Mejora de la anonimización de texto mediante explicabilidad orientada al riesgo de re-identificación

Benet Manzanares-Salor
Universitat Rovira i Virgili, CYBERCAT
Tarragona, España
benet.manzanares@urv.cat

David Sánchez
Universitat Rovira i Virgili, CYBERCAT
Tarragona, España
david.sanchez@urv.cat

***Resumen*—Este es un resumen extendido del artículo publicado en [1]. La anonimización de texto es una tarea compleja realizada habitualmente por expertos humanos de forma manual, cosa que conlleva costes significativos. Los mecanismos automáticos propuestos en la literatura se basan principalmente en *reconocimiento de entidades nombradas* (NER), una técnica reconocida como insuficiente en términos de protección de la privacidad. Para abordar esta limitación se propone PETRE, un método diseñado para mejorar la protección de la privacidad que proporciona garantías empíricas contra la re-identificación. Este emplea técnicas de explicabilidad para detectar y enmascarar iterativamente los términos que causan el mayor riesgo de re-identificación hasta que se alcanza el nivel de k-anonimato deseado. Los experimentos demuestran que PETRE es capaz de mejorar de manera significativa y consistente la protección de la privacidad de las anonimizaciones basadas en NER, a la vez que muestran las limitaciones de la anonimización manual.**

***Index Terms*—Protección de la privacidad, Modelos de lenguaje, Explicabilidad, Anonimización de texto, Re-identificación**

I. Introducción

La sociedad digital actual requiere grandes volúmenes de información, siendo el texto una de las fuentes más comunes y ricas en información. No obstante, los datos textuales a menudo contienen información sensible sobre individuos (como historiales médicos o casos judiciales), creando un conflicto directo con regulaciones de privacidad como el *reglamento general de protección de datos (GDPR)* [2] en Europa. Para mitigar estos riesgos, los gestores de datos recurren a la anonimización, consistente en el enmascaramiento irreversible de los datos para que no puedan atribuirse a los sujetos a los que se refieren.

Mientras que la anonimización de datos estructurados (tabulares) es un campo maduro con modelos formales como k-anonimato [3] y la privacidad diferencial [4], la anonimización de datos no estructurados como el texto es una tarea mucho más compleja. La ambigüedad inherente y la dependencia del contexto del lenguaje exigen un amplio conocimiento del dominio para identificar la información que facilita la re-identificación. Debido a estas dificultades, la anonimización de texto se realiza a la práctica principalmente de forma manual [5], [6].

Para automatizar o asistir este proceso, se han propuesto mecanismos basados en el *reconocimiento de entidades nombradas (NER)* [7], [8], [9]. Estos métodos utilizan modelos de aprendizaje automático para detectar tipos predefinidos de entidades (por ejemplo, nombres o lugares), que se asumen como facilitadores de la re-identificación. Sin embargo, este enfoque presenta serias deficiencias, ya que proporciona una protección insuficiente, carece de una garantía de privacidad y no incorpora un mecanismo para configurar el grado de protección deseado. Por estas razones, los métodos basados en NER pueden servir como un punto de partida para la anonimización manual, no son una solución completa.

I-A. Contribuciones

Considerando las limitaciones del NER, se propone *PETRE (privacy enhancement for text via risk-oriented explainability)*, un método fundamentado en la privacidad que busca mejorar las anonimizaciones basadas en NER imponiendo una reducción del riesgo de re-identificación definida por el usuario. Sus principales contribuciones incluyen:

- **Generalidad:** El método es aplicable a cualquier método de anonimización de partida, independientemente de su principio de funcionamiento.
- **Privacidad configurable:** Permite configurar de forma intuitiva el nivel de privacidad deseado, mediante la especificación de los sujetos a proteger y del riesgo de re-identificación máximo permitido, expresado en términos de k-*anonimato probabilístico*.
- **Personalización por sujeto:** A diferencia del k-anonimato estándar, el conocimiento de fondo para evaluar los riesgos puede ser específico para cada sujeto, reduciendo el enmascaramiento innecesario.
- **Anonimización iterativa:** El riesgo de re-identificación se revalúa después de cada enmascaramiento hasta alcanzar el umbral de privacidad, minimizando el sobre-enmascaramiento.
- **Garantía empírica:** Al ser guiado por el riesgo observado, el resultado anonimizado presenta un riesgo empírico de re-identificación $\leq 1/k$, eliminando la necesidad de evaluaciones manuales *a posteriori*.
- **Automatización:** El flujo de trabajo es totalmente automático, sin requerir intervención humana.

