MultiLeg: Dataset for Text Sanitisation in Less-resourced Languages

Rinalds Vīksna, Inguna Skadiņa, Roberts Rozis

MOTIVATION

- Text sanitization aims to lessen the risk of disclosing personally identifying information (PII) while keeping the text useful for the downstream task.
- Text sanitization describes a process, that transforms documents through edit operations such as hiding specific text spans or replacing them with different values.
- A common 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, followed by the decision, which text spans to transform and how.

Original	**Jon Snow was attacked by an eagle**
NERC	**Jon Snow PERSON was attacked by an eagle**
Obfuscated	**Person 1 was attacked by an eagle**

MOTIVATION

- Named entity recognition and classification task aims to detect named entities and to classify these entities into appropriate categories.
- 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, with little work done on less-resourced languages, such as Polish, Estonian, or Latvian.

- In this work, we present a multilingual, parallel dataset manually labeled with semantic categories useful for the removal of personally identifying information.

The Dataset

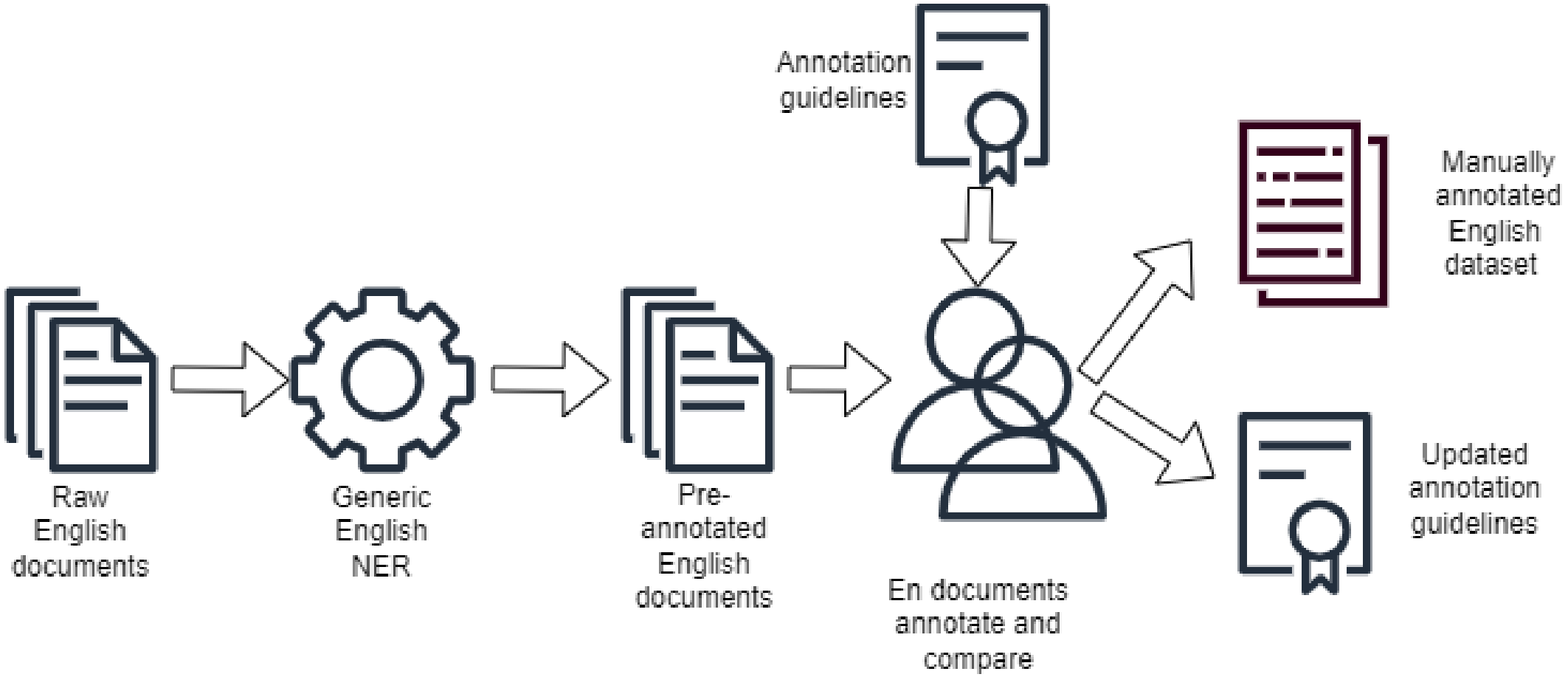
- A diverse set of 60 documents (Judgements, Applications, Requests for a preliminary ruling, Opinions, and Orders) from the Court of Justice of the European Union.
- The documents, available in 24 EU languages, were selected from the years 2019-2022. We selected a subset of documents in 8 languages (Danish, English, Estonian, Finnish, Lithuanian, Latvian, Polish, and Swedish) for further processing.
- Selected documents were converted into plain text format and segmented into sentences.
- Documents are aligned on sentence level.
- Finally, we split the dataset into training and evaluation sets of 50 (2456 segments) and 10 (626 segments) documents.

