Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 4466–4477
Marseille, 11–16 May 2020

CODE ALLTAG 2.0 — A Pseudonymized German-Language Email Corpus

Elisabeth Eder[1] Ulrike Krieg-Holz[1] Udo Hahn[2]
[1]Institut für Germanistik, Alpen-Adria-Universität Klagenfurt, Klagenfurt, Austria;
[2]Jena University Language & Information Engineering (JULIE) Lab, Friedrich-Schiller-Universität Jena, Jena, Germany
{elisabeth.eder | ulrike.krieg-holz}@aau.at | udo.hahn@uni-jena.de

Abstract

The vast amount of social communication distributed over various electronic media channels (tweets, blogs, emails, etc.), so-called user-generated content (UGC), creates entirely new opportunities for today's NLP research. Yet, data privacy concerns implied by the unauthorized use of these text streams as a data resource are often neglected. In an attempt to reconciliate the diverging needs of unconstrained raw data use and preservation of data privacy in digital communication, we here investigate the automatic recognition of privacy-sensitive stretches of text in UGC and provide an algorithmic solution for the protection of personal data via pseudonymization. Our focus is directed at the de-identification of emails where personally identifying information does not only refer to the sender but also to those people, locations, dates, and other identifiers mentioned in greetings, boilerplates and the content-carrying body of emails. We evaluate several de-identification procedures and systems on two hitherto non-anonymized German-language email corpora (CODE ALLTAG$_{S+d}$ and CODE ALLTAG$_{XL}$), and generate fully pseudonymized versions for both (CODE ALLTAG 2.0) in which personally identifying information of all social actors addressed in these mails has been camouflaged (to the greatest extent possible).

Keywords: email corpus, de-identification, pseudonymization, data privacy, German language resource

1. Introduction

With the rapidly increasing adoption of electronic interaction platforms, we observe an unprecedented upsurge of digitally transmitted private communication and exploding volumes of so-called user-generated contents (UGC). As a major characteristic of these new communication habits, a sender's individual email, post, comment is distributed to an often (very) large number of addressees—the recipients of an email, followers in social media, other users of a platform etc.. Hence, hitherto private communication becomes intentionally public.

In response to these changes, digital (social) media communication has become a major focus of research in NLP. Yet there seems to be a lack of awareness among NLP researchers that the exploitation of natural language data from such electronic communication channels, whether for commercial, administrative or academic purposes, has to comply with binding legal regulations (Wilson et al., 2016). Dependent on each country's legislation system, different rules for privacy protection in raw text data are enforced (cf., e.g., two recent analyses for the US (Mulligan et al., 2019) and the EU (Hoofnagle et al., 2019)). Even privacy-breach incidents in a legal grey zone can be harmful for the actors involved (including NLP researchers).

This dilemma is evidenced dramatically in the so-called AOL search data leak.[1] In August of 2006, American Online (AOL) made a large query log collection freely accessible on the Internet for a limited time. The data were extracted over three months from their search engine to support academic research. The collection included roughly 650k users issuing 20 million queries *without* any significant anonymization of individual users. The result of this release, among others, was the disclosure of private information for a number of AOL users. The most troubling aspect of the data leak was the ease by which single individuals could be pinpointed in the logs. Even ignoring social security, driver license, and credit card numbers, the *New York Times* demonstrated the ability to unlock the identity of a real user.[2] The outline of this incident and countermeasures against this privacy crash are reported by Adar (2007) from whom we adopted the case description as well. While query logs from search engines might still be at the lower end of the vulnerability chain for data privacy (still, with drastic implications (Jones et al., 2007)), UGC bundled in freely distributed corpora is clearly at its higher end, since real names of persons, their gender, age, or locations, etc. are dispersed all over such documents.[3] Surprisingly, despite its high relevance for NLP operating on UGC, the topic of data privacy has long been neglected by the mainstream of NLP research. While it has always been of utmost importance for medical NLP (Meystre, 2015), it has received almost no attention in NLP's non-medical camp for a long time (for two early exceptions, cf. Rock (2001); Medlock (2006)).