Los experimentos demuestran que PETRE reduce el riesgo de re-identificación de cualquier anonimización de partida y compite con métodos de vanguardia sin sus costes asociados.

II. Fundamentos

II-A. Anonimización

La anonimización consiste en transformar los datos de forma irreversible para impedir la identificación de los sujetos. Una vez adecuadamente anonimizados, los datos dejan de ser personales y quedan fuera del alcance de marcos legales como la GDPR [2].

Este proceso introduce distorsión y reduce la utilidad de los datos, implicando un balanceo entre privacidad y usabilidad [10], [9].

En este contexto, los atributos de los datos se clasifican en: *identificadores* (p. ej., nombre o DNI), que permiten la identificación directa; *cuasi-identificadores* (p. ej., fecha + lugar de nacimiento + empleo), que posibilitan la re-identificación cuando se combinan con conocimiento de fondo, es decir, información disponible públicamente sobre los individuos; *atributos confidenciales* (p. ej., diagnósticos o salarios), que contienen información sensible; y *atributos no confidenciales*, que no implican riesgo ni revelan datos sensibles.

La anonimización elimina los identificadores y transforma solo los cuasi-identificadores necesarios para garantizar el nivel de privacidad requerido, a la vez que preserva la máxima utilidad que la protección permita.

II-B. k-Anonimato

El k-anonimato [3] es un modelo de privacidad ampliamente utilizado en datos estructurados, que exige que cada individuo sea indistinguible de al menos otros $k-1$ respecto a sus cuasi-identificadores. Esto impone un límite superior en la probabilidad de re-identificación:

$$\text{Probabilidad de re-identificación} \leq \frac{1}{k} \tag{1}$$

El parámetro k permite ajustar el equilibrio entre la protección de la privacidad y la utilidad de los datos. El k-anonimato normalmente se implementa mediante supresión y generalización de atributos hasta formar registros idénticos en conjuntos de tamaño mínimo k.

El k-anonimato probabilístico [11] relaja este requisito manteniendo el mismo límite de riesgo, pero sin necesidad de generar registros indistinguibles. Esta propiedad lo hace más adecuado para texto, donde imponer el requisito indistinguibilidad puede mermar significativamente la utilidad.

II-C. Anonimización de texto basada en NER

Los enfoques más comunes de anonimización automática se basan en NER, y asumen que los (cuasi-)identificadores corresponden a tipos específicos de entidades nombradas (personas, lugares, etc.). Estos métodos detectan un conjunto predefinido de NE y las eliminan o substituyen por su tipo (PERSONA, LUGAR).

Sin embargo, presentan limitaciones importantes:

- **Premisa cuestionable, protección insuficiente:** Limitar los (cuasi-)identificadores a entidades nombradas no ofrece una protección robusta, ya que cualquier término relacionado con un individuo y disponible públicamente puede permitir re-identificación. Como consecuencia, muchos términos quedan sin enmascarar y persiste un riesgo elevado.

