Privacy- and Utility-Preserving NLP with Anonymized Data: A case study of Pseudonymization

Olexandr Yermilov[1]*, **Vipul Raheja**[2] , **Artem Chernodub**[2]
[1]Ukrainian Catholic University, Applied Sciences Faculty, [2]Grammarly
oleksandr.yermilov@ucu.edu.ua,
{vipul.raheja,artem.chernodub}@grammarly.com

Abstract

This work investigates the effectiveness of different pseudonymization techniques, ranging from rule-based substitutions to using pre-trained Large Language Models (LLMs), on a variety of datasets and models used for two widely used NLP tasks: text classification and summarization. Our work provides crucial insights into the gaps between original and anonymized data (focusing on the pseudonymization technique) and model quality and fosters future research into higher-quality anonymization techniques to better balance the trade-offs between data protection and utility preservation. We make our code, pseudonymized datasets, and downstream models publicly available.[1]

1 Introduction

With the advances in artificial intelligence and data-hungry machine learning systems, privacy and compliant data governance have become increasingly important. Text documents, such as emails, court rulings, customer service chats, interview transcripts, and patient records, frequently contain personally identifiable information (PII), such as mentions of persons, locations, organizations, etc. While the collection and use of text data is necessary for improving products or services, conducting research, or providing personalized recommendations, it has to be done in a safe, responsible and compliant way.

However, access to text data becomes a challenge where data containing personally identifiable mentions is involved. Although it is a widely accepted notion that no data is truly anonymous and is said to be an unattainable target (Rocher et al., 2019), pseudonymization, on the other hand, is recognized by the GDPR as one of the ways (and a requirement) to reduce risks of re-identification of a data subject (European Commission, 2016). Following Eder et al. (2022), we define *pseudonymization* as recognizing entities bearing privacy-sensitive information, and their replacement by realistic substitutes.

Original	Sarah works at The Times in London with Rachel and David.
Sanitized	PERSON_1 works at ORGANIZATION_1 in LOCATION_1 with PERSON_2 and PERSON_3.
Pseudonymized	Sophie works at Manchester Evening News in Manchester with Emma and Tom.

Table 1: While the primary focus of our work is Pseudonymization, we use Sanitization as a baseline for comparison. Different types of underlines correspond to different categories of entities to be pseudonymized.

With the right implementation and safeguards, pseudonymization can be a useful technique for protecting the privacy of individuals while still enabling data-driven technological advances, such as NLP research, enabling researchers to work with sensitive data, while reducing data privacy risks. However, there is a risk that quality of texts can often be compromised by techniques such as pseudonymization, which can not only negatively affect downstream NLP tasks and analyses, it can also reduce the utility of anonymized data for other research. It is noteworthy that while privacy and utility-preserving NLP has been a crucial topic in the medical domain, it has been largely overlooked in mainstream NLP research, barring a few recent works (Section 2). The quality of clinical texts can often be compromised by de-identification. Therefore, in this work, we investigate the effectiveness of pseudonymization as a technique for working with NLP models. Specifically, we consider three different systems for pseudonymization:

1. **NER**, which uses named entity recognition

*Work done as an intern at Grammarly.
[1]https://github.com/olexandryermilov/privacy-preserving-nlp

Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023), pages 232–241

(NER) models to detect text spans containing PII, and then uses a knowledge graph to replace the detected spans;
2. **Seq2Seq**, which formulates the task of pseudonymization as a sequence-to-sequence (Seq2Seq) transformation, using an encoder-decoder model;
3. **LLM**, which leverages the zero-shot and few-shot learning capabilities of large, pre-trained language models (LLMs) for performing the task of pseudonymization.

We then use the aforementioned systems to pseudonymize training datasets for two widely-used NLP tasks: text classification and summarization, and evaluate the performance of models (trained on these pseudonymized datasets) on downstream tasks. Through our analyses, we provide crucial insights into the effectiveness of different pseudonymization techniques for data anonymization, and their effect on downstream NLP tasks, from a privacy and utility perspective. Finally, we make our code, pseudonymized datasets, and downstream models publicly available to foster future research into privacy- and utility-preserving NLP.

2 Related Work

Pseudonymization has predominantly been researched in Clinical NLP up until recently, focusing on NLP techniques on how to replace PII such as named entities in medical texts, across different languages. For English medical texts, Sweeney (1996) was one of the first pseudonymization systems, followed by numerous works such as Sweeney et al. (2005); Uzuner et al. (2007); Neamatullah et al. (2008); Meystre et al. (2010); Kushida et al. (2012); Carrell et al. (2013); Sánchez et al. (2013); Meystre (2015); Dernoncourt et al. (2017); Liu et al. (2017); Iwendi et al. (2020).

