Avances, desafíos y limitaciones en el uso de herramientas de anonimización automática de texto, audio, imágenes y vídeo[1]

Advances, challenges and limitations in the use of automatic anonymization tools for text, audio, images and video.

Jorge Rico Ródenas, María Viana Rozas, María José Vilella Sánchez, Lara De Santos Tabares, Fernando Espinoza Cuadros, Ricardo Heredia Joya, Silvia Rodríguez-Jiménez, Jaime Cortón-González, Pablo Regodón-Cerezo, Rubén Pérez Ramón y Joan Orti Navarro
Miembros de Sigma Cognition

Fecha de recepción: 15.01.2026
Fecha de aceptación: 22.06.2026
Fecha de publicación en línea: octubre 2026

Resumen:

La anonimización automática de datos no estructurados, como imágenes, vídeos, grabaciones de audio o documentos de texto libre, ha experimentado en los últimos años avances significativos debido en gran medida a la mejora en los modelos de Inteligencia Artificial, dedicados al reconocimiento de entidades, personas y objetos. En este artículo se presenta, en primer lugar, una breve introducción al ámbito de la anonimización en general y, en particular, a la anonimización de datos no estructurados, también denominados multimodales o multimedia.
Posteriormente se ofrece un análisis comparativo de herramientas comerciales y de código abierto que llevan a cabo tareas de anonimización sobre distintos tipos de datos. Por último, se muestran los avances, retos y dificultades en la automatización de la anonimización de textos, del habla, de la voz o de la imagen, identificadas en la bibliografía y en el marco de un proyecto de investigación que llevan a cabo los autores.

Palabras clave:

Anonimización; multimodalidad; datos no estructurados; modelos de lenguaje; inteligencia artificial generativa.

Abstract:

The automatic anonymisation of unstructured data, such as images, videos, audio recordings or free text documents, has seen significant advances in recent years, largely due to improvements in Artificial Intelligence models dedicated to the recognition of entities, people and objects. This article begins with a brief introduction to the field of anonymisation in general and, in particular, to the anonymisation of unstructured data, also known as multimodal or multimedia data. Subsequently, a comparative analysis of commercial and open-source tools that perform anonymisation tasks on different types of data is provided. Finally, the advances, challenges and difficulties in the automation of text, speech, voice and image anonymisation, identified in the literature and in the context of a research project carried out by the authors, are presented.

Keywords:

Anonymisation; multimodality; unstructured data; language models; generative artificial intelligence.

[1] Este proyecto está cofinanciado por los fondos Next Generation EU del Plan de Recuperación, Transformación y Resiliencia, gestionado por Red.es y el Centro de Referencia de Espacios de Datos (CRED) del Ministerio para la Transformación Digital y de la Función Pública.

Sumario:

1. Introducción. **2.** Los datos no estructurados. **3.** Análisis comparativo de herramientas. **4.** Desafíos de la anonimización de datos multimodales. **4.1.** Modelos generativos y extracción de información. **4.2.** Escasez de datos anotados. **4.3.** Entidades en contextos multilingües y multimodales. **4.4.** Más allá de las entidades: autoría y estilo. **4.5.** Datos personales en el habla. **4.6.** La voz como dato personal. **4.7.** Anonimización y autorreconocimiento. **4.8.** Datos personales en contenido visual. **4.9.** La IA generativa como estrategia de seudonimización. **4.10.** La sincronización de voz e imagen en vídeos anonimizados. **5.** Conclusiones.

Summary:

1. *Introduction.* ***2.*** *Unstructured data.* ***3.*** *Comparative analysis of tools.* ***4.*** *Challenges in the anonymization of multimodal data.* ***4.1.*** *Generative models and information extraction.* ***4.2.*** *Scarcity of annotated data.* ***4.3.*** *Entities in multilingual and multimodal contexts.* ***4.4.*** *Beyond entities: authorship and style.* ***4.5.*** *Personal data in speech.* ***4.6.*** *Voice as personal data.* ***4.7.*** *Anonymisation and self-recognition.* ***4.8.*** *Personal data in visual content.* ***4.9.*** *Generative AI as a pseudonymization strategy.* ***4.10.*** *Synchronization of voice and image in anonymized videos.* ***5.*** *Conclusions.*

1. Introducción

El crecimiento exponencial del Big Data, de la conectividad celular (3G en adelante), del internet de las cosas (IoT por sus siglas en inglés), la inteligencia artificial (IA) y los servicios en la nube ha llevado a un aumento sin precedentes en la recopilación de datos tanto en organismos públicos como en entidades privadas. Esto es consecuencia directa de la digitalización de la sociedad actual, en la cual casi todos los servicios tienen ya una presencia online casi obligatoria. Ante este panorama, la protección de la privacidad individual se ha convertido en una prioridad tanto para gobiernos como para organizaciones privadas, lo que ha impulsado la evolución de herramientas y técnicas de anonimización. Actualmente, muchas aplicaciones de IA requieren el uso de datos que no contengan información personal, es decir, que no permitan la identificación de una persona por su imagen, voz o datos, como ocurre en aplicaciones de automoción y en numerosos casos de uso industriales, públicos y comerciales.

Por otro lado, el uso de los datos se ha vuelto una parte fundamental en la mejora de la operativa y en los resultados de prácticamente cualquier empresa, ya que los datos constituyen el alimento de los modelos de aprendizaje automático y su calidad y cantidad resultan determinantes. En este contexto, el interés por la conservación y almacenamiento de datos personales ha crecido tanto para los individuos como para las organizaciones que deben cumplir con marcos legales como el Reglamento General de Protección de Datos (GDPR) en la Unión Europea. Este establece principios como la minimización de datos y la limitación de la finalidad (art. 5), así como la obligación de implementar medidas técnicas y organizativas adecuadas, incluyendo la anonimización o seudonimización cuando sea posible (arts. 25 y 32). De forma análoga, la Ley de Portabilidad y Responsabilidad de Seguros Médicos (HIPAA) en Estados Unidos establece criterios específicos para la desidentificación de datos de salud (45 CFR 164.514), incluyendo métodos como *Safe Harbor* o *Expert Determination*, permitiendo su uso sin restricciones cuando dejan de considerarse información personal identificable. Más allá de la legislación, la evolución de la regulación de la privacidad refleja una creciente sensibilización ante los riesgos asociados a la digitalización y la recolección masiva de datos personales[2], que hoy en día incluyen también medios multimodales como grabaciones de voz, imágenes, geolocalización o señales biométricas. Durante las décadas pasadas, la rápida adopción de infraestructuras basadas en la nube y la mercantilización de los datos de comportamiento[3] han ampliado la tensión entre privacidad y utilidad, lo que ha puesto de manifiesto la insuficiencia de los marcos legales por sí solos y la necesidad de soluciones técnicas que complementen los esfuerzos regulatorios.

Los casos recientes de violación de datos personales como los sufridos por Facebook (extracción ilícita de datos de unos 87 millones de perfiles en el caso *Cambridge Analytica*), Yahoo! (brecha masiva que afectó a más de 500 millones de cuentas) o Netflix (exposición de datos vinculados a contenidos y usuarios en incidentes de terceros), por citar los más conocidos, justifican por sí solos la necesidad de esta tecnología. En este contexto, los procesos de anonimización han ido evolucionando y adaptándose a las necesidades de una sociedad cada vez más comprometida con la seguridad de los individuos. No obstante, la industria se enfrenta actualmente a varios desafíos clave, entre ellos:

[2] Solove, Daniel J. *Understanding Privacy*. Harvard university press, 2010.

[3] Zuboff, Shoshana. "The Age of Surveillance Capitalism". In *Social Theory Re-Wired*, 203–13. Routledge, 2023, https://doi.org/10.4324/9781003320609.

- Encontrar un equilibrio entre utilidad y privacidad, dado que un mayor grado de anonimización puede reducir o incluso destruir el valor analítico de los datos.
- Reducir el riesgo de la reidentificación, que aflora al combinar múltiples conjuntos de datos.
- Desarrollar sistemas de anonimización en tiempo real, algo especialmente crítico en sistemas de IoT o de monitorización continua.

De entre todos los sectores, el de la salud es probablemente el más sensible en términos de privacidad de datos, dado el carácter altamente sensible de la información médica. En este **ámbito** se emplean modelos híbridos que combinan k-anonimidad con reglas basadas en privacidad diferencial[4], lo que permite la publicación de conjuntos de datos sintéticos representativos sin comprometer la identidad de los pacientes, como ocurre en algunas soluciones desarrolladas por plataformas como IBM Explorys o Google Health. A pesar de ello, se han documentado casos de reidentificación[5] al combinar datos clínicos anonimizados con información genética u otras fuentes externas, lo que evidencia la necesidad de mejorar las metodologías actuales. En respuesta, las tendencias recientes en anonimización apuntan hacia una mayor automatización, el uso de la inteligencia artificial, la generación de datos sintéticos mediante modelos como las GAN[6] y la integración del principio de privacidad por diseño a lo largo de todo el ciclo de vida de los sistemas de información.

2. Los datos no estructurados

La mayoría de los procesos de anonimización se aplican sobre registros de bases de datos o documentos de texto estructurados, es decir, documentos tabulados (por ejemplo, hojas de cálculo o documentos CSV) en los que cada columna representa un tipo de dato privado distinto: el nombre, la edad, el teléfono, la profesión, etc. En este tipo de documentos es relativamente sencillo aplicar técnicas de anonimización, puesto que los datos ocupan posiciones predecibles dentro del documento. Algunas de estas técnicas son: el intercambio o aleatorización de datos entre distintas apariciones del mismo valor; la generalización o sustitución de datos concretos por otros más genéricos; o la sustitución del dato privado por una etiqueta, una serie de asteriscos, una clave criptográfica o incluso un sinónimo generado de forma artificial. La *Guía básica de anonimización de la Autoridad de Protección de Datos de Singapur*[7], traducida y publicada por la Agencia Española de Protección de Datos, contiene recomendaciones sobre la anonimización de este tipo de datos, y la herramienta que distribuye la misma agencia implementa alguna de esas funciones para ser usadas con el software Microsoft Excel.

En los documentos de texto no estructurado, o texto libre, por otro lado, no es posible predecir la posición en la que aparecerán los datos personales, sino que hay que detectarlos previamente mediante técnicas de reconocimiento de entidades nombradas (NER, por sus siglas en inglés). Una vez detectada la información personal, podemos aplicar aproximadamente las mismas técnicas de anonimización que se describen para los documentos estructurados.

Para cada tipo de dato privado conviene un tipo de estrategia distinta. Así, por ejemplo, para anonimizar la edad se suele sustituir el valor real por un rango y para anonimizar números de cuenta se emplea habitualmente el enmascaramiento parcial del número por asteriscos. Por su parte, en los documentos audiovisuales también hay aspectos que revelan la identidad de las personas que se deben anonimizar. En el audio puede haber información personal como nombres, direcciones o teléfonos que se suelen anonimizar insertando silencios o tonos puros. No obstante, la voz también es un rasgo que permite la identificación de las personas. Las técnicas más utilizadas para anonimizar la voz van desde la distorsión de la señal de voz mediante la adición de ruido o modificación de aspectos acústicos, hasta la resíntesis mediante conversores de texto a voz. Una técnica de anonimización de la voz es la ofuscación de atributos de locutor, que consiste en suprimir características de la voz que revelan información sobre la edad, sexo, o procedencia del hablante sin eliminar por completo otros aspectos que permiten asociar la voz a una única persona. Por último, en la imagen y el vídeo, algunas de las técnicas más utilizadas tradicionalmente para anonimizar

[4] El Emam, Khaled, and Luk Arbuckle. *Anonymizing Health Data: Case Studies and Methods to Get You Started*. O'Reilly Media, Inc., 2013.

