NERaseText: Sensitive-Aware Text Sanitization under Differential Privacy

Félix Melo[1,3] **Luis Miranda**[2,3] **Matías Toro**[1,3]
Federico Olmedo[1,3] Jocelyn Dunstan[2,3]
[1]Universidad de Chile [2]Pontificia Universidad Católica de Chile [3]Instituto Milenio Fundamento de los Datos
{mtoro, federico.olmedo, felix.melo}@dcc.uchile.cl
{lmirandn,jdunstan}@uc.cl

Abstract

The widespread adoption of large language models (LLMs) has increased the need for privacy-preserving text processing techniques. Existing differentially private text sanitization methods apply uniform privacy parameters across all vocabulary elements, failing to recognize that certain words carry different levels of sensitive information. We propose **NERaseText**, a Named Entity Recognition-assisted differentially private text sanitization framework that dynamically allocates privacy budgets based on the sensitivity level of individual words. Our approach achieves comparable results to a state-of-the-art framework, while also providing tighter privacy guarantees to sensitive words and utilizing a lower privacy budget.

1 Introduction

The rapid proliferation of large language models (LLMs) and their deployment in external services have fundamentally transformed how organizations process text data [4, 7]. However, this adoption has raised privacy concerns due to potential adversarial attacks and stricter regulations [15, 13]. Rather than relying on common anonymization methods, differential privacy (DP) [9] offers a rigorous framework for protecting free-text data, providing strong mathematical guarantees while preserving data utility [16].

Among existing differentially private solutions for protecting Personally Identifiable Information (PII), text sanitization has emerged as a particularly practical approach [10, 18, 6]. Unlike methods requiring model training (e.g., Differentially Private Stochastic Gradient Descent DP-SGD [1, 19]), text sanitization preprocesses documents to remove sensitive information while preserving semantic structure, making the resulting text ready for use with external services [14, 3, 12].

Existing sanitization methods typically treat all tokens as equally sensitive, though some PIIs (e.g., names, emails) clearly carry greater privacy risks. Frequency-based adaptations [18] add more noise to rare words under the assumption of higher sensitivity, but frequency alone does not reliably identify sensitive content and would discriminate between rare and common PIIs, potentially affecting underrepresented and overrrepresented populations disproportionately. A semantically grounded alternative is Named Entity Recognition (NER), which has proven effective in detecting sensitive entities regardless of frequency [5, 17].

Furthermore, several existing methods [18, 6, 11] introduce ad-hoc privacy notions, making them difficult to compare with other approaches and leaving unclear what exactly is being protected. To address this limitation, we adopt Local Differential Privacy (LDP), a widely accepted framework that provides strong theoretical guarantees [9, 10]. LDP ensures that individuals privatize their data before sharing it, removing the need for a trusted curator.[9, 10].

Proceedings of the Latinx in AI Workshop @ NeurIPS-25.

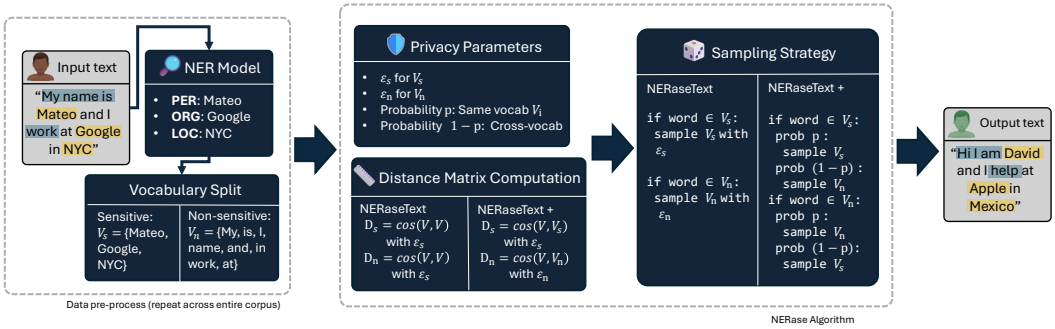

Figure 1: NERaseText workflow: NER model identifies sensitive entities (highlighted in orange) to partition vocabulary into V_s and V_n. Distance matrices are computed with privacy parameters $\epsilon_s < \epsilon_n$. NERaseText samples within vocabulary boundaries while NERaseText+ enables probabilistic cross-vocabulary sampling controlled by parameter p.