The techniques proposed in related works range from simply replacing the detected text spans by a placeholders, pseudonyms or synthetic surrogates using lists, lexical substitution such as synonyms or hypernyms, or knowledge bases (Lison et al., 2021; Pilán et al., 2022). Relatedly, techniques such as *C*-sanitize (Sánchez and Batet, 2016), *t*-plausibility (Anandan et al., 2012), and more recently, Yue et al. (2021) have proposed frameworks for privacy-aware and -preserving document sanitization and pseudonymization.

While numerous recent works such as the aforementioned ones have investigated the topic of pseudonymization, our work comes closest to Lamboltshammer et al. (2019); Obeid et al. (2019); Berg et al. (2020); Vakili et al. (2022) and Liu et al. (2023), which focus on analyzing different techniques of data anonymization or pseudonymization, and their effect on downstream tasks. However, our work differs from those since they focus on different domains, different tasks, and different techniques.

3 Pseudonymization Systems

The general architecture of a pseudonymization system consists of two steps, where they first recognize entities bearing PII (detection), and the second sub-system their replacement by realistic substitutes (replacement). For this work, we restrict our analysis to three predominant categories of named entities: `PERSON (PER)`, `LOCATION (LOC)`, and `ORGANIZATION (ORG)`. Using this general framework, we describe the three types of systems that are used in our experiments:

3.1 NER-based Pseudonymization (NER-PS)

The NER-based system uses an off-the-shelf Named Entity Recognition (NER) system to first detect spans of named entities that belong to the aforementioned categories. We use two publicly available NER systems for the first stage: spaCy[2] and FLAIR[3]. The Spacy NER is a fine-tuned RoBERTa model (Liu et al., 2019), whereas the FLAIR NER is a LSTM-CRF model based on Flair embeddings (Akbik et al., 2018).

The detected named entity spans are then replaced with named entities having similar characteristics, such as gender and language of origin (as described in Wikidata) for `PER`s, and so on. We first generate a list of replacement candidates, and then randomly sample a single item from this list under some predefined constraints (details in A.1).

We refer to the two NER-based systems as **NER-PS$_{\text{(SPACY)}}$** and **NER-PS$_{\text{(FLAIR)}}$**.

3.2 Seq2Seq Pseudonymization (Seq2Seq-PS)

The Seq2Seq-based system was developed by fine-tuning a BART-`base` model (Lewis et al., 2020) on a parallel corpus of pseudonymized texts (created using the NER-PS system). An important thing to note is that this system does not exactly fit the

[2]We use spaCy v3.5.1: `spacy.io/usage/v3-5`
[3]We use FLAIR v0.12.2: `github.com/flairNLP/flair`

Task	Dataset name	train size	dev size	test size	domain	metrics
Summarization	CNN/DM (Nallapati et al., 2016)	286,817	13,368	11,487	news	ROUGE-1/2/L
Text classification	IMDB (Maas et al., 2011)	25,000	N/A	25,000	movie reviews	F-score

Table 2: Details of the evaluated downstream tasks.

two-step process outlined above, as it performs the full task in a single-step text-to-text transformation.

Specifically, we developed two variants of this system using the same NER models as the NER-PS. We refer to the two Seq2Seq-PS variants as **Seq2Seq-PS$_{(SPACY)}$**, **Seq2Seq-PS$_{(FLAIR)}$**, depending on which NER-PS system was used to create the parallel training data for training the system.

3.3 LLM Pseudonymization (LLM-PS)

Following the aforementioned two-step architecture, the LLM-based system is based on a sequential chain of two LLMs: GPT-3[4] (Brown et al., 2020) and ChatGPT[5]. For the first step, we extract named entities using GPT-3 with a 1-shot prompt (details in Appendix A.3), and then perform 1-shot pseudonymization on the extracted named entities using ChatGPT.

We chose GPT-3 to perform the detection step of the architecture due to the fact it works much faster on big paragraphs of text (characterized by both text classification and summarization tasks). Despite being considerably slow, we chose ChatGPT (GPT-3.5) for the replacement step, since the size of the input text is much smaller for the replacement sub-task, and we observed better qualitative performance with this model compared to GPT-3.