[5] Racine, V. 2017. "Big Data & Genetic Privacy: Re-identification of Anonymized Data."

[6] Xu, Lei, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. "Modeling Tabular Data Using Conditional Gan". Advances in Neural Information Processing Systems 32 (2019).

[7] AEPD. "Guía básica de Anonimización de la Autoridad de Protección de Datos de Singapur." (https://www.aepd.es/documento/guia-basica-anonimizacion.pdf, 2022, accedida en diciembre de 2025).

rostros, matrículas o letreros son el emborronamiento, el pixelado o el tapado con un recuadro negro. Desde la irrupción de la IA generativa se añade a estas técnicas la generación sintética de la imagen.

3. Análisis comparativo de herramientas

Existe una amplia gama de herramientas existentes para la anonimización de datos. Las soluciones de código abierto, como ARX, ofrecen gran flexibilidad y control, especialmente útiles en contextos académicos o de investigación. Estas herramientas permiten aplicar múltiples modelos de anonimización y generar informes de riesgo de reidentificación. Por otro lado, las plataformas comerciales, como Nymiz o Gallio PRO, están diseñadas para integrarse con sistemas empresariales complejos y ofrecen capacidades avanzadas como anonimización en tiempo real, seguimiento de cumplimiento legal, auditorías automáticas y control de acceso granular.

Una evaluación comparativa de estas herramientas muestra que, mientras las opciones comerciales son más robustas en escenarios de producción a gran escala, las herramientas de código abierto siguen siendo una alternativa válida y poderosa para aplicaciones de menor escala.

En la Tabla 1 presentamos un benchmarking detallado donde se resumen las características más importantes de algunas de las herramientas existentes especializadas en la anonimización de texto, imagen, audio, vídeo o de carácter multimodal. Este benchmarking se estructura en torno a dos ejes fundamentales: las técnicas empleadas y los objetos a identificar categorizados según la naturaleza del dato. El término "objetos" hace referencia exclusivamente a elementos visuales susceptibles de anonimización (p. ej. caras, matrículas), mientras que "entidades" designa las unidades semánticas de información personal identificable en textos, lo que incluye la transcripción de audios (p. ej. nombres, direcciones). Los datos de la tabla reflejan la información aportada de forma pública por los creadores de cada herramienta.

ARX	https://arx.deidentifier.org
Objetos/Entidades	Sexo, edad, raza, estado civil, educación, país de nacimiento, empleo, salario
Técnicas	Generalización, sustitución
Idiomas	Multilingüe
Dominios	Texto
Brighter AI	https://brighter.ai
Objetos/Entidades	Caras, cuerpos, metadatos, matrículas
Técnicas	Deep Natural, difuminado, enmascaramiento
Dominios	Imagen, vídeo
CaseGuard	https://caseguard.com
Objetos/Entidades	Caras, matrículas, pantallas, cuadernos, cuerpos, vehículos, nombres, direcciones, teléfonos, números de ID, tarjetas de crédito, emails
Técnicas	Edición con IA, difuminación, enmascaramiento, generación de transcripciones
Idiomas	Multilingüe
Dominios	Texto, imagen, audio, vídeo
Delphix	https://www.perforce.com/products/delphix
Objetos/Entidades	Nombres, emails, direcciones, tarjetas de crédito
Técnicas	Enmascaramiento
Idiomas	Inglés
Dominios	Texto
Egonym	https://egonym.com
Objetos/Entidades	Caras, metadatos
Técnicas	IA Generativa, difuminado
Dominios	Imagen

ElevenLabs	https://elevenlabs.io/es
Técnicas	Conversión texto-voz, conversión voz-texto, IA conversacional, doblaje, clonación
Idiomas	Multilingüe
Dominios	Audio
Eraseid	https://piktid.com
Objetos/Entidades	Caras, edades, peinados, expresiones, etnias
Técnicas	IA Generativa
Dominios	Imagen
Facit	https://facit.ai/
Objetos/Entidades	Caras, coches, nombres, emails, números de teléfono
Técnicas	Enmascaramiento, difuminado, seudonimización
Idiomas	Inglés
Dominios	Vídeo, texto
Gallio PRO	https://gallio.pro/es
Objetos/Entidades	Caras, matrículas
Técnicas	Desenfoque, difuminado
Dominios	Imagen, vídeo
Lingvanex	https://lingvanex.com
Objetos/Entidades	Nombres, lugares, organizaciones, direcciones, fechas, edades
Técnicas	Enmascaramiento, seudonimización, generalización, perturbación
Idiomas	109 idiomas
Dominios	Texto
Nijta	https://www.nijta.com/es
Técnicas	Anonimización con entidades PII, síntesis
Idiomas	5 idiomas
Dominios	Audio
Nymiz	https://nymiz.com
Objetos/Entidades	Direcciones, códigos, seguros de verificación, monedas, fechas, IBAN, direcciones IP, localizaciones geográficas, NIF, NIE
Técnicas	Seudonimización, sustitución
Idiomas	Español y 102 idiomas más
Dominios	Texto
Pangeanic	https://pangeanic.com/use-cases/mapa
Objetos/Entidades	Bancos e instituciones financieras, compañías de seguros, editoriales, defensa y aplicación de la ley, gobiernos, administraciones públicas
Técnicas	Sustitución, enmascaramiento, eliminación
Idiomas	25 idiomas
Dominios	Texto
Presidio	https://microsoft.github.io/presidio
Objetos/Entidades	Personas, lugares, tarjetas de crédito, crypto, carné de conducir, pasaporte, número bancario
Técnicas	Enmascaramiento, sustitución, sintetización, subrayado, hash y encrypt
Idiomas	Multilingüe
Dominios	Texto

Stick to Digital	https://stdd.es
Objetos/Entidades	Nombres, organizaciones, teléfonos, direcciones, emails, IBAN, datos médicos, enfermedades.
Técnicas	Ofuscación, sustitución
Idiomas	Multilingüe
Dominios	Texto

Tabla 1: Análisis de herramientas de anonimización.

4. Desafíos de la anonimización de datos multimodales

La variedad de herramientas presentadas en la sección anterior pone de manifiesto los avances obtenidos en los últimos años en la automatización de la anonimización. Estos avances se deben principalmente a mejoras en los modelos de detección de entidades, objetos y personas en cualquier fuente y en la capacidad de la Inteligencia Artificial Generativa para procesar y generar texto, vídeo, voz o imagen. Pese a estos avances, sigue habiendo una serie de dificultades y desafíos para la automatización de la anonimización en documentos multimodales o multimedia, como los que se exponen a continuación.

4.1. Modelos generativos y extracción de información

Los Grandes Modelos de Lenguaje (LLM según sus siglas en inglés) como GPT en sus distintas versiones, Gemini o Llama, han demostrado ser extraordinariamente eficaces para tareas que implican generación de texto, como las respuestas a preguntas, el resumen automático, la conversión de lenguaje natural a código, etc. No obstante, en tareas que implican clasificación o etiquetado de texto, como el reconocimiento de entidades nombradas (NER), estos LLM suelen tener niveles de acierto inferiores a modelos más pequeños basados en aprendizaje supervisado, como aquellos entrenados mediante ajuste fino a partir de modelos de tipo BERT[8], una arquitectura de redes neuronales diseñada específicamente para comprender el contexto bidireccional del lenguaje y que, mediante ajuste fino sobre datos anotados, puede especializarse en tareas concretas con alta precisión.

Uno de los principales problemas de utilizar los LLM para la detección de entidades son las alucinaciones que inevitablemente producen estos modelos[9]. Aunque se restrinja al máximo la salida del modelo mediante instrucciones muy precisas, ocasionalmente devuelve textos que no se ajustan al formato esperable. En muchos casos, la respuesta incluye texto adicional (p. ej., el razonamiento llevado a cabo para obtener la respuesta, la respuesta traducida a otro idioma o codificada en un lenguaje de programación). La salida del LLM requiere ser procesada para eliminar estos errores y para extraer o corregir la posición de las entidades, ya que estos modelos no son especialmente buenos contando palabras o calculado posiciones dentro de un texto[10].

Para mitigar estas limitaciones pueden emplearse estrategias manuales, semiautomáticas o completamente automatizadas. Algunos errores recurrentes, como la inclusión de razonamientos intermedios dentro de la respuesta, pueden corregirse mediante procedimientos automáticos relativamente simples. Sin embargo, las alucinaciones no presentan un patrón estable, de modo que los métodos automáticos más estrictos pueden eliminar también información relevante. Por este motivo, la revisión manual continúa siendo recomendable, aunque en la práctica suele resultar inviable cuando el volumen de datos es demasiado elevado.

A pesar de estas limitaciones, investigaciones recientes muestran que, en determinados dominios, como el médico[11], los modelos de NER basados en LLM pueden tener niveles similares o superiores

[8] Zhu, Yuanjing, and Yunan Liu. "LLM-NER: Advancing Named Entity Recognition with Lora+ Fine-Tuned Large Language Models". In 2025 11th International Conference on Computing and Artificial Intelligence (ICCAI), 364–68. IEEE, 2025, https://10.1109/ICCAI66501.2025.00063.

[9] Huang, Lei, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, et al. "A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions". ACM Transactions on Information Systems 43, no. 2 (2025): 1–55, https://10.1145/3703155.

[10] Fu, Tairan, Raquel Ferrando, Javier Conde, Carlos Arriaga, and Pedro Reviriego. "Why Do Large Language Models (LLMs) Struggle to Count Letters?" arXiv Preprint arXiv:2412. 18626, 2024, https://doi.org/10.48550/arXiv.2412.18626.

[11] Majid, Iyad, Vaibhav Mishra, Rohith Ravindranath, and Sophia Y. Wang. "Evaluating the Performance of Large Language

que los basados en modelos de tipo BERT cuando se utilizan las estrategias de *few-shot*, proporcionando al modelo algunos ejemplos de anotación y cadenas de pensamientos (CoT, por sus siglas en inglés).

No obstante, muchas de estas comparaciones se realizan sobre tareas de NER que incluyen únicamente cuatro categorías (persona, organización, localización y miscelánea), lo que reduce considerablemente las posibilidades de error. Al tratarse de clases conceptualmente bien diferenciadas, la identificación resulta más sencilla y se obtienen tasas de acierto más elevadas. Además, la mayor parte de estos estudios se centra en el inglés como lengua principal, sin extender el análisis a otros idiomas, lo cual evidencia un claro sesgo lingüístico. Este fenómeno, ampliamente documentado en el ámbito del procesamiento del lenguaje natural[12], se debe a la mayor disponibilidad de datos de entrenamiento en inglés frente a otros idiomas. Como consecuencia, los resultados obtenidos en este idioma no siempre se generalizan a lenguas con menos recursos, donde el rendimiento de los modelos puede disminuir de forma significativa.

Mediante un corpus interno[13] elaborado por los autores de esta investigación se evaluaron distintos LLM de código abierto, dos LLM genéricos (*DeepSeek* y *Qwen*), y otros específicamente entrenados para tareas de extracción de información como Gollie[14] o Guidex[15]. La evaluación de este tipo de modelos consistió en comparar una serie de documentos anotados manualmente con esos mismos textos anotados por un sistema automático. Las métricas resultantes de esa comparación permitieron observar la precisión, la exhaustividad, o la puntuación F1, un valor calculado a partir de estos dos últimos.

