MultiLeg: Dataset for Text Sanitisation in Less-resourced Languages

Rinalds Vīksna[1,2], Inguna Skadiņa[1,2], Roberts Rozis[1]
[1] Tilde, Vienības gatve 75a, Riga, Latvia
[2] Faculty of Computing, University of Latvia, Raiņa bulv. 29, Riga, Latvia
{Firstname.Lastname}@tilde.lv

Abstract

Text sanitization is the task of detecting and removing personal information from the text. While it has been well-studied in monolingual settings, today, there is also a need for multilingual text sanitization. In this paper, we introduce MultiLeg: a parallel, multilingual named entity (NE) dataset consisting of documents from the Court of Justice of the European Union annotated with semantic categories suitable for text sanitization. The dataset is available in 8 languages, and it contains 3082 parallel text segments for each language. We also show that the pseudonymized dataset remains useful for downstream tasks.

Keywords: text sanitization, legal domain, multilingual, named entities

1. Introduction

The need for textual data is ever-increasing in machine learning. However, a large part of textual data contains private information and thus cannot be safely used. Several NLP techniques such as text anonymization, text de-identification, and pseudonymization address this problem. While text anonymization, fully compliant with GDPR[1], is a very difficult goal to achieve (Weitzenboeck et al., 2022), text sanitization aims to lessen the risk of disclosing personally identifying information (PII) while keeping the text useful for the downstream task.

In this paper we use text sanitization as a broad term, describing a process, that transforms documents through edit operations such as hiding particular text spans or replacing them with different values (Papadopoulou et al., 2022) in order to protect PII. The standard approach to text sanitization starts with a named entity recognition and classification (NERC) to obtain a list of text spans that may need to be obfuscated (Papadopoulou et al., 2022), followed by the decision which text spans to transform and how.

Named entity recognition and classification task aims to detect named entities and to classify these entities into appropriate categories. This task was introduced in Message Understanding Conference-6 (Merchant et al., 1996) aiming at the recognition of time (and date), numeric (percentages and money), and named (people, organizations, locations) entities.

Systems used for text de-identification use a broader set of categories describing various types of PII, such as contact information, ID numbers, ethnicity, profession, age, sex, workplace, family status and relations, and others. Most of existing studies on text sanitization focus on English or Spanish languages (Juez-Hernandez et al., 2023), with little work done on less-resourced languages, such as Polish (Oleksy et al., 2021), Estonian (TEXTA, 2022), or Latvian (Skadina et al., 2022). Moreover, available multilingual de-identification NE data sets are small, e.g., the MAPA dataset contains only 12 annotated documents (Arranz et al., 2022).

In this paper, we present a multilingual, parallel dataset manually labeled with semantic categories useful for the removal of personally identifying information. The dataset[2] consists of 60 documents collected from the Court of Justice of the European Union in 8 languages (English, Danish, Estonian, Finnish, Lithuanian, Latvian, Polish, and Swedish), while experiments presented in this paper are performed for 6 languages (Estonian and Finnish excluded due to late availability).

2. Related Work

Text sanitization approaches can be divided into two large groups - (1) a NER-based approach, where from a set of identified entities some or all are selected for sanitization and (2) a privacy-preserving data publishing (PPDP) approach which operates with an explicit account of disclosure risk and anonymizes documents by enforcing a privacy model (Pilán et al., 2022). The PPDP approach is commonly used for database record anonymization (Domingo-Ferrer et al., 2016), but when applied to

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

[2] https://github.com/tilde-nlp/MultiLeg-dataset

LREC-COLING 2024, pages 11776–11782

text anonymization, has a number of limitations and scalability issues (Lison et al., 2021).

The NER-based approach has been widely used to detect and sanitize Protected Health Information (PHI) in clinical texts (Meystre et al., 2010), (Ribeiro et al., 2023) motivated by the US HIPAA[3] act, and in the legal domain (Arranz et al., 2022), (Oksanen et al., 2022), motivated by the need to share court records with public.

