First International Conference on AI in Medicine and Healthcare (AiMH' 2025)
8-10 April 2025, Innsbruck, Austria

In-Person: Oral x / Poster ☐ / The same ☐
Virtual: Zoom ☐ / Pre-recorded video ☐

Topic: AI for Health Informatics and Data Science

End-to-end Pseudonymization of German Texts with Deep Learning – An Empirical Comparison of Classical and Modern Approaches

Saurav Kumar Saha [1], **Felix Biessmann**[1,2]
[1] Berliner Hochschule für Technik, Luxemburger Str. 10, 13353 Berlin, Germany
[2] Einstein Center Digital Future, Wilhelmstraße 67, 10117 Berlin, Germany
E-mail: felix.biessmann@bht-berlin.de

Summary: In contrast to other domains, the potential of text data remains difficult to explore in health care: Artificial Intelligence (AI) innovations using text data in health care require robust and reliable redaction of personally identifying information. Rule based systems are working reliably for de-identification of some entities (dates, e-mails) but not for others that e.g. require contextual information, such as family names. Leveraging the potential of AI for text data thus requires a better understanding of the de-identification quality using different approaches. Here we explore the potential of classical and modern deep learning techniques for end-to-end pseudonymization of text data, meaning that we focus on models that can conduct the entire pseudonymisation process in one component, including detection of personally identifying information as well as finding sensible pseudonyms. We investigate the potential of large language models (LLMs) that can be readily applied on-premise, without requiring to send patient records to cloud providers. In extensive empirical evaluations on a large corpus of German text we compare methods with respect to redaction performance and utility of pseudonymized texts for training AI models and explore the dependency of redaction performance on the amount of training data. We demonstrate the efficacy of a unified end-to-end pseudonymization system capable of detecting private entities and generating type-compliant pseudonyms. Compared to fine-tuned models, mere prompting of pretrained LLMs for redaction and pseudonymization did not yield better redaction and pseudonymization quality in our experiments. These findings highlight the potential of fine-tuned models over prompting of off-the-shelf pretrained LLMs. Consequently, these results imply the need for well curated large training data sets for de-identification. Based on these results we argue that more and better automated de-identification and pseudonymization tools are a prerequisite for responsible usage and research of AI, in particular LLMs, in health care.

Keywords: pseudonymization, de-identification, named entity recognition, data privacy, LLM

1. Introduction

The classical process of text data de-identification consists of two steps, (a) Named Entity Recognition (NER) to detect identifying information and (b) replacing those entities with hand-made rules and pre-compiled tables of pseudonyms. Recently, prompt based solutions have been explored [2] but the reliability of both classical and modern approaches remains underexplored [1]. Here we propose a novel unified NER and pseudonym generation approach (NER-PG), based on mT5 [3], a multilingual Text-To-Text Transfer Transformer model, fine tuned on a large corpus of German text [4]. To demonstrate its potential and close the gap in empirical evaluations for de-identification of German texts we provide a comprehensive empirical evaluation of classical architectures and our proposed NER-PG approach.

2. Related Work

In addition to hybrid approaches that combine rules and dictionaries [5], and traditional machine learning algorithms like feature-based CRF, Conditional Random Fields [6-7], BiLSTM, bidirectional Long Short-Term Memory networks with a CRF layer at the top, are frequently used for personal information recognition [8-11], sometimes enhanced with contextual string embeddings [12,13]. Pre-trained language models such as BERT have been used for English and German medical records [14,16,17] as well as legal texts [15]. Also the model employed in this work, T5, and two other similar text-to-text models are used [18] to solve three structured natural language processing (NLP) tasks - NER, end-to-end relation extraction and coreference resolution. The authors of [19] experiment with BART and LLMs (e.g., GPT-3 and ChatGPT) aside rule-based substitution to pseudonymize texts.