Como se observa en la Tabla 2, el mejor rendimiento corresponde al modelo supervisado entrenado específicamente en el marco de este proyecto. Este modelo ha sido entrenado con la librería Flair[16] y utiliza como arquitectura subyacente un modelo basado en BERT adaptado al dominio y afinado para reconocer 12 entidades diferentes, lo que contribuye a su mayor precisión.

Modelo	Precisión	Exhaustividad	F1
Basado en BERT	0.87	0.87	0.88
DeepSeek	0.72	0.62	0.63
Guidex	0.73	0.54	0.62
Gollie	0.75	0.50	0.60
Qwen	0.77	0.47	0.54

Tabla 2: Evaluación de modelos.

En cuanto a los LLM evaluados, se seleccionaron DeepSeek y Qwen, dos modelos de referencia de acceso abierto, por dos motivos principales. En primer lugar, ambos ofrecen un amplio soporte multilingüe, que incluye los idiomas relevantes para el proyecto, lo que resulta esencial para garantizar la aplicabilidad de los experimentos. En segundo lugar, hay trabajos que evidencian su buen rendimiento en diversas tareas de PLN, incluyendo extracción de información o reconocimiento de NER[17].

En comparación, estos LLM generales no alcanzan niveles competitivos, ya que el modelo supervisado los supera en casi un 20% en la métrica F1, que se usa habitualmente para evaluar esta tarea.

Models for Named Entity Recognition in Ophthalmology Clinical Free-Text Notes". In AMIA Annual Symposium Proceedings, 2024:778, 2025.

[12] Papadimitriou, I., Lopez, K., & Jurafsky, D. (2022, October 11). Multilingual BERT has an accent: Evaluating English influences on fluency in multilingual models. arXiv.org. https://arxiv.org/abs/2210.05619

[13] Corpus formado por 1731 textos en español con una media de 54 palabras cada uno. El 70 por ciento son textos del ámbito de la salud recopilados de la publicación de libre acceso SINC (https://www.agenciasinc.es) y el resto son menajes cortos de tipo SMS o email creados por anotadores de Sigma Cognition con información personal ficticia. Todos los textos han sido anotados en las categorías que aparecen en la tabla 3.

[14] Sainz, Oscar, Iker García-Ferrero, Rodrigo Agerri, Oier Lopez de Lacalle, German Rigau, and Eneko Agirre. "Gollie: Annotation Guidelines Improve Zero-Shot Information-Extraction". arXiv Preprint arXiv:2310.03668, 2023, https://doi.org/10.48550/arXiv.2310.03668.

[15] De La Fuente, Neil, Oscar Sainz, Iker García-Ferrero, and Eneko Agirre. "GUIDEX: Guided Synthetic Data Generation for Zero-Shot Information Extraction". arXiv Preprint arXiv:2506. 00649, 2025, https://doi.org/10.48550/arXiv.2506.00649.

[16] https://flairnlp.github.io

[17] Zhan, Z., Zhou, S., Zhou, H., Deng, J., Hou, Y., Yeung, J., & Zhang, R. (2025, March 1). *An evaluation of DeepSeek Models in Biomedical Natural Language Processing*. arXiv.org. https://arxiv.org/abs/2503.00624

Además, las salidas generadas por estos debieron ser revisadas de forma semiautomática, dado que con frecuencia no seguían el esquema establecido en el *prompt*, lo cual dificultaba la alineación entre sus predicciones y el conjunto de prueba durante la evaluación.

Los resultados sirven también para observar una fuerte variabilidad según la categoría de entidad: aquellas con un patrón formal estable, como direcciones de correo electrónico, nombres propios o entidades geopolíticas, alcanzan precisiones superiores al 90%, mientras que categorías conceptualmente más abiertas, como rol, afiliación o enfermedad, descienden en ocasiones por debajo del 20%, lo que evidencia que la naturaleza de la categoría condiciona de forma decisiva el rendimiento de los modelos.

En la Tabla 3 se muestran los resultados obtenidos durante la prueba con DeepSeek, en la que se puede percibir una clara diferencia entre las categorías numéricas y aquellas que no siguen un patrón tan establecido.

Etiqueta	Precisión	Exhaustividad	F1
ADDRESS	0.26	0.30	0.28
AGE	0.21	0.44	0.28
DISEASE	0.64	0.56	0.60
EMAIL	0.90	0.94	0.92
GPE	0.64	0.83	0.72
ID_NUM	0.80	0.96	0.87
NAME	0.93	0.87	0.90
NORP	0.86	0.36	0.51
NUM_BANK	0.21	0.86	0.33
ORG	0.40	0.61	0.48
PHONE	0.98	0.95	0.97
ROLE	0.72	0.28	0.40

Tabla 3: Resultados obtenidos con DeepSeek, divididos por etiqueta.

En entidades con formatos muy definidos, como números de teléfono, identificadores nacionales o cuentas bancarias, siguen utilizándose reglas basadas en expresiones regulares, capaces de ofrecer precisiones cercanas al 100% cuando están bien especificadas, aunque con menor exhaustividad que los métodos de aprendizaje automático debido a la gran variabilidad con la que estas entidades pueden aparecer en el texto.

4.2. Escasez de datos anotados

Los modelos monolingües basados en BERT, por su parte, tienen como inconveniente la necesidad de ser entrenados con datos anotados de calidad. Este tipo de datos no siempre son fáciles de conseguir, dado que los documentos de partida contienen información de carácter personal y no suelen formar parte de los repositorios públicos. Para evitar esta dificultad, una solución habitual consiste en recurrir a la augmentación de datos: un conjunto de técnicas que permiten generar de manera artificial nuevos ejemplos de entrenamiento. Entre estas técnicas se encuentran, por ejemplo, la traducción automática, la retro traducción o la sustitución de términos por sinónimos, que ayudan a incrementar la variabilidad lingüística sin necesidad de recopilar nuevos datos sensibles. Estas aproximaciones permiten ampliar el corpus manteniendo, en la medida de lo posible, la estructura narrativa o semántica del texto original.

Entre los desafíos específicos de la anonimización multimodal y multilingüe, destaca la escasa disponibilidad de datos de partida. En numerosos proyectos, el corpus inicial se reduce a un conjunto limitado de documentos que no permiten realizar un entrenamiento robusto ni generalizable. Incluso cuando se trabaja con un único idioma, si este no es el inglés, es frecuente encontrar problemas de calidad, como anotaciones inconsistentes, textos mal traducidos o errores derivados de la falta de recursos lingüísticos adecuados. Esta situación se agrava, como se mencionó antes, especialmente en lenguas con menos recursos, como el euskera o el gallego, que carecen de grandes corpus anotados y cuyos datos disponibles suelen ser reducidos, desequilibrados o no estandarizados.

En escenarios con escasez de datos, la augmentación se ha consolidado como una estrategia esencial para generar ejemplos adicionales de entrenamiento y facilitar el desarrollo de modelos competitivos en distintas lenguas y modalidades. Estas técnicas permiten crear nuevas

instancias a partir de datos reales mediante transformaciones como la sustitución de sinónimos, la modificación aleatoria de la estructura del texto, la generación de paráfrasis con modelos generativos, la traducción y retro traducción o el seudoetiquetado de datos no anotados. Su objetivo principal es ampliar la diversidad del corpus manteniendo, en la medida de lo posible, la coherencia semántica del contenido original.

Diversos trabajos han demostrado que la augmentación puede mejorar de forma notable la robustez de los modelos NER en contextos de bajos recursos. Sin embargo, también destacan limitaciones importantes: la generación automática puede producir textos poco naturales o incoherentes, introducir sinónimos inadecuados para el dominio o generar traducciones excesivamente libres. Para mitigar el ruido sintético, propuestas como el *meta-reweighting mechanism* de Wu et al[18] buscan estimar la fiabilidad de los ejemplos generados y reducir el impacto de los menos fiables. Además, algunos sistemas de augmentación, como *TransWebEdu*[19], dependen del inglés como lengua intermedia, lo que introduce sesgos tipológicos que afectan especialmente a lenguas morfológicamente ricas, como el euskera, comprometiendo la naturalidad del texto y la consistencia de las etiquetas.

Estos problemas resultan especialmente críticos en tareas de anonimización, donde la preservación del contexto semántico y la consistencia del etiquetado son fundamentales. Transformaciones como la paráfrasis o la traducción pueden desplazar, alterar o eliminar entidades personales, generando patrones erróneos que el modelo podría aprender. Aun así, la augmentación sigue siendo una herramienta valiosa en entornos multilingües y multimodales con pocos datos, ya que permite compensar la falta de corpus anotados, reducir la dependencia de datos sensibles y mejorar la capacidad de generalización, siempre que se aplique con mecanismos de control y supervisión que limiten la aparición de ruido, sesgos y pérdidas de información relevante.

4.3. Entidades en contextos multilingües y multimodales

Como hemos visto en las secciones anteriores, los modelos monolingües basados en entrenamiento supervisado o ajuste fino de modelos tipo BERT siguen siendo hoy en día una opción interesante para la detección de entidades en general y de entidades con información de identificación personal (PII) en particular, ya que son modelos mucho más ligeros, y a menudo, más fiables que los LLM. No obstante, para ser utilizados en contextos multilingües necesitan un sistema previo de detección de idioma.

Modelos de detección de idioma como *LangId*[20] o *FastText*[21] tienen niveles de acierto por encima del 90 por ciento en distintos corpus de referencia utilizando apenas 50 caracteres. Ahora bien, en los documentos multimodales no siempre se encuentran contextos tan extensos, especialmente en el texto reconocido en la pantalla mediante técnicas de reconocimiento óptico de caracteres (OCR). El texto reconocido puede estar fragmentado en distintos objetos o incluso parcialmente ocluido. Las imágenes y los vídeos también pueden presentar textos manuscritos susceptibles de contener entidades con información personal. En este caso, aunque actualmente los modelos basados en redes convolucionales consiguen niveles de acierto por encima del 90 por ciento, circunstancias como la mala iluminación, el solapamiento de caracteres o errores de escritura pueden provocar errores de reconocimiento e impedir la detección y por tanto el ocultamiento de información personal.

Es importante destacar que la gran mayoría de modelos necesitan de un detector de texto inicial que de alguna manera seleccione el área donde se encuentra el texto para que los modelos de reconocimiento puedan funcionar correctamente. Esta limitación viene dada por el tipo de datasets donde se entrenan los modelos de reconocimiento, donde solamente se encuentran palabras con claridad en vez de en situaciones complicadas reales. Además, este tipo de modelos, presentan limitaciones a la hora de reconocer caracteres manuscritos o digitales de alfa-

[18] Wu, Linzhi, Pengjun Xie, Jie Zhou, Meishan Zhang, Ma Chunping, Guangwei Xu, and Min Zhang. "Robust Self-Augmentation for Named Entity Recognition with Meta Reweighting". In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 4049–60, 2022, https://10.18653/v1/2022.naacl-main.297.

[19] Wang, Josiah, Josiel Figueiredo, and Lucia Specia. "MultiSubs: A Large-Scale Multimodal and Multilingual Dataset". In Proceedings of the Thirteenth Language Resources and Evaluation Conference, 6776–85, 2022. https://arxiv.org/abs/2103.01910.

[20] Lui, Marco, and Timothy Baldwin. "Langid. Py: An off-the-Shelf Language Identification Tool". In Proceedings of the ACL 2012 System Demonstrations, 25–30, 2012.

