Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 741–752
Marseille, 20-25 June 2022

"Beste Grüße, *Maria Meyer*" — Pseudonymization of Privacy-Sensitive Information in Emails

Elisabeth Eder[1] Michael Wiegand[1] Ulrike Krieg-Holz[1] Udo Hahn[2]
[1] Alpen-Adria-Universität Klagenfurt, Klagenfurt, Austria
[2] Friedrich-Schiller-Universität Jena, Jena, Germany
{elisabeth.eder | michael.wiegand | ulrike.krieg-holz}@aau.at udo.hahn@uni-jena.de

Abstract

The exploding amount of user-generated content has spurred NLP research to deal with documents from various digital media formats (tweets, chats, emails, etc.). Using these texts as language resources implies complying with legal data privacy regulations. To protect the personal data of individuals and preclude their identification, we employ pseudonymization. More precisely, we identify those text spans that carry information revealing an individual's identity (e.g., names of persons, locations, phone numbers, or dates) and subsequently substitute them with synthetically generated surrogates. Based on CODE ALLTAG, a German-language email corpus, we address two tasks. The first task is to evaluate various architectures for the automatic recognition of privacy-sensitive entities in raw data. The second task examines the applicability of pseudonymized data as training data for such systems since models learned on original data cannot be published for reasons of privacy protection. As outputs of both tasks, we, first, generate a new pseudonymized version of CODE ALLTAG compliant with the legal requirements of the General Data Protection Regulation (GDPR). Second, we make accessible a tagger for recognizing privacy-sensitive information in German emails and similar text genres, which is trained on already pseudonymized data.

Keywords: pseudonymization, data privacy, email corpus, German language resources, named entity recognition

1. Introduction

With the rapidly increasing adoption of social media platforms, we observe an upsurge of digitally transmitted private communication and exploding volumes of so-called user-generated content (UGC). Responding to this fundamental move in communication habits world-wide, digital (social) media communication has become a major focus of research in NLP.

From a user perspective though, intentionally sharing personal opinions, stances and attitudes via social media also leaves footprints behind that can be used for demographic profiling, further social grouping activities based on the analysis of properties users (unintentionally) disclose in their digital discourse (Kosinski et al., 2013; Volkova et al., 2015). Despite the evident relevance of privacy protection for digital media data, including the threat of re-identification of individual authors, this topic has long been neglected by mainstream NLP research. While it has always been of utmost importance for medical NLP (Meystre, 2015), it has received almost no attention in non-medical NLP domains for a long time (except for two early works by Rock (2001) and Medlock (2006)). In response to data privacy legislation (see, for example, two recent analyses of rules targeting privacy protection in data for the US (Mulligan et al., 2019) and the EU (Hoofnagle et al., 2019)), many more NLP studies nowadays address the protection of individual data privacy (Li et al., 2018; Coavoux et al., 2018; Elazar and Goldberg, 2018; Mosallanezhad et al., 2019; Friedrich et al., 2019; Feng et al., 2020; Huang et al., 2020); for a recent survey, see Lison et al. (2021).

In general, two steps are required to eliminate privacy-sensitive information from raw text data:

First, the text spans containing **privacy-sensitive data**, such as names, phone numbers, IDs, etc., have to be detected. We treat this task as a named entity recognition (NER) problem, but the types of privacy-sensitive entities differ from those commonly focused on (e.g., PER(son), LOC(action), ORG(anization)). Therefore, we call this step ***pi*** **(*p*rivacy-sensitive *i*nformation) recognition** and likewise denote the relevant entity categories as ***pi*** **entities**.

Second, the recognized pi entities have to be substituted—either by some artificial, meaningless code (e.g., *'xxx'*), the named entity type of the respective string (e.g., [PHONE]), or a randomly generated alternative instance from the same privacy type (e.g., the female person name *'Irene'* is mapped to *'Maria'*). The latter approach is called **surrogate generation** and due to its constructive nature preserves crucial discriminative information, linguistic fluency and contextual clues. Challenges for various entity types arising from this transformation step are discussed by Stubbs et al. (2015b).

The term **pseudonymization** subsumes recognizing entities bearing privacy-sensitive information (pi recognition) and their replacement by realistic substitutes (surrogate generation). Figure 1 illustrates the basic workflow for email pseudonymization.[1]

[1]There is much confusion about proper terminology use in this field (cf. Garfinkel (2015)), primarily due to mixing up linguistic string manipulation issues (the NLP perspective) with the legal topic of re-identifying individual persons (e.g.,

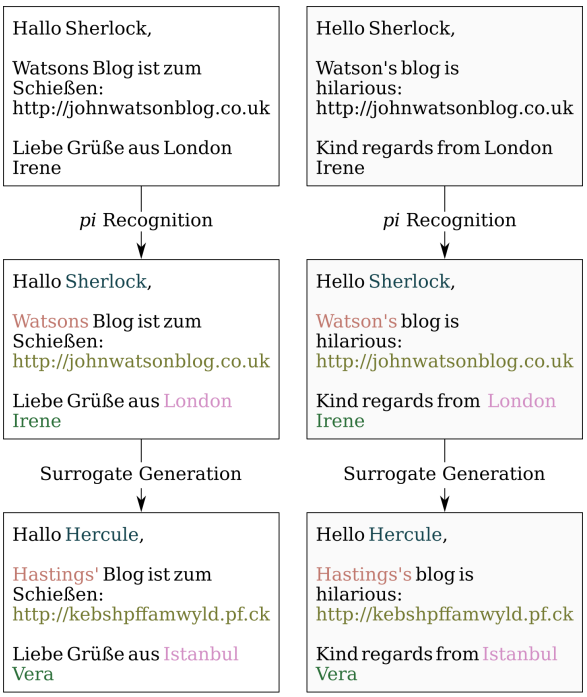

Figure 1: Pseudonymization workflow from an original email via the recognition of privacy-sensitive information (*pi*) to its pseudonymized form containing synthetic substitutes for different *pi* entity types (German original (left), English translation (right))

This paper focuses primarily on the *pi* recognition step, but it also considers pseudonymized data where *pi* entities have been substituted with surrogates. Based on the German-language email corpus CODE ALLTAG (Krieg-Holz et al., 2016) (described in Section 3), we address two tasks. The first task evaluates different deep learning architectures for ***pi* recognition in original text data** (Section 4) as a prerequisite to generate pseudonymized data for the second task; this step gives rise to a new version of CODE ALLTAG with improved pseudonymization. Since the public release of models trained on non-modified raw data carries the risk of leaking privacy-sensitive information,[2] the second task tries to avoid such issues by using already **pseudonymized data to train *pi* recognition models** (Section 5) for the purpose of developing a tagger for identifying *pi* entities that can be made publicly available.

Summarizing, our contributions are as follows:

- an evaluation of different deep learning architectures for recognizing privacy-sensitive entities in German emails,
- an examination of the applicability of pseudonymized data for *pi* recognition in German emails,
- a tagger for recognizing privacy-sensitive information in German emails and similar text genres,
- a new version of the German-language email corpus CODE ALLTAG with improved pseudonymization compliant with the legal requirements of the General Data Protection Regulation (GDPR; Hintze (2018)).[3]

2. Related Work

Due to strict legal requirements a plethora of work on pseudonymization with its two steps, *pi* recognition and surrogate generation, has been carried out in clinical NLP (for a recent survey, cf. Yogarajan et al. (2020)). Only very recently this topic has gained the interest of the general NLP community as well.