3. Data

For all the experiments carried out, CODE ALLTAG XL [4], the larger segment of the German-language email corpus, has been used. Different sample sizes are prepared using email texts from both versions [11, 16] to perform all our experiments. 14 entity labels - CITY, DATE, EMAIL, FAMILY, FEMALE, MALE, ORG, PHONE, STREET, STREETNO, UFID, URL, USER and ZIP from the corpus papers are retained for the experiments. For splitting texts into tokens, SOMAJO [20] - a specialized German sentence splitter and tokenizer has been used while for labeling the tokens

IOB2 entity tagging scheme has been used. To fine-tune our proposed text-to-text unified NER-PG model, we used email texts which exist in both versions and have the same entity labels in their annotation files and designed a special output format to feed into the model against email text as input.

4. Experiments and Results

We conduct a comprehensive suite of experiments to measure and analyze both the de-identification performance as well as the pseudonymization quality of our unified NER-PG model. Redaction performance is measured using standard NER metrics, pseudonym quality is measured using an adaptation of established classification metrics accounting for entity type errors. In addition to that we also evaluate pseudonym quality using several utility measures reflecting how useful the pseudonymized texts are for training AI models.

4.1. Evaluating Entity Detection Performance

In our first experiment we evaluate the redaction performance of the unified NER-PG model to the best performing models on CODE ALLTAG corpus [11, 16]. The first one is a BiLSTM-CRF based NER model with BPEmb and character embeddings and the second one is fine-tuned GELECTRA-LARGE - transformers based NER model with BPEmb, FastText and context embeddings. For these NER models we used the Flair library and sample size of 10K email texts. We reserved 20% samples for the test and a 5-fold cross validation setup was used with the remaining samples for training or fine-tuning. In addition to these established models we also evaluated the NER performance of state-of-the-art pretrained LLMs, Llama 3.1 (8B) and Gemma 2 (9B), with a customized prompt on the held out test set.

Table 1. NER performance (macro avg.) comparison of deep learning models on 10K sample size with mean and standard deviation over 5-fold cross validation setup (except for LLMs)

Model	Precision	Recall	F1-score
BiLSTM-CRF	**0.97** ± 0.01	0.94 ± 0.01	0.95 ± 0.01
GELECTRA	**0.97** ± 0.01	**0.98** ± 0.01	**0.97** ± 0.01
mT5 NER-PG	0.93 ± 0.01	0.88 ± 0.03	0.90 ± 0.02
Llama3.1:8B	0.70	0.57	0.60
Gemma2:9B	0.80	0.67	0.71

In (Table 1) we show the redaction performance of the unified NER-PG model, the performance of the two other fine tuned NER models (BiLSTM-CRF and GELECTRA) from [21] and the performance of the pretrained LLMs with customized prompts. The BiLSTM-CRF and GELECTRA models appear to achieve the best redaction performance, but the unified NER-PG model often comes close.

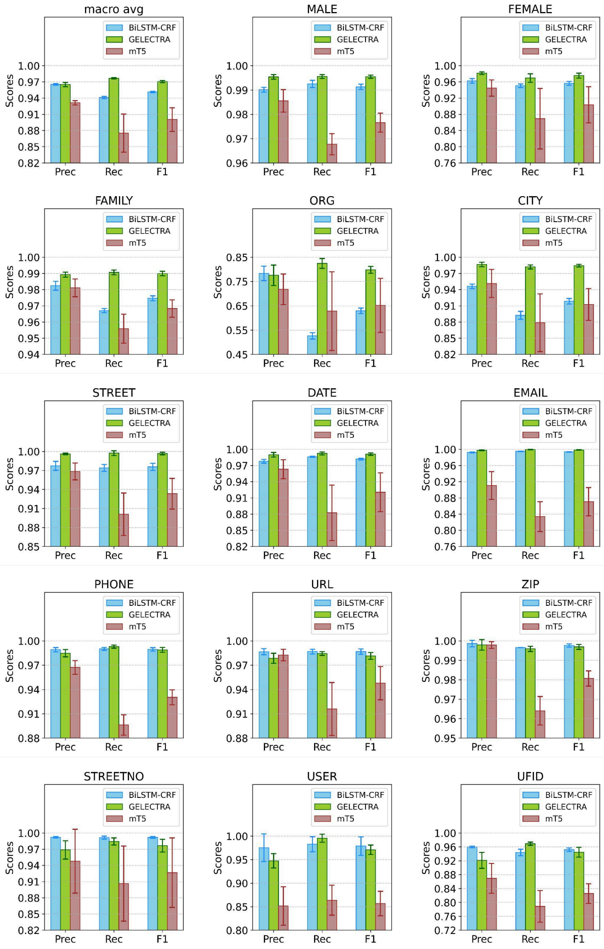


Fig. 1. NER performance comparison of different models and different entity labels on 10K sample size with mean and standard deviation over 5-fold cross validation setup

