Anonymise: A Tool for Multilingual Document Pseudonymisation

Rinalds Vīksna[1,2] and **Inguna Skadiņa**[1,2]
[1] Tilde, Vienības gatve 75a, Riga, Latvia
[2] Faculty of Science and Technology, University of Latvia, Raiņa bulv. 19, Riga, Latvia
{Firstname.Lastname}@tilde.lv

Abstract

According to the EU legislation, documents containing personal information need to be anonymized before public sharing. However, manual anonymisation is a time-consuming and costly process. Thus, there is a need for a robust text de-identification technique that accurately identifies and replaces personally identifiable information. This paper introduces the *Anonymise* tool, a system for document de-identification. The tool accepts various text document types (e.g., MS Word or plain-text), de-identifies personal information, and saves the de-identified document in its original format. The tool employs a modular architecture, integrating list-based matching, regular expressions and deep-learning-based named entity recognition to detect text spans for redaction. Our evaluation results demonstrate high recall rates, making *Anonymise* a reliable solution for ensuring no sensitive information is left exposed. The tool can be accessed through a user-friendly web-based interface or API, offering flexibility for both individual and large-scale document processing needs. [1] By automating document de-identification with high accuracy and efficiency, *Anonymise* presents a reliable solution for ensuring compliance with EU privacy regulations while reducing the time and cost associated with manual anonymisation.

1 Introduction

Due to legislative changes (GDPR[2] in the EU, HIPAA[3] in the US), it is essential to remove sensitive and personally identifiable information (PII) from the text before sharing it publicly or using it to train AI solutions. Manual text de-identification is a laborious process, while being far from perfect. For example, Neamatullah et al. (2008) found that the recall of manual de-identification can range from 0.63 to 0.94, with a mean of 0.81).

Numerous studies explore named entity recognition (NER) tools to create a text de-identification workflow (Ribeiro et al., 2023; Arranz et al., 2022; Giorgi and Bader, 2019). However, most of the existing systems do not offer an end-to-end ability to de-identify documents and in many cases they are tailored for a single language (Catelli et al., 2020; Syed et al., 2022).

In this demonstration paper, we introduce the Anonymise tool, a system designed for semi-automated multilingual document de-identification. The tool allows users to perform Word or plain-text document de-identification using multiple PII replacement strategies. It offers a web interface, where the user may review the de-identified documents and edit the de-identified entities, and an API to de-identify a large number of documents.

2 Related Work

Personally identifiable information can be reduced to a set of named entity categories (Catelli et al., 2020). The text de-identification task then consists of labeling and removing text spans containing such categories.

Driven by the need to share court records with the public, the legal domain has received significant attention from researchers and practitioners. The Text Anonymization Benchmark (TAB) corpus (Pilán et al., 2022), based on court cases from the European Court of Human Rights, explicitly marks text spans that need to be masked to consider the document anonymised. In the MAPA project (Arranz et al., 2022), a multilingual dataset based on court documents from the Court of Justice of the European Union has been annotated using a

[1] The system is available online at https://tilde.ai/document-anonymisation/

[2] https://eur-lex.europa.eu/eli/reg/2016/679/oj

[3] https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html

Proceedings of Recent Advances in Natural Language Processing,pages 1327–1332
Varna, Sep 8–10, 2025
https://doi.org/10.26615/978-954-452-098-4-154

fine-grained named entity hierarchy. A multilingual parallel dataset, manually labeled with semantic categories useful for the removal of personally identifiable information, was also created from the documents of the Court of Justice of the European Union (Vīksna and Skadiņa, 2024).

Early systems for text de-identification used gazetteers in combination with regular expressions and fixed rules (Meystre et al., 2010), while more recent solutions focus on deep learning methods and are based on pre-trained transformer language models (PLM), such as BERT (Devlin et al., 2019) XLM-R (Conneau et al., 2020) or recently LLMs, such as ChatGPT/GPT-4 (Laskar et al., 2023; Liu et al., 2023). In practical applications, ensemble methods, combining rule-based, dictionary-based and PLM or LLM-based methods, are used to achieve best-in-class NER performance (Arranz et al., 2022; Murugadoss et al., 2024).