2.1. Recognition of Privacy-Sensitive Information

Progress in recognizing privacy-sensitive information can mainly be attributed to various English-language-based shared tasks in clinical NLP, which focus on 18 different entity types bearing privacy-sensitive information, so-called *Protected Health Information* (PHI) (Uzuner et al., 2007; Stubbs et al., 2015a; Stubbs et al., 2017). Besides hybrid approaches combining rules, dictionaries and classical machine learning algorithms (Norgeot et al., 2020), Recurrent Neural Networks (RNN), especially bidirectional Long Short-Term Memory Networks (Bi-LSTM), and Conditional Random Fields (CRF) are widely used for *pi* recognition as evidenced by the works of Dernoncourt et al. (2017), Liu et al. (2017) or Lange et al. (2020), among others (for surveys, see Leevy et al. (2020) and Yogarajan et al. (2020)). Recently, also transformer-based architectures found their way into this field. Johnson et al. (2020) fine-tune BERT (Devlin et al., 2019) for *pi* recognition on English medical records and García-Pablos et al. (2020) for Spanish clinical notes.

Work outside the clinical domain is rarer despite the undeniable relevance of privacy protection in non-medical genres. Several studies describe efforts to hide personal information in short text messages (SMS). Treurniet et al. (2012) (for Dutch) and Chen and Kan (2013) (for English and Mandarin) automatically recognize a handful of different *pi* entity types in text messages but do not describe the methods for the automatic recognition of these entities in detail (Chen and Kan

based on persistent mapping tables). We propose a distinction definitionally relevant for NLP: We consider de-identification a hypernym for constructive pseudonymization and destructive anonymization (the latter replacing entity mentions with artificial codes or entity type labels, which removes discriminative information).

[2]Carlini et al. (2021), e.g., demonstrate for GPT-2 (trained on huge fragments of the public Internet) (Radford et al., 2019) that an adversary can perform a training data extraction attack to recover individual training examples and thus harvest privacy-sensitive data by querying the language model.

[3]Our pseudonymization efforts must be reassessed regularly to account for technological developments that might allow re-identification of individuals (Kamocki et al., 2018).

(2013) only mention the use of regular expressions). Panckhurst (2013) proposes (for French) a semi-automatic anonymization process, with automatic pre-annotation of privacy-sensitive items by a dictionary-based anonymization tool (SEEK&HIDE; Accorsi et al. (2012)) and subsequent manual validation and correction by humans. Within this project, Patel et al. (2013) compare the unsupervised SEEK&HIDE system with a supervised learning approach using decision trees to combine the outcome of both approaches.

Adams et al. (2019) detect *pi* entities in human-computer chat data with a CRF classifier, while Jensen et al. (2021) automatically recognize five categories of *pi* mentions in job postings. The latter work also compares Bi-LSTMs and transformer models for this task.

Targeting emails, Minkov et al. (2005) identify personal names using a CRF classifier as well. Eder et al. (2020) use Bi-LSTMs with a CRF on top to recognize a set of 15 different privacy-sensitive named entity types, which they specially designed for pseudonymization. While we adopted their *pi* categories, the current work, to the best of our knowledge, is the first to use transformers for identifying *pi* entities in emails.

2.2. Surrogate Generation

The second step of the pseudonymization workflow, surrogate generation, has been examined by Carrell et al. (2013) with the 'Hiding In Plain Sight' approach: By replacing detected privacy-sensitive chunks with realistic synthetic surrogates the few identifiers 'leaked' are difficult to distinguish from the synthetic surrogates. In order to evaluate this claim, Carrell et al. (2019) investigate how a malicious attacker performing a parrot attack could expose leaked *pi* entities and find experimental evidence that such an attack can attenuate but not eliminate the protective effect of pseudonymization. This increased protection effect of surrogate generation is a major advantage for pseudonymization over anonymization.

For English medical texts, SCRUB (Sweeney, 1996) was one of the first surrogate generation systems followed by work from Uzuner et al. (2007), Yeniterzi et al. (2010), Deléger et al. (2014), Stubbs et al. (2015b), Stubbs and Uzuner (2015) and Chen et al. (2019). Similar procedures based on rules and dictionaries were used for Danish (Pantazos et al., 2011), Swedish (Alfalahi et al., 2012) and Spanish (Lima Lopez et al., 2020) clinical datasets, as well as for a learner corpus of Swedish (Megyesi et al., 2018; Volodina et al., 2020).

For German, Eder et al. (2019) propose a comparable surrogate generation system developed for emails followed by an adaptation to German medical texts (Lohr et al., 2021). While Eder et al. (2019) also evaluate the quality of the pseudonymized data by comparing *pi* recognition performances on original and pseudonymized text, they do not release a model trained on such pseudonymized data to automatically identify *pi* entities, though.

Rather than utilizing rules and dictionaries, Friedrich et al. (2019) replace *pi* mentions in medical English texts with close neighbors in embedding space (i.e., words with similar word embeddings). However, they point out the risk of inferring the original information through overlapping neighbor spaces. Therefore, they propose transforming text non-reversibly into a non-interpretable vector space representation instead of hiding privacy-sensitive entities in the raw text data. This adversarially learned representation of medical text then allows privacy-preserving sharing of training data for systems recognizing *pi* entities. Thus, they avoid the surrogate generation step completely. This strategy bears similarities to another privacy-enhancing approach based on federated learning (Yang et al., 2019). In federated learning, models are trained locally and only gradients are shared (see, e.g., Basu et al. (2021; Hathurusinghe et al. (2021; Jana and Biemann (2021)). However, Hitaj et al. (2017) show that even this distributed privacy protection scheme is vulnerable to attacks. Since we aim at sharing the entire dataset alongside a tagger for *pi* entities, these approaches do not match our requirements in any case.

3. Data

Our work is based on CODE ALLTAG (Krieg-Holz et al., 2016), a text corpus composed of two non-overlapping collections of emails. The larger portion, CODE ALLTAG$_{XL}$, was extracted from various archived USENET newsgroups and consists of 1.5 million German-language emails that merely underwent rudimentary data cleansing. This huge data set is complemented by a much smaller set of 1,390 German-language emails, CODE ALLTAG$_S$, collected on the basis of voluntary email donation and consent for publication, if properly de-identified.

Privacy-sensitive text spans in the complete document set of CODE ALLTAG$_S$ as well as in 1,000 randomly picked emails from CODE ALLTAG$_{XL}$ were manually annotated with privacy-sensitive information entity types (Eder et al., 2019; Eder et al., 2020). These *pi* entity types were specially designed for the pseudonymization of emails and, thus, differ from those commonly used in the news-centric NER community, with its focus on persons (PER), locations (LOC) and organizations (ORG). The latter named entity types are insufficient in light of data privacy considerations and pseudonymization, which require finer and better-targeted type granularities. We list the *pi* entity types for email pseudonymization in Table 1 (for more details, cf. Eder et al. (2019)).