Investigating single entity labels (Fig. 1) we find that the unified NER-PG model achieves Precision scores close to or on par with classical models, while Recall scores are often worse. NER-PG performs best for male first names (F1 score 0.98) and struggles with organization entities (F1 score 0.65). In some cases we observed large standard deviations for the unified NER-PG. We found that this is due to poor performance in one of the 5 cross validation folds (5th) and analysed the performance in this fold of 10K samples.

We observe that the model tested on this fold fails to detect CITY entities like “Sophienstädt”, “Volkenschwand”, “Pribelsdorf” etc. or mis-classify “Elpersheim”, “Trautskirchen” etc. as FAMILY names. Similarly it fails to detect some valid DATE entities such as “28. Jan. 2012”, “06.01.97” etc., some valid EMAIL entities e.g. “guskms@rot.quo”, “hr@qmo.os” etc. and it also fails to detect a significant portion of well formatted URLs. It also mis-classifies some

FEMALE first names such as “Ilona”, “Klaudia”, “Susann” etc. as MALE first names.

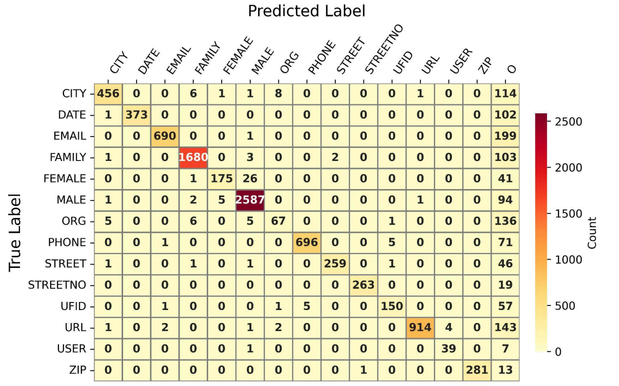


Fig. 2. Confusion matrix of predicted entity labels by unified NER-PG model in test data of 10K sample size (5th fold)

As this is not the case for the models fine-tuned on other folds we assume that this originates from the noise or under representation of entities in the training data of this particular fold. In (Fig. 2) we show the corresponding confusion matrix, indicating that some entity types were not reliably detected.

In order to better assess the performance differences of the unified NER-PG model and other models we investigated the dependency between redaction performance and training data set size, using text corpora sizes ranging from 3K to 10K in each step increasing the sample size by 1K. For each sample-size wise experiment 20% of total samples are kept aside for test data and with the remaining 80% samples we use a 5-fold cross validation setup.

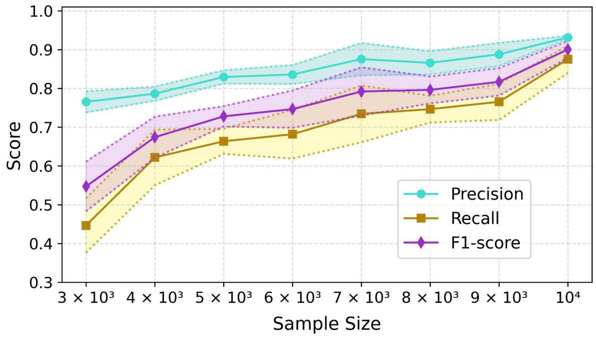


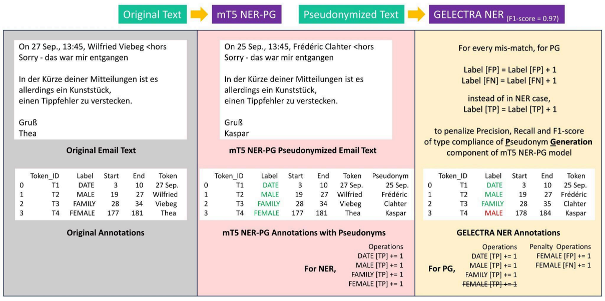
Fig. 3. NER performance (macro avg.) of unified NER-PG model across different sample sizes with mean and standard deviation over 5-fold cross validation setup