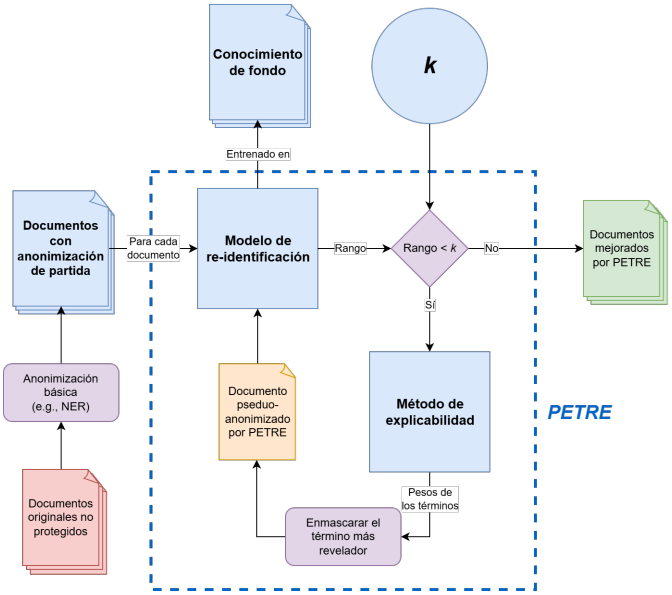


Figura 1. Arquitectura de PETRE

- **Falta de garantía:** No proporcionan una garantía de privacidad, por lo que el nivel de protección solo puede evaluarse *ex post*.
- **Incapacidad de configuración:** No incluyen parámetros que permitan controlar el nivel de privacidad y, por tanto, ajustar la relación entre protección y utilidad deseada.

III. Metodología

PETRE parte de texto anonimizado deficientemente (a partir de NER, por ejemplo) y lo protege de forma automática e iterativa hasta alcanzar un nivel de privacidad predefinido por el usuario en términos de k-anonimato probabilístico.

El usuario debe especificar los documentos a procesar, el valor deseado de k y el conocimiento de fondo. PETRE evalúa iterativamente el riesgo de re-identificación, y utiliza explicabilidad para detectar y enmascarar progresivamente los términos más reveladores hasta que la probabilidad de re-identificación sea $\leq 1/k$. Este diseño iterativo, detallado en la Figura 1, garantiza que solo se enmascaren los términos necesarios, evitando el sobre-enmascaramiento.

Los componentes centrales de PETRE son:

- **Documentos anonimizados de partida:** El conjunto de documentos de entrada, resultado de una anonimización básica (por ejemplo, NER), etiquetados con la identidad del individuo a proteger.
- **Conocimiento de fondo:** Una colección de documentos públicos identificados utilizados para entrenar el modelo de re-identificación. El solapamiento entre este conocimiento y los documentos de partida determina el riesgo; un conocimiento de fondo más escaso reduce el enmascaramiento necesario para un determinado k.
- **Modelo de re-identificación:** Un clasificador entrenado con el conocimiento de fondo que predice la identidad del individuo. Evalúa el nivel de privacidad de un documento generando una distribución de probabilidad sobre las identidades posibles. El riesgo se evalúa a partir del rango (posición) de la identidad correcta en esa distribución.

- k **(umbral de privacidad):** Un entero mayor que uno que define el nivel de riesgo máximo permitido $(1/k)$ y establece la condición de parada del proceso iterativo. El proceso se detiene cuando el rango predicho es $Rango \geq k$. Un valor de $k \geq 2$ garantiza que ningún individuo pueda ser re-identificado de forma única, aunque se suelen desear valores superiores que impliquen menor riesgo (por ejemplo, $k > 10$).
- **Método de explicabilidad:** Un módulo que estima la influencia (pesos) de los términos del documento en la probabilidad predicha para la identidad correcta. Pesos positivos indican que un término contribuye a la re-identificación, y serán éstos los que se prioricen para enmascaramiento en cada iteración.

III-A. Formalización

El conjunto de documentos de partida se define como $A = \{(D_1, I_1), \ldots, (D_n, I_n)\}$, donde cada documento D_i está asociado a una identidad I_i. El modelo de re-identificación $\Gamma(d)$ devuelve una distribución de probabilidad $P = (p_1, \ldots, p_n)$, y la re-identificación se mide mediante la posición de p_i en dicha distribución, denotada como $Rango(d, i, \Gamma)$. El límite del riesgo se deriva de la siguiente manera:

$$Rango(d, i, \Gamma) = k \rightarrow \text{Riesgo de re-identificación}(d) \leq 1/k \quad (2)$$

Esto corresponde al k-anonimato probabilístico, donde el riesgo queda acotado en $1/k$ sin requerir indistinguibilidad entre documentos.