3.1. Original Data

For the following experiments addressing the *pi* recognition on original data (task 1) and testing data for the *pi* recognition using pseudonymized data (task 2), we employed emails in their original, i.e., unpseudonymized, form. We used the entire CODE

pi Entity Type	Abbreviation
family names	FAMILY
female given names	FEMALE
male given names	MALE
organizations	ORG
user names	USER
city names	CITY
zip codes	ZIP
street names	STREET
street numbers	STREETNO
dates	DATE
passwords	PASS
unique formal identifiers	UFID
email addresses	EMAIL
phone numbers	PHONE
URLs	URL

Table 1: Privacy-sensitive information (pi) entity types relevant for emails and similar text genres

ALLTAG$_S$ corpus, which includes 8,866 pi entities, for the pi model selection (see Section 4.2) since it contains a higher variety of pi categories and more personal information, in general. Additionally, we utilized 1,000 randomly chosen emails from CODE ALLTAG$_{XL}$, with 3,226 pi entities altogether, to evaluate the resulting classifier on the much noisier CODE ALLTAG$_{XL}$ corpus (Section 4.3). Table 2 gives a quantitative breakdown of both corpora in their original shape.

	Original CODE ALLTAG	
	S	***XL***$_{1k}$
emails	1,390	1,000
tokens	152,309	94,646
pi entities	8,866	3,226
pi tokens	13,120	3,944
pi annotation	manual	manual
structure	well-curated	noisy
origin	personally donated	public archive

Table 2: Properties and privacy-sensitive information of the original versions of CODE ALLTAG$_S$ (denoted as *S*) and the 1,000 email sample from CODE ALLTAG$_{XL}$ (denoted as *XL*$_{1k}$)

3.2. Pseudonymized Data

For the experiments with pseudonymized data discussed in Section 5, we used two different corpus slices from CODE ALLTAG$_{XL}$. First, we replaced the manually annotated 1k emails from CODE ALLTAG$_{XL}$[4] with automatically produced type-preserving substitutes using the surrogate generation system from Eder et al. (2019).[5] This system maintains temporal orderings and coreferences between pi entities on the global document level.[6] While we kept this document-level consistency for the final pseudonymized corpus, for *XL*$_{1k}$, which serves as training data for pi taggers, we employed the system at the local sentence level only. Thus, we replaced the same pi entities originating from different sentences of the same document (e.g., multiple mentions of *'Irene'*) with different surrogates to further prevent re-identification.

Since this set of pseudonymized sentences originating from the 1k emails picked from CODE ALLTAG$_{XL}$ contains a lower amount of pi entities (about one third only compared to CODE ALLTAG$_S$), we further enlarged the training data of the pseudonymized *XL*$_{1k}$ with another 2,000 randomly picked emails from CODE ALLTAG$_{XL}$ that were automatically tagged for pi entities. This way, we tried to compensate for the lower density of pi entities and higher noise in that corpus. We obtained those pi entities by applying the best pi recognition model determined in Section 4. Then, we again replaced the pi entities the model found with surrogates as described above. This corpus version is called *XL*$_{2k}$. It contains over 10k automatically recognized and pseudonymized pi entities. Table 3 offers some descriptive statistics for both corpus variants.[7]

	Pseudonymized CODE ALLTAG	
	XL$_{1k}$	***XL***$_{2k}$
emails	1,000	2,000
tokens	94,485	196,607
pi entities	3,226	10,896
pi tokens	3,780	13,203
pi annotation	manual	automatic

Table 3: Properties and privacy-sensitive information of the pseudonymized *XL*$_{1k}$ and the automatically tagged 2k emails from CODE ALLTAG$_{XL}$ (*XL*$_{2k}$)

We did not take a pseudonymized version of CODE ALLTAG$_S$ into account for two reasons. First, the donors of the emails of CODE ALLTAG$_S$ agreed to a publication only after reliable and complete pseudonymization. By releasing a pi recognition model trained on these emails, we would lose the advantage of the pseudonymization approach that any potentially missed privacy-sensitive information is hard to distinguish from synthetic surrogates. Second, CODE

[4]The pseudonymized version of *XL*$_{1k}$ underwent some further normalization steps for this phase in order to provide a cleaner model.

[5]Available under `https://github.com/ee-2/SurrogateGeneration`. In order to prevent any potential leakage, we substituted the provided lexicons of surrogate candidates with slightly different ones.

[6]The surrogate generation system replaces the same *pi* entities with the same surrogates, i.e., it substitutes all mentions of *'Irene'* with *'Maria'* throughout the document. For more details see Eder et al. (2019).

[7]Note that the token sizes of the original and the pseudonymized version of *XL*$_{1k}$ differ because entities do not get substituted by the same amount of tokens in every case.

ALLTAG$_S$ also contains emails not written by the donors as part of email threads. Since consent was only gathered from the individual donor and not from other people involved in the donated email conversations, we will not distribute these emails at all.

4. Task 1: Recognition of Privacy-Sensitive Entities Using Original Data

The first task deals with the recognition of privacy-sensitive entities in our original data that are not pseudonymized. We start by comparing different architectures for identifying *pi* entities based on the original version of the well-curated CODE ALLTAG$_S$ corpus. Next, we evaluate the selected model architecture on the noisier CODE ALLTAG$_{XL}$ (represented by XL_{1k}). We use the outcomes of this examination to automatically tag *pi* entities in the entire CODE ALLTAG$_{XL}$ and, thus, produce a pseudonymized version of it. Likewise, we generate training data for task 2, namely the 2k mails of XL_{2k}. We summarize task 1 in Figure 2.

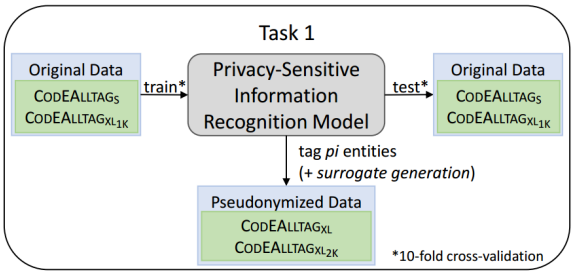

Figure 2: Task 1 trains and evaluates *pi* recognition models on original data; the final model (together with a surrogate generation step) is used to produce pseudonymized data

4.1. Setup

We used FLAIR (Akbik et al., 2019), a state-of-the-art NLP and text embedding library, as a framework for our sequence tagging experiments. For sentence segmentation and tokenization, we employed SOMAJO (Proisl and Uhrig, 2016) because it is specially designed for German web and social media texts. All evaluations are based on *pi* entities represented by the BIO annotation scheme (*'B'* preceding a token's tag stands for the *Beginning* of an entity, *'I'* (*Inside*) for its continuation, and *'O'* (*Outside*) for tokens that do not belong to any *pi* entity). We used the weighted average of precision, recall and F_1 score as evaluation measures.

Model Architecture. We compared the commonly used bidirectional LSTM (Hochreiter and Schmidhuber, 1997), with a CRF on top (Huang et al., 2015), with fine-tuning transformers, each of them with a linear classifier as last layer to predict tags.[8] For hyperparameter settings, we followed the recommendations and defaults for NER within the FLAIR architecture as described by Schweter and Akbik (2020).