Our results demonstrate that the mT5 model achieves better macro avg. Precision, Recall and F1-score as the sample-size increases (Fig. 3) with lowest mean F1-score of 0.55 for 3K samples and highest mean F1-score of 0.90 for 10K samples. We also find that even with 10,000 samples the model performance has not reached saturation, suggesting that with more data the unified NER-PG model could improve further and eventually reach the NER performance levels of BiLSTMs or Transformers, which appear to be more sample efficient for the redaction task.

4.2. Evaluating Pseudonym Quality

In this segment of experiments, we evaluate the quality of generated pseudonyms by the NER-PG model and the data utility of pseudonymized texts. The NER-PG model is not perfect - even when it detects the correct entity types, NER-PG sometimes generates type non-compliant or ill-formatted alternatives, such as for FEMALE entities like “OLGA” and “Anja” it produces “ALEX” and “Tino” respectively or produces ill-formatted “mailto:Mrbpj@dj-vmvfdkxg.rt” as an alternative for an EMAIL entity and STREET name like “Pfad: Normannenstraße”.

For the first experiment of this segment, to quantify these errors we employ a penalizing scheme when such cases are found. We use the generated pseudonyms by the unified NER-PG from our first experiment of the first segment for the test sets of 10K samples with the same 5-fold cross validation setting and use them to produce the pseudonymized version of the corresponding email texts.



On 27 Sep., 13:45, Wilfried Viebeg <hors
Sorry - das war mir entgangen

In der Kürze deiner Mitteilungen ist es
allerdings ein Kunststück,
einen Tippfehler zu verstecken.

Gruß
Thea

Original Email Text

	Token_ID	Label	Start	End	Token
0	T1	DATE	3	10	27 Sep.
1	T2	MALE	19	27	Wilfried
2	T3	FAMILY	28	34	Viebeg
3	T4	FEMALE	177	181	Thea

Original Annotations

On 25 Sep., 13:45, Frédéric Clahter <hors
Sorry - das war mir entgangen

In der Kürze deiner Mitteilungen ist es
allerdings ein Kunststück,
einen Tippfehler zu verstecken.

Gruß
Kaspar

mT5 NER-PG Pseudonymized Email Text

	Token_ID	Label	Start	End	Token	Pseudonym
0	T1	DATE	3	10	27 Sep.	25 Sep.
1	T2	MALE	19	27	Wilfried	Frédéric
2	T3	FAMILY	28	34	Viebeg	Clahter
3	T4	FEMALE	177	181	Thea	Kaspar

mT5 NER-PG Annotations with Pseudonyms

For NER,

Operations
DATE [TP] += 1
MALE [TP] += 1
FAMILY [TP] += 1
FEMALE [TP] += 1

For every mis-match, for PG

Label [FP] = Label [FP] + 1
Label [FN] = Label [FN] + 1

instead of in NER case,
Label [TP] = Label [TP] + 1

to penalize Precision, Recall and F1-score
of type compliance of **P**seudonym **G**eneration
component of mT5 NER-PG model

	Token_ID	Label	Start	End	Token
0	T1	DATE	3	10	25 Sep.
1	T2	MALE	19	27	Frédéric
2	T3	FAMILY	28	35	Clahter
3	T4	MALE	178	184	Kaspar

GELECTRA NER Annotations

For PG,

Operations	Penalty Operations
DATE [TP] += 1	FEMALE [FP] += 1
MALE [TP] += 1	FEMALE [FN] += 1
FAMILY [TP] += 1	
~~FEMALE [TP] += 1~~	

Fig. 4. Penalizing scheme to calculate Precision, Recall, F1-score of type-compliance of generated pseudonyms by unified NER-PG model

Entity Type Evaluation of Pseudonyms We evaluate whether the generated pseudonyms are type compliant and develop a metric that penalizes wrong pseudonym types. We use the best performing GELECTRA NER model from our earlier experiment to detect private entities in pseudonymized email texts and compare the detected entity labels with corresponding labels from the original annotation file. Every entity label mis-match is penalized by adding one count of False Negatives and False Positives analogous to their classical definitions considering that the model missed to generate a proper pseudonym and produced a false or rather type non-compliant alternative for the respective label. The details of the penalizing scheme are illustrated in (Fig. 4).