Si el $Rango$ está por debajo de la k fijada, se requiere enmascaramiento. La función de explicabilidad $\xi(d, i, \Gamma)$ genera los pesos $(w_1, \ldots, w_m)$ asociados a los términos $(t_1, \ldots, t_m)$ de d, donde los pesos positivos señalan influencia directa a la re-identificación. Entonces, el término obteniendo el mayor peso se enmascara.

El documento resultante se revalúa, repitiendo iterativamente el proceso hasta alcanzar un $Rango \geq k$.

IV. Experimentos

A continuación se detallan los experimentos realizados para la validación y análisis de PETRE como método de mejora de anonimización de texto. El código y datos se encuentran disponibles en el siguiente repositorio de GitHub: https://github.com/BenetManzanaresSalor/PETRE.

IV-A. Conjunto de datos

Los experimentos se realizaron sobre un conjunto de 553 resúmenes de biografías de la Wikipedia seleccionadas aleatoriamente [12]. Estos resúmenes se anonimizaron manualmente, cosa que utilizaremos como referencia. Hasta dónde sabemos, este es el mayor corpus de propósito general anonimizado por humanos disponible públicamente.

IV-B. Implementación

El conocimiento de fondo y el valor de k de PETRE se ajustaron para igualar el nivel de protección de la anonimización humana. Los propios resúmenes de cada artículo actúan como conocimiento de fondo, replicando la información disponible para los expertos humanos y representando un *peor escenario* habitual en la evaluación de la privacidad. Para determinar k, se utilizó el modelo de re-identificación propuesto en [13] con dicho conocimiento de fondo. El 81 % de los documentos anonimizados manualmente obtuvieron un $rango = 1$, lo que sugiere que los expertos buscaban un nivel de anonimización no mayor a $k = 2$ (el mínimo que evita la re-identificación). Por ello, se adopta $k = 2$ en los experimentos posteriores. El artículo original incluye análisis adicionales sobre cómo varían los resultados con distintos valores de conocimiento de fondo y k. Finalmente, como método de explicabilidad se utilizó SHAP [14].

IV-C. Anonimización de partida

Evaluamos el efecto de PETRE sobre un conjunto de cuatro métodos basados en NER reportados en la literatura [15], [16], [13], [17], [18]. Tres de ellos se basan en el paquete de anotación Standford NER [19], diferenciándose en los tipos de entidades detectadas: `St.NER3`, `St.NER4` y `St.NER7`. El cuarto corresponde al proyecto de código abierto de Microsoft denominado `Presidio`[1].

Adicionalmente, evaluamos dos estrategias orientadas a la *publicación de datos con preservación de privacidad (PPDP)*: `Word2Vec` [15] y k`-anonymity greedy` [12]. Dichos enfoques no dependen de conjuntos fijos de entidades predefinidas, y representan el estado del arte en cuanto a protección de texto [10], [13], [17], [18]. No obstante, presentan limitaciones prácticas, como por ejemplo, depender de recursos externos complejos de obtener, como grafos de conocimiento.

IV-D. Resultados

Para evaluar el alineamiento entre las anonimizaciones automáticas y manuales, empleamos las métricas *recall* y *precisión*, habituales en la literatura [16], [15], [10].

Recall representa el porcentaje de términos identificadores marcados por los expertos humanos que también fueron detectados por el método automático, lo cual refleja el nivel de protección alcanzado. Como se aprecia en la figura 2, PETRE mejora el *recall* de forma significativa y consistente en todas las anonimizaciones de partida. Esto es particularmente evidente en aquellas basadas en NER, que parten de un valor más bajo. La mejora permite situarlas en niveles comparables a los métodos orientados a PPDP, siendo PETRE más sencillo de implementar gracias a su menor complejidad y la ausencia de dependencias externas. Asimismo, PETRE logra aumentar (aunque más levemente) el *recall* de estos últimos métodos.