Embeddings. We used the German language model GELECTRA-LARGE based on the ELECTRA architecture as transformer model, which yields the best results for several German NER and document classification tasks (Chan et al., 2020).[9] Similar to Schweter and Akbik (2020) for NER, we also experimented with concatenating non-contextual word embeddings to the transformer-based embeddings produced for each word. For that, we used FASTTEXT word embeddings (Grave et al., 2018) based on COMMON CRAWL and, to handle out-of-vocabulary lexicalizations, BPEMB[10] subword embeddings (Heinzerling and Strube, 2018) based on Byte Pair Encoding (BPE) (Sennrich et al., 2016).

Document Context. We tested the inclusion of the document context for sequence tagging as proposed by Schweter and Akbik (2020). For each sentence, they pass the sentence itself as well as 64 subtokens from the left and the right context to the transformer. Consequently, embeddings of the tokens of the sentence are then calculated taking this surrounding context into account. Their approach is called FLERT.

Learning from NER Tasks. Recognizing privacy-sensitive entities in emails shares many similarities with the standard NER task. The ORG category is congruent, whereas PER and LOC constitute coarser-grained supercategories requiring finer-grained *pi* entity types for our task (e.g., PER(son) subsumes FEMALE, MALE, FAMILY and USER from Table 1). Therefore, we further examined whether our use case might profit from employing models already fine-tuned on NER tasks by further fine-tuning them on our task. For better comparability of the results, we only used the basic setting for the transformer architecture and did not take the document context or any non-contextual word embeddings into account. We switched to multilingual XLM-ROBERTA-LARGE transformers (Conneau et al., 2020) for this experiment because several such models are already fine-tuned on NER tasks and are publicly available. We compared a pure (i.e., not fine-tuned for a specific task) XLM-ROBERTA-LARGE transformer (denoted as XLM-R in the following) with two fine-tuned versions of these embeddings. The first version (XLM-R GERMAN)[11] is fine-tuned on the German CONLL data (Tjong Kim Sang and De Meulder, 2003) for the recognition of PER, ORG, LOC and MISC. The second model (XLM-R

[8] A CRF on top of the transformer embeddings yielded slightly worse results.

[9] We also experimented with GBERT-BASE, GBERT-LARGE and GELECTRA-BASE (all from Chan et al. (2020)) but found them to perform worse.

[10] We took the 100-dimensional BPEMB with vocabulary size 100,000.

[11] https://huggingface.co/xlm-roberta-large-finetuned-conll03-german

Model Architecture	(Concatenated) Embeddings	Document Context	Prec	Rec	F_1
LSTM+CRF	BPEMB + Character (Eder et al. (2020))		85.00	79.11	81.44
LSTM+CRF	GELECTRA-LARGE		86.66	86.01	86.22
LSTM+CRF	GELECTRA-LARGE + FASTTEXT + BPEMB		87.26	86.60	86.79
LSTM+CRF	GELECTRA-LARGE	✓	81.21	87.10	87.01
LSTM+CRF	GELECTRA-LARGE + FASTTEXT + BPEMB	✓	87.45	86.79	86.95
Transformer	GELECTRA-LARGE		87.14	87.64	87.31
Transformer	GELECTRA-LARGE + FASTTEXT + BPEMB		86.80	**88.44**	87.56
Transformer	GELECTRA-LARGE	✓	87.73	87.58	87.52
Transformer	GELECTRA-LARGE + FASTTEXT + BPEMB	✓	**88.10**	88.27	**88.09**

Table 4: Evaluation of different setups for *pi* recognition on CODE ALLTAG$_S$ (10-fold cross-validation)

HRL)[12] is not only trained on German CONLL but also on NER training data (with the categories PER, ORG and LOC) from another nine high-resource languages, which may be beneficial for our task.

4.2. Performance on CODE ALLTAG$_S$

Table 4 shows the results of the different taggers for a 10-fold cross-validation on CODE ALLTAG$_S$. All configurations clearly outperform the best model from Eder et al. (2020) who used BPEMB combined with character embeddings within the LSTM+CRF architecture.[13] In general, the LSTM+CRF-based models yielded worse results than the transformer models. The transformer version with GELECTRA-LARGE, FASTTEXT and BPEMB embeddings which takes document context into account produced the best results. It exceeds the model from Eder et al. (2020) significantly with $p < 0.005$. (We used the two-sided Wilcoxon signed-rank test on F_1 for calculating significance.) It also achieved significantly better results than a LSTM+CRF based on GELECTRA-LARGE with $p < 0.005$, and a LSTM+CRF with GELECTRA-LARGE in combination with non-contextual embeddings with $p < 0.05$. In comparison, the results for LSTM+CRF models incorporating the context of sentences are not significantly lower. Contrasting the best model with the other transformer-based architectures, we found no significant differences. Obviously, taking the document context into account, as well as concatenating non-contextual word embeddings, i.e., FASTTEXT and BPEMB embeddings, does not improve the performance of transformer models significantly for our task.

Our error analysis revealed that ORGs are among the hardest categories to recognize for the models. Therefore, it seems reasonable to take advantage of embeddings already fine-tuned on NER data. Although these data are limited to news articles, they contain more ORG entities from which the models could learn.

Table 5 depicts the outcomes of experiments with embeddings fine-tuned on NER data. It shows that employing the XLM-R GERMAN model fine-tuned on German data or XLM-R HRL embeddings fine-tuned on 10 different languages did not yield significant performance differences. The pure XLM-R model, which reached similar results as the GELECTRA-LARGE model, returned even slightly, but not significantly, better results. Likewise, results for the ORG category did not improve. Therefore, we may conclude that against our assumption identifying privacy-sensitive entities on our data does not profit from models that have already been fine-tuned on NER tasks.

Embeddings	Prec	Rec	F_1
XLM-R GERMAN*	86.83	87.04	86.85
XLM-R HRL*	87.13	87.22	87.11
XLM-R	**87.21**	87.43	87.25
GELECTRA-LARGE	87.14	**87.64**	**87.31**

Table 5: Results for fine-tuning models already fine-tuned on NER tasks (marked with '*') and pure models for *pi* recognition on CODE ALLTAG$_S$ (10-fold cross-validation)

4.3. Performance on CODE ALLTAG$_{XL}$

We also evaluated the best model (the fine-tuned transformer with GELECTRA-LARGE, FASTTEXT and BPEMB embeddings that includes document context) on CODE ALLTAG$_{XL}$ since we wanted to publish the pseudonymized version of this dataset, as well as a *pi* recognition model based on such data. Table 6 depicts the results of this evaluation. When applying a model trained on CODE ALLTAG$_S$ (denoted as *S*) to the 1k emails from CODE ALLTAG$_{XL}$ (denoted as *XL*$_{1k}$), performance drops to an F_1 score of 76.85. This comes as no surprise since CODE ALLTAG$_{XL}$ is a lot noisier than CODE ALLTAG$_S$. The results for training and testing on the 1,000 mails from CODE ALLTAG$_{XL}$

[12] https://huggingface.co/Davlan/xlm-roberta-large-ner-hrl

[13] Note that we split the cross-validation folds based on emails, while Eder et al. (2020) split on lines, which explains the differences to their reported results. Thus, we avoided having *pi* entities that appear multiple times in one document in training and testing data, resulting in lower performance scores.

in a 10-fold cross-validation setup, which are even worse, also show that *pi* recognition is more difficult on this corpus segment. XL_{1k} contains fewer entities and some categories appear only rarely (e.g., email addresses, phone numbers, street names or ZIP codes), which means they are more difficult to learn. Therefore, we joined both corpora to benefit from the higher amount and diversity of entities in CODE ALLTAG$_S$. We merged both corpora for training while testing only on the third of CODE ALLTAG$_{XL}$ left out from training for a 3-fold cross-validation. Hence, the test set is similar for all three evaluations on XL_{1k}. This setting reached an F_1 score of 78.37 and thus indeed improved performance.