This naïve perspective is beginning to change these days with the ever-growing importance of ethical concerns related to the processing of social media texts (Thomas et al., 2017; Scantamburlo and Pelillo, 2016; Flick, 2016; De Choudhury and De, 2014; Grodzinsky and Tavani, 2010). However, there are currently no systematic actions taken to hide personally sensitive information from downstream applications when dealing with chat, SMS, tweet,

[1] Briefly described in `https://en.wikipedia.org/wiki/AOL_search_data_leak`, accessed on Nov. 24, 2019.

[2] `https://www.nytimes.com/2006/08/09/technology/09aol.html`, last accessed Nov. 24, 2019.

[3] In a famous study, Sweeney (2000) showed, e.g., that 87% (or roughly 50%) of the population in the US could be uniquely identified based only on three data items, namely, 5-digit ZIP (or symbolic name of the place of residence), gender, and date of birth, using 1990 U.S. Census data.

blog or email raw data.[4] Since this attitude also faces legal implications, a quest for the protection of individual data privacy has been raised and, in the meantime, finds active response in the most recent work of the NLP community (Li et al., 2018; Coavoux et al., 2018; Elazar and Goldberg, 2018), including the design of privacy-preserving Data Management Plans compliant with EU's General Data Protection Regulation (GDPR) (Kamocki et al., 2018).

In general, two basic approaches to eliminate privacy-bearing data from raw text data can be distinguished. The first one, *anonymization*, identifies instances of relevant privacy categories (e.g., person names or dates) and replaces sensitive strings by some artificial code (e.g., *'xxx'*). This *masking* approach might be appropriate to eliminate privacy-bearing data in the medical world, but is likely to be inappropriate for most NLP applications since crucial discriminative information and contextual clues will be erased by such a scrubbing procedure.

The second approach, *pseudonymization*, preserves such valuable information by *substituting* privacy-bearing text strings with randomly generated alternative synthetic instances from the same privacy type (e.g., the person name *'Suzanne Walker'* is mapped to *'Caroline Snyder'*).[5] As a common denominator, the term *de-identification* subsumes both, anonymization and pseudonymization.[6]

The focus of this paper will be on the identification of instances of relevant privacy categories based on a structured type system of privacy categories (essential for both approaches) *and* a privacy type-compliant substitution of the original text mentions. We will demonstrate our approach on two variants of CODE ALLTAG (Krieg-Holz et al., 2016), a German-language email corpus introduced at LREC 2016 that, at that time, could not be made publicly available due to possible privacy leakage. The current version, CODE ALLTAG 2.0, is basically a pseudonymized variant of CODE ALLTAG, now publicly available at `https://github.com/codealltag`.

We start with a discussion of related work in Section 2 and then introduce the semantic types we consider as relevant carriers of personal information in emails in Section 3. Next, we provide an overview of the email corpus our experiments are based on in Section 4, including a brief description of manual annotation activities for gold standards. In Section 5, we turn to the description and evaluation of different approaches to recognizing privacy-sensitive information, a task we treat primarily as a named entity recognition problem. In Section 6, we present the new pseudonymized version of CODE ALLTAG, CODE ALLTAG 2.0. After that, we conclude with a summary of our main contributions and an outlook into future work in Section 7.

2. Related Work

The main thrust of work on de-identification has been performed for clinical NLP.[7] Main drivers of progress in this field were two challenge tasks within the context of the I2B2 (Informatics for Integrating Biology & the Bedside) initiative[8] which focused on 18 different types of Protected Health Information (PHI) categories as required by US legislation (HIPAA).[9] The first of these challenge tasks was launched in 2006 for 889 hospital discharge summaries (Uzuner et al., 2007). The second was run in 2014 and addressed an even broader set of PHI categories (Stubbs et al., 2015a). The