Once a set of entities for de-identification is identified, an obfuscation method should be chosen according to the needs of the users.

Text de-identification systems may include a linker component to produce global pseudonyms, consistent on a document level. The idea is to chain three processes: a named entity recognition, an entity linking, and a substitution engine (Francopoulo and Schaub, 2020).

Another approach, adversarial anonymisation (Staab et al., 2024), uses LLMs to iteratively paraphrase text until no PII can be inferred from it.

3 System Design

The Anonymise tool consists of four distinct components (see Figure 1):

- **Frontend (GUI)** is the part of the application facing the user and running in a web browser. The frontend allows the user to upload a document, select a pseudonym generation strategy, review results, and download the processed document.
- **Rest API** module provides access to the tool in a machine-readable way and serves as a backend for the frontend part of the tool. This module also converts document to and from the internal format used by the tool.
- **Named entity recognizer** (NER) module identifies PII spans in the document. For PII detection it uses a fine-tuned pre-trained language model together with a list of regular expressions.
- **Pseudonymisation module** replaces the entities detected by the NER with suitable pseudonyms.

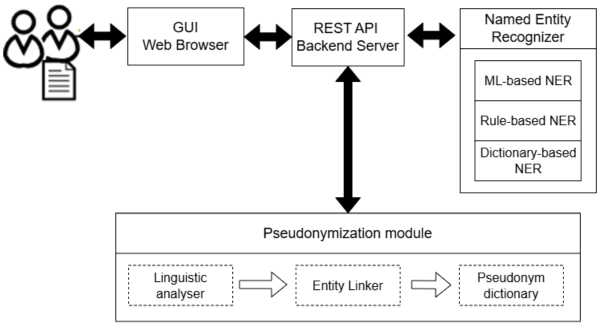


Figure 1: Main components of the Anonymise de-identification tool.

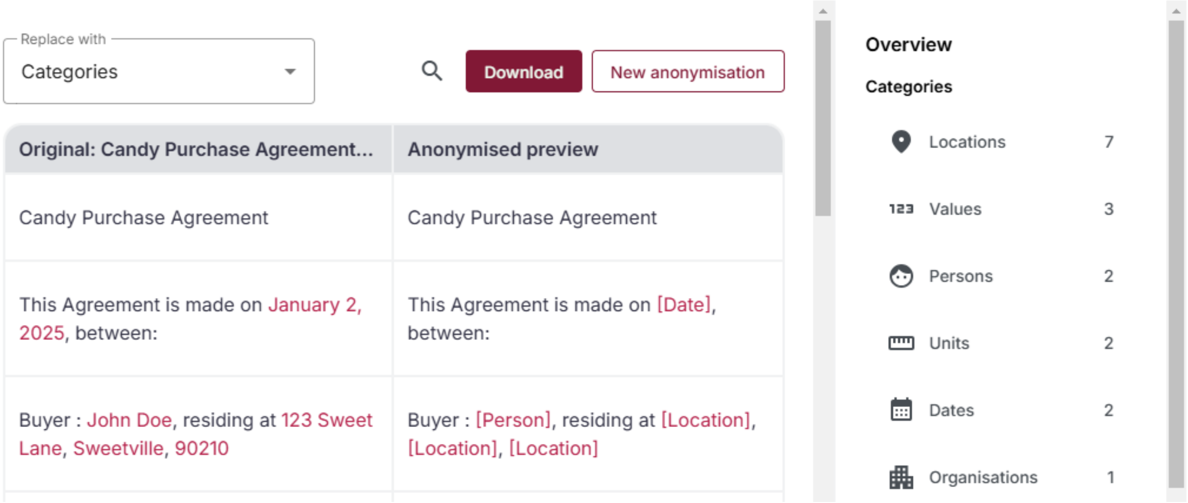
We implement the Anonymise tool as a Dockerized web application that allows the user to upload documents, process/anonymise, review result, and download the de-identified version of the documents.