Early systems for text sanitization used NER datasets in combination with regular expressions, fixed rules, and gazetteers (Meystre et al., 2010), while in the recent solutions, the focus has shifted to machine learning and in particular deep learning methods using transformer models, such as BERT (Devlin et al., 2019) XLM-R (Conneau et al., 2020) or large language models, such as ChatGPT/GPT-4 (Laskar et al., 2023; Liu et al., 2023).

Several de-identification datasets for the medical domain are available in English (Stubbs and Uzuner, 2015) or Spanish (Marimon et al., 2019). The datasets are annotated using entity categories derived from HIPAA guidelines, which include direct identifiers, such as person names, IDs, and contact information, and indirect identifiers such as date, location, profession, and age. The entities in the i2b2/UTHealth(Stubbs and Uzuner, 2015) corpus are annotated in a fine-grained manner, to enable easier replacement with appropriate pseudonyms, or to allow users to select a subset of entities to hide. For example, the distinction between DOCTOR and PATIENT allows to de-identify patient names, while keeping doctor names.

One of the first datasets produced for text de-identification is the ITAC corpus (Medlock, 2006). It consists of 2500 personal email messages in English, annotated with predefined entity classes (Person, Location, Organizations, Addresses, Titles, ID codes, ethnic terms, reference codes, usernames, and passwords). Due to the sensitive nature of this data, it is pseudonymized using hybrid semi-supervised and manual processes to replace entities without changing their nature. JobStack corpus (Jensen et al., 2021) is based on Job postings in the English language from StackOverflow, annotated with Name, Location, Organization, Contact, and Profession categories. TAB benchmark (Pilán et al., 2022) dataset Pilán et al. (2022) consists of ECHR court cases in English annotated with Person, Code, Locations, Organizations, Demographic, Datetime, Quantity, and Misc entity types. Additionally, a set of confidential attributes (Belief, Politics, Sex, Ethnic, and Health) is annotated. The confidential attributes are typically unknown to an external attacker and, thus usually not seen as quasi-identifiers, however in the TAB benchmark they are annotated to prevent sensitive attribute disclosure. Each entity mention contains its offset in the text, semantic type, identifier type (direct, quasi, or no need to mask), and a unique identifier for the entity it refers to. This approach allows evaluation of anonymization performance with respect to a single person whose attributes should be anonymized. The MAPA project provides datasets intended for training a multilingual de-identification system for health and legal domains. The legal part of the dataset, collected from eur-lex (Arranz et al., 2022), is available in 24 languages. It consists of over 2000 sentences manually annotated using fine-grained 3-level annotation schema. The medical part of the dataset consists of 485 clinical cases in French translated into other 23 languages using machine translation and restoring annotation tags.

3. Creation of Parallel Dataset

We selected a diverse set of 60 documents (Judgements, Applications, Requests for a preliminary ruling, Opinions, and Orders) from the Court of Justice of the European Union. The documents, available in 24 EU languages, were selected from the years 2019-2022. We selected a subset of documents in 8 languages (Danish, English, Estonian, Finnish, Lithuanian, Latvian, Polish, and Swedish) for further processing.

At first, selected documents were converted into plain text format and segmented into sentences. Then, segmented documents were aligned to obtain parallel documents. In rare cases, when the content was missing due to the inaccurate translation, blank lines were inserted. The total segment count for this dataset is 3082 segments in each language, with English having extra segments at the end of some documents (in the English version signatures are kept, while in other language versions term "signatures" is used). Finally, we split the dataset into training and evaluation sets of 50 (2456 segments) and 10 (626 segments) documents.

3.1. Annotation Schema

HIPAA[4] provides a list of identifiers that need to be anonymized for a document to be considered anonymized in the U.S. This approach is closer to the de-identification paradigm.

GDPR[5] defines personal data as any information relating to an identified or identifiable natural per-

[3] https://www.hhs.gov/hipaa/index.html

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

[5] https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679

son. Thus any data that may be used to identify a natural person, such as direct identifiers or indirect identifiers, needs to be pseudonymized in order to consider the document as anonymized.