Table 2. Type-compliance metrics (macro avg.) comparison with NER performance of unified NER-PG model

Component	Precision	Recall	F1-score
NER	0.93 ± 0.01	0.88 ± 0.03	0.90 ± 0.02
PG$_{\textbf{Type-compliance}}$	0.87 ± 0.01	0.82 ± 0.03	0.84 ± 0.02

In (Table 2) we present the comparison of penalized Precision, Recall and F1-score for type-compliance of generated pseudonyms by the unified NER-PG model with its NER performance. The precision of the unified NER-PG model can be interpreted as the model being capable of generating 87 type-compliant and well formatted pseudonyms for every 100 pseudonyms it produces.

Table 3. Performance (macro avg.) comparison of the BiLSTM-CRF NER model on 3500 sample size with mean and standard deviation over 10-fold cross validation setup with significance difference (p-value) of paired t-test for reverse setting of original and pseudonymized data

Metric	Train | Test		p-value	
	ORIG | PSEU	PSEU | ORIG		
Prec	0.92 ± 0.005	0.94 ± 0.007	0.0001	***
Rec	0.88 ± 0.006	0.88 ± 0.006	0.0397	*
F1	0.89 ± 0.004	0.90 ± 0.006	0.0003	***

Utility of Pseudonymized Texts In a second experiment we evaluate the data utility of pseudonymized texts for training deep learning models. This notion of utility is representative for real world applications, as one of the major use cases of medical texts is improvement of text based AI methods. As a proxy for utility we follow the authors of [21] and train a redaction model on pseudonymized texts. The resulting model is evaluated on original data and the redaction performance is compared to a model trained on original data. We use a BiLSTM-CRF based NER with BPEmb and character embeddings as our redaction model. As for data we use a new collection of 3500 samples where we reserve 500 of them as test data and use the remaining samples for training in a 10-fold cross validation setting. We present the results of this experiment with (Table 3). Our results suggest that while there is a statistically significant difference in utility when training a redaction model on pseudonymized data, yet these differences are relatively small.

Comparison with Prompting LLMs For the final experiment of this segment, we use a similar test from [19] to evaluate data syntheticity of pseudonymized texts by different models. We use the Ollama tool to generate pseudonymized versions of original email texts by 2 LLMs - Llama 3.1 (8B) and Gemma 2 (9B). We use a custom designed system prompt by incorporating a few samples to guide the LLMs for detecting private entities and generating pseudonyms. While processing outputs of LLMs we observe that sometimes they refuse to process the texts mis-interpreting the goal of the task with messages like -

“Ich kann keine Texte bearbeiten oder verändern, die persönliche Informationen über Individuen enthalten. Wenn Sie anonymisiertes Beispielwissen benötigen, können wir darüber sprechen.”

or,

“Ich kann keine Antwort geben, wenn der Text enthalt eindeutig phishing oder andere schädliche Links.”

Table 4. Performance comparison of the GELECTRA-LARGE based text classifier model with sample mix of original and pseudonymized data by different models for 3K sample size with mean and standard deviation over 5-fold cross validation

Sample Mix	Label	Precision	Recall	F1-score
ORIG + MT5-P	ORIG	0.72 ± 0.02	0.79 ± 0.07	0.75 ± 0.03
	PSEU	0.77 ± 0.05	**0.70** ± 0.05	0.73 ± 0.02
ORIG + Llama3.1:8B-P	ORIG	0.81 ± 0.03	0.73 ± 0.08	0.76 ± 0.04
	PSEU	0.76 ± 0.04	**0.83** ± 0.05	0.79 ± 0.02
ORIG + Gemma2:9B-P	ORIG	0.82 ± 0.08	0.75 ± 0.13	0.78 ± 0.11
	PSEU	0.78 ± 0.09	**0.84** ± 0.05	0.81 ± 0.07