4 Experiments

In this section, we experimentally evaluate the considered pseudonymization methods. First, we evaluate the negative impact of pseudonymization on the downstream tasks' quality. Next, we compare the privacy preservation quality of different pseudonymization methods. Finally, we evaluate the consistency and privacy-preservation characteristics of pseudonymized texts through a text syntheticity detection experiment.

4.1 Downstream Tasks Performance

Since pseudonymization may introduce additional noise into the processed data, we evaluate the impact of various pseudonymization methods on target dataset quality for the respective downstream tasks. We first pseudonymize the texts for two downstream tasks: Summarization and Text Classification (Table 2), using the aforementioned methods, and then train and evaluate the trained models on their respective task-specific metrics.

	Summarization			Classification
	ROUGE-1	ROUGE-2	ROUGE-L	F-score
Original text	**42.82**	**20.13**	**36.33**	**88.42**
NER-S$_{(SPACY)}$	41.59	19.17	29.07	87.65
NER-S$_{(FLAIR)}$	39.05	17.52	27.43	87.88
NER-PS$_{(SPACY)}$	**41.93**	**19.38**	**29.36**	88.06
NER-PS$_{(FLAIR)}$	40.25	18.04	27.97	88.14
S2S-PS$_{(SPACY)}$	39.1	17.23	26.96	88.10
S2S-PS$_{(FLAIR)}$	36.04	15.07	24.73	88.13
LLM-PS	38.62	16.57	26.34	**88.15**

Table 3: Results of downstream evaluation tasks: summarization (left) and text classification (right). The smaller the gap with the original text, the better the utility is preserved.

For training, we fine-tune the `bart-base`[6] (Lewis et al., 2020) for the Summarization task, and `bert-base-cased`[7] (Devlin et al., 2019) for the Text Classification task. In both scenarios, we train the models for three epochs using *AdamW* optimization (Loshchilov and Hutter, 2017) with learning rate $\alpha = 2 * 10^{-5}$, and batch size 8.

For evaluation, as a baseline, we use the quality obtained with the original (non-pseudonymized) texts using the same training process to make sure the difference in metrics is caused only by the difference in the training datasets. Also, as an additional baseline, we compare the results of pseudonymization with two NER-based sanitizations (Table 1 for reference) denoted by **NER-S$_{(SPACY)}$** and **NER-S$_{(FLAIR)}$**. The sanitization method is the same as NER-PS (Section 3.1) except that the detected named entities are replaced with enumerated placeholders, e.g. `PERSON_1`, `LOCATION_2`, and `ORGANIZATION_3`, instead of Wikidata-based named entities. Evaluation results on both the downstream tasks are presented in Ta-

[4]We use `text-curie-001` as the GPT-3 model.
[5]We use `gpt-turbo-3.5` as the GPT-3.5 model.
[6]https://huggingface.co/facebook/bart-base
[7]https://huggingface.co/bert-base-cased

ble 3. We observe that NER-based pseudonymization achieves the best results for the summarization task, and approaches with spaCy as the underlying NER system show better results compared to FLAIR. These results are related to the fact that FLAIR is a better NER system, which results in making more changes to the original text and introducing more noise into the dataset. This is further compounded with LLM-PS, as it performs a greater amount of edits, thus, forcing the summarization model to learn different patterns than the original dataset, leading to lower ROUGE scores.

For the classification task, all pseudonymization approaches show similar results, although using FLAIR as the underlying system results in better classification performance compared to spaCy. The difference in task formulations explains this small difference between methods: sentiment classification mostly relies on words with positive/negative sentiment, not on the named entities in the text (although, named entities might associate with positive/negative sentiment more than others (Batra and Rao, 2010), resulting in a correlation between them and sentiment of the text). Hence, pseudonymization might have a very limited effect on the task-specific performance. On the other hand, the summarization task is more sensitive to any errors introduced by the NER/Replacement models, as any false positives or false negatives might lead to inconsistent entity mentions and entity relationships, leading to a corruption in the data, which might be further learned by the summarization model.

4.2 Privacy Preservation

Another risk with pseudonymization is that some named entities will still remain non-anonymized. To estimate these risks of false negatives, we evaluate our methods of pseudonymization on a standard NER benchmark: The English CoNLL-2003 test set (Tjong Kim Sang and De Meulder, 2003). We pseudonymize the dataset, and compare the resulting texts to the originals. We measure the percentage of named entities of each type in the original texts that get leaked into the pseudonymized texts.

