**Etiquetado de IA**

Introduccion: desafios

desinformacion, deepfakes, manipulacion, etc

impacto en politica, educacion, medios y redes sociales.

que es etiquetado en ia

añadir al contenido una etiqueta para hacer saber al usuario que no es contenido generado por una persona sino por una ia.

segun isibilidad:

* visible: marca de agua
* metadatos: marcas de agua digitales. manipulables si no se protegen bien.

metodo:

* manual
* automatico: herramienta lo incluye por defecto
* deteccion: identificar si lo ha hecho una ia

por que es necesario?

* transparencia y confianza
* cumplimiento normativo
* proteccion contra abusos
* etica en contextos sensibles

tecnologias

* watermarking en origen: marcas invisibles incrustadas durante la generacion
* atribucion y trazabilidad: metadatos estructurados
* deteccion posterior

etiquetado en origen

visibles

caracteristicas:

* inserción en espacio latente (donde el usuario no lo pueda ver)
* marcas persistentes
* no afectan a la calidad

ejemplos:

* synthId (google)

Google Deepmind

* se puede detectar despues de distintas modificaciones

synthid - texto

* cada token tiene una probabilidad para añadirle watermarking
* cada palabra tiene probabilidad de tener watermarking
* probabilidad de la combinación
* el texto queda igual de bueno pero esas palabras estan alteradas para formar parte del watermarking

symthid - audio

* se modifican ciertos espectogramas para que se note qye estan modificados para generar la marca de agua
* inaudible para el ser humano
* estables delante de compresion
* ser detectables
* se convierte el audio a espectograma y en éste se busca con cnn o red bayesiana la marca de agua.

synthid - image/video

* se marca cada uno de los fotogramas del video
* en las imagenes se marca la imagen en si

metadatos, firmas digitales y trazabilidad

se aplica criptografia o a los metadatos.

se aplica hash para firmar digitalmente.

trazabilidad: cadenas de confianza, se hace hash de los metadatos y firma digital y se añade a la cadena. asi se pued seguir el historial como una blockchain.

C2PA - content provenance and authenticity

permite asociar informacion de procedencia aun archivo de medios. todo queda firmado digitalmente para que cualquier persona o sistema pueda verificarlo.

deteccion posterior

una vez hecho el etiquetado, se analiza usando modelos XAI entrenados. detectan patrones sinteticos estdisticos.

polygraf

detecta texto generado por ia

clasificacion + XAI

analisis estructural → deteccion semantica → clasificadores entrenados → informe con probabilidad y evidencias

implicaciones eticas

transparencia: protege la autonomía del usuario

derechos de autor

obras generadas por ia no estan protegidas legalmente

importante para sabe rel el origen en uso fraudulento.

opinion publica

65% de la gente cree que se deberia etiquetar a la ia

retos y limitaciones

deteccion automatica poco fiable

no existe un estandar internacional nificado

no todos los dispositivos reconocen los mismos metadatos

Futuro

detectar patrones automáticamente para saber si es ia o no, algoritmos para detectarlo.

educacion digital y conciencia publica, informarse

AI act europeo plantea esta iniciativa para etiquetarlas