In our work, we annotate all identifiers specified by HIPAA and add semantic categories that could be used to indirectly identify a person:

- **PER** - A person's name, surname, initial, nickname, alias, username
- **IDNUM** - Various ID numbers and codes, such as passport numbers, driver's licenses, vehicle license numbers, fax, telephone numbers, and similar unique identifiers

and indirect identifiers:

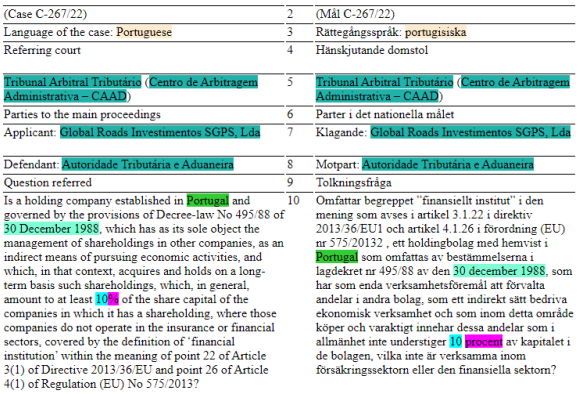
- **LOC** - Address strings, countries, regions, cities, streets, famous buildings or other places
- **ORG** - Specific (named) organization: a company, institution, or association of two or more people having a particular purpose
- **URL** - An URL address, or a hostname
- **DATE** - A day, month, or a year
- **AMOUNT_VALUE** - Usually used together with AMOUNT_UNIT to indicate some numeric attribute, such as age, height, percentages, or money
- **AMOUNT_UNIT** - A determinate quantity (as of length, time, heat, or value) adopted as a standard of measurement
- **NATIONALITY** - Nationality, ethnicity, language, or citizenship of a person
- **PROFESSION** - Job titles or ranks of a person
- **TITLE** - Named art, creative works, events

3.2. Annotation Process

The annotation process consists of several steps. At first, the English part of the dataset is pre-annotated using a 4-class English NER[6] that ships with the Flair library (Schweter and Akbik, 2020). Then, this pre-annotated dataset is annotated by two experts using the Docanno tool (Nakayama et al., 2018). Annotations produced by two human experts are compared and any disagreement is resolved according to the guidelines using the 6-step method by Oortwijn et al. (2021).

The annotated English dataset is used to train a multilingual NER model by using the Flair library to fine-tune the multilingual XLM-RoBERTa model. The multilingual XLM-RoBERTa model fine-tuned on English data is able to perform NE tagging in other languages via transfer learning. This NER model is then used to pre-annotate datasets for other languages.

Since datasets in all languages are parallel on a segment level, we created a script to allow quick review of the datasets. The script reads two or more parallel texts in different languages (e.g., English and Finnish) or versions (e.g., before and after pseudonymization) and converts them into HTML with appropriate markup (see Figure 1).



(Case C-267/22)	2	(Mål C-267/22)
Language of the case: Portuguese	3	Rättegångsspråk: portugisiska
Referring court	4	Hänskjutande domstol
Tribunal Arbitral Tributário (Centro de Arbitragem Administrativa – CAAD)	5	Tribunal Arbitral Tributário (Centro de Arbitragem Administrativa – CAAD)
Parties to the main proceedings	6	Parter i det nationella målet
Applicant: Global Roads Investimentos SGPS, Lda	7	Klagande: Global Roads Investimentos SGPS, Lda
Defendant: Autoridade Tributária e Aduaneira	8	Motpart: Autoridade Tributária e Aduaneira
Question referred	9	Tolkningsfråga
Is a holding company established in Portugal and governed by the provisions of Decree-law No 495/88 of 30 December 1988, which has as its sole object the management of shareholdings in other companies, as an indirect means of pursuing economic activities, and which, in that context, acquires and holds on a long-term basis such shareholdings, which, in general, amount to at least 10% of the share capital of the companies in which it has a shareholding, where those companies do not operate in the insurance or financial sectors, covered by the definition of 'financial institution' within the meaning of point 22 of Article 3(1) of Directive 2013/36/EU and point 26 of Article 4(1) of Regulation (EU) No 575/2013?	10	Omfattar begreppet ''finansiellt institut'' i den mening som avses i artikel 3.1.22 i direktiv 2013/36/EU1 och artikel 4.1.26 i förordning (EU) nr 575/20132 , ett holdingbolag med hemvist i Portugal som omfattas av bestämmelserna i lagdekret nr 495/88 av den 30 december 1988, som har som enda verksamhetsföremål att förvalta andelar i andra bolag, som ett indirekt sätt bedriva ekonomisk verksamhet och som inom detta område köper och varaktigt innehar dessa andelar som i allmänhet inte understiger 10 procent av kapitalet i de bolagen, vilka inte är verksamma inom försäkringssektorn eller den finansiella sektorn?