Building on these insights, we propose **NERaseText**, a novel approach that combines LDP with NER-based sensitivity detection to achieve sensitivity-aware text sanitization. By dynamically allocating different privacy budgets to sensitive and non-sensitive words ($\epsilon_s < \epsilon_n$), our approach effectively reduces the total privacy cost per document compared to uniform methods while maintaining strong protection for sensitive entities. Our method is evaluated on two datasets containing sensitive information, demonstrating that with rigorous privacy levels, we can maintain competitive utility on downstream tasks while providing stronger privacy guarantees for sensitive entities compared to uniform approaches.

2 Background

2.1 Differential Privacy and Local Differential Privacy

Differential Privacy (DP) [9] provides a rigorous mathematical framework for privacy protection. A randomized mechanism M satisfies ϵ-differential privacy if for all neighboring datasets D and D' differing by at most one record, and for all possible outputs S:

$$\Pr[M(D) \in S] \leq e^{\epsilon} \cdot \Pr[M(D') \in S] \tag{1}$$

Local Differential Privacy (LDP) [8] extends this concept by requiring each individual to privatize their own data locally before sharing it with any data collector. A randomized mechanism M satisfies ϵ-LDP if for all pairs of input values x, x' and all possible outputs y:

$$\Pr[M(x) = y] \leq e^{\epsilon} \cdot \Pr[M(x') = y] \tag{2}$$

In the context of text sanitization, LDP ensures that an adversary cannot distinguish between any two input words based on the sanitized output, providing strong word-level privacy guarantees without requiring trust in a central data curator.

3 Method

While NER can detect sensitive words, it cannot itself perform privacy-preserving rewriting, as it is deterministic and lacks the probabilistic sampling required for differential privacy. To enable this, we combine NER outputs with embedding-based distance computations for word replacement.

NERaseText is an NER-assisted, differentially private text sanitization framework which is composed of sensitivity detection via NER, embedding mapping with distance computation, and sensitivity-aware replacement (Figure 1). It uses two privacy parameters: ϵ_s for sensitive words and ϵ_n for non-sensitive words, with $\epsilon_s < \epsilon_n$ to enforce stricter privacy for sensitive entities while preserving utility elsewhere. We introduce two variants: NERaseText, which operates within a unified vocabulary space to maintain semantic coherence, and NERaseText+, which employs probabilistic cross-vocabulary sampling for stronger privacy through vocabulary mixing.

3.1 Sensitivity Detection and Embedding Mapping

We first run a NER model on the target corpus to identify sensitive words in predefined categories (Persons, Organizations, Locations) with confidence $p \geq t$, where t is a threshold parameter controlling how strictly a word is classified into a category. For our experiments, we employ Flair's framework pre-trained NER model [2], which provides state-of-the-art performance for English entity recognition. This partitions the vocabulary V into disjoint sets: sensitive words V_s and non-sensitive words $V_n = V \setminus V_s$. While we focus on these three entity types, future work could extend coverage to dates, phone numbers, medical terms, and financial information.
Then, we generate word-to-index mappings and construct the embedding lists: `all_embeddings` for the unified vocabulary, `sensitive_embeddings` and `normal_embeddings` with corresponding index mappings. The NER process can be applied either on the corpus vocabulary for efficiency or on the entire embedding vocabulary for broader coverage.

3.2 Distance Computation and Word Replacement

NERaseText computes distance matrices with cosine similarity: D_s (sensitive embeddings) and D_n (non-sensitive embeddings) using privacy parameters ϵ_s and ϵ_n respectively. NERaseText+ adds probabilistic cross-vocabulary sampling: words sample from their own category with probability p, otherwise from the opposite category. The parameters ϵ_s, ϵ_n, and p enable fine-tuning of the privacy-utility trade-off.

3.3 Privacy Analysis

Both NERaseText variants satisfy word-level Local Differential Privacy, providing theoretical guarantees for individual word protection. Unlike related approaches [18], our method provides principled privacy guarantees based on semantically meaningful sensitivity categories.

NERaseText Privacy Guarantee: For any two input words x, x' and output word y, NERaseText satisfies $(\epsilon_n \cdot c)$-LDP where c is the maximum embedding distance, ensuring consistent privacy protection across the vocabulary.

NERaseText+ Privacy Guarantee: The probabilistic cross-vocabulary sampling introduces additional privacy bounds dependent on the sampling parameter p, with the overall privacy guarantee determined by $\max(\epsilon_s \cdot c + \ln(\frac{1-p}{p}), \epsilon_n \cdot c + \ln(\frac{1-p}{p}), \epsilon_n \cdot c)$.