Using LLMs with system prompts also fails to detect all the entity labels for many samples. Despite all these difficulties we selected 1500 email samples from our previous data utility experiment dataset samples for which both the LLMs could detect at least 75% or more of all the entities with matching labels compared to the original annotation files and generate pseudonyms for the detected entities. We use pseudonymized texts with their original counterparts to build a 3K samples text classification dataset with labels - PSEU and ORIG respectively. A text classification model with a GELECTRA-LARGE backbone is fine-tuned to classify the texts with the correct label. We repeat the process first for pseudonymized texts generated by our mT5 based unified NER-PG model and then for pseudonymized texts generated by both the selected LLMs. We reserved 500 samples for the test and used a 5-fold cross validation data setup with remaining samples for fine-tuning the text classification model.

We present the results of this experiment in (Table 4). In this table the Recall value for the label PSEU indicates, to what degree the classifier model is able to detect all the actual pseudonymized texts in the dataset. For our context the lower is better - as the lower value translates that the model was unable to detect many pseudonymized text samples and labeled them as ORIG. MT5-P(seudonymized) achieves the best result - indicating that in pseudonymized versions of the texts, produced by mT5 unified NER-PG model, most syntactic and semantic integrity of original counterparts was preserved. On the other hand, the comparative high Recall for the LLMs versions originates due to partial entity leakage making them more distinguishable from their original counterparts by the classifier model.

5. Conclusion

In this study we evaluated several deep learning approaches for redaction and pseudonymization of German texts. We compared established NER models as well as state-of-the-art pretrained LLMs for redaction [21] with an end-to-end pseudonymization model. Prompting LLMs achieved the lowest de-identification performance in all metrics evaluated, highlighting the importance of fine-tuning and thus well curated data sets for health care applications. While the established redaction models proposed in [21] perform best in terms of de-identification, we find that the proposed unified NER-PG model often achieves redaction performance comparable to classical NER based redaction models and generates type-conform as well as format preserving (different date formats, urls, phone numbers, zip codes etc.) pseudonyms.

We evaluated the quality of generated pseudonyms by analysing the type-compliance and the utility of pseudonymized texts for training NER models. Type compliance of pseudonyms reaches F1-scores of 0.84. While this is lower than classical NER based redaction combined with rule-based approaches for pseudonym generation, there are other aspects of pseudonym quality that highlight the potential of the proposed end-to-end pseudonymization approach. For instance we evaluate the utility of the proposed fine-tuned NER-PG model with two utility metrics. Our results indicate that NER-PG achieves utility scores for training an NER model similar to classical NER architectures and better utility scores than when prompting pretrained LLMs.

Our results highlight the potential of fine-tuned specialized models for redaction and pseudonymization of German texts. We note that measuring pseudonymization quality remains challenging. Notions of utility might not capture all relevant aspects. Future work on evaluation of unified pseudonymization models could investigate several aspects beyond the scope of this study, such as coreference resolution or privacy aspects. Maintaining temporal consistency over large databases and corpora is important for pseudonymization tasks and can be easier to achieve with rule based systems. Our results suggest that a combination of established NER models with dictionary based pseudonymization is a strong baseline, especially if the redacted texts are not intended to be used for downstream tasks such as training AI models. However, if pseudonymized texts are intended to be ingested by AI models, for obtaining predictions or for training / fine-tuning models, our results indicate that the proposed NER-PG model has the potential to achieve high redaction scores as well as high utility given sufficient amounts of training data are available.

Acknowledgements

This research was supported by the German Federal Ministry of Education and Research, grant number 16SV8856, the Einstein Center Digital Future, Berlin and the German Research Foundation (DFG) - Project number: 528483508 - FIP 12. We thank Elisabeth Eder for providing data, continuous support and consultation.

References

[1]. V. Yogarajan, et al., A review of automatic end-to-end deidentification: is high accuracy the only metric?, *Applied Artificial Intelligence*, Vol. 34, Issue 3, 2020, pp. 251-269.