Figure 1: Reference HTML screenshot.

Reference HTML files are provided to annotators working on the rest of the language versions for reference.

For the non-English datasets, one annotator (of native or professional proficiency) performed the annotation process using the provided instructions and reference HTML that displayed the already annotated documents in English. The task of the annotators for non-English languages was to label the text according to the annotation schema, using English data as examples.

After annotation of a dataset, a new version of reference HTML files is produced to review annotations and ensure consistency. Any disagreement between different language versions is reviewed according to the annotation guidelines.

3.3. De-identification

After the annotation step, we produce a de-identified version of the dataset by replacing PER entities with appropriate substitutes.

For substitution, the following procedure is applied. First, from the entire corpus, all PER spans are extracted, deduplicated, and clustered by (sur)name(s). Then, for each cluster replacements are selected manually in an attempt to preserve the gender and nationality of the PER entity. The name and surname pseudonyms for replacement were selected in English from Wikipedia, taking the

[6] https://huggingface.co/flair/ner-english-large

	da	en	lt	lv	pl	sv
Unit	209	213	211	211	212	207
Value	210	213	211	212	212	211
Date	1168	1163	1174	1159	1160	1142
ID	17	18	18	18	18	18
LOC	834	854	817	819	851	650
NAT	73	74	98	75	74	73
ORG	3002	2966	2878	2886	2906	3072
PER	565	607	578	580	578	581
PROF	429	430	384	467	474	472
TITLE	43	43	43	43	43	42
URL	5	5	5	5	5	5
Total	6555	6586	6417	6475	6533	6473

Table 1: Number of entities per language and entity type.

most popular names from European regions. Finally, PER entity tokens were manually inflected by annotators to match the inflection of the original form if it differs from the lemma.

Although we applied de-identification only to PER entities [7], a similar procedure could be applied to other entity types as well.

3.4. Statistics of Found Entities

Table 1 provides statistics about identified entities per entity type and language. We can see that translations are not completely parallel, since the number of entities differs between languages. Entities of type ORG and PER may be substituted with pronouns if they reappear. The translation may also skip some mentions of entities if it is clear from the context what is being discussed. Lithuanian versions of documents typically contain more repeated mentions of nationality entities in the scope of a single paragraph, but less repeated profession mentions. Swedish documents contain comparatively fewer LOC entities. In Swedish, LOC entities, which in other languages would be separated, could be part of compound words, which are not labeled, e.g., "unionsmarknaden" - "the Union market". In general, the number of entities in different languages does not differ significantly - the largest number of annotated entities is in English (6586), and the smallest in Lithuanian (6417) - a difference of 2.5%.

[7]The requirement to de-identify person names was given by the Data Protection Officer. No further de-identification was applied to the dataset, as the publications related to judicial proceedings before the Court of Justice are open data and necessary anonymization is already applied by the Court of Justice.

Model\Data	original	de-identified
original	91.39 ± 0.25	91.55 ± 0.18
de-identified	91.16 ± 0.11	91.23 ± 0.16

Table 2: F_1 score and standard deviation for NER models trained on original and de-identified data, evaluated on original and de-identified test sets.

NER	en	lt	lv	pl	sv	da	multi
en	**84**	62	61	64	74	77	70
lt	41	**86**	82	62	58	65	65
lv	47	79	**89**	59	60	64	65
pl	44	69	72	**85**	60	64	65
sv	55	58	61	66	**85**	83	67
da	49	56	61	63	82	**86**	65
multi	**91**	**93**	**92**	**93**	**90**	**92**	**92**

Table 3: Evaluation of trained NER models. Rows: models trained on respective train sets, columns: the result of the evaluation (F_1 scores) on the test set in a given language. multi is the concatenated data set of monolingual data.