ANNOTATION

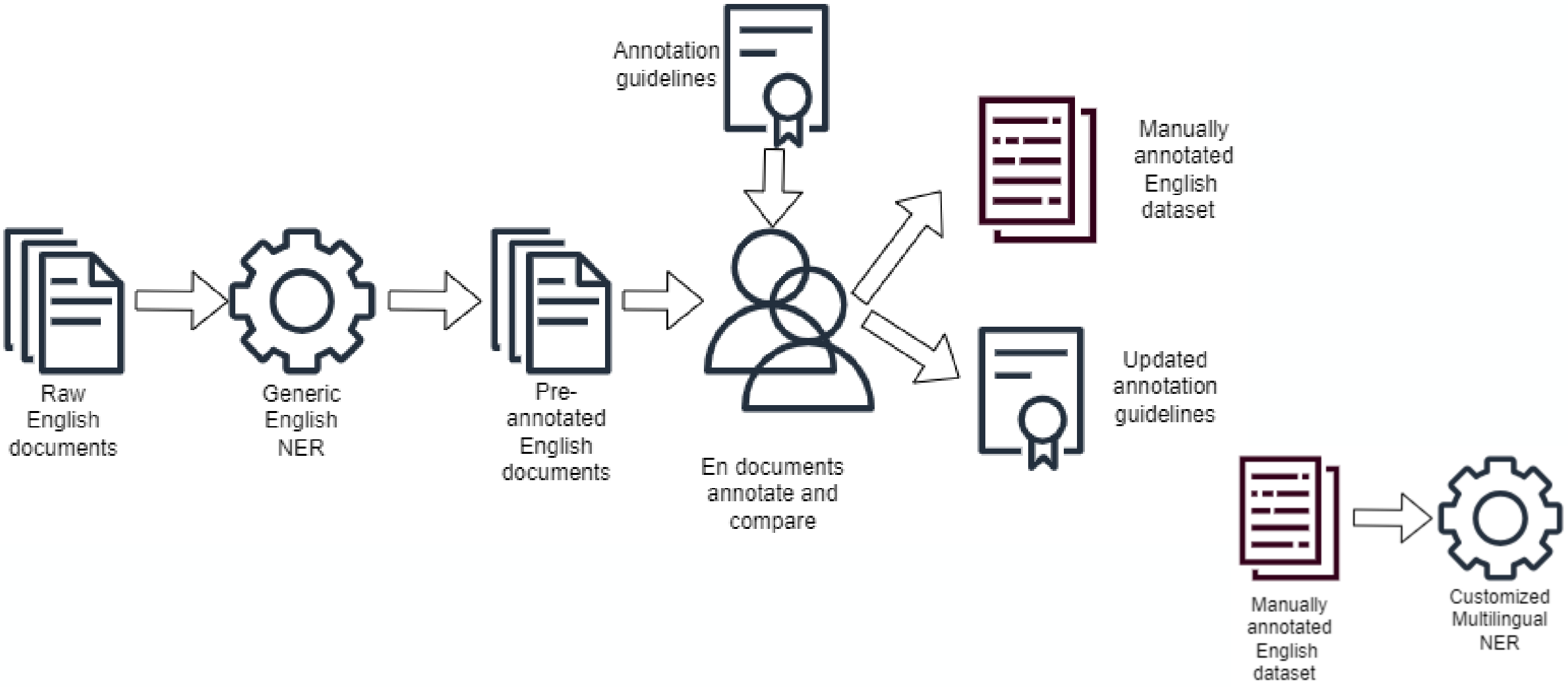
In our work, we annotate all identifiers specified by HIPAA and add semantic categories that could be used to indirectly identify a person:

- Person (name, surname, nicknames, initials, usernames, aliases)
- IDs (driver, passport, vehicle, telephone and similar codes)
- Locations (address, famous locations or buildings)
- Organizations
- URLs
- Dates
- Amounts (describing attributes or properties of a person or something related to a person)
- Nationalities
- Professions
- Titles (of named art, creative works or events)

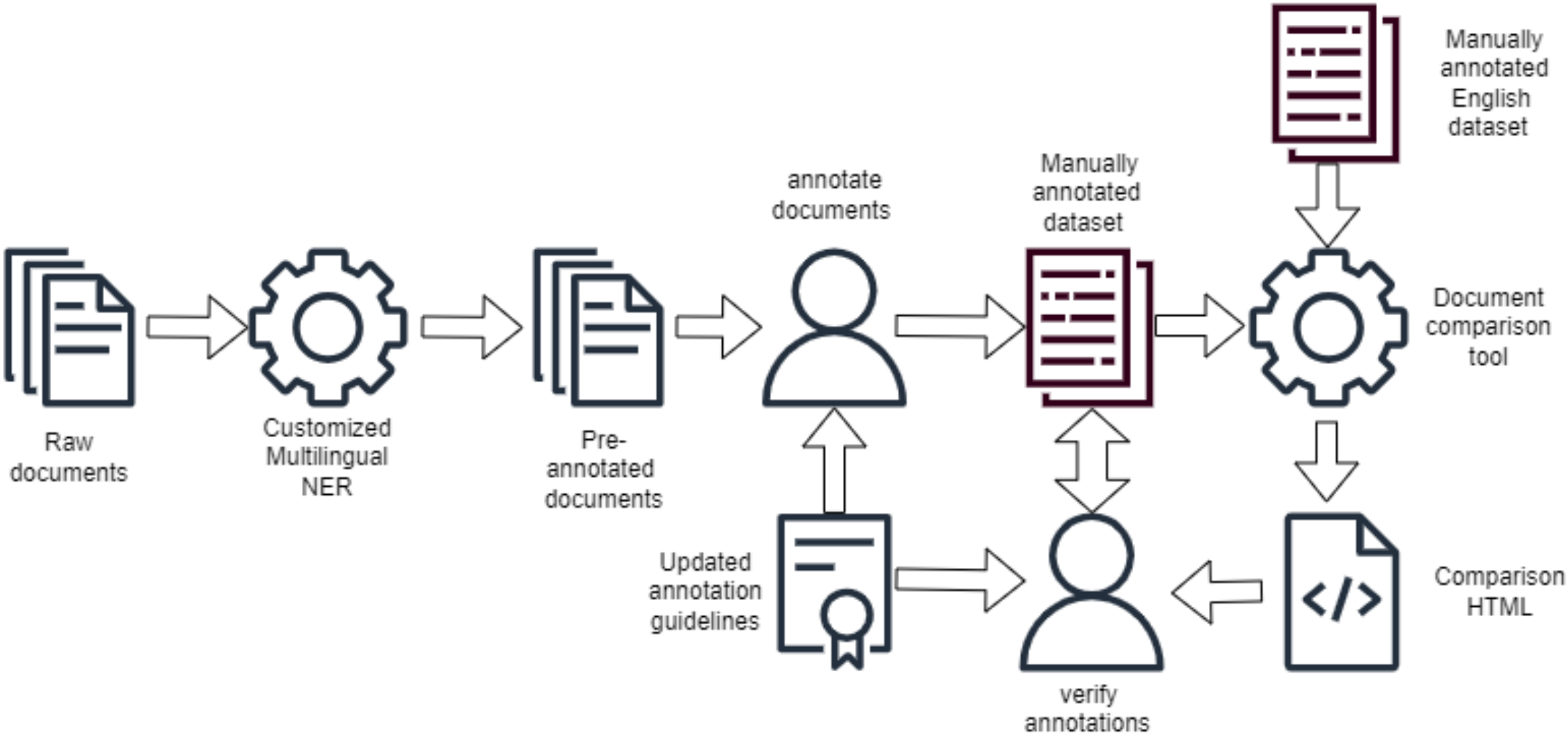
ANNOTATION


Annotation guidelines
Raw English documents
Generic English NER
Pre-annotated English documents
En documents annotate and compare
Manually annotated English dataset
Updated annotation guidelines

ANNOTATION


Annotation guidelines
Raw English documents
Generic English NER
Pre-annotated English documents
En documents annotate and compare
Manually annotated English dataset
Updated annotation guidelines
Manually annotated English dataset
Customized Multilingual NER

ANNOTATION


Raw documents
Customized Multilingual NER
Pre-annotated documents
annotate documents
Manually annotated dataset
Manually annotated English dataset
Document comparison tool
Updated annotation guidelines
verify annotations
Comparison HTML

ANNOTATION COMPARISON HTML

Side-by-side review of annotations
Quickly find diferences in annotations

(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?

ANNOTATION STATS

Number of entities per language and entity type

	da	en	lt	lv	pl	sv	et	fi	Total
Unit	211	213	211	211	212	207	204	209	1678
Value	213	213	211	212	212	211	205	207	1684
Date	1166	1163	1174	1158	1160	1142	1152	1198	9313
IDNUM	19	20	20	20	20	20	19	20	158
LOC	817	859	824	821	854	667	778	797	6417
NAT	73	74	98	75	74	73	72	76	615
ORG	2989	2956	2870	2880	2904	3066	2852	2962	23479
PER	582	607	577	580	578	581	579	580	4665
PROF	429	429	384	466	474	472	465	462	3581
TITLE	43	43	43	43	44	42	43	43	344
URL	5	5	5	5	5	5	5	5	40
Total	6547	6582	6417	6471	6537	6487	6374	6559	51974

De-identification