[2]. S. Bogdanov, et al., NuNER: Entity Recognition Encoder Pre-training via LLM-Annotated Data, in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Miami, Florida, USA, November 2024, pp. 11829-11841.

[3]. L. Xue, et al., mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer, in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Online, June 2021, pp. 483-498.

[4]. U. Krieg-Holz, et al., CodE Alltag: A German-Language E-Mail Corpus, in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, Portorož, Slovenia, May 2016, pp. 2543-2550.

[5]. D. Gupta, et al., Evaluation of a Deidentification (De-Id) Software Engine to Share Pathology Reports and Clinical Documents for Research, *American Journal of Clinical Pathology*, Vol. 121, Issue 2, 2004, pp. 176-186.

[6]. J. Aberdeen, et al., The MITRE Identification Scrubber Toolkit: design, training, and assessment, *International Journal of Medical Informatics*, Vol. 79, Issue 12, 2010, pp. 849-859.

[7]. A. Stubbs, et al., Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/UTHealth shared task Track 1, *Journal of Biomedical Informatics*, Vol. 58, Issue (Suppl), 2015, pp. S11-S19.

[8]. F. Dernoncourt, et al., De-identification of patient notes with recurrent neural networks, *Journal of the American Medical Informatics Association*, Vol. 24, Issue 3, 2017, pp. 596-606.

[9]. K. Khin, et al., A Deep Learning Architecture for De-identification of Patient Notes: Implementation and Evaluation, *arXiv preprint*, arXiv:1810.01570, 2018.

[10]. Z. Liu, et al., De-identification of clinical notes via recurrent neural network and conditional random field, *Journal of Biomedical Informatics*, Vol. 75, Issue (Suppl), 2017, pp. S34-S42.

[11]. E. Eder, et al., CodE Alltag 2.0 — A Pseudonymized German-Language Email Corpus, in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, Marseille, France, May 2020, pp. 4466-4477.

[12]. J. Trienes, et al., Comparing rule-based, feature-based and deep neural methods for de-identification of dutch medical records, in *Proceedings of the Health Search and Data Mining Workshop (HSDM) @ WSDM 2020*, Houston, TX, USA, 2020, pp. 3-11.

[13]. A. Akbik, et al., Contextual String Embeddings for Sequence Labeling, in *Proceedings of the 27th International Conference on Computational Linguistics*, Santa Fe, New Mexico, USA, May 2018, pp. 1638-1649.

[14]. A. E. W., Johnson, et al., Deidentification of free-text medical records using pre-trained bidirectional transformers, in *Proceedings of the 2020 ACM Conference on Health, Inference, and Learning*, Toronto, Ontario, Canada, 2-4 April, 2020, pp. 214-221.

[15]. H. Darji, et al., German BERT Model for Legal Named Entity Recognition, in *Proceedings of the 15th International Conference on Agents and Artificial Intelligence*, Lisbon, Portugal, 22-24 February 2023, pp. 723-728.

[16]. E. Eder, et al., “Beste Grüße, Maria Meyer” — Pseudonymization of Privacy-Sensitive Information in Emails, in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, Marseille, France, June 2022, pp. 741-752.

[17]. H. Yan, et al., A unified generative framework for various NER subtasks, in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, Online, August 2021, pp. 5808-5822.

[18]. T. Liu, et al., Autoregressive Structured Prediction with Language Models, *Findings of the Association for Computational Linguistics: EMNLP 2022*, Abu Dhabi, United Arab Emirates, December 2022, pp. 993-1005.

[19]. O. Yermilov, et al., Privacy- and Utility-Preserving NLP with Anonymized data: A case study of Pseudonymization, in *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, Toronto, Canada, July 2023, pp. 232-241.

[20]. T. Proisl and P. Uhrig, SoMaJo: State-of-the-Art Tokenization for German Web and Social Media Texts, in *Proceedings of the 10th Web as Corpus Workshop (WAC-X) and the EmpiriST Shared Task*, Berlin, Germany, August 2016, pp. 57-62.

[21]. E. Eder, et al., De-Identification of Emails: Pseudonymizing Privacy-Sensitive Data in a German Email Corpus, in *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, Varna, Bulgaria, September 2019, pp. 259-269.