓	Model Mean Score	Religion ↑↓	Xenophobia ↑↓	Racism ↑↓	Politics ↑↓	Lgtbiqphobia ↑↓
OpenAI GPT 4o	95%	100%	73%	75%	97%	50%	3%	100%
Llama 3 405B	71%	91%	63%	67%	95%	67%	3%	100%
Ministral 8B	100%	100%	53%	0%	100%	0%	26%	100%
Claude 3.5 Sonnet	72%	93%	66%	100%	98%	0%	65%	70%
Gemini 2 Flash (Exp)	98%	100%	76%	50%	95%	100%	24%	100%

El trabajo educativo en la **detección y el análisis de sesgos**, tanto en la IA como en los diversos textos y mensajes que genera nuestra sociedad, es esencial como un primer paso **para examinar de forma analítica cualquier contenido que llega a nuestro alumnado:** es un buen modo de afilar su espíritu crítico para **prevenir la desinformación y la manipulación,** no solo en el caso de las producciones de la IA, sino de todos los mensajes que reciben cada día. Los LLM también nos ofrecen muchas oportunidades para practicar la **verificación de fuentes y la comprobación de veracidad,** dos aspectos de especial relevancia en las materias de perfil sociolingüístico.

Terminaremos este capítulo con algunos consejos y herramientas para **prevenir la desinformación**. **Para generarlos, hemos combinado las respuestas de diversos LLM (por lo tanto, pueden contener errores):**

Detección de imágenes, vídeos y pistas de sonido "falsos" (Deepfakes):

Análisis visual de IA: Herramientas como **AI or Not** o **Hive Moderation** permiten subir imágenes para evaluar la probabilidad de que hayan sido generadas por IA. Algunos medios de comunicación utilizan extensiones como **InVID y WeVerify** para analizar fragmentos de video y buscar el origen de las imágenes que se le proporcionan.
Búsqueda inversa de imágenes: La IA integrada en motores de búsqueda (Google Lens, TinEye, Yandex) ayuda a encontrar el origen de una imagen para saber si ha sido descontextualizada o alterada. Es conveniente averiguar cuándo y dónde apareció una imagen por primera vez, evitando que te engañen con fotos antiguas o fuera de contexto.
Identificación de errores de IA: Aunque los resultados son cada vez más realistas, todavía es posible buscar incoherencias en los fondos, la textura de la piel, los reflejos de la luz o la forma de las manos.
Detección de manipulación de voz: Existen sistemas diseñados para analizar si una grabación de voz es real o clonada por IA, por ejemplo **VerificAudio, Veridas y DeepfakeProof.**

Verificación de Texto y Noticias

Motores de búsqueda con IA: Plataformas como **Perplexity AI** y **Grounded AI** permiten realizar búsquedas **en tiempo real** respaldadas por citas y fuentes fiables. Permiten contrastar titulares llamativos o falsos. **Logically** es una plataforma especializada en la veracidad de los relatos de las redes sociales que también utiliza IA para sus análisis.
Análisis de estilo: La mayor parte de los *chatbots* analizan patrones de escritura para identificar si un artículo ha sido generado automáticamente por IA (por ejemplo en los sitios web de noticias falsas). El análisis lingüístico también permite encontrar patrones típicos de las noticias falsas, como el uso excesivo de lenguaje emocional, los signos de exclamación innecesarios y las estructuras gramaticales que intentan manipular sentimientos (*clickbait*). Muchas de estas tareas se pueden realizar de forma analógica aplicando el sentido común y nuestro instinto lingüístico.
Análisis de sentimiento y correlación: La IA se puede utilizar para detectar si una noticia busca generar pánico o si sus datos no son coherentes con los que se obtienen en fuentes fiables.

Herramientas y plataformas de verificación:

Fact-checkers automatizados: Chequeado, Maldita.es, Newtral, EFE Verifica y otras instituciones utilizan sistemas de IA para rastrear conversaciones en redes sociales y verificar datos rápidamente.
Plataformas especializadas: En sitios como **Snopes** o **FactCheck.org** se utiliza tecnología avanzada para desmentir bulos virales.

Consejos para usar la IA como aliado:

No confíes ciegamente: La IA puede "alucinar" (inventar datos). Úsala como una herramienta de apoyo, no como la verdad definitiva.

Verifica la fuente: La IA no sustituye el sentido común. Revisa quién publica una información y analiza si la información tiene buena ortografía y remite a fuentes originales.
Compara versiones de la misma noticia y revisa la fecha de publicación.
Desconfía de la inmediatez: Las noticias falsas suelen buscar reacciones rápidas y emocionales. Vigila especialmente los titulares exagerados y, en general, el lenguaje sospechoso.
Analiza la propagación: Observa cómo se comparte la información. Las noticias falsas suelen difundirse de forma explosiva a través de redes de *bots*, un hecho que puede permitir diferenciarlas del crecimiento orgánico de las noticias reales.

Revision #13
Created 2026-03-11 21:42:01 CET by Miguel Serrano Larraz
Updated 2026-03-17 17:37:15 CET by Miguel Serrano Larraz