4. Evaluation

In order to study the impact of the de-identification strategy, we trained two NER models: using the original dataset and using de-identified data set. We evaluate them on the original and de-identified test set. NER models are trained using the FLAIR toolkit and XLM-R model. Training is done for 20 epochs using batch size 8. The training and evaluation are done 3 times and the average and standard deviations are calculated. Evaluation results are presented in Table 2. The observed NER performance drop, when using an NER model trained on the de-identified dataset, is less than the standard deviation. All further work is done using the pseudonymized dataset.

6 monolingual NER models are trained using the respective de-identified dataset, and one multilingual model is trained using the combined dataset. All models are evaluated on monolingual test sets, and on combined test set (multilingual). The process is repeated 4 times and the averaged results are shown in Table 3. Monolingual models show good performance only when evaluated on a test set of the same language as training data. However, the performance of the multilingual model surpasses any of the monolingual models on the respective test sets, achieving a F_1 score of 92.

We evaluate previously trained best English and multilingual NER models on the TAB (Pilán et al., 2022) benchmark dataset using TAB evaluation script.[8] The evaluation results together with the results of the TAB Longformer (with a window size

[8]https://github.com/NorskRegnesentral/text-anonymization-benchmark

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**
en	0.916	0.506	0.894	0.479	0.458
multi	**0.930**	0.508	**0.933**	0.479	0.448

Table 4: Evaluation results for English and multilingual models on TAB test set compared to TAB baseline and Longformer model. R_{di+qi} - Token-level recall on all identifiers; ER_{di} - Entity-level recall on direct identifiers; ER_{qi} - Entity-level recall on quasi-identifiers; P_{di+qi} - Weighted, token-level precision on all identifiers; WP_{di+qi} - Weighted, mention-level precision on all identifiers.

of 4,096 and a label weight of (10,1)) model and Presidio (+ORG) are shown in Table 4. The multilingual NER model evaluated against the TAB ECHR test set shows good recall on all identifiers and quasi-identifiers (R_{di+qi} and ER_{qi}), while recall of direct identifiers (CODE) is poor. The TAB performs sanitization with respect to a single selected person. The MultiLeg NER-based approach performs sanitization to all persons mentioned in the document. This leads to poor Mention-level precision when evaluated on the TAB ECHR test set (WP_{di+qi}). We prioritized recall over precision to remove as much sensitive information as possible. Thus, precision scores are comparable to Presidio, while recall scores outperform both Presidio and TAB.

5. Conclusions

We have presented MultiLeg: a multilingual, manually annotated NE dataset tailored for text sanitization use. The dataset consists of publicly available documents, and we have pseudonymized person names to comply with personal data protection requirements. We show that the pseudonymized dataset remains useful for downstream tasks.

The dataset is released in 8 languages (English, Danish, Estonian, Finnish, Lithuanian, Latvian, Polish, and Swedish), while experiments presented in this paper are preformed for 6 languages (Estonian and Finnish excluded due to late availability). The dataset contains 3082 parallel text segments per language. Compared to the only other available multilingual dataset for text sanitization found in the literature (MAPA dataset), this dataset is larger, aligned on segment level, and annotated with single-level annotations.

While no system may claim complete compatibility with GDPR, a system trained on this data would help perform the sanitization by highlighting PII and other relevant entities. The multilinguality of this dataset allows training NER systems that could process text in multiple languages using a single model, or train models for less-resourced languages such as the Baltic languages.

Limitations

Although our dataset is larger than available multilingual datasets, it is considerably smaller (per-language) than monolingual NER datasets. Since hyper-parameter tuning on training large models is computationally very costly, mostly default parameters were used in our experiments.

Ethics Statement

Our work fully complies with the ACL Code of Ethics[9]. We use only publicly available datasets and relatively low compute amounts while conducting our experiments to enable reproducibility. All human data annotators were fairly compensated in accordance with market rates.

Acknowlegements

The work described in this paper is performed in the H2020 project STARLIGHT ("Sustainable Autonomy and Resilience for LEAs using AI against High priority Threats"). This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 101021797.