3.1 User Interface

An example snapshot of the user interface is shown in Figure 2. The user interface consists of the document upload screen and the editor. The editor view contains two main sections – the overview section on the right side and an editor section consisting of two text columns on the left side (see Figure 2). The left column of the editor section shows the original text with the identified entities highlighted, and the right column shows the output of de-identification tool. Each mention of an identified entity can be deleted and its category or pseudonym can be changed. A new entity mention may also be added here, by selecting the corresponding text span and assigning NE category to it. The de-identified document is available for download, with the content shown in the ”Anonymised preview” column.

3.2 Named Entity Recognition

An ensemble NER consists of three components: a NER model, a rule-based module using regular expressions (RegEx) with context words, and a list module. The NER model is built using the Flair framework (Schweter and Akbik, 2020) by fine-tuning the multilingual XLM-R model (Conneau et al., 2020) on the MultiLeg dataset (Vīksna and Skadiņa, 2024). It supports all the entity types presented in the MultiLeg dataset: Person, ID number,



Original: Candy Purchase Agreement...	Anonymised preview
Candy Purchase Agreement	Candy Purchase Agreement
This Agreement is made on January 2, 2025, between:	This Agreement is made on [Date], between:
Buyer : John Doe, residing at 123 Sweet Lane, Sweetville, 90210	Buyer : [Person], residing at [Location], [Location], [Location]

Overview

Categories

Locations	7
Values	3
Persons	2
Units	2
Dates	2
Organisations	1

Figure 2: The user interface of the tool.

Location, Organization, URL, Date, Amounts, Nationality, and Profession. The RegEx module is implemented using Presidio (Mendels et al., 2018) with custom recognizers for Estonian, Latvian, and Lithuanian ID numbers and document codes. The List module contains a list of pre-defined strings which, if found in the text, are labeled as sensitive entities. This list is intended to be editable and extendable by the user.

3.3 Pseudonymisation

The tool supports three methods of obfuscation: masking PIIs with asterisks, suppression using a category name, and pseudonymisation using grammatically suitable substitutes. The masking method simply masks each entity mention with a preset number of asterisks, while suppression replaces each entity mention with its category name.

The pseudonymisation using grammatically suitable substitutes is challenging for morphologically rich languages, as the tool should provide the pseudonym in the correct form, e.g., inflection, gender, and number. The Stanza (Qi et al., 2020) library is used to determine the grammatical features of each token. Since Stanza provides a separate model for each language, the main language of the sentence is determined using fasttext (Joulin et al., 2016). Using a threshold of 0.33, a list of possible languages is obtained, and if the primary language of the document declared by the user is among them, it is used; otherwise, the language with the highest score is assumed to be the primary language of the sentence. For multi-word entity mentions, grammatical features of the headword are used. The pseudonyms are selected from a pre-defined list of candidate entities. The list of substitutes was obtained from the WMT monolingual news datasets (Kocmi et al., 2024) by applying the NER and Stanza parser. Extracted named entities are stored together with their grammatical information as potential pseudonym candidates.

4 Evaluation

The Anonymise tool has been evaluated using the MultiLeg test-set (Vīksna and Skadiņa, 2024) for four languages: English, Estonian, Latvian, and Lithuanian. For comparison, the same documents were also de-identified using a text de-identification system from the EC NLP Services[4] and the MAPA multilingual legal model[5]. For evaluation standard NER measures (precision (P), Recall (R), and F_1 score) were used. Results of the evaluation are summarized in Table 1.

Both MAPA and the EC-hosted tool show poor recall scores, as they miss most instances of some entity types (Codes, nationalities, amounts) in this dataset. The Code category is particularly important because it alone allows the identification of the document in question. Even if all other entities are de-identified, the attacker would still be able to identify the persons mentioned by the case number. Both tools perform considerably better on English data, compared to other languages. Conversely, the Anonymise tool performs equally in all 4 languages. Anonymise tool shows high recall, detecting over 90% of the sensitive entities, however, precision on this dataset is comparatively lower. Anonymise also detects and hides entities not labeled as sensitive in the test set, such as EC directive codes and