We observe that NER-based approaches show better results than Seq2Seq approaches, and FLAIR approaches show better results compared to their spaCy equivalents (Table 4), which confirms the observations of the previous experiment. Similar to the observations in Section 4.1, the former observation is related to the fact that the errors present in NER systems are propagated into the Seq2Seq approaches due to the way they were trained.

	PER	ORG	LOC	Mean
NER-PS$_{(SPACY)}$	23.00	37.9	**19.48**	27.23
NER-PS$_{(FLAIR)}$	**2.48**	**10.09**	21.55	**10.23**
Seq2Seq-PS$_{(SPACY)}$	70.14	78.68	79.74	75.67
Seq2Seq-PS$_{(FLAIR)}$	14.82	36.65	65.76	36.03
LLM-PS	34.36	33.09	40.36	35.53

Table 4: Results of privacy preservation experiment on CoNLL-2003 test set. We report the False Negative Rate for each type of named entity. Lower is better.

	Precision	Recall	F-score
NER-PS$_{(SPACY)}$	99.12	97.86	98.49
NER-PS$_{(FLAIR)}$	98.68	95.96	97.30
Seq2Seq-PS$_{(SPACY)}$	99.94	99.76	99.85
Seq2Seq-PS$_{(FLAIR)}$	99.61	98.41	99.01
LLM-PS	**85.61**	**66.92**	**75.12**

Table 5: Results of text syntheticity detection experiment. Lower is better.

4.3 Text Syntheticity Detection

As mentioned above, pseudonymization may corrupt relationships and alignment among named entities and other artifacts in the text. For example, the United States never had a president named "John Smith." Due to such contextual distortions, pseudonymization can negatively affect the quality of processed texts in hard-to-predict ways.

To estimate the degree to which pseudonymized texts are similar to natural ones, we carry out a text syntheticity detection experiment. We combine original and pseudonymized texts from the Summarization task to a single dataset, and train a text classification model with the goal of detecting pseudonymized texts from their non-pseudonymized counterparts, using the same model and settings as for the Text Classification task (Section 4.1). The results are presented in Table 5.

LLM-PS shows the best results for this experiment, which are about an order of magnitude better than replacement-based pseudonymization methods. We observe that it is happening because in LLM-rewritten texts, named entities are in better agreement with the context, making it the best-performing system for preserving the syntactic and semantic integrity of the original text.

5 Conclusions

We investigate the effectiveness of pseudonymization for NLP research with privacy-sensitive data. We develop three different approaches for this task, and evaluate them from three aspects: downstream task performance (on two downstream tasks: text summarization and text classification), privacy preservation, and text syntheticity detection. We find that the proposed approaches have pros and cons for pseudonymization, so one must chose what task and objective (privacy vs. utility) is the most important for them. NER-based systems with FLAIR perform the best for privacy preservation and downstream task performance, whereas the LLM-based system shows the best results for preserving the integrity of the text.

Limitations

While we endeavor in this work to shed light on the impact of various pseudonymization techniques, we recognize a major limitation of our work – especially the LLM-based pseudonymization approach. Using closed-source LLMs may not be an acceptable solution for many settings since it requires sending a (potentially sensitive) text to a third-party API, which, in the absence of appropriate legal safeguards and responsible-use agreements, defeats the purpose of privacy preservation.

There are some more technical limitations of the work, such as the following:

- While this is a problem that affects sensitive texts in all languages, all the experiments were conducted for data in the English language only.
- LLMs are highly sensitive to prompts, as well as the number and ordering of examples provided for few-shot learning. In this work, we experimented with a limited number of prompts for LLM-PS due to API cost constraints.
- For the data privacy detection experiment, the FLAIR NER system was trained using the CoNLL-2003 dataset, which might affect its performance for privacy protection tasks. This may also apply to GPT-3 and ChatGPT models as the authors do not state specifically on which data they were trained.
- We considered only a limited part of named entity types, specifically, `PERSON (PER)`, `LOCATION (LOC)`, and `ORGANIZATION (ORG)`, whereas it is well understood that PII encompasses a much broader range of data types (eg. dates, phone numbers, etc.). We also do not consider sentiments associated with named entities used for substitution in the downstream task of text classification.

We plan to address these in future work.

Ethics Statement

User data privacy and data anonymization, are sensitive, and very important matters. Through this work, we try to dive deeper into the challenges and opportunities of using pseudonymization as a technique to strike a suitable tradeoff between privacy- and utility preservation. The goal of this work is to expose the strengths and limitations of different techniques and their implications. The datasets, knowledge bases, and models that we work with have been publicly released for many years. All of these artifacts are considered to be in the public sphere from a privacy perspective. We do not make any recommendations on using these on public or private datasets without proper due diligence for privacy, security, legal, and compliance measures.