[21] Joulin, Armand, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. "Fasttext. Zip: Compressing Text Classification Models". arXiv Preprint arXiv:1612. 03651, 2016, https://doi.org/10.48550/arXiv.1612.03651.

betos no latinos debido a la escasez de datos de entrenamiento representando esos alfabetos.

Recientemente, se le ha ido dando más importancia a las capacidades OCR de los LLM y VLM (modelos de lenguaje visual) hasta el punto de tener varios *benchmarks* de conversión de imágenes a formato JSON, como se muestra en el *OmniAI OCR Benchmark*[22], o tareas de reconocimiento de texto, extracción de información clave en documentos o reconocimiento de expresiones matemáticas manuscritas como se puede ver en el *OpenVLM Leaderboard*[23]. En todos ellos, destaca la combinación de *Qwen2.5*[24] como modelo de lenguaje con *InternViT2.5*[25] como modelo de visión, consiguiendo posiciones muy notables en todas las categorías.

4.4. Más allá de las entidades: autoría y estilo

La anonimización de texto está mayoritariamente focalizada en la detección y ocultación de entidades que contienen información personal. Ahora bien, varios estudios muestran que estas medidas son insuficientes para anonimizar un texto, ya que persisten rasgos lingüísticos capaces de desvelar al autor[26]. Métodos como *character trigrams + KNN*[27] o *Writeprints + SVM*[28] ponen de manifiesto que la ortografía, la sintaxis o la semántica actúan como identificadores de identidad difíciles de eliminar.

Las técnicas de *author obfuscation* buscan alterar de forma controlada el estilo de escritura para dificultar la atribución automática de autoría sin comprometer la legibilidad del texto. Para ello, aplican transformaciones en los niveles ortográfico, sintáctico y semántico, modificando aspectos como la puntuación, las mayúsculas, ciertos patrones léxicos o combinaciones gramaticales que pueden funcionar como huellas distintivas de cada autor.

El análisis de estos niveles muestra que los modelos de atribución se apoyan en múltiples rasgos, desde elementos superficiales hasta patrones más profundos. Algunas particularidades, como la gestión de la puntuación o la presencia recurrente de determinados errores ortográficos, resultan especialmente reveladoras y actúan como marcas estilométricas muy estables, capaces de distinguir a unos autores de otros. De manera similar, ciertas preferencias léxicas o combinaciones gramaticales frecuentes también contribuyen a delimitar perfiles de escritura. Al manipular selectivamente estos elementos, es posible reducir de manera significativa la capacidad de los modelos para identificar regularidades que permitan atribuir un texto a un escritor concreto.

En conjunto, estas observaciones apuntan a que la alteración controlada de rasgos estilométricos constituye una estrategia eficaz para disminuir la precisión de los modelos de atribución y reforzar la protección de la identidad textual. La mejora y futura integración de estas técnicas en sistemas de anonimización comerciales o de código abierto podría ofrecer a los usuarios un nivel adicional de privacidad en contextos donde la escritura funciona como un dato identificable.

4.5. Datos personales en el habla

Las entidades no solo pueden aparecer en el texto o en la imagen; también los locutores de los documentos audiovisuales producen información personal. El reconocimiento de entidades

[22] getomni-ai. 2025, "GitHub - Getomni-Ai/Benchmark: OCR Benchmark." GitHub, consultado el 13-1-2026. https://github.com/getomni-ai/benchmark.

[23] Huggingface.co. 2025, "Open VLM Leaderboard - a Hugging Face Space by Opencompass." 2024. Huggingface, consultado el 13-1-2026, https://huggingface.co/spaces/opencompass/open_vlm_leaderboard.

[24] Qwen: An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, et al. "Qwen2.5 Technical Report". arXiv [Cs.CL], arXiv. http://arxiv.org/abs/2412.15115, 2025, https://doi.org/10.48550/arXiv.2412.15115.

[25] Chen, Zhe, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, et al. "Expanding Performance Boundaries of Open-Source Multimodal Models with Model, Data, and Test-Time Scaling". arXiv Preprint arXiv:2412. 05271, https://arxiv.org/abs/2412.05271, 2024, https://doi.org/10.48550/arXiv.2412.05271.

[26] Pissarra, David, Isabel Curioso, João Alveira, Duarte Pereira, Bruno Ribeiro, Tomás Souper, Vasco Gomes, André Carreiro, and Vitor Rolla. "Unlocking the Potential of Large Language Models for Clinical Text Anonymization: A Comparative Study". In Proceedings of the Fifth Workshop on Privacy in Natural Language Processing, 74–84, 2024. https://doi.org/10.18653/v1/2024.privatenlp-1.8.

[27] Juola, Patrick. "Detecting Stylistic Deception". In Proceedings of the Workshop on Computational Approaches to Deception Detection, 91–96, 2012. https://www.aclweb.org/anthology/W12-0414.pdf.

[28] Afroz, Sadia, Michael Brennan, and Rachel Greenstadt. "Detecting Hoaxes, Frauds, and Deception in Writing Style Online". In 2012 IEEE Symposium on Security and Privacy, 461–75. IEEE, 2012.https://doi.org/10.1109/sp.2012.34.

PII en el habla depende no solo de la calidad de los modelos de detección de entidades, sino también de los modelos de transcripción automática del habla (ASR), ya que el texto sobre el que se aplican los modelos de detección de entidades se obtiene previamente a partir de las transcripciones generadas mediante estos modelos. En consecuencia, los errores en la transcripción tienen un impacto relevante en la tarea de reconocimiento de entidades[29]. Por ejemplo, si se transcriben incorrectamente palabras claves como nombres propios o lugares, los modelos de detección de entidades pueden no reconocerlas o clasificarlas erróneamente. Dentro del marco de la anonimización de información sensible, esto tiene un impacto importante, ya que una entidad no detectada puede dar lugar a la pérdida de privacidad.

Otras de las limitaciones que pueden presentar las transcripciones generadas por modelos ASR es que carecen de señales lingüísticas importantes, tales como puntuación o uso de mayúsculas, lo que dificulta aún más la tarea de los modelos de detección de entidades. Estos problemas se acentúan en escenarios multilingües, donde los errores de transcripción son más frecuentes.

En resumen, la calidad de los modelos ASR desempeña un papel fundamental en el rendimiento global del sistema. Emplear modelos robustos y adaptados al dominio resulta clave para mejorar la fiabilidad de la detección de entidades y así garantizar una anonimización efectiva.

4.6. La voz como dato personal

No solo el contenido hablado en un audio puede contener información personal: la voz en sí se considera un dato personal según la GDPR (arts. 4.1, 4.14 y 9.1). La señal de voz contiene diferentes fuentes de variabilidad. Una de ellas es la variabilidad asociada a la identidad del hablante, que está asociada a patrones en la señal de voz, en concreto a rasgos paralingüísticos: timbre, tono, volumen y ritmo de la voz. El hallazgo de esta asociación ha permitido desarrollar sistemas de caracterización de hablante en la literatura, en concreto modelos que permiten predecir atributos del hablante a partir de estos rasgos paralingüísticos extraídos de la señal de la voz, tales como edad-sexo[30], patologías de la voz[31], acento, y emoción.

Ante el avance en el desarrollo de tecnologías de perfilado e identificación del hablante, surge la preocupación acerca de la preservación de la privacidad del hablante ante el posible mal uso de la información sensible que puede ser extraída a partir de muestras de voz. Bajo la regulación de protección de datos europea, la supresión de toda información que pueda ser extraída a partir de la voz es crítica. Es en este punto donde cobra protagonismo la anonimización de la voz, cuyo objetivo es enmascarar la identidad del hablante en la señal de voz de manera que la señal anonimizada suene totalmente diferente a la señal original, pero manteniendo a la vez la inteligibilidad de la señal de habla.

En la literatura podemos encontrar dos tipos de soluciones de la anonimización: física y lógica[32]. La anonimización física consiste en perturbar la señal de voz ya sea mediante la adición de ruido o modificación de parámetros acústicos (velocidad, amplitud, volumen)[33]. En el caso de la anonimización lógica, el objetivo se centra en modificar la identidad del hablante en la señal original. Entre las soluciones encontramos: sistemas de transformación de voz[34], centradas en la modificación de aspectos del habla tales como emoción, acento, o identidad; sistemas de conversión

[29] Szymański, Piotr, Lukasz Augustyniak, Mikolaj Morzy, Adrian Szymczak, Krzysztof Surdyk, and Piotr Żelasko. "Why Aren't We NER yet? Artifacts of ASR Errors in Named Entity Recognition in Spontaneous Speech Transcripts". In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 1746–61, 2023, https://10.18653/v1/2023.acl-long.98.

[30] Lastow, Fredrik, Edwin Ekberg, and Pierre Nugues. "Language-Agnostic Age and Gender Classification of Voice Using Self-Supervised Pre-Training". In 2022 Swedish Artificial Intelligence Society Workshop (SAIS), 1–9. IEEE, 2022. https://doi.org/10.1109/SAIS55783.2022.9833071.

[31] Moro-Velazquez, Laureano, Jorge A. Gomez-Garcia, Julian D. Arias-Londoño, Najim Dehak, and Juan I. Godino-Llorente. "Advances in Parkinson's Disease Detection and Assessment Using Voice and Speech: A Review of the Articulatory and Phonatory Aspects". Biomedical Signal Processing and Control 66 (2021): 102418. https://doi.org/10.1016/J.BSPC.2021.102418.

[32] Fang, Fuming, Xin Wang, Junichi Yamagishi, Isao Echizen, Massimiliano Todisco, Nicholas Evans, and Jean-Francois Bonastre. "Speaker Anonymization Using X-Vector and Neural Waveform Models". arXiv Preprint arXiv:1905. 13561, 2019, https://doi.org/10.21437/ssw.2019-28.

[33] Jin, Q., Toth, A. R., Schultz, T., & Black, A. W. (2009). Speaker de-identification via voice transformation. *Proceedings of the 2009 IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU 2009*, 529–533. https://doi.org/10.1109/ASRU.2009.5373356

[34] Jin, Q., Toth, A. R., Schultz, T., & Black, A. W. (2009). Speaker de-identification via voice transformation. *Proceedings of the 2009 IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU 2009*, 529–533. https://doi.org/10.1109/ASRU.2009.5373356

de voz[35], el cual consiste en modificar el timbre de voz del locutor original por el de otro locutor distinto; y sistemas basados en supresión de características de locutor[36], cuya estrategia consiste en aplicar técnicas basadas en aprendizaje adversario para suprimir atributos relacionadas a la identidad del hablante tales como sexo, edad o acento de las representaciones acústicas, pero manteniendo la utilidad de la voz anonimizada, además de la inteligibilidad y naturalidad del contenido hablado.

El continuo avance en modelos generativos ha permitido dar un salto en calidad en la anonimización de los audios y, por tanto, una mayor inteligibilidad del contenido hablado, además de alcanzar una mayor precisión en la ofuscación de la identidad del locutor. No obstante, todas las soluciones desarrolladas, además de *benchmarks* y marcos de evaluación[37], en el contexto de la anonimización de la voz solo contemplan el escenario donde un solo hablante o locutor interviene en el audio, resultando un modelo de anonimización menos generalista en escenarios con múltiples hablantes (escenario que es más próximo a un caso de uso real de la anonimización de la voz). Ante esto, han empezado a surgir iniciativas que involucran no solo desarrollos o mejoras de los sistemas de anonimización, sino marcos de evaluación[38] que permiten medir el alcance de estos sistemas en escenarios multihablantes empleando métricas objetivas estándar y corpus de evaluación.