Training	Testing	Prec	Rec	F_1
S	XL_{1k}	77.11	77.15	76.85
XL_{1k}	XL_{1k}	72.61	75.52	73.83
S & XL_{1k}	XL_{1k}	76.73	80.17	78.37
S	*S*	88.10	88.27	88.09

Table 6: Evaluation on 1k emails from CODE ALLTAG$_{XL}$ (denoted as XL_{1k}) of models trained on CODE ALLTAG$_S$ (denoted as *S*) or XL_{1k} as well as on a merger of both (denoted as *S* & XL_{1k}); training and testing data were kept strictly disjoint while the test data (XL_{1k}) were comparable.[14]

We applied the model trained on CODE ALLTAG$_S$ and XL_{1k} for *pi* recognition on the entire CODE ALLTAG$_{XL}$ corpus and substituted the recognized *pi* entities with surrogates keeping coreferences between entities across each email intact. Thereby, we can provide a new and better pseudonymized version of this corpus compared to the previous version (Eder et al., 2020).

5. Task 2: Recognition of Privacy-Sensitive Entities Using Pseudonymized Data

Since we want to share a *pi* recognition model for emails and related genres, task 2 evaluates the applicability of pseudonymized data as training data for recognizing privacy-sensitive information. For that purpose, we compared two different settings. First, we took a smaller pseudonymized corpus based on manually annotated *pi* entities. For the second setting, we enlarged such costly to generate data by pseudonymized texts where *pi* entities had been automatically recognized. We trained taggers on both types of pseudonymized data (described in Section 3.2) and evaluated them on the original corpora introduced in Section 3.1. Figure 3 summarizes task 2.

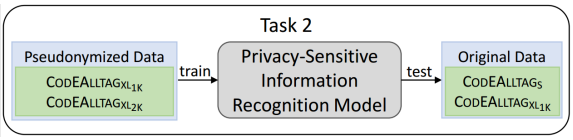

Figure 3: Task 2 evaluates models that were trained on pseudonymized data on original data

5.1. Setup

For this task and the final model, we did not include the document context (the FLERT approach) for the recognition *pi* entities. This is because shuffling sentences and not taking any context surrounding these sentences into account should prevent the leakage of private information even more and, therefore, make re-identification even harder. Instead, we only employed the GELECTRA-LARGE transformer with FASTTEXT and BPEMB embeddings, a setting with no significant performance differences compared to utilizing document context on our data (see Table 4). Similar to task 1, we evaluated on *pi* entities represented by the BIO annotation scheme and used the weighted average of precision, recall and F_1 score.

5.2. Performance

Table 7 shows the results for employing pseudonymized corpora (the pseudonymized versions of XL_{1k} and XL_{2k}) as training data for *pi* recognition models and applying them to raw text (the original versions of CODE ALLTAG$_S$ and XL_{1k}). Not surprisingly, performance decreased for models trained on these camouflaged data (see Yeniterzi et al. (2010), Deléger et al. (2014), Friedrich et al. (2019), Eder et al. (2019) and Berg et al. (2019) for similar findings, in some studies to a lesser degree, though). Applying a model trained on pseudonymized XL_{1k} to the original CODE ALLTAG$_S$ only reached an F_1 score of 46.54. Note that applying a model trained on XL_{1k} to CODE ALLTAG$_S$ reached lower results, in general (yielding an F_1 score of 57.77 trained on its original version). When we merged the pseudonymized version of XL_{1k} with XL_{2k}, performance increased considerably up to 65.07 F_1, thus outperforming the original setting.

For testing on the original XL_{1k}, which is noisier and contains fewer entities than CODE ALLTAG$_S$, we used a 10-fold cross-validation procedure. This means that we trained models on $9/10$ of the pseudonymized data and applied it to the remaining original and unpseudonymized part for all 10 different folds.

Employing pseudonymized XL_{1k} solely on its original version yielded 56.64 F_1. Combining both pseudonymized corpora XL_{1k} and XL_{2k}, again, improved results notably, reaching an F_1 score of 61.12.

[14]For training and testing on XL_{1k} as well as for training and testing on *S*, we used 10-fold cross-validation. For *S* merged with XL_{1k}, we performed a 3-fold cross-validation solely on the latter corpus. Thus, we tested on each third of XL_{1k} while we trained on the other $\frac{2}{3}$ combined with *S*. For training on *S* and testing on XL_{1k}, we trained a model on *S* three times, tested these models on XL_{1k} and evaluated the three runs.

Training	Testing	Prec	Rec	F_1
Pseudonymized Data	Original Data			
XL_{1k}	S	65.15	43.39	46.54
XL_{1k} & XL_{2k}	S	80.84	62.97	65.07
Original Data				
XL_{1k}	S	71.19	59.63	57.77
Pseudonymized Data	Original Data			
XL_{1k}	XL_{1k}	69.27	50.38	56.64
XL_{1k} & XL_{2k}	XL_{1k}	76.26	54.34	61.12
Original Data				
XL_{1k}	XL_{1k}	72.16	74.11	72.68

Table 7: Results for training on pseudonymized data of the smaller, manually annotated XL_{1k} and XL_{2k} with a larger quantity of automatically recognized pi entities while testing on the original emails of S as well as of XL_{1k} (10-fold cross-validation), in comparison to training and testing on original data

Still, this setting did not outperform the original one, which achieved 72.68 F_1.

We ascribe these inferior results of models trained on pseudonymized data especially to the substitutions for organizations, cities, URLs and email addresses. The surrogate generation system commonly replaces organizations and city names with rather infrequent names, while URLs and email addresses are substituted with randomly generated characters. Therefore, recognizing organizations and, to a lesser degree, city names, URLs and email addresses tends to be particularly difficult (e.g., email addresses rewritten as strings of random characters in pseudonymized data are often confused with female, male or family names on original data probably due to their character similarity).

Yet, we may conclude that enlarging the pseudonymized version of the smaller and manually annotated XL_{1k} with XL_{2k}, which contributes a large number of automatically detected pi entities, turns out to be beneficial for our task. However, these results should be taken with caution since we evaluated on the same dataset (yet not on the same sentences) due to the lack of complementary data. We plan to account for this shortcoming in future work.