Another risk is that pseudonymization may corrupt the names of people, organizations, and locations and state them in an inappropriate context and therefore produce offensive texts.

6 Acknowledgements

We express our gratitude to Oleksii Molchanovskyi, Viktor Zamaruiev, Max Gubin, Dmytro Lider, the Ukrainian Catholic University, and Grammarly for providing support and computational resources. To our communities: While we are writing this, our homeland Ukraine continues to resist the unprovoked Russian invasion. We are grateful to everyone who defends Ukraine, declares support to the people of Ukraine, and is sending aid. Thank you!

References

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Balamurugan Anandan, Chris Clifton, Wei Jiang, Mummoorthy Murugesan, Pedro Pastrana-Camacho, and Luo Si. 2012. T-plausibility: Generalizing words to desensitize text. *Trans. Data Privacy*, 5(3):505–534.

Siddharth Batra and D.T.V Dharmajee Rao. 2010. Entity based sentiment analysis on twitter.

Hanna Berg, Aron Henriksson, and Hercules Dalianis. 2020. The impact of de-identification on downstream named entity recognition in clinical text. In *Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis*, pages 1–11, Online. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

David Carrell, Bradley Malin, John Aberdeen, Samuel Bayer, Cheryl Clark, Ben Wellner, and Lynette Hirschman. 2013. Hiding in plain sight: use of realistic surrogates to reduce exposure of protected health information in clinical text. *Journal of the American Medical Informatics Association*, 20(2):342–348.

Franck Dernoncourt, Ji Young Lee, Özlem Uzuner, and Peter Szolovits. 2017. De-identification of patient notes with recurrent neural networks. *J. Am. Medical Informatics Assoc.*, 24(3):596–606.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Elisabeth Eder, Michael Wiegand, Ulrike Krieg-Holz, and Udo Hahn. 2022. "beste grüße, maria meyer" — pseudonymization of privacy-sensitive information in emails. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 741–752, Marseille, France. European Language Resources Association.

European Commission. 2016. 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) (Text with EEA relevance).

Celestine Iwendi, Syed Atif Moqurrab, Adeel Anjum, Sangeen Khan, Senthilkumar Mohan, and Gautam Srivastava. 2020. N-sanitization: A semantic privacy-preserving framework for unstructured medical datasets. *Computer Communications*, 161:160–171.

Clete A Kushida, Deborah A Nichols, Rik Jadrnicek, Ric Miller, James K Walsh, and Kara Griffin. 2012. Strategies for de-identification and anonymization of electronic health record data for use in multicenter research studies. *Medical care*, 50(Suppl):S82.

Thomas J. Lampoltshammer, Lörinc Thurnay, and Gregor Eibl. 2019. Impact of anonymization on sentiment analysis of twitter postings. *Data Science – Analytics and Applications*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Pierre Lison, Ildikó Pilán, David Sanchez, Montserrat Batet, and Lilja Øvrelid. 2021. 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)*, pages 4188–4203, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Zengjian Liu, Buzhou Tang, Xiaolong Wang, and Qingcai Chen. 2017. De-identification of clinical notes via recurrent neural network and conditional random field. *J. of Biomedical Informatics*, 75(S):S34–S42.

Zhengliang Liu, Xiaowei Yu, Lu Zhang, Zihao Wu, Chao Cao, Haixing Dai, Lin Zhao, Wei Liu, Dinggang Shen, Quanzheng Li, Tianming Liu, Dajiang Zhu, and Xiang Li. 2023. Deid-gpt: Zero-shot medical text de-identification by gpt-4.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Nuno Mamede, Jorge Baptista, and Francisco Dias. 2016. Automated anonymization of text documents. In *2016 IEEE Congress on Evolutionary Computation (CEC)*, pages 1287–1294.

Stephane M. Meystre. 2015. *De-identification of Unstructured Clinical Data for Patient Privacy Protection*, pages 697–716. Springer International Publishing, Cham.

Stephane M Meystre, F Jeffrey Friedlin, Brett R South, Shuying Shen, and Matthew H Samore. 2010. Automatic de-identification of textual documents in the electronic health record: a review of recent research. *BMC medical research methodology*, 10(1):1–16.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

Ishna Neamatullah, Margaret M Douglass, Li-Wei H Lehman, Andrew Reisner, Mauricio Villarroel, William J Long, Peter Szolovits, George B Moody, Roger G Mark, and Gari D Clifford. 2008. Automated de-identification of free-text medical records. *BMC medical informatics and decision making*, 8(1):1–17.