Precisión corresponde al porcentaje de términos detectados automáticamente que coinciden con los identificadores establecidos por los expertos. La figura 3 muestra que PETRE afecta negativamente este valor, reduciéndolo aproximadamente al 50 %. Esto puede justificarse considerando que el 81 % de las anonimizaciones manuales obtuvieron un $rango = 1$, lo que evidencia una protección insuficiente. En consecuencia, PETRE se ha visto obligado a enmascarar numerosos (cuasi-)identificadores no contemplados en la anonimización manual para cumplir con el requisito $k = 2$.

V. Conclusiones y trabajo futuro

PETRE representa una solución práctica y configurable para mejorar la anonimización de texto, especialmente para los métodos basados en NER, que son los más utilizados en la

[1] https://github.com/microsoft/presidio/releases/tag/2.2.354

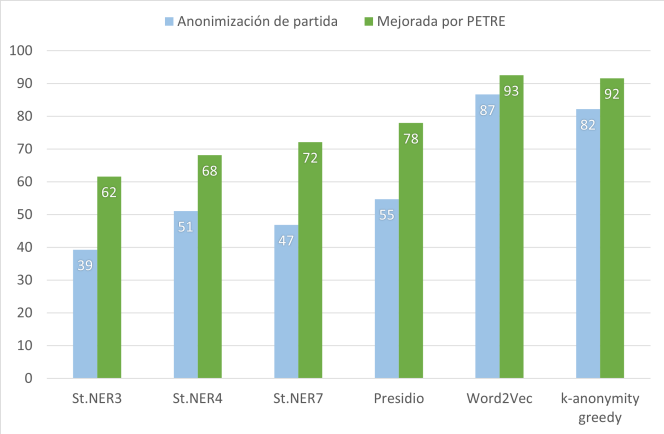


Figura 2. Recall antes y después de PETRE

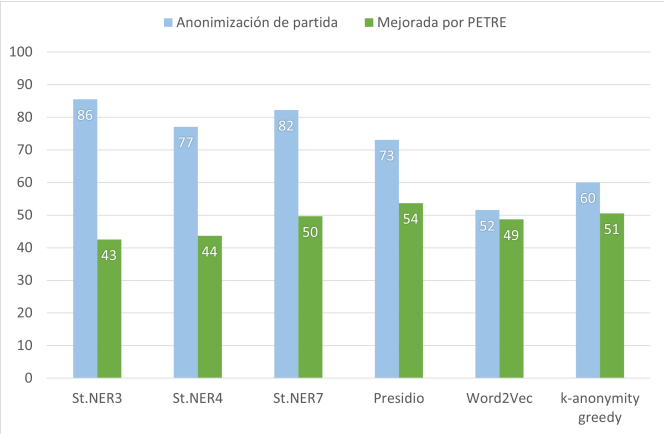


Figura 3. Precisión antes y después de PETRE

práctica pero que carecen de garantías. Los resultados experimentales confirman que PETRE mejora significativamente y consistentemente el nivel de privacidad de las anonimizaciones de partida.

La evaluación del riesgo empírico de PETRE también ha puesto de manifiesto deficiencias en la protección lograda por la anonimización manual.

El principal inconveniente de la propuesta es el tiempo de ejecución causado por la naturaleza iterativa del proceso y el elevado coste computacional del método de explicabilidad (SHAP[14]), siendo esta la operación más costosa. Como trabajo futuro, se pretende investigar métodos de explicabilidad alternativos con un menor coste computacional.

Agradecimientos

Este trabajo ha sido financiado por la Generalitat de Catalunya (Premio ICREA Acadèmia a D. Sánchez, y ayuda 2021SGR-00115), el MCIN/AEI bajo la ayuda PID2024-157271NB-I00 "CLEARING-IT", y los fondos NextGenerationEU/PRTR de la Unión Europea a través del INCIBE (proyecto "HERMES" y Cátedra de Ciberseguridad INCIBE-URV). B. Manzanares-Salor cuenta con el apoyo del Gobierno de España bajo la ayuda FPU23/01785.

Referencias

[1] B. Manzanares-Salor and D. Sánchez, "Enhancing text anonymization via re-identification risk-based explainability," *Knowledge-Based Systems*, vol. 310, p. 112945, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S095070512401579X

[2] European Commission, "Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation)," 2016. [Online]. Available: https://eur-lex.europa.eu/eli/reg/2016/679/oj

[3] P. Samarati, "Protecting respondents identities in microdata release," *IEEE Transactions on Knowledge and Data Engineering*, vol. 13, no. 6, pp. 1010–1027, 2001.