6. Conclusion

In this study, we addressed the problem of protecting privacy-sensitive information (pi) in emails with pseudonymization as a two-step procedure. First, entities carrying privacy-sensitive information are recognized. Second, these text spans are substituted with automatically generated, naturally appearing surrogates.

We treated the recognition of privacy-sensitive entities primarily as a NER problem using 15 distinct categories. We evaluated various deep learning architectures for this task on the German-language email corpus CODE ALLTAG. To the best of our knowledge, this is the first study in which experiments with fine-tuning transformers and learning from standard named entity tasks are conducted to recognize privacy-sensitive entities in emails.

We found that fine-tuning transformers yields better results than employing LSTMs, whereas concatenating non-contextual word embeddings to the transformer embeddings and including the contextual surroundings of a sentence does not lead to significantly better results. Also, fine-tuning transformers already fine-tuned on NER tasks did not improve performance scores. However, we showed that the best system outperforms previous approaches. Hence, we can provide a new version of CODE ALLTAG with improved pseudonymization compared to the version previously available.

As releasing models trained on original, non-pseudonymized data carries the risk of leaking privacy-sensitive information, we further examined the applicability of using pseudonymized data as training data for pi recognition systems. Enlarging pseudonymized data based on manually annotated pi entities with emails where models trained on original data automatically identified pi entities indeed improved performance. When applying systems trained on pseudonymized data to original texts, results are still inferior compared to training on original data, though. Consequently, we will focus on more realistic surrogate generation approaches in future work.

Employing 3,000 pseudonymized emails with about 290,000 tokens and over 14,000 manually as well as automatically detected pi entities as training data, we can provide a tagger for privacy-sensitive information recognition in German-language emails. This model may be beneficial not only for pi recognition and pseudonymization of emails but also for similar privacy-sensitive text genres where openly available taggers for pi entities are rare (e.g., tweets or chats). The pre-trained model may serve as a good starting point to facilitate further de-identification efforts, e.g., as a pre-tagging device for speeding up subsequent manual annotation.

Both the tagger for privacy-sensitive information recognition and the pseudonymized version of CODE ALLTAG are available under `https://github.com/codealltag/`.

7. Bibliographical References

Accorsi, P., Patel, N., Lopez, C., Panckhurst, R., and Roche, M. (2012). SEEK&HIDE: anonymising a French SMS corpus using natural language processing techniques. *Lingvisticæ Investigationes*, 35(2):163–180.

Adams, A., Aili, E., Aioanei, D., Jonsson, R., Mickelsson, L., Mikmekova, D., Roberts, F., Valencia, J. F., and Wechsler, R. (2019). ANONYMATE: a toolkit for anonymizing unstructured chat data. In *Proceedings of the Workshop on NLP and Pseudonymisation @ NODALIDA 2019*, pages 1–7. Linköping Electronic Press.

Akbik, A., Bergmann, T., Blythe, D., Rasul, K., Schweter, S., and Vollgraf, R. (2019). FLAIR: an easy-to-use framework for state-of-the-art NLP. In *NAACL-HLT 2019 — Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Demonstrations Session*, pages 54–59. Association for Computational Linguistics.

Alfalahi, A., Brissman, S., and Dalianis, H. (2012). Pseudonymisation of personal names and other PHIs in an annotated clinical Swedish corpus. In *BioTxtM 2012 — Proceedings of the 3rd Workshop on Building and Evaluating Resources for Biomedical Text Mining @ LREC 2012*, pages 49–54.

Basu, P., Singha Roy, T., Naidu, R., Muftuoglu, Z., Singh, S., and Mireshghallah, F. (2021). Benchmarking differential privacy and federated learning for BERT models. In *ML4data 2021 — Proceedings of the Workshop on Machine Learning for Data: Automated Creation, Privacy, Bias @ ICML 2021*.

Berg, H., Chomutare, T., and Dalianis, H. (2019). Building a de-identification system for real Swedish clinical text using pseudonymised clinical text. In *LOUHI 2019 — Proceedings of the 10th International Workshop on Health Text Mining and Information Analysis @ EMNLP-IJCNLP 2019*, pages 118–125. Association for Computational Linguistics (ACL).

Carlini, N., Tramèr, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, Ú., Oprea, A., and Raffel, C. (2021). Extracting training data from large language models. In *Proceedings of the 30th USENIX Security Symposium*, pages 2633–2650. The USENIX Association.

Carrell, D. S., Malin, B. A., Aberdeen, J. S., Bayer, S., Clark, C., Wellner, B., and Hirschman, L. (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.

Carrell, D. S., Cronkite, D. J., Li, M. R., Nyemba, S., Malin, B. A., Aberdeen, J. S., and Hirschman, L. (2019). The machine giveth and the machine taketh away: a parrot attack on clinical text deidentified with Hiding in Plain Sight. *Journal of the American Medical Informatics Association*, 26(12):1536–1544.

Chan, B., Schweter, S., and Möller, T. (2020). German's next language model. In *COLING 2020 — Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796.

Chen, T. and Kan, M.-Y. (2013). Creating a live, public Short Message Service corpus: the NUS SMS corpus. *Language Resources and Evaluation*, 47(2):299–335.

Chen, A., Jonnagaddala, J., Nekkantti, C., and Liaw, S.-T. (2019). Generation of surrogates for de-identification of electronic health records. In *MEDINFO 2019 — Proceedings of the 17th World Congress on Medical and Health Informatics: Health and Wellbeing e-Networks for All*, number 264 in Studies in Health Technology and Informatics, pages 70–73, Amsterdam etc. IOS Press.

Coavoux, M., Narayan, S., and Cohen, S. B. (2018). Privacy-preserving neural representations of text. In *EMNLP 2018 — Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1–10. Association for Computational Linguistics (ACL).

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L. S., and Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In *ACL 2020 — Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451. Association for Computational Linguistics (ACL).

Deléger, L., Lingren, T., Ni, Y., Kaiser, M., Stoutenborough, L., Marsolo, K., Kouril, M., Molnar, K., and Solti, I. (2014). Preparing an annotated gold standard corpus to share with extramural investigators for de-identification research. *Journal of Biomedical Informatics*, 50:173–183.

Dernoncourt, F., Lee, J. Y., Uzuner, Ö., and Szolovits, P. (2017). De-identification of patient notes with recurrent neural networks. *Journal of the American Medical Informatics Association*, 24(3):596–606.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. N. (2019). BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT 2019 — 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. Association for Computational Linguistics (ACL).

Eder, E., Krieg-Holz, U., and Hahn, U. (2019). De-identification of emails: pseudonymizing privacy-sensitive data in a German email corpus. In *RANLP 2019 — Proceedings of the 12th International Conference on "Recent Advances in Natural Language Processing:" Natural Language Processing in a*

Deep Learning World, pages 259–269, Shoumen, Bulgaria. Incoma Ltd.

Eder, E., Krieg-Holz, U., and Hahn, U. (2020). CODE ALLTAG 2.0: a pseudonymized German-language email corpus. In *LREC 2020 — Proceedings of the 12th International Conference on Language Resources and Evaluation*, pages 4466–4477. European Language Resources Association (ELRA).