Jihad S. Obeid, Paul M. Heider, Erin R. Weeda, Andrew J. Matuskowitz, Christine M. Carr, Kevin Gagnon, Tami L. Crawford, and Stéphane M. Meystre. 2019. Impact of de-identification on clinical text classification using traditional and deep learning classifiers. *Studies in health technology and informatics*, 264:283 – 287.

Anthi Papadopoulou, Pierre Lison, Lilja Øvrelid, and Ildikó Pilán. 2022. Bootstrapping text anonymization models with distant supervision. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4477–4487, Marseille, France. European Language Resources Association.

Ildikó Pilán, Pierre Lison, Lilja Øvrelid, Anthi Papadopoulou, David Sánchez, and Montserrat Batet. 2022. The Text Anonymization Benchmark (TAB): A Dedicated Corpus and Evaluation Framework for Text Anonymization. *Computational Linguistics*, 48(4):1053–1101.

Luc Rocher, Julien M Hendrickx, and Yves-Alexandre De Montjoye. 2019. Estimating the success of re-identifications in incomplete datasets using generative models. *Nature communications*, 10(1):1–9.

James P Sweeney, Keith S Portell, James A Houck, Reginald D Smith, and John J Mentel. 2005. Patient note deidentification using a find-and-replace iterative process. *Journal of Healthcare Information Management: JHIM*, 19(3):65–70.

Latanya Sweeney. 1996. Replacing personally-identifying information in medical records, the scrub system. In *Proceedings of the AMIA annual fall symposium*, page 333. American Medical Informatics Association.

David Sánchez and Montserrat Batet. 2016. C-sanitized: A privacy model for document redaction and sanitization. *Journal of the Association for Information Science and Technology*, 67(1):148–163.

David Sánchez, Montserrat Batet, and Alexandre Viejo. 2013. Automatic general-purpose sanitization of textual documents. *IEEE Transactions on Information Forensics and Security*, 8(6):853–862.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Özlem Uzuner, Yuan Luo, and Peter Szolovits. 2007. Evaluating the state-of-the-art in automatic de-identification. *Journal of the American Medical Informatics Association*, 14(5):550–563.

Thomas Vakili, Anastasios Lamproudis, Aron Henriksson, and Hercules Dalianis. 2022. Downstream task performance of BERT models pre-trained using automatically de-identified clinical data. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4245–4252, Marseille, France. European Language Resources Association.

Denny Vrandečić. 2012. Wikidata: A new platform for collaborative data collection. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12 Companion, page 1063–1064, New York, NY, USA. Association for Computing Machinery.

Xiang Yue, Minxin Du, Tianhao Wang, Yaliang Li, Huan Sun, and Sherman S. M. Chow. 2021. Differential privacy for text analytics via natural text sanitization. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3853–3866, Online. Association for Computational Linguistics.

A Training Details

A.1 NER Pseudonymization (NER-PS)

As part of the two-step pseudonymization pipeline, for both **NER-PS$_{(SPACY)}$** and **NER-PS$_{(FLAIR)}$** systems, we leverage Wikidata for the second step – generation of candidates for replacement.

Following some prior works (Mamede et al., 2016; Papadopoulou et al., 2022), we sample the replacements candidates from Wikidata[8] (Vrandečić,

[8] https://www.wikidata.org/wiki/Wikidata:Main_Page

2012), a knowledge graph where *objects* (entities) are linked together by *properties*. We consider specific membership properties, namely `instance_of` (P31), `subclass_of` (P279), and `part_of` (P361), indicating a hierarchical association from specific to general.

Given an entity mention that needs to be replaced, we first find a leaf node in the graph that matches the given entity mention. Then, we traverse the graph to extract sibling nodes via the hierarchical associations, and generate replacement candidates based on additional filters. For instance, we filter `PERSON` entity candidates with ones that have the same gender and language of origin. For `ORGANIZATION` entities, similar industry and country; and for `LOCATION` entities, similar location type and country. We then random sample a single item from this list of filtered candidates under the aforementioned constraints.

A.2 Seq2Seq Pseudonymization (Seq2Seq-PS)

We fine-tune `bart-base`[9] (Lewis et al., 2020) for Seq2Seq models. We train the models for three epochs using *AdamW* optimization (Loshchilov and Hutter, 2017) with the learning rate $\alpha = 2 * 10^{-5}$, the batch size is 8. Training corpus was sampled from the Wikipedia articles and has size of 19M samples.