[4] https://language-tools.ec.europa.eu/NLPServices/NLP

[5] https://mapa-demo.pangeamt.com/

System	EN			ET			LT			LV			ALL		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
EC tool	0.86	0.32	0.46	0.63	0.28	0.39	0.74	0.28	0.40	0.70	0.32	0.44	0.73	0.29	0.42
MAPA	0.84	0.31	0.45	0.61	0.28	0.39	0.74	0.28	0.40	0.69	0.31	0.43	0.72	0.30	0.42
Anonymise	0.79	0.93	**0.86**	0.77	0.94	**0.85**	0.78	0.93	**0.85**	0.79	0.93	**0.85**	0.78	0.93	**0.85**

Table 1: Evaluation results on MultiLeg dataset.

European institutions. High recall in combination with lower precision means that part of the labeled entities are false positives (entities wrongly labeled for de-identification). For the de-identification task, high recall is more important than high precision because the primary goal is to ensure that no sensitive information is left unidentified and exposed. This means that the cost of a false negative is high (the sensitive data remains identifiable), while the inclusion of entity linking and the graphical editor allows the user to remove the false positives easily, making the cost of false positives low.

Extrinsic evaluation of the Anonymise tool was performed using the TAB benchmark dataset and evaluation script (Pilán et al., 2022). The evaluation results are shown in Table 2. The Anonymise tool evaluated against the TAB ECHR test set shows good recall on all identifiers and quasi-identifiers. It is also second to the Longformer trained on this dataset on the task of detecting direct identifiers. On this dataset, precision is low, as the tool identifies categories not included in the dataset for de-identification.

System	R_{di+qi}	ER_{di}	ER_{qi}	P_{di+qi}	WP_{di+qi}
Presidio	0.782	0.463	0.802	0.542	0.609
TAB	0.919	**1.000**	0.916	**0.836**	**0.850**
EC tool	0.796	0.500	0.781	0.726	0.680
MAPA	0.809	0.500	0.795	0.709	0.654
Anonymise	**0.950**	0.827	**0.940**	0.551	0.456

Table 2: Evaluation results of the tools on TAB test set. R_{di+qi} - Token-level recall; ER_{di} - Entity-level recall on direct identifiers; ER_{qi} - Entity-level recall on quasi-identifiers; P_{di+qi} - Weighted, token-level precision; WP_{di+qi} - Weighted, mention-level precision.

Runtime/resource metrics The tool consists of several modules, the most resource-intensive of which are the Flair-based NER module, the Linguistic analyser, and the backend that handles file conversion. The demo version of the tool, set up on a Kubernetes cluster using a 4-core CPU with 16 GB of RAM, can process 100 files containing 4917 segments of text in 22 minutes.

5 Conclusions

In this paper, we presented the Anonymise tool, a Dockerized web application that allows semi-automated document de-identification.

Anonymise can different types of documents, present the content marked with various PII categories to the user, suggest and apply a range of pseudonyms, and save the de-identified text into source document format.

The tool is able to detect 12 broad entity types covering both direct and indirect identifiers. Evaluation shows that tool achieves a high recall of over 0.93 and F_1 score of 0.85 on MultiLeg dataset and token-level recall R_{di+qi} of 0.95 on TAB test set. Anonymise can be adapted to other languages or domains by using appropriate NER models and re-building the pseudonym dictionaries. The tool includes a configuration page, where the user may (un)select entity types for de-identification. A possible improvement may be the inclusion of confidence thresholds, allowing the user to tune the recall vs. precision balance. The tool can be used through API to de-identify a large number of documents, or through web interface, where the user can review the de-identified documents and edit entities.

Acknowledgments

This research has been supported by the ICT Competence Centre (www.itkc.lv) within the project 2.5 *Daudzvalodīgs dokumentu anonimizācijas un pseidonimizācijas mākslīgā intelekta risinājums* (Multilingual artificial intelligence solution for document anonymization and pseudonymization) of EU Structural funds, ID no 5.1.1.2.i.0/1/22/A/CFLA/008. This work was supported by the ”Language Technology Initiative” project (No.2.3.1.1.i.0/1/22/I/CFLA/002), funded by the European Union Recovery and Resilience Mechanism Investment and the National Development Plan.