4 Experiments

4.1 Experimental Setup

For a fair evaluation between frameworks, we used local differential privacy (LDP) as the benchmark for privacy comparison, with all downstream tasks performed under the same conditions.

Baseline Methods: We compare against SANTEXT [18], an embedding-based method applying uniform privacy parameters. Evaluation uses pre-trained GloVe embeddings on sentiment analysis (SST-2) and natural language inference (QNLI) tasks. We also include non-private baselines as performance upper bounds.

We evaluate NERaseText on standard text datasets under multiple privacy settings, using pre-trained GloVe embeddings. Sanitized texts are tested on sentiment analysis (SST-2) and natural language inference (QNLI).

Flair's NER model (ner-english-large) identifies entities with confidence threshold t = 0.3. Privacy parameters $\epsilon \in 1, 2, 4, 8$ are applied per word, (total budget scales with each individual use of epsilon), with cross-sampling probability $1 - p = 0.3$ for NERaseText+. The per-word budget is $\epsilon_n \cdot c$, and since $1 - p < 0.5$, the same applies to NERaseText+. The total epsilon of the corpus is the maximum epsilon between all its documents.

Cosine similarity defines the embedding distance matrix. With LDP, the privacy parameter is scaled by the maximum possible distance c, which equals 1 for cosine similarity, minimizing budget inflation.

4.2 Results & Discussion

Table 1 shows the performance of models trained on sanitized texts compared to original texts. NERaseText maintains competitive performance while providing stronger privacy guarantees for sensitive entities.

Method	Benchmark Accuracy (%)		Total Privacy Budget ϵ	
	SST-2	QNLI	SST-2	QNLI
Original (No Privacy)	92.43	90.87	0	0
SANTEXT ($\epsilon = 1$)	49.66	52.48	53	431
SANTEXT ($\epsilon = 2$)	50.46	53.32	106	862
SANTEXT ($\epsilon = 4$)	49.66	52.24	212	1724
SANTEXT ($\epsilon = 8$)	49.42	52.9	424	3448
NERaseText (ϵ_n=1, ϵ_s=0.5)	51.48	52.5	47.5	373
NERaseText (ϵ_n=2, ϵ_s=1)	50.34	53.4	95	746
NERaseText (ϵ_n=4, ϵ_s=2)	50.34	53.5	190	1492
NERaseText (ϵ_n=8, ϵ_s=4)	49.54	52.3	380	2984
NERaseText+ (ϵ_n=1, ϵ_s=0.5, p=0.7)	**51.5**	52.9	**43**	345.5
NERaseText+ (ϵ_n=2, ϵ_s=1, p=0.7)	47.13	51.9	92	708
NERaseText+ (ϵ_n=4, ϵ_s=2, p=0.7)	51.8	**53.6**	180	**1410**
NERaseText+ (ϵ_n=8, ϵ_s=4, p=0.7)	48.9	53.6	360	2812

Table 1: Downstream task performance on sanitized texts. Higher scores indicate better utility preservation.

NERaseText maintains performance comparable to SANTEXT, slightly better in some cases, on sentiment and QNLI tasks, while reducing the total practical privacy budget (e.g., 43 vs. 53 at $\epsilon = 1$ for SST-2 and 1410 vs 1724 at $\epsilon = 4$ for QNLI). This shows that sensitivity-aware allocation provides better protection for sensitive entities without harming downstream accuracy. NERaseText+ offers additional probabilistic guarantees, trading some utility for stronger privacy. Overall, the results validate that privacy can be applied selectively without degrading model usefulness.

Theoretical analysis confirms that both NERaseText variants satisfy local differential privacy [9], with privacy bounds dependent on the maximum embedding distance and the chosen epsilon values. The sensitivity-aware approach provides stronger guarantees for high-risk entities while allowing better utility preservation for non-sensitive content.

5 Conclusions & Future Work

We introduced NERaseText, a sensitivity-aware text sanitization method under LDP. By leveraging NER, our framework dynamically allocates stricter privacy budgets to sensitive PIIs while maintaining utility for non-sensitive ones. Our results show that NERaseText achieves comparable downstream accuracy to prior work while reducing overall practical privacy costs and strengthening protection of high-risk entities. The NERaseText+ variant further extends this approach by incorporating probabilistic replacement mechanisms that provide additional privacy guarantees at the cost of some utility degradation.