4.7. Anonimización y autorreconocimiento

A pesar de los avances descritos en el apartado anterior, las técnicas de anonimización de voz no son siempre lo suficientemente robustas para ocultar la identidad de los hablantes. En la práctica, suelen persistir riesgos residuales de reidentificación y, sobre todo, de vinculabilidad entre grabaciones del mismo hablante. Por ello, además de la evaluación frente a sistemas automáticos, es relevante considerar un escenario complementario que es el autorreconocimiento, es decir, la capacidad del propio hablante para identificarse en una señal transformada.

En muchos escenarios, la anonimización de voz se implementa mediante modificaciones directas de la señal, por ejemplo, transformaciones acústicas como la modificación de la frecuencia fundamental o el desplazamiento de formantes. Estas intervenciones se utilizan mayormente por su simplicidad y porque suelen preservar en buena medida la inteligibilidad. Sin embargo, su robustez es limitada, ya que la identidad vocal no depende de un único rasgo, sino de un conjunto de indicios distribuidos en distintos niveles de la señal, por lo que ese tipo de intervenciones parciales tienden a dejar información residual suficiente para reconocer al hablante. Incluso enfoques más avanzados basados en el clonado de voz permiten un cierto nivel de autorreconocimiento.

En este contexto, literatura reciente[39] resalta que no es suficiente la alteración de un componente aislado, puesto que la prosodia puede actuar como una fuente robusta de información sobre el hablante. En el marco de un proyecto realizado por los autores de este artículo, pendiente de publicación, se llevó a cabo un experimento de percepción para comprobar hasta qué punto los hablantes son capaces de reconocer su propia voz sometida a distintas técnicas de anonimización. Los resultados sugieren que las transformaciones más simples no impiden el autorreconocimiento, mientras que enfoques más complejos basados en resíntesis condicionada a un pseudolocutor lo dificultan de manera más notable, aunque sin eliminarlo por completo, dado que persisten en la señal modificada rasgos prosódicos, melódicos o rítmicos muy particulares de los hablantes que les permiten reconocerse a sí mismos.

[35] Alegre, F., Soldi, G., & Evans, N. (2014). Evasion and obfuscation in automatic speaker verification. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 749–753, https://doi.org/10.1109/ICASSP.2014.6853696

[36] Perero-Codosero, J. M., Espinoza-Cuadros, F. M., & Hernández-Gómez, L. A. (2022). X-vector anonymization using autoencoders and adversarial training for preserving speech privacy. *Computer Speech & Language*, *74*, 101351. https://doi.org/10.1016/J.CSL.2022.101351

[37] Panariello, Michele, Natalia Tomashenko, Xin Wang, Xiaoxiao Miao, Pierre Champion, Hubert Nourtel, Massimiliano Todisco, Nicholas Evans, Emmanuel Vincent, and Junichi Yamagishi. "The Voiceprivacy 2022 Challenge: Progress and Perspectives in Voice Anonymisation". *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024, https://10.1109/TASLP.2024.3430530

[38] Miao, Xiaoxiao, Ruijie Tao, Chang Zeng, and Xin Wang. "A Benchmark for Multi-Speaker Anonymization". *IEEE Transactions on Information Forensics and Security*, 2025, https://10.1109/TIFS.2025.3556345

[39] Champion, P. (2023). Anonymizing speech: Evaluating and designing speaker anonymization techniques. *arXiv preprint arXiv:2308.04455*.

4.8. Datos personales en contenido visual

La identificación y anonimización de información personal en la imagen es también un requisito necesario para el cumplimiento de las normativas europea y americana. Esta tarea se lleva a cabo en dos fases: una primera fase de detección basada en visión por computador y una segunda de procesamiento en la que ocurre la transformación de la zona de interés para ocultar información personal.

En la fase de detección se localizan en la imagen de forma precisa rostros, personas, objetos como matrículas de vehículos, documentos de identidad, direcciones de domicilios y otra información que permita identificar a las personas. Entre las herramientas utilizadas para llevar a cabo esta tarea destacan los siguientes detectores generalistas en tiempo real[40], entrenados ambos con el corpus *COCO*[41]:

- *YOLO* (*You Only Look Once*)[42], una familia de modelos de detección de objetos de código abierto, con una implementación práctica muy orientada al usuario.
- *RT-DETR* (*Real Time Detection Transformer*)[43] desarrollada por investigadores de la Universidad de *Beijing*, que utiliza una arquitectura basada en transformadores.

En la fase de procesamiento, se realiza la transformación de la zona donde se encuentra el elemento de interés de manera que los datos puedan seguir siendo utilizados para cualquier fin, como por ejemplo el entrenamiento y evaluación de modelos, sin comprometer el rendimiento ni la identidad de los individuos.

Los algoritmos de procesamiento de imagen más utilizados para esta tarea[44] cuando la prioridad es la privacidad son el difuminado, el pixelado o la aplicación de una máscara negra. En cambio, cuando se busca un equilibrio con la utilidad posterior de los datos, se aplican técnicas de IA generativa como la sustitución para crear datos sintéticos realistas en las zonas anonimizadas utilizando, por ejemplo, *variational autoencoders* (VAE)[45], *generative adversarial networks* (GAN)[46], *diffusion models*[47], *transformers*[48], o *vision language models* (*VLM*)[49].

Las técnicas clásicas agilizan los tiempos de inferencia debido al bajo coste computacional de aplicación a cambio de distorsionar en mayor medida el contenido original de la imagen, mientras que las técnicas generativas tienen más coste computacional y mayor tiempo de inferencia a cambio de preservar mejor el contenido original de la imagen.

[40] Lee, Jun Ha, and Su Jeong You. "Balancing Privacy and Accuracy: Exploring the Impact of Data Anonymization on Deep Learning Models in Computer Vision". *IEEE Access* 12 (2024): 8346–58.https://doi.org/10.1109/access.2024.3352146.

[41] Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P. & Zitnick, C. L. (2014, September). Microsoft coco: Common objects in context. In *European conference on computer vision* (pp. 740-755). Cham: Springer International Publishing.

[42] Redmon, Joseph, Santosh Divvala, Ross Girshick, et al. 2015. "You Only Look Once: Unified, Real-Time Object Detection." *arXiv* 794: arXiv:1506.02640. https://doi.org/10.48550/ARXIV.1506.02640.

[43] Lv, Wenyu, Yian Zhao, Qinyao Chang, et al. 2024. "RT-DETRv2: Improved Baseline with Bag-of-Freebies for Real-Time Detection Transformer." *arXiv*, July 24, arXiv:2407.17140. https://doi.org/10.48550/ARXIV.2407.17140.

[44] Birnstill, Pascal, Daoyuan Ren, and Jürgen Beyerer. "A User Study on Anonymization Techniques for Smart Video Surveillance". In *2015 12th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 1–6. IEEE, 2015. https://doi.org/10.1109/AVSS.2015.7301805.

[45] Kingma, Diederik P., Max Welling, Diederik P. Kingma, et al. 2013. "Auto-Encoding Variational Bayes." *arXiv* 12 (1): arXiv:1312.6114. https://doi.org/10.48550/ARXIV.1312.6114.

[46] Goodfellow, Ian J., Jean Pouget-Abadie, Mehdi Mirza, et al. 2014. "Generative Adversarial Networks." arXiv, no. 2017: arXiv:1406.2661. https://doi.org/10.48550/ARXIV.1406.2661.

[47] Ho, Jonathan, Ajay Jain, Pieter Abbeel, et al. 2020. "Denoising Diffusion Probabilistic Models." *arXiv*, December 16, arXiv:2006.11239. https://doi.org/10.48550/ARXIV.2006.11239.

[48] Dosovitskiy, Alexey, Lucas Beyer, Alexander Kolesnikov, et al. 2020. "An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale." arXiv, arXiv:2010.11929.

[49] Bordes, Florian, Richard Yuanzhe Pang, Anurag Ajay, et al. 2024. *An Introduction to Vision-Language Modeling*. May 27. https://arxiv.org/pdf/2405.17247.

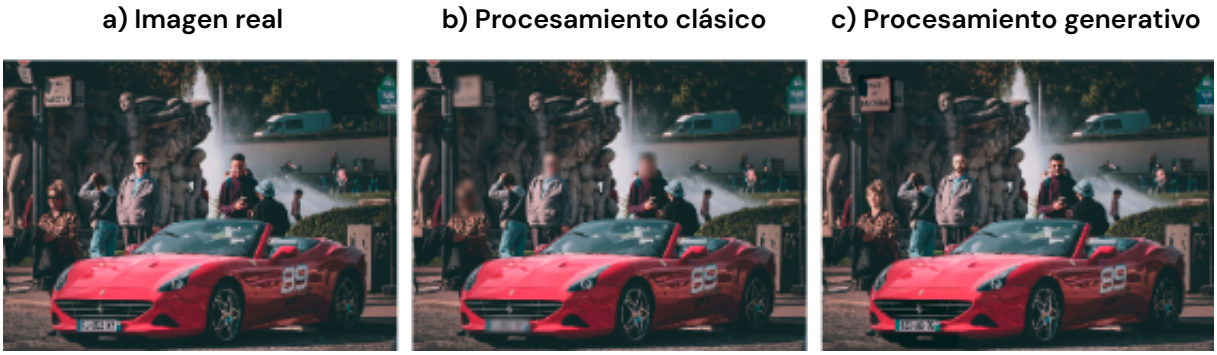


Figura 1. Ejemplo representativo de las dos tipologías de anonimización sobre una imagen (a). Por un lado, el procesamiento clásico donde se distorsiona la imagen ocultando el contenido de interés (b), que en este caso serían las matrículas, las caras y el texto de los letreros. Por otro lado, con las técnicas generativas (c) vemos cómo aparte de la ocultación hay una sustitución del contenido de interés de manera que la imagen original no pierde el contexto visual. Imagen original extraída de *Pexels* bajo licencia de libre uso[50].

La búsqueda del equilibrio entre la utilidad y la privacidad ha motivado diversas investigaciones centradas en la anonimización visual automática y escalable, analizando el impacto para el entrenamiento de modelos de visión artificial[51]. En este sentido son interesantes estudios como *Fântomas*[52], donde se evalúa la reversibilidad de las técnicas de anonimización de rostros, o estudios donde se evalúa la usabilidad de los datos después de la anonimización para llegar a una solución de compromiso entre preservar la identidad protegida y conservar la capacidad de detectar acciones o comportamientos de las personas[53].

En esta línea, investigaciones como *NullFace*[54] y *Reverse Personalization*[55] exploran la anonimización por sustitución condicionada y localizada, donde solo se cambian los atributos de los rostros clave para no poder reconocerlos, pero conservando el contexto del entorno y de los atributos que no son de carácter personal.

En resumen, la detección y el procesamiento de los datos personales en contenido visual mediante pixelado, difuminado o sustitución es crucial para garantizar la protección de la privacidad. De esta forma se protege la información confidencial o personal mediante técnicas que permitan enmascarar exclusivamente la información sensible, de forma irreversible o reversible, manteniendo aquella que ayude o sea de interés para la tarea a realizar posteriormente con los datos. Las diversas investigaciones han demostrado que el método de anonimización más adecuado varía según la tarea de visión por computador, la arquitectura del modelo y las clases a detectar. Esa variación responde principalmente a la dificultad con la que se pueda detectar el elemento de interés en el vídeo y al uso que se le quiera dar a posteriori, dependiendo de si se quiere simplemente ocultar la información sin importar el contenido original o si además se quiere reconstruir el contexto visual de la imagen para darle algún otro uso una vez que el archivo cumple los requisitos de anonimización.