References

Victoria Arranz, Khalid Choukri, Montse Cuadros, Aitor García Pablos, Lucie Gianola, Cyril Grouin, Manuel Herranz, Patrick Paroubek, and Pierre Zweigenbaum. 2022. MAPA project: Ready-to-go open-source datasets and deep learning technology to remove identifying information from text documents. 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 64–72, Marseille, France. European Language Resources Association.

Rosario Catelli, Francesco Gargiulo, Valentina Casola, Giuseppe De Pietro, Hamido Fujita, and Massimo Esposito. 2020. Crosslingual named entity recognition for clinical de-identification applied to a covid-19 italian data set. *Applied Soft Computing*, 97:106779.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

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, Stroudsburg, PA, USA. Association for Computational Linguistics.

Gil Francopoulo and Léon-Paul Schaub. 2020. Anonymization for the GDPR in the Context of Citizen and Customer Relationship Management and NLP. In *workshop on Legal and Ethical Issues (Legal2020)*, pages 9–14, Marseille, France. LREC2020, ELRA.

John M Giorgi and Gary D Bader. 2019. Towards reliable named entity recognition in the biomedical domain. *Bioinformatics*, 36(1):280–286.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, Mariya Shmatova, Steinthór Steingrímsson, and Vilém Zouhar. 2024. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46, Miami, Florida, USA. Association for Computational Linguistics.

Md Tahmid Rahman Laskar, M Saiful Bari, Mizanur Rahman, Md Amran Hossen Bhuiyan, Shafiq Joty, and Jimmy Huang. 2023. A systematic study and comprehensive evaluation of ChatGPT on benchmark datasets. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 431–469, Toronto, Canada. Association for Computational Linguistics.

Zhengliang Liu, Yue Huang, Xiaowei Yu, Lu Zhang, Zihao Wu, Chao Cao, Haixing Dai, Lin Zhao, Yiwei Li, Peng Shu, Fang Zeng, Lichao Sun, 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.

Omri Mendels, Coby Peled, Nava Vaisman Levy, Sharon Hart, Tomer Rosenthal, Limor Lahiani, et al. 2018. Microsoft Presidio: Context aware, pluggable and customizable pii anonymization service for text and images.

Stephane Meystre, F Friedlin, Brett South, Shuying Shen, and Matthew Samore. 2010. Automatic de-identification of textual documents in the electronic health record: A review of recent research. *BMC medical research methodology*, 10:70.

Karthik Murugadoss, Saivikas Killamsetty, Deeksha Doddahonnaiah, Nakul Iyer, Michael Pencina, Jeffrey Ferranti, John Halamka, Bradley A. Malin, and Sankar Ardhanari. 2024. Scaling text de-identification using locally augmented ensembles. *medRxiv*.

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.

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.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Bruno Ribeiro, Vitor Rolla, and Ricardo Santos. 2023. INCOGNITUS: A toolbox for automated clinical notes anonymization. In *Proceedings of the 17th*

Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, pages 187–194, Dubrovnik, Croatia. Association for Computational Linguistics.

Stefan Schweter and Alan Akbik. 2020. FLERT: Document-level features for named entity recognition.

Robin Staab, Mark Vero, Mislav Balunović, and Martin Vechev. 2024. Large language models are advanced anonymizers.

Mahanazuddin Syed, Kevin Sexton, Melody Greer, Shorabuddin Syed, Joseph VanScoy, Farhan Kawsar, Erica Olson, Karan Patel, Jake Erwin, Sudeepa Bhattacharyya, Meredith Zozus, and Fred Prior. 2022. Deidner model: A neural network named entity recognition model for use in the de-identification of clinical notes. In *Proceedings of the 15th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2022) - HEALTHINF*, pages 640–647. INSTICC, SciTePress.

Rinalds Vīksna and Inguna Skadiņa. 2024. Multi-Leg: Dataset for text sanitisation in less-resourced languages. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11776–11782, Torino, Italia. ELRA and ICCL.