Elazar, Y. and Goldberg, Y. (2018). Adversarial removal of demographic attributes from text data. In *EMNLP 2018 — Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 11–21. Association for Computational Linguistics (ACL).

Feng, Q., He, D., Liu, Z., Wang, H., and Choo, K.-K. R. (2020). SECURENLP: a system for multi-party privacy-preserving natural language processing. *IEEE Transactions on Information Forensics and Security*, 15:3709–3721.

Friedrich, M., Köhn, A., Wiedemann, G., and Biemann, C. (2019). Adversarial learning of privacy-preserving text representations for de-identification of medical records. In *ACL 2019 — Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5829–5839. Association for Computational Linguistics (ACL).

García-Pablos, A., Perez, N., and Cuadros, M. (2020). Sensitive data detection and classification in Spanish clinical text: experiments with BERT. In *LREC 2020 — Proceedings of the 12th International Conference on Language Resources and Evaluation*, pages 4486–4494. European Language Resources Association (ELRA).

Garfinkel, S. L. (2015). De-identification of personal information. Technical Report NISTIR 8053, National Institute of Standards and Technology (NIST), U.S. Department of Commerce, Gaithersburg/MD, October.

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., and Mikolov, T. (2018). Learning word vectors for 157 languages. In *LREC 2018 — Proceedings of the 11th International Conference on Language Resources and Evaluation*, pages 3483–3487. European Language Resources Association (ELRA).

Hathurusinghe, R., Nejadgholi, I., and Bolic, M. (2021). A privacy-preserving approach to extraction of personal information through automatic annotation and federated learning. In *PrivateNLP 2021 — Proceedings of the 3rd Workshop on Privacy in Natural Language Processing @ NAACL-HLT 2021*, pages 36-45. Association for Computational Linguistics (ACL).

Heinzerling, B. and Strube, M. (2018). BPEMB: tokenization-free pre-trained subword embeddings in 275 languages. In *LREC 2018 — Proceedings of the 11th International Conference on Language Resources and Evaluation*, pages 2989–2993. European Language Resources Association (ELRA).

Hintze, M. (2018). Viewing the GDPR through a de-identification lens: a tool for compliance, clarification, and consistency. *International Data Privacy Law*, 8(1):86–101.

Hitaj, B., Ateniese, G., and Perez-Cruz, F. (2017). Deep models under the GAN: information leakage from collaborative deep learning. In *CCS '17 — Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pages 603–618. Association for Computing Machinery (ACM).

Hochreiter, S. and Schmidhuber, H. J. (1997). Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Hoofnagle, C. J., van der Sloot, B., and Zuiderveen Borgesius, F. (2019). The European Union General Data Protection Regulation: what it is and what it means. *Information & Communications Technology Law*, 28(1):65–98.

Huang, Z., Xu, W., and Yu, K. (2015). Bidirectional LSTM-CRF models for sequence tagging. arXiv preprint arXiv:1508.01991.

Huang, Y., Song, Z., Chen, D., Li, K., and Arora, S. (2020). TEXTHIDE: tackling data privacy for language understanding tasks. In *Findings of the Association for Computational Linguistics — EMNLP 2020*, pages 1368–1382. Association for Computational Linguistics (ACL).

Jana, A. and Biemann, C. (2021). An investigation towards differentially private sequence tagging in a federated framework. In *PrivateNLP 2021 — Proceedings of the 3rd Workshop on Privacy in Natural Language Processing @ NAACL-HLT 2021*, pages 30–35. Association for Computational Linguistics (ACL).

Jensen, K. N., Zhang, M., and Plank, B. (2021). Deidentification of privacy-related entities in job postings. In *NoDaLiDa 2021 — Proceedings of the 23rd Nordic Conference on Computational Linguistics*, number 45 in NEALT Proceedings Series, pages 210-221. Linköping University Electronic Press.

Johnson, A. E. W., Bulgarelli, L., and Pollard, T. J. (2020). Deidentification of free-text medical records using pre-trained bidirectional transformers. In *CHIL 2020 — Proceedings of the 2020 ACM Conference on Health, Inference, and Learning*, pages 214–221. Association for Computing Machinery (ACM).

Kamocki, P., Mapelli, V., and Choukri, K. (2018). Data Management Plan (DMP) for language data under the new General Data Protection Regulation (GDPR). In *LREC 2018 — Proceedings of the 11th International Conference on Language Resources and Evaluation*, pages 135–139. European Language Resources Association (ELRA).

Kosinski, M., Stillwell, D. J., and Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the*

National Academy of Sciences of the United States of America, 110(15):5802–5805.

Krieg-Holz, U., Schuschnig, C., Matthies, F., Redling, B., and Hahn, U. (2016). CODE ALLTAG: a German-language e-mail corpus. In *LREC 2016 — Proceedings of the 10th International Conference on Language Resources and Evaluation*, pages 2543–2550. European Language Resources Association (ELRA).

Lange, L., Adel, H., and Strötgen, J. (2020). Closing the gap: joint de-identification and concept extraction in the clinical domain. In *ACL 2020 — Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6945--6952. Association for Computational Linguistics (ACL).

Leevy, J. L., Khoshgoftaar, T. M., and Villanustre, F. (2020). Survey on RNN and CRF models for de-identification of medical free text. *Journal of Big Data*, 7(1):#73 (73:1–73:22).

Li, Y., Baldwin, T., and Cohn, T. (2018). Towards robust and privacy-preserving text representations. In *ACL 2018 — Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, volume 2: Short Papers, pages 25–30. Association for Computational Linguistics (ACL).

Lima Lopez, S., Perez, N., García-Sardiña, L., and Cuadros, M. (2020). HITZALMED: anonymisation of clinical text in Spanish. In *LREC 2020 — Proceedings of the 12th International Conference on Language Resources and Evaluation*, pages 7038–7043. European Language Resources Association (ELRA).

Lison, P., Pilán, I., Sanchez, D., Batet, M., and Øvrelid, L. (2021). Anonymisation models for text data: state of the art, challenges and future directions. In *ACL-IJCNLP 2021 — Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics & 11th International Joint Conference on Natural Language Processing*, volume 1: Long Papers, pages 4188–4203. Association for Computational Linguistics (ACL).

Liu, Z., Tang, B., Wang, X., and Chen, Q. (2017). De-identification of clinical notes via recurrent neural network and conditional random field. *Journal of Biomedical Informatics*, 75(Supplement):S34–S42.

Lohr, C., Eder, E., and Hahn, U. (2021). Pseudonymization of PHI items in German clinical reports. In *Public Health and Informatics. MIE 2021 — Proceedings of the 31st Conference on Medical Informatics in Europe*, number 281 in Studies in Health Technology and Informatics, pages 273–277, Amsterdam etc. IOS Press.

Medlock, B. (2006). An introduction to NLP-based textual anonymisation. In *LREC 2006 — Proceedings of the 5th International Conference on Language Resources and Evaluation*, pages 1051–1056. European Language Resources Association (ELRA).