4.9 La IA generativa como estrategia de seudonimización

Los LLM se han consolidado como herramientas altamente eficaces para el procesamiento y la generación automática de texto, destacando por su capacidad de comprensión del lenguaje natural y razonamiento automático. Estas propiedades los hacen adecuados para la anonimización textual, donde estudios recientes indican que los LLM pueden ofrecer niveles de protección superiores a

[50] *Pexels*. Consultado el 13-01-2026, https://www.pexels.com/photo/luxurious-red-vehicle-parked-on-cobb-14119764/.

[51] Hukkelås, Håkon, and Frank Lindseth. 'Does Image Anonymization Impact Computer Vision Training?' In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 140–50, 2023. https://arxiv.org/abs/2306.05135.

[52] Todt, Julian, Simon Hanisch, and Thorsten Strufe. 'Fantômas: Understanding Face AnonymizationReversibility'. *arXiv Preprint arXiv:2210. 10651*, 2022. https://doi.org/10.48550/arxiv.2210.10651.

[53] Aslam, Nazia, and Kamal Nasrollahi. 'Balancing Privacy and Action Performance: A Penalty-Driven Approach to Image Anonymization'. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 729–38, 2025. https://doi.org/10.48550/arxiv.2504.14301.

[54] Kung, Han-Wei, Tuomas Varanka, Terence Sim, and Nicu Sebe. "NullFace: Training-Free Localized Face Anonymization". *arXiv Preprint arXiv:2503. 08478*, 2025.https://doi.org/10.48550/arxiv.2503.08478.

[55] Kung, Han-Wei, Tuomas Varanka, and Nicu Sebe. "Reverse Personalization". *arXiv Preprint arXiv:2512. 22984*, 2025. https://doi.org/10.48550/arxiv.2512.22984.

métodos clásicos y manuales[56][57]

La arquitectura típica de un sistema de seudonimización incluye dos fases: detección de datos sensibles y sustitución por seudónimos. La sustitución puede realizarse mediante diccionarios con selección aleatoria dentro de unas categorías o etiquetas genéricas, lo cual mantiene el control sobre los resultados, pero puede afectar a la naturalidad del texto y complicar análisis posteriores de unidades léxicas complejas[58]. Otros sistemas combinan ambas fases en un solo paso mediante modelos *Seq2Seq* preentrenados para la tarea, aunque su rendimiento suele ser inferior a los métodos basados en *NER*[59].

La IA generativa permite reemplazar entidades PII por seudónimos plausibles manteniendo coherencia contextual[60] mediante herramientas como *Pseudollm*[61]. En este trabajo, Hou et al. proponen en primer lugar generar candidatos de sustitución que mantengan coherencia semántica, sean plausibles, y no permitan reconstruir los datos originales. Los sustitutos pueden seleccionarse mediante muestreo aleatorio, más simple pero menos coherente, o mediante un LLM, más contextual pero más costoso computacionalmente.

La fase de reemplazo puede realizarse de tres formas: reemplazo directo, que solo sustituye palabras pero puede generar incoherencias; reemplazo guiado por *prompts*, donde el LLM aplica los sustitutos predefinidos y devuelve resultados más naturales, aunque depende del modelo y del *prompt*; y reemplazo integrado en la generación controlada, en el que el LLM genera candidatos y los inserta directamente en el texto en un paso, logrando mayor coherencia contextual pero con mayor complejidad de implementación.

A pesar de sus ventajas, los LLM propietarios presentan limitaciones, como el riesgo de privacidad al enviar datos sensibles a terceros. En un estudio reciente llevado a cabo por los autores de este artículo utilizando *Qwen3-Next* (un LLM de código abierto) se ha observado que la seudonimización es viable en textos multilingües, pero persisten problemas con ciertos tipos de datos (especialmente numéricos), confusiones semánticas y un rendimiento desigual en lenguas con menos recursos. Los LLM ofrecen un enfoque prometedor para la seudonimización automática, combinando protección de datos con coherencia y naturalidad del texto, aunque aún requieren mejoras en precisión y seguridad para su aplicación generalizada.

En el ámbito de la anonimización de imagen estática y video, este tipo de tecnologías generativas permiten eliminar la información de carácter personal, como los rostros, las matrículas de los coches, direcciones de las señales o incluso texto sobreimpreso en la imagen, conservando a su vez la coherencia espacial y contextual de esos mismos elementos. Trabajos como, por ejemplo, *DynamicFace*[62] o *DiffFace*[63] destacan por su capacidad para ocultar la identidad personal preservando rasgos faciales, consiguiendo así una mayor coherencia temporal y espacial en el ámbito de vídeo. Los modelos generativos con capacidades de "*inpainting*" como el modelo *FLUX* 1 desarrollado por *Black Forest*[64] y *Stable Diffusion*[65], permiten un enfoque agnóstico al elemento a sustituir mediante

[56] Patsakis, Constantinos, and Nikolaos Lykousas. "Man vs the Machine in the Struggle for Effective Text Anonymisation in the Age of Large Language Models". *Scientific Reports* 13, no. 1 (2023): 16026, https://doi.org/10.1038/s41598-023-42977-3.

[57] Wang, Tao, Yushu Zhang, Shuren Qi, Ruoyu Zhao, Zhihua Xia, and Jian Weng. "Security and Privacy on Generative Data in Aigc: A Survey". *ACM Computing Surveys* 57, no. 4 (2024): 1–34, https://doi.org/10.1145/3703626.

[58] Lison, Pierre, Ildikó Pilán, David Sanchez, Montserrat Batet, and Lilja Øvrelid. "Anonymisation Models for Text Data: State of the Art, Challenges and Future Directions". In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 4188–4203, 2021, https://10.18653/v1/2021.acl-long.323.

[59] Yermilov, Oleksandr, Vipul Raheja, and Artem Chernodub. "Privacy-and Utility-Preserving NLP with Anonymized Data: A Case Study of Pseudonymization". *arXiv Preprint arXiv:2306. 05561*, 2023, https://doi.org/10.48550/arXiv.2306.05561.

[60] Hou, Shilong, Ruilin Shang, Zi Long, Xianghua Fu, and Yin Chen. "A General PseudonymizationFramework for Cloud-Based Llms: Replacing Privacy Information in Controlled Text Generation". *arXiv Preprint arXiv:2502. 15233*, 2025, https://doi.org/10.48550/arXiv.2502.15233.

[61] Bienati, Arianna. "Pseudollm: A Multilingual LLM-Enhanced Tool for Context-Aware Text Pseudonymization". In *Workshop" Corpus Linguistics 2040: Which Data, Which Methods, Which Models?"*, 2025.

[62] Wang, Runqi, Yang Chen, Sijie Xu, et al. 2025. "DynamicFace: High-Quality and Consistent Face Swapping for Image and Video Using Composable 3D Facial Priors." *arXiv*, July 7, arXiv:2501.08553. https://doi.org/10.48550/ARXIV.2501.08553.

[63] Kim, Kihong, Yunho Kim, Seokju Cho, et al. 2022. "DiffFace: Diffusion-Based Face Swapping with Facial Guidance." *arXiv*, December 27, arXiv:2212.13344. https://doi.org/10.48550/ARXIV.2212.1334

[64] Forest Labs, Black, Stephen Batifol, Andreas Blattmann, et al. 2025. "FLUX.1 Kontext: Flow Matching for In-Context Image Generation and Editing in Latent Space." *arXiv*, June 24, arXiv:2506.15742. https://doi.org/10.48550/ARXIV.2506.15742

[65] Podell, Dustin, Zion English, Kyle Lacey, et al. 2023. "SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis." *arXiv*, July 4, arXiv:2307.01952. https://doi.org/10.48550/ARXIV.2307.01952

prompts y máscaras. Estos modelos ajustan dinámicamente el prompt según detecciones o interacciones externas, facilitando tareas como la anonimización de texto mediante la combinación de modelos de reconocimiento y seudonimización para generar prompts específicos que fuercen la sustitución.

Los métodos clásicos de anonimización de voz se basan en métodos que aplican alteraciones en la señal de la voz, por ejemplo, distorsión de la señal mediante modificación del pitch o cambio en la velocidad del habla, o adición de ruido. Sin embargo, estos métodos presentan dos inconvenientes: mayor degradación de la inteligibilidad del audio anonimizado, y poca robustez frente a la reidentificación debido a que el proceso puede ser revertido. En contraste, los métodos basados en IA generativa, en concreto los modelos generativos adversarios *GAN*[66] y *DANN*[67], han permitido modificar la voz de una persona a otra diferente preservando el contenido lingüístico, pero suprimiendo rasgos biométricos identificables. Uno de los enfoques más usados gracias a su efectividad en suprimir rasgos biométricos en la voz son los métodos basados en redes neuronales profundas que permiten mapear características acústicas de una voz original hacia un espacio latente y luego reconstruir una nueva señal de voz con atributos distintos a la original[68] (por ejemplo, género, edad y acento). De esta forma, se obtienen voces sintéticas que mantienen la naturalidad y la inteligibilidad, pero que no están asociadas a la identidad del individuo original. Asimismo, los modelos de *transformers*[69] han mejorado significativamente la síntesis de voz, aumentando la robustez de los sistemas de anonimización tanto en privacidad como en inteligibilidad, reduciendo el riesgo de reidentificación e incrementando la naturalidad de la voz.

4.10 La sincronización de voz e imagen en videos anonimizados

Uno de los desafíos técnicos más complejos en la anonimización de contenido audiovisual es mantener la sincronización labial (*lip-sync)* cuando se sustituye el audio original por voz sintética o anonimizada. Este problema surge especialmente en contextos donde se requiere preservar tanto la privacidad del hablante como la naturalidad y credibilidad del contenido visual. La sincronización de labios requiere que el audio generado mantenga no solo el contenido semántico del mensaje original, sino también su ritmo, pausas, prosodia y cadencia. Cuando estos elementos no están alineados con los movimientos faciales del vídeo, el resultado puede resultar incongruente a nivel visual, lo que reduce la calidad del material anonimizado.

En el estado del arte se encuentran soluciones comerciales de *lip-sync*, tales como *Sieve*[70], y *RaskAI*[71], así como soluciones de código como *Musetalk*[72] y *LatentSync*[73]. En el caso de las soluciones comerciales, estas presentan limitaciones en cuanto a falta de modularidad al ser soluciones cerradas que no permiten ajustar independientemente cada componente del *pipeline*, y dependencia de interfaces de programación de aplicaciones propietarias, lo que restringe el control sobre aspectos como la calidad, privacidad y costes. Por el contrario, las soluciones de código abierto ofrecen una mayor flexibilidad y capacidad de adaptación.

El mayor desafío técnico en este flujo de trabajo es el alineamiento temporal: el audio generado debe respetar las pausas, el ritmo y la cadencia del original para que los movimientos labiales del vídeo coincidan con la nueva pista de audio. Esto requiere no solo un sintetizador de voz (TTS) de alta calidad, sino también técnicas de post-procesamiento del audio para ajustar la velocidad, insertar silencios y controlar la entonación. En pruebas realizadas con vídeos cortos en cinco idiomas (español, inglés, francés, italiano, portugués), se ha demostrado que es posible lograr sincronización labial aceptable utilizando flujos de trabajo modulares de código abierto, tanto en escenarios de anonimización de voz como en la sustitución de información personal. No obstante, persisten

[66] Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, et al. 2020. "Generative Adversarial Networks." *Communications of the ACM* 63 (11): 139–44. https://doi.org/10.1145/3422622

[67] Ganin, Yaroslav, Evgeniya Ustinova, Hana Ajakan, et al. 2015. "Domain-Adversarial Training of Neural Networks." *Advances in Computer Vision and Pattern Recognition* 17 (9783319583464): 189–209. https://doi.org/10.1007/978-3-319-58347-1_10

[68] Aloufi, Ranya, Hamed Haddadi, and David Boyle. 2020. "Privacy-Preserving Voice Analysis via Disentangled Representations." *CCSW 2020 - Proceedings of the 2020 ACM SIGSAC Conference on Cloud Computing Security Workshop*, September 4, 1–14. https://doi.org/10.1145/3411495.34213

[69] Vaswani, Ashish, Google Brain, Noam Shazeer, et al. 2017. "Attention Is All You Need." *Advances in Neural Information Processing Systems*

[70] "Sieve: Video datasets for frontier AI", accedido enero de 2026, https://www.sieve.a

[71] "Rask AI: Traduce vídeos con voces realistas de IA", accedido enero de 2026, https://www.rask.ai/.