This study has been supported by the EU Recovery and Resilience Facility project "Language Technology Initiative" (No 2.3.1.1.i.0/1/22/I/CFLA/002).

6. Bibliographical 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*

[9] https://www.aclweb.org/portal/content/acl-code-ethics

Conference, pages 64–72, Marseille, France. European Language Resources Association.

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.

Josep Domingo-Ferrer, David Sánchez, and Jordi Soria-Comas. 2016. *Database Anonymization: Privacy Models, Data Utility, and Microaggregation-based Inter-model Connections*, volume 8. Springer Cham.

Kristian Nørgaard Jensen, Mike Zhang, and Barbara Plank. 2021. De-identification of privacy-related entities in job postings. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics*, United States. Association for Computational Linguistics. NoDaLiDa 2021 ; Conference date: 31-05-2021.

Rodrigo Juez-Hernandez, Lara Quijano-Sánchez, Federico Liberatore, and Jesús Gómez. 2023. Agora: An intelligent system for the anonymization, information extraction and automatic mapping of sensitive documents. *Applied Soft Computing*, 145:110540.

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.

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.

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.

Montserrat Marimon, Aitor Gonzalez-Agirre, Ander Intxaurrondo, Heidy Rodriguez, Jose Lopez Martin, Marta Villegas, and Martin Krallinger. 2019. Automatic de-identification of medical texts in spanish: the meddocan track, corpus, guidelines, methods and evaluation of results. In *IberLEF@ SEPLN*, pages 618–638.

Ben Medlock. 2006. An introduction to NLP-based textual anonymisation. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Roberta Merchant, Mary Ellen Okurowski, and Nancy Chinchor. 1996. The multilingual entity task (MET) overview. In *TIPSTER TEXT PROGRAM PHASE II: Proceedings of a Workshop held at Vienna, Virginia, May 6-8, 1996*, pages 445–447, Vienna, Virginia, USA. Association for Computational Linguistics.

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.

Hiroki Nakayama, Takahiro Kubo, Junya Kamura, Yasufumi Taniguchi, and Xu Liang. 2018. doccano: Text annotation tool for human. Software available from https://github.com/doccano/doccano.

Arttu Oksanen, Eero Hyvönen, Minna Tamper, Jouni Tuominen, Henna Ylimaa, Katja Löytynoja, Matti Kokkonen, and Aki Hietanen. 2022. An Anonymization Tool for Open Data Publication of Legal Documents.

Marcin Oleksy, Norbert Ropiak, and Tomasz Walkowiak. 2021. Automated anonymization of text documents in polish. *Procedia Computer Science*, 192:1323–1333. Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 25th International Conference KES2021.

Yvette Oortwijn, Thijs Ossenkoppele, and Arianna Betti. 2021. Interrater disagreement resolution: A systematic procedure to reach consensus in annotation tasks. In *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 131–141, Online. Association for Computational Linguistics.

Anthi Papadopoulou, Yunhao Yu, Pierre Lison, and Lilja Øvrelid. 2022. Neural text sanitization with explicit measures of privacy risk. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 217–229, Online only. Association for Computational Linguistics.

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.

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.

Inguna Skadina, Baiba Saulite, Ilze Auzina, Normunds Gruzitis, Andrejs Vasiljevs, Raivis Skadins, and Mārcis Pinnis. 2022. Latvian language in the digital age: The main achievements in the last decade. *Baltic Journal of Modern Computing*, 10(3):490–503.

Amber Stubbs and Ozlem Uzuner. 2015. Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/uthealth corpus. *Journal of biomedical informatics*, 58S.

TEXTA. 2022. Anonymizing identifying information in court cases / case study.

Emily M Weitzenboeck, Pierre Lison, Malgorzata Cyndecka, and Malcolm Langford. 2022. The GDPR and unstructured data: is anonymization possible? *International Data Privacy Law*, 12(3):184–206.

7. Language Resource References

Ildikó Pilán and Pierre Lison and Lilja Øvrelid and Anthi Papadopoulou and David Sánchez and Montserrat Batet. 2022. *The Text Anonymization Benchmark (TAB): A Dedicated Corpus and Evaluation Framework for Text Anonymization*. Published on github.