Megyesi, B., Granstedt, L., Johansson, S., Prentice, J., Rosén, D., Schenström, C.-J., Sundberg, G., Wirén, M., and Volodina, E. (2018). Learner corpus anonymization in the age of GDPR: insights from the creation of a learner corpus of Swedish. In *NLP4CALL 2018 — Proceedings of the 7th Workshop on NLP for Computer Assisted Language Learning*, number 152 in Linköping Electronic Conference Proceedings, pages 47–56.

Meystre, S. M. (2015). De-identification of unstructured clinical data for patient privacy protection. In Aris Gkoulalas-Divanis et al., editors, *Medical Data Privacy Handbook*, pages 697–716. Springer, Cham, Switzerland.

Minkov, E., Wang, R. C., and Cohen, W. W. (2005). Extracting personal names from email: applying named entity recognition to informal text. In *HLT-EMNLP 2005 — Proceedings of the Human Language Technology Conference & 2005 Conference on Empirical Methods in Natural Language Processing*, pages 443–450. Association for Computational Linguistics (ACL).

Mosallanezhad, A., Beigi, G., and Liu, H. (2019). Deep reinforcement learning-based text anonymization against private-attribute inference. In *EMNLP-IJCNLP 2019 — Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing & 9th International Joint Conference on Natural Language Processing*, pages 2360–2369. Association for Computational Linguistics (ACL).

Mulligan, S. P., Freeman, W. C., and Linebaugh, C. D. (2019). Data protection law: an overview. Technical Report CRS Report R45631, Congressional Research Service.

Norgeot, B., Muenzen, K., Peterson, T. A., Fan, X., Glicksberg, B. S., Schenk, G., Rutenberg, E., Oskotsky, B., Sirota, M., Yazdany, J., Schmajuk, G., Ludwig, D., Goldstein, T., and Butte, A. J. (2020). Protected health information filter (PHILTER): accurately and securely de-identifying free-text clinical notes. *npj Digital Medicine*, 3:#57.

Panckhurst, R. (2013). A large SMS corpus in French: from design and collation to anonymisation, transcoding and analysis. *Procedia Social and Behavioral Sciences*, 95:96–104.

Pantazos, K., Lauesen, S., and Lippert, S. (2011). De-identifying an EHR database: anonymity, correctness and readability of the medical record. In *User Centred Networked Health Care. MIE 2011 — Proceedings of the 23rd Conference of the European Federation of Medical Informatics*, number 169 in Studies in Health Technology and Informatics, pages 862–866, Amsterdam etc. IOS Press.

Patel, N., Accorsi, P., Inkpen, D. Z., Lopez, C., and Roche, M. (2013). Approaches of anonymisation of an SMS corpus. In *Computational Linguistics and Intelligent Text Processing. CICLing 2013 — Proceedings of the 14th International Conference*

on Computational Linguistics and Intelligent Text Processing. Part I, number 7816 in Lecture Notes in Computer Science (LNCS), pages 77–88, Berlin, Heidelberg. Springer.

Proisl, T. and Uhrig, P. (2016). SoMaJo: state-of-the-art tokenization for German Web and social media texts. In *WAC-X — Proceedings of the 10th Web as Corpus Workshop and the EmpiriST Shared Task @ ACL 2016*, pages 57–62. Association for Computational Linguistics (ACL).

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners. Technical report, OpenAi.

Rock, F. (2001). Policy and practice in the anonymisation of linguistic data. *International Journal of Corpus Linguistics*, 6(1):1–26.

Schweter, S. and Akbik, A. (2020). FLERT: document-level features for named entity recognition. arXiv preprint arXiv:2011.06993.

Sennrich, R., Haddow, B., and Birch, A. (2016). Neural machine translation of rare words with subword units. In *ACL 2016 — Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, volume 1: Long Papers, pages 1715–1725. Association for Computational Linguistics (ACL).

Stubbs, A. and Uzuner, Ö. (2015). Annotating longitudinal clinical narratives for de-identification: the 2014 i2b2/UTHealth corpus. *Journal of Biomedical Informatics*, 58(Supplement):S20–S29.

Stubbs, A., Kotfila, C., and Uzuner, Ö. (2015a). Automated systems for the de-identification of longitudinal clinical narratives: overview of 2014 i2b2/ UTHealth Shared Task Track 1. *Journal of Biomedical Informatics*, 58(Supplement):S11–S19.

Stubbs, A., Uzuner, Ö., Kotfila, C., Goldstein, I., and Szolovits, P. (2015b). Challenges in synthesizing surrogate PHI in narrative EMRs. In Aris Gkoulalas-Divanis et al., editors, *Medical Data Privacy Handbook*, pages 717–735. Springer, Cham, Switzerland.

Stubbs, A., Filannino, M., and Uzuner, Ö. (2017). De-identification of psychiatric intake records: overview of 2016 CEGS NGRID Shared Tasks Track 1. *Journal of Biomedical Informatics*, 75(Supplement):S4–S18.

Sweeney, L. (1996). Replacing personally-identifying information in medical records, the Scrub system. In *AMIA '96 — Proceedings of the 1996 AMIA Annual Fall Symposium (formerly SCAMC). Beyond the Superhighway: Exploiting the Internet with Medical Informatics*, pages 333–337, Philadelphia/PA. Hanley & Belfus.

Tjong Kim Sang, E. F. and De Meulder, F. (2003). Introduction to the CoNLL-2003 Shared Task: Language-independent Named Entity Recognition. In *CoNLL 2003 — Proceedings of the 7th Conference on Computational Natural Language Learning @ HLT-NAACL 2003*, pages 142–147. Association for Computational Linguistics (ACL).

Treurniet, M., De Clercq, O., van den Heuvel, H., and Oostdijk, N. H. J. (2012). Collecting a corpus of Dutch SMS. In *LREC 2012 — Proceedings of the 8th International Conference on Language Resources and Evaluation*, pages 2268–2273. European Language Resources Association (ELRA).

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

Volkova, S., Bachrach, Y., Armstrong, M., and Sharma, V. (2015). Inferring latent user properties from texts published in social media. In *AAAI-IAAI '15 — Proceedings of the 29th AAAI Conference on Artificial Intelligence & 27th Innovative Applications of Artificial Intelligence Conference*, pages 4296–4297, Palo Alto/CA. AAAI Press.

Volodina, E., Mohammed, Y. A., Derbring, S., Matsson, A., and Megyesi, B. (2020). Towards privacy by design in learner corpora research: a case of on-the-fly pseudonymization of Swedish learner essays. In *COLING 2020 — Proceedings of the 28th International Conference on Computational Linguistics*, pages 357–369.

Yang, Q., Liu, Y., Chen, T., and Tong, Y. (2019). Federated machine learning: concept and applications. *ACM Transactions on Intelligent Systems and Technology*, 10(2):#12 (12:1–12:19).

Yeniterzi, R., Aberdeen, J. S., Bayer, S., Wellner, B., Hirschman, L., and Malin, B. A. (2010). Effects of personal identifier resynthesis on clinical text de-identification. *Journal of the American Medical Informatics Association*, 17(2):159–168.

Yogarajan, V., Pfahringer, B., and Mayo, M. (2020). A review of automatic end-to-end de-identification: is high accuracy the only metric? *Applied Artificial Intelligence*, 34(3):251–269.