[72] "MuseTalk", accedido enero de 2026, https://github.com/TMElyralab/MuseTalk.

[73] "LatentSync", accedido enero de 2026, https://github.com/bytedance/LatentSync.

desafíos en la naturalidad prosódica y en la estabilidad del alineamiento en secuencias largas o con múltiples hablantes, lo que hace necesario avanzar en el control prosódico fino en TTS, el alineamiento automático de pausas, la evaluación perceptual con usuarios y el soporte multilingüe robusto para idiomas con menos recursos.

5. Conclusiones

En este artículo hemos repasado al estado del arte de la anonimización automática de datos no estructurados, como texto libre, audio, imagen o vídeo, poniendo el foco en los avances y desafíos que encontramos tanto en la bibliografía como en el marco de un proyecto de investigación que llevan a cabo los autores. Como conclusiones generales podemos destacar las siguientes.

La anonimización de datos multimodales cada vez va a ser más necesaria. En un mundo cada vez más dependiente de la tecnología, la compartición de datos en entornos seguros, como los espacios de datos, va a requerir de sistemas que automaticen la anonimización de datos estructurados y no estructurados. La economía basada en el dato que promueven varios gobiernos europeos parte de que las instituciones y empresas compartan sus datos. La anonimización es una manera, no la única, de garantizar que la información se comparte de forma segura respetando la privacidad de las personas.

El análisis comparativo realizado muestra además que existe una amplia variedad de herramientas especializadas en anonimización de texto, imagen, audio y vídeo. Las soluciones de código abierto destacan por su flexibilidad y adaptabilidad en entornos de investigación, mientras que las plataformas comerciales ofrecen mayores capacidades de integración, automatización y escalabilidad en escenarios empresariales. Esta diversidad confirma que no existe una solución universal y que la elección de las herramientas depende del dominio, el tipo de dato y el contexto de uso.

Los métodos basados en IA generativa conviven con enfoques más tradicionales en la anonimización de datos. Las técnicas clásicas, apoyadas en aprendizaje automático o reglas, siguen siendo relevantes por su eficiencia y porque en dominios específicos pueden ofrecer mejores resultados con menor coste computacional. Por su parte, los modelos generativos han ampliado notablemente las capacidades de anonimización, especialmente en escenarios complejos y multimodales. Sin embargo, no sustituyen por completo a los métodos tradicionales, sino que suelen integrarse en soluciones híbridas donde cada técnica se aplica según el tipo de dato, su sensibilidad y el uso previsto. Esta combinación se está consolidando como la opción más eficaz para equilibrar privacidad, utilidad de los datos y coste de procesamiento.

Los datos anotados de calidad siguen siendo escasos. Especialmente en lenguas consideradas de bajos recursos, es difícil encontrar repositorios con información privada que pueda ser utilizada para el entrenamiento o evaluación de modelos de detección de entidades, personas u objetos. Como consecuencia de ello los modelos en estos idiomas tienen menor rendimiento, lo cual expone a sus hablantes a una mayor vulnerabilidad. En este escenario, las técnicas de augmentación de datos se han convertido en una herramienta clave para ampliar corpus de entrenamiento, aunque todavía presentan limitaciones relacionadas con la introducción de ruido, sesgos o inconsistencias semánticas.

La dificultad de la anonimización depende del tipo de entidad y su contexto. Los resultados analizados evidencian, además, que la dificultad de la anonimización no depende únicamente del tipo de modelo utilizado, sino también de la naturaleza de las entidades y del contexto en el que aparecen. Mientras que entidades con estructuras formales bien definidas (como correos electrónicos, números de identificación o direcciones) pueden ser detectadas con alta precisión mediante reglas o patrones, otras categorías más ambiguas o contextuales (como roles, enfermedades o relaciones semánticas complejas) presentan un rendimiento significativamente inferior. Esta heterogeneidad confirma la necesidad de enfoques adaptativos y sensibles al dominio, capaces de ajustar dinámicamente las estrategias de detección y ocultación.

Más allá de las entidades explícitas, la anonimización debe considerar también otros elementos identificativos menos evidentes. En texto, diferentes estudios muestran que el estilo de escritura puede actuar como una huella de autoría, incluso cuando las entidades personales han sido eliminadas. Las técnicas de ofuscación estilométrica constituyen, por tanto, así como una línea de investigación prometedora para reforzar la privacidad textual mediante la modificación controlada de rasgos lingüísticos.

En los datos audiovisuales, la anonimización presenta retos adicionales por la interacción entre modalidades. En imagen y vídeo todo depende de detectar correctamente los elementos sensi-

bles, y existe un equilibrio entre técnicas rápidas pero destructivas (como el difuminado) y métodos generativos más costosos pero que preservan mejor el contexto. En audio, la calidad de la transcripción condiciona la detección de entidades, y además la voz sigue siendo un dato identificativo difícil de anonimizar por completo. Por último, en vídeo, mantener la coherencia entre audio e imagen, especialmente la sincronización labial, añade complejidad al proceso.

Como líneas futuras, será importante mejorar los modelos multimodales para anonimizar texto, audio e imagen de forma conjunta, crear mejores métricas para evaluar el riesgo de reidentificación en situaciones reales e incorporar mecanismos de privacidad en tiempo real. Además, será necesario avanzar en recursos multilingües de calidad e integrar estas soluciones de forma segura y eficiente en entornos reales.

Por último, conviene recordar que, si bien esta tecnología tiene como objetivo proteger derechos de las personas, concretamente su derecho a la privacidad, también puede ser usada de forma espuria para fines ilícitos que requieran ocultar la identidad propia, suplantar la identidad de otra persona o generar una imagen falsa de la misma. Estos casos de uso no están recogidos en este estudio; no obstante, resulta imprescindible tenerlos en cuenta a la hora de desarrollar tecnología desde una perspectiva ética y responsable.

Bibliografía

Afroz, Sadia, Michael Brennan, and Rachel Greenstadt. "Detecting Hoaxes, Frauds, and Deception in Writing Style Online". In 2012 IEEE Symposium on Security and Privacy, 461–75. IEEE, 2012. https://doi.org/10.1109/sp.2012.34.

Alegre, F., Soldi, G., & Evans, N. (2014). Evasion and obfuscation in automatic speaker verification. ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, 749–753, https://doi.org/10.1109/ICASSP.2014.6853696

Aloufi, Ranya, Hamed Haddadi, and David Boyle. 2020. "Privacy-Preserving Voice Analysis via Disentangled Representations." CCSW 2020 - Proceedings of the 2020 ACM SIGSAC Conference on Cloud Computing Security Workshop, September 4, 1–14. https://doi.org/10.1145/3411495.34213

Aslam, Nazia, and Kamal Nasrollahi. 'Balancing Privacy and Action Performance: A Penalty-Driven Approach to Image Anonymization'. In Proceedings of the Computer Vision and Pattern Recognition Conference, 729–38, 2025. https://doi.org/10.48550/arxiv.2504.14301.

Bienati, Arianna. "Pse(udollm: A Multilingual LLM-Enhanced Tool for Context-Aware Text Pseudonymization". In Workshop" Corpus Linguistics 2040: Which Data, Which Methods, Which Models?", 2025.

Birnstill, Pascal, Daoyuan Ren, and Jürgen Beyerer. "A User Study on Anonymization Techniques for Smart Video Surveillance". In 2015 12th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), 1–6. IEEE, 2015. https://doi.org/10.1109/AVSS.2015.7301805.

Bordes, Florian, Richard Yuanzhe Pang, Anurag Ajay, et al. 2024. An Introduction to Vision-Language Modeling. May 27. https://arxiv.org/pdf/2405.17247.

Champion, P. (2023). Anonymizing speech: Evaluating and designing speaker anonymization techniques. arXiv preprint arXiv:2308.04455.

Chen, Zhe, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, et al. "Expanding Performance Boundaries of Open-Source Multimodal Models with Model, Data, and Test-Time Scaling". arXiv Preprint arXiv:2412. 05271, https://arxiv.org/abs/2412.05271, 2024, https://doi.org/10.48550/arXiv.2412.05271.

De La Fuente, Neil, Oscar Sainz, Iker García-Ferrero, and Eneko Agirre. "GUIDEX: Guided Synthetic Data Generation for Zero-Shot Information Extraction". arXiv Preprint arXiv:2506. 00649, 2025, https://doi.org/10.48550/arXiv.2506.00649.

Dosovitskiy, Alexey, Lucas Beyer, Alexander Kolesnikov, et al. 2020. "An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale." arXiv, arXiv:2010.11929.

El Emam, Khaled, and Luk Arbuckle. Anonymizing Health Data: Case Studies and Methods to Get You Started. O'Reilly Media, Inc., 2013.

Fang, Fuming, Xin Wang, Junichi Yamagishi, Isao Echizen, Massimiliano Todisco, Nicholas Evans, and Jean-Francois Bonastre. "Speaker Anonymization Using X-Vector and Neural Waveform Models". arXiv Preprint arXiv:1905. 13561, 2019, https://doi.org/10.21437/ssw.2019-28.

Forest Labs, Black, Stephen Batifol, Andreas Blattmann, et al. 2025. "FLUX.1 Kontext: Flow Matching for In-Context Image Generation and Editing in Latent Space." arXiv, June 24, arXiv:2506.15742. https://doi.org/10.48550/ARXIV.2506.15742

Fu, Tairan, Raquel Ferrando, Javier Conde, Carlos Arriaga, and Pedro Reviriego. "Why Do Large Language Models (LLMs) Struggle to Count Letters?" arXiv Preprint arXiv:2412. 18626, 2024, https://doi.org/10.48550/arXiv.2412.18626.

Ganin, Yaroslav, Evgeniya Ustinova, Hana Ajakan, et al. 2015. "Domain-Adversarial Training of Neural Networks." Advances in Computer Vision and Pattern Recognition 17 (9783319583464): 189–209. https://doi.org/10.1007/978-3-319-58347-1_10

getomni-ai. 2025, "GitHub - Getomni-Ai/Benchmark: OCR Benchmark." GitHub, consultado el 13-1-2026. https://github.com/getomni-ai/benchmark.

Goodfellow, Ian J., Jean Pouget-Abadie, Mehdi Mirza, et al. 2014. "Generative Adversarial Networks." arXiv, no. 2017: arXiv:1406.2661. https://doi.org/10.48550/ARXIV.1406.2661.

Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, et al. 2020. "Generative Adversarial Networks." Communications of the ACM 63 (11): 139–44. https://doi.org/10.1145/3422622

Ho, Jonathan, Ajay Jain, Pieter Abbeel, et al. 2020. "Denoising Diffusion Probabilistic Models." arXiv, December 16, arXiv:2006.11239. https://doi.org/10.48550/ARXIV.2006.11239.