A.3 LLM Pseudonymization (LLM-PS)

Table 6 shows the prompts we have used for calls to GPT-3 and ChatGPT models. In the first prompt, we are giving the example of extracting named entities (specifically, persons, organizations, and locations) from a small paragraph of text. In the second prompt, we are giving the task as a system message and give examples of changing named entities (again, persons, organizations and location) to named entities of the same type. These prompts can be extended to include named entities of other types. However, this approach should be taken with appropriate caution, as it can also change other parts of the text since single-shot GPT-3 might treat other words in the text as named entities. For example, in sample 11165 from IMDB train set, this is the named entities GPT-3 parse from the text: `Friday the 13th, Bernie, old man, family, Slashers` and here is the pseudonymized response from ChatGPT: `Halloween, Nancy, young woman, relatives, Killers`. As we can observe, parts of the request which are not named entities changed in a completely different way: `"family"` was changed to a synonym word `"relatives"`, while `"old man"` was changed to an antonym `"young woman"`.

B Data Examples

Table 7 shows examples of pseudonymization parts of different samples. We can notice the poor performance of S2S-PS$_{(\text{SPACY})}$ and preservation of context in LLM-PS generated text.

[9] https://huggingface.co/facebook/bart-base

Stage	Model	Illustrative Prompt(s) / API calls
NER	GPT-3	Find all the locations, names and organizations in the following text. Write them separated by commas: **Text**: Daniel worked in Google for five years before moving from America to France. Daniel is now working with Emma in Danone and living in Paris. **Answer**: Daniel, Google, America, France, Emma, Danone, Paris. **Text**: <text-to-anonymize> **Answer**: <response-from-API>
Replacement	ChatGPT	{ "role": "system", "content": "Change following named entities using different named entities of the same type." }, { "role": "user", "content": "Africa, James Potter, Google, Poland, Lily Jameson, Danone" }, { "role": "assistant", "content": "Asia, John Lennon, Microsoft, Germany, Anna Smith, Starbucks", }, { "role": "user", "content": <entities-to-pseudonymize> }, { "role": "assistant", "content": <response-from-API>, },

Table 6: Illustrative prompts for single-shot named entity recognition and replacement tasks for the LLM-PS System.