[4] C. Dwork, "Differential privacy," in *International Colloquium on Automata, Languages and Programming*, vol. 4052. Springer, 2006, Conference Proceedings, pp. 1–12.

[5] E. Bier, R. Chow, P. Golle, T. H. King, and J. Staddon, "The rules of redaction: Identify, protect, review (and repeat)," *IEEE Security & Privacy*, vol. 7, no. 6, pp. 46–53, 2009.

[6] S. Lima-López, E. Farré-Maduell, L. Gasco Sánchez, A. López Rueda, S. Frid, A. Conesa, X. Pastor, and M. Krallinger, "Carmen-i: Anonymization protocol for clinical reports in english," Zenodo, Tech. Rep., 2023.

[7] Y. Huang, Z. Song, D. Chen, K. Li, and S. Arora, "Texthide: Tackling data privacy in language understanding tasks," in *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, 2020, pp. 1368–1382.

[8] F. Hassan, J. Domingo-Ferrer, and J. Soria-Comas, "Anonymization of unstructured data via named-entity recognition," in *International Conference on Modeling Decisions for Artificial Intelligence*. Mallorca, Spain: Springer, 2018, Conference Proceedings, pp. 296–305.

[9] I. Pilán, B. Manzanares-Salor, D. Sánchez, and P. Lison, "Truthful text sanitization guided by inference attacks," *Applied Soft Computing*, vol. 185, p. 114013, 2025.

[10] P. Lison, I. Pilán, D. Sánchez, M. Batet, and L. 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)*, 2021, pp. 4188–4203.

[11] J. Soria-Comas and J. Domingo-Ferrer, "Probabilistic k-anonymity through microaggregation and data swapping," in *2012 IEEE International Conference on Fuzzy Systems*, 2012, pp. 1–8.

[12] A. Papadopoulou, P. Lison, L. Øvrelid, and I. Pilán, "Bootstrapping text anonymization models with distant supervision," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, Jun. 2022, pp. 4477–4487.

[13] B. Manzanares-Salor, D. Sánchez, and P. Lison, "Evaluating the disclosure risk of anonymized documents via a machine learning-based re-identification attack," *Data Min. Knowl. Discov.*, vol. 38, no. 6, pp. 4040–4075, 2024. [Online]. Available: https://doi.org/10.1007/s10618-024-01066-3

[14] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, vol. 30. Association for Computing Machinery, 2017, Conference Proceedings, pp. 4768–4777.

[15] F. Hassan, D. Sánchez, and J. Domingo-Ferrer, "Utility-preserving privacy protection of textual documents via word embeddings," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, pp. 1058–1071, 2023.

[16] I. Pilán, P. Lison, L. Ørelid, A. Papadopoulou, D. Sánchez, and M. Batet, "The text anonymization benchmark (tab): A dedicated corpus and evaluation framework for text anonymization," *Computational Linguistics*, vol. 48, no. 8, pp. 1053–1101, 2022.

[17] B. Manzanares-Salor, D. Sánchez, and P. Lison, "Automatic evaluation of disclosure risks of text anonymization methods," in *Privacy in Statistical Databases - International Conference, PSD 2022, Paris, France, September 21-23, 2022, Proceedings*, ser. Lecture Notes in Computer Science, J. Domingo-Ferrer and M. Laurent, Eds., vol. 13463. Springer, 2022, pp. 157–171.

[18] B. Manzanares-Salor and D. Sánchez, "A comparative analysis, enhancement and evaluation of text anonymization with pre-trained large language models," *Expert Systems with Applications*, vol. 297, p. 129474, 2026. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417425030908

[19] C. D. Manning, M. Surdeanu, J. Bauer, J. R. Finkel, S. Bethard, and D. McClosky, "The stanford corenlp natural language processing toolkit," in *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Baltimore, Maryland, USA, 2014, pp. 55–60.