We will address several key limitations and opportunities for improvement. First, PIIs extend beyond the LOC/ORG/PER categories studied here, requiring more comprehensive NER models for other domain-specific sensitive content. Second, we will evaluate the framework on multilingual datasets and diverse domains to assess generalizability beyond English text. Third, we will systematically analyze broader parameter ranges for epsilon values and threshold tuning to better understand the privacy-utility trade-off space. Finally, we will extend text quality measurement beyond downstream task performance to include linguistic quality metrics such as MAUVE scores and semantic coherence analysis.

Our work contributes to making privacy-preserving NLP more nuanced and effective, potentially enabling broader adoption of differential privacy in sensitive text processing applications.

References

[1] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, CCS '16, page 308–318, New York, NY, USA, 2016. Association for Computing Machinery.

[2] A. Akbik, T. Bergmann, D. Blythe, K. Rasul, S. Schweter, and R. Vollgraf. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59, 2019.

[3] Anthropic. Anthropic. `https://www.anthropic.com/`, 2025. Accessed: 2025-09-09.

[4] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA, 2021. Association for Computing Machinery.

[5] O. Bridal, T. Vakili, and M. Santini. Cross-Clinic De-Identification of Swedish Electronic Health Records: Nuances and Caveats. In *Proceedings of the Workshop on Ethical and Legal Issues in Human Language Technologies and Multilingual De-Identification of Sensitive Data In Language Resources within the 13th Language Resources and Evaluation Conference*, pages 49–52, Marseille, France, June 2022. European Language Resources Association.

[6] S. Chen, F. Mo, Y. Wang, C. Chen, J.-Y. Nie, C. Wang, and J. Cui. A customized text sanitization mechanism with differential privacy. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5747–5758, Toronto, Canada, July 2023. Association for Computational Linguistics.

[7] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, and (...). The llama 3 herd of models, 2024.

[8] J. C. Duchi, M. I. Jordan, and M. J. Wainwright. Local privacy, data processing inequalities, and statistical minimax rates, 2014.

[9] C. Dwork. Differential privacy. In M. Bugliesi, B. Preneel, V. Sassone, and I. Wegener, editors, *Automata, Languages and Programming*, pages 1–12, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.

[10] O. Klymenko, S. Meisenbacher, and F. Matthes. Differential privacy in natural language processing the story so far. In O. Feyisetan, S. Ghanavati, P. Thaine, I. Habernal, and F. Mireshghallah, editors, *Proceedings of the Fourth Workshop on Privacy in Natural Language Processing*, pages 1–11, Seattle, United States, July 2022. Association for Computational Linguistics.

[11] L. Lyu, X. He, and Y. Li. Differentially private representation for NLP: Formal guarantee and an empirical study on privacy and fairness. In T. Cohn, Y. He, and Y. Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2355–2365, Online, Nov. 2020. Association for Computational Linguistics.

[12] Meta AI. Meta ai. `https://ai.meta.com/`, 2025. Accessed: 2025-09-09.

[13] Office of the Federal Register, National Archives and Records Administration. Public law 104 - 191 - health insurance portability and accountability act of 1996, 1996.

[14] OpenAI. Openai. `https://openai.com/`, 2025. Accessed: 2025-09-09.

[15] The European Parliament and the Council of the European Union. General data protection regulation, 2016.

[16] Y. Wang, Q. Wang, L. Zhao, and C. Wang. Differential privacy in deep learning: Privacy and beyond. *Future Generation Computer Systems*, 148:408–424, 2023.

[17] X. Yang, T. Lyu, Q. Li, C.-Y. Lee, J. Bian, W. R. Hogan, and Y. Wu. A study of deep learning methods for de-identification of clinical notes in cross-institute settings. *BMC Medical Informatics and Decision Making*, 19(5):232, Dec. 2019.

[18] X. Yue, M. Du, T. Wang, Y. Li, H. Sun, and S. S. M. Chow. Differential privacy for text analytics via natural text sanitization. In *Findings, ACL-IJCNLP 2021*, 2021.

[19] X. Yue, H. Inan, X. Li, G. Kumar, J. McAnallen, H. Shajari, H. Sun, D. Levitan, and R. Sim. Synthetic text generation with differential privacy: A simple and practical recipe. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1321–1342, Toronto, Canada, July 2023. Association for Computational Linguistics.