	Text Classification	Text Summarization
Original	Does it get any uglier than this? The only good thing in this movie was Natassia Malthe, with her stunning Norwegian beauty. God, I wish Michael Ironside and the DeLuise brothers would stop accepting dumb roles in dumb movies! I mean, at least SeaQuest was nice! I know Mr. Ironside from a lot of movies, he has acted in 164 movies at this date!! It's true that he was rarely in a major role, but still!	By . Chris Waugh . Pep Guardiola will never be sacked as Bayern Munich head coach according to the Bundesliga champions' chairman. Karl-Heinz Rummenigge was questioned about whether or not he was worried that many of Bayern's German World Cup-winning stars had yet to return to pre-season training when he made the claim. German newspaper Welt am Sonntag carried an article on Sunday claiming Guardiola's side could struggle this season due to the tiring World Cup campaign. VIDEO Scroll down to watch Pep Guardiola lose it with a journalist and get soaked in beer. Bayern Munich chairman Karl-Heinz Rummenigge says the club will 'never' sack boss Pep Guardiola.
NER-PS$_{(SPACY)}$	Does it get any uglier than this? The only good thing in this movie was Boeing Gap, with her stunning Norwegian beauty. God, I wish Lakshmi Kevin and the Hector brothers would stop accepting dumb roles in dumb movies! I mean, at least EGL was nice! I know Mr. Dani from a lot of movies, he has acted in 164 movies at this date!! It's true that he was rarely in a major role, but still!	By . Nikki Scott . *Pep Guardiola* will never be sacked as ASV Cham Engelbert Strauss head coach according to the Bundesliga champions' chairman. KunzKuppuswamyMarkus Kul was questioned about whether or not he was worried that many of ASV Cham Engelbert Strauss's German World Cup-winning stars had yet to return to pre-season training when he made the claim. German newspaper Modernine TV Hub Omnicare carried an article on Sunday claiming Gentek's side could struggle this season due to the tiring World Cup campaign. VIDEO Scroll down to watch *Pep* Gentek lose it with a journalist and get soaked in beer. ASV Cham Engelbert Strauss chairman KunzKuppuswamyMarkus Kul says the club will 'never' sack boss Xavier Gentek.
NER-PS$_{(FLAIR)}$	Does it get any uglier than this? The only good thing in this movie was Delcine Fleak, with her stunning Norwegian beauty. *Elmore*, I wish Nicolas Loveridge and the Perreira brothers would stop accepting dumb roles in dumb movies! I mean, at least *SeaQuest* was nice! I know Mr. Catala from a lot of movies, he has acted in 164 movies at this date!! It's true that he was rarely in a major role, but still!	By . Robin Kloss . Jesús Lascurain will never be sacked as BSV Kickers Emden head coach according to the Bundesliga champions' chairman. Peyush Herwig was questioned about whether or not he was worried that many of *Duchy of Saxe-Weimar-Eisenach's* German World Cup-winning stars had yet to return to pre-season training when he made the claim. German newspaper Der Angriff carried an article on Sunday claiming *Lascurain's* side could struggle this season due to the tiring World Cup campaign. VIDEO Scroll down to watch Jesús Lascurain lose it with a journalist and get soaked in beer. BSV Kickers Emden chairman Peyush Herwig says the club will 'never' sack boss Jesús Lascurain
S2S-PS$_{(SPACY)}$	Does it get any uglier than this? The only good thing in this movie was *Natassia Malthe*, with her stunning Norwegian beauty. God, I wish Alistair D'Alessandro and the *DeLuise* brothers would stop accepting dumb roles in dumb movies! I mean, at least *SeaQuest* was nice! I know Mr. Suryanarayan from a lot of movies, he has acted in 164 movies at this date!! It's true that he was rarely in a major role, but still.	By . Floor Blythe . *Pep Guardiola* will never be sacked as *Bayern Munich* head coach according to the Bundesliga champions' chairman. *Karl-Heinz Rummenigge* was questioned about whether or not he was worried that many of *Bayern*'s German World Cup-winning stars had yet to return to pre-season training when he made the claim. German newspaper *Welt am Sonntag* carried an article on Sunday claiming *Guardiola*'s side could struggle this season due to the tiring World Cup campaign. VIDEO Scroll down to watch *Pep Guardiola* lose it with a journalist and get soaked in beer. *Bayern Munich* chairman Jörn-*Heinz Rummenigge* says the club will 'never' sack boss *Pep Guardiola*.
S2S-PS$_{(FLAIR)}$	Does it get any uglier than this? The only good thing in this movie was Jyotirmoye Dhanraj, with her stunning Norwegian beauty. God, I wish Alvan Kostas and the Sivaramakrishna brothers would stop accepting dumb roles in dumb movies! I mean, at least *SeaQuest* was nice! I know Mr. Sankar from a lot of movies, he has acted in 164 movies at this date!! It's true that he was rarely in a major role, but still!	By . Helge Kowalczyk . Raghuvinder Cárdenas will never be sacked as TSV Heiligheim head coach according to the Bundesliga champions' chairman. Gertrudin Günther was questioned about whether or not he was worried that many of *SV Altenburg*'s German World Cup-winning stars had yet to return to pre-season training when he made the claim. German newspaper *Welt am Sonntag* carried an article on Sunday claiming Cárdenas's side could struggle this season due to the tiring World Cup campaign. VIDEO Scroll down to watch *Gijsbertus* Cárdenas lose it with a journalist and get soaked in beer . TSV *Heiligenburg* chairman Gertrudin *Schleicher* says the club will 'never' sack boss *Gijsbertus* Cárdenas.
LLM-PS	Does it get any uglier than this? The only good thing in this movie was Maria Olsen, with her stunning Norwegian beauty. God, I wish Tricia Helfer and the Hemsworth brothers would stop accepting dumb roles in dumb movies! I mean, at least Battlestar Galactica was nice! I know Mr. *Ironside* from a lot of movies, he has acted in 164 movies at this date !! It's true that he was rarely in a major role, but still!	By. Amanda Wilson. Jürgen Klopp will never be sacked as Borussia Dortmund head coach according to the Bundesliga champions' chairman. Franz Beckenbauer was questioned about whether or not he was worried that many of *Bayern*'s UEFA Champions League-winning stars had yet to return to off-season preparations when he made the claim . German newspaper *Welt am Sonntag* carried an article on Sunday claiming Klopp's side could struggle this season due to the tiring World Cup campaign. VIDEO Scroll down to watch Jürgen Klopp lose it with a journalist and get soaked in beer. Borussia Dortmund chairman Franz Beckenbauer says the club will 'never' sack boss Jürgen Klopp.

Table 7: Examples of Pseudonymization