- After the annotation step, we produce a deidentified 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 were selected manually 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 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.

De-identification

- In order to study the impact of the de-identification strategy, we trained two NER models - one using the original dataset and another using de-identified data set. We evaluate them on the original and de-identified test set.

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

F1 score and standard deviation for NER models trained on original and de-identified data, evaluated on original and de-identified test sets.

SOME RESULTS AND OBSERVATIONS

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**

Evaluation of trained NER models.

Rows: models trained on respective train sets,

Columns: the result of the evaluation (F1 scores) on the test set in each language.

Multi is the concatenated data set of monolingual data.

SOME RESULTS AND OBSERVATIONS

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

We evaluate previously trained best English and multilingual NER models on the TAB (Pilán et al., 2022) benchmark dataset using TAB evaluation script.

The multilingual NER model evaluated against the TAB ECHR test set shows good recall on all identifiers and quasi-identifiers (Rdi+qi and ERqi), while recall of direct identifiers (CODE) is poor.

Further, TAB performs sanitization for a single selected person, while NER-based approach performs sanitization to all persons mentioned in the document. This leads to poor precision scores when evaluated on the TAB ECHR test set (WPdi+qi).

CONCLUSION

- 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)

- 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.

- Available at https://github.com/tilde-nlp/MultiLeg-dataset

THANK YOU FOR YOUR ATTENTION!

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 work 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).