Hou, Shilong, Ruilin Shang, Zi Long, Xianghua Fu, and Yin Chen. "A General PseudonymizationFramework for Cloud-Based Llms: Replacing Privacy Information in Controlled Text Generation". arXiv Preprint arXiv:2502. 15233, 2025, https://doi.org/10.48550/arXiv.2502.15233.

Huang, Lei, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, et al. "A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions". ACM Transactions on Information Systems 43, no. 2 (2025): 1–55, https://10.1145/3703155.

Huggingface.co. 2025, "Open VLM Leaderboard - a Hugging Face Space by Opencompass." 2024. Huggingface, consultado el 13-1-2026, https://huggingface.co/spaces/opencompass/open_vlm_leaderboard.

Hukkelås, Håkon, and Frank Lindseth. 'Does Image Anonymization Impact Computer Vision Training?' In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 140–50, 2023. https://arxiv.org/abs/2306.05135.

Jin, Q., Toth, A. R., Schultz, T., & Black, A. W. (2009). Speaker de-identification via voice transformation. Proceedings of the 2009 IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU 2009, 529–533. https://doi.org/10.1109/ASRU.2009.5373356

Jin, Q., Toth, A. R., Schultz, T., & Black, A. W. (2009). Speaker de-identification via voice transformation. Proceedings of the 2009 IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU 2009, 529–533. https://doi.org/10.1109/ASRU.2009.5373356

Joulin, Armand, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. "Fasttext. Zip: Compressing Text Classification Models". arXiv Preprint arXiv:1612. 03651, 2016, https://doi.org/10.48550/arXiv.1612.03651.

Juola, Patrick. "Detecting Stylistic Deception". In Proceedings of the Workshop on Computational Approaches to Deception Detection, 91–96, 2012. https://www.aclweb.org/anthology/W12-0414.pdf.

Kim, Kihong, Yunho Kim, Seokju Cho, et al. 2022. "DiffFace: Diffusion-Based Face Swapping with Facial Guidance." arXiv, December 27, arXiv:2212.13344. https://doi.org/10.48550/ARXIV.2212.1334

Kingma, Diederik P., Max Welling, Diederik P. Kingma, et al. 2013. "Auto-Encoding Variational Bayes." arXiv 12 (1): arXiv:1312.6114. https://doi.org/10.48550/ARXIV.1312.6114.

Kung, Han-Wei, Tuomas Varanka, and Nicu Sebe. "Reverse Personalization". arXiv Preprint arXiv:2512. 22984, 2025. https://doi.org/10.48550/arxiv.2512.22984.

Kung, Han-Wei, Tuomas Varanka, Terence Sim, and Nicu Sebe. "NullFace: Training-Free Localized Face Anonymization". arXiv Preprint arXiv:2503. 08478, 2025.https://doi.org/10.48550/arxiv.2503.08478.

Lastow, Fredrik, Edwin Ekberg, and Pierre Nugues. "Language-Agnostic Age and Gender Classification of Voice Using Self-Supervised Pre-Training". In 2022 Swedish Artificial Intelligence Society Workshop (SAIS), 1–9. IEEE, 2022. https://doi.org/10.1109/SAIS55783.2022.9833071.

Lee, Jun Ha, and Su Jeong You. "Balancing Privacy and Accuracy: Exploring the Impact of Data Anonymization on Deep Learning Models in Computer Vision". IEEE Access 12 (2024): 8346–58.https://doi.org/10.1109/access.2024.3352146.

Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P. & Zitnick, C. L. (2014, September). Microsoft coco: Common objects in context. In European conference on computer vision (pp. 740-755). Cham: Springer International Publishing.

Lison, Pierre, Ildikó Pilán, David Sanchez, Montserrat Batet, and Lilja Øvrelid. "AnonymisationModels for Text Data: State of the Art, Challenges and Future Directions". In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 4188–4203, 2021, https://10.18653/v1/2021.acl-long.323.

Lui, Marco, and Timothy Baldwin. "Langid. Py: An off-the-Shelf Language Identification Tool". In Proceedings of the ACL 2012 System Demonstrations, 25–30, 2012.

Lv, Wenyu, Yian Zhao, Qinyao Chang, et al. 2024. "RT-DETRv2: Improved Baseline with Bag-of-Freebies for Real-Time Detection Transformer." arXiv, July 24, arXiv:2407.17140. https://doi.org/10.48550/ARXIV.2407.17140.

Majid, Iyad, Vaibhav Mishra, Rohith Ravindranath, and Sophia Y. Wang. "Evaluating the Performance of Large Language Models for Named Entity Recognition in Ophthalmology Clinical Free-Text Notes". In AMIA Annual Symposium Proceedings, 2024:778, 2025.

Miao, Xiaoxiao, Ruijie Tao, Chang Zeng, and Xin Wang. "A Benchmark for Multi-Speaker Anonymization". IEEE Transactions on Information Forensics and Security, 2025, https://10.1109/TIFS.2025.3556345

Moro-Velazquez, Laureano, Jorge A. Gomez-Garcia, Julian D. Arias-Londoño, Najim Dehak, and Juan I. Godino-Llorente. "Advances in Parkinson's Disease Detection and Assessment Using Voice and Speech: A Review of the Articulatory and Phonatory Aspects". Biomedical Signal Processing and Control 66 (2021): 102418. https://doi.org/10.1016/J.BSPC.2021.102418.

Panariello, Michele, Natalia Tomashenko, Xin Wang, Xiaoxiao Miao, Pierre Champion, Hubert Nourtel, Massimiliano Todisco, Nicholas Evans, Emmanuel Vincent, and Junichi Yamagishi. "The Voiceprivacy 2022 Challenge: Progress and Perspectives in Voice Anonymisation". IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2024, https://10.1109/TASLP.2024.3430530

Papadimitriou, I., Lopez, K., & Jurafsky, D. (2022, October 11). Multilingual BERT has an accent: Evaluating English influences on fluency in multilingual models. arXiv.org. https://arxiv.org/abs/2210.05619

Patsakis, Constantinos, and Nikolaos Lykousas. "Man vs the Machine in the Struggle for Effective Text Anonymisation in the Age of Large Language Models". Scientific Reports 13, no. 1 (2023): 16026, https://doi.org/10.1038/s41598-023-42977-3.

Perero-Codosero, J. M., Espinoza-Cuadros, F. M., & Hernández-Gómez, L. A. (2022). X-vector anonymization using autoencoders and adversarial training for preserving speech privacy. Computer Speech & Language, 74, 101351. https://doi.org/10.1016/J.CSL.2022.101351

Pexels. Consultado el 13-01-2026, https://www.pexels.com/photo/luxurious-red-vehicle-parked-on-cobb-14119764/.

Pissarra, David, Isabel Curioso, João Alveira, Duarte Pereira, Bruno Ribeiro, Tomás Souper, Vasco Gomes, André Carreiro, and Vitor Rolla. "Unlocking the Potential of Large Language Models for Clinical Text Anonymization: A Comparative Study". In Proceedings of the Fifth Workshop on Privacy in Natural Language Processing, 74–84, 2024. https://doi.org/10.18653/v1/2024.privatenlp-1.8.

Podell, Dustin, Zion English, Kyle Lacey, et al. 2023. "SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis." arXiv, July 4, arXiv:2307.01952. https://doi.org/10.48550/ARXIV.2307.01952

Qwen: An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, et al. "Qwen2.5 Technical Report". arXiv [Cs.CL], arXiv. http://arxiv.org/abs/2412.15115, 2025, https://doi.org/10.48550/arXiv.2412.15115.

Racine, V. 2017. "Big Data & Genetic Privacy: Re-identification of Anonymized Data."

Redmon, Joseph, Santosh Divvala, Ross Girshick, et al. 2015. "You Only Look Once: Unified, Real-Time Object Detection." arXiv 794: arXiv:1506.02640. https://doi.org/10.48550/ARXIV.1506.02640.

Sainz, Oscar, Iker García-Ferrero, Rodrigo Agerri, Oier Lopez de Lacalle, German Rigau, and Eneko Agirre. "Gollie: Annotation Guidelines Improve Zero-Shot Information-Extraction". arXiv Preprint arXiv:2310.03668, 2023, https://doi.org/10.48550/arXiv.2310.03668.

Solove, Daniel J. Understanding Privacy. Harvard university press, 2010.

Szymański, Piotr, Lukasz Augustyniak, Mikolaj Morzy, Adrian Szymczak, Krzysztof Surdyk, and Piotr Żelasko. "Why Aren't We NER yet? Artifacts of ASR Errors in Named Entity Recognition in Spontaneous Speech Transcripts". In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 1746–61, 2023, https://10.18653/v1/2023.acl-long.98.

Todt, Julian, Simon Hanisch, and Thorsten Strufe. 'Fantômas: Understanding Face Anonymization Reversibility'. arXiv Preprint arXiv:2210.10651, 2022. https://doi.org/10.48550/arxiv.2210.10651.

Vaswani, Ashish, Google Brain, Noam Shazeer, et al. 2017. "Attention Is All You Need." Advances in Neural Information Processing Systems

Wang, Josiah, Josiel Figueiredo, and Lucia Specia. "MultiSubs: A Large-Scale Multimodal and Multilingual Dataset". In Proceedings of the Thirteenth Language Resources and Evaluation Conference, 6776–85, 2022. https://arxiv.org/abs/2103.01910.

Wang, Runqi, Yang Chen, Sijie Xu, et al. 2025. "DynamicFace: High-Quality and Consistent Face Swapping for Image and Video Using Composable 3D Facial Priors." arXiv, July 7, arXiv:2501.08553. https://doi.org/10.48550/ARXIV.2501.08553.

Wang, Tao, Yushu Zhang, Shuren Qi, Ruoyu Zhao, Zhihua Xia, and Jian Weng. "Security and Privacy on Generative Data in Aigc: A Survey". ACM Computing Surveys 57, no. 4 (2024): 1–34, https://doi.org/10.1145/3703626.

Wu, Linzhi, Pengjun Xie, Jie Zhou, Meishan Zhang, Ma Chunping, Guangwei Xu, and Min Zhang. "Robust Self-Augmentation for Named Entity Recognition with Meta Reweighting". In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 4049–60, 2022, https://10.18653/v1/2022.naacl-main.297.

Xu, Lei, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. "Modeling Tabular Data Using Conditional Gan". Advances in Neural Information Processing Systems 32 (2019).

Yermilov, Oleksandr, Vipul Raheja, and Artem Chernodub. "Privacy-and Utility-Preserving NLP with Anonymized Data: A Case Study of Pseudonymization". arXiv Preprint arXiv:2306. 05561, 2023, https://doi.org/10.48550/arXiv.2306.05561.

Zhan, Z., Zhou, S., Zhou, H., Deng, J., Hou, Y., Yeung, J., & Zhang, R. (2025, March 1). An evaluation of DeepSeek Models in Biomedical Natural Language Processing. arXiv.org. https://arxiv.org/abs/2503.00624

Zhu, Yuanjing, and Yunan Liu. "LLM-NER: Advancing Named Entity Recognition with Lora+ Fine-Tuned Large Language Models". In 2025 11th International Conference on Computing and Artificial Intelligence (ICCAI), 364–68. IEEE, 2025, https://10.1109/ICCAI66501.2025.00063.

Zuboff, Shoshana. "The Age of Surveillance Capitalism". In Social Theory Re-Wired, 203–13. Routledge, 2023, https://doi.org/10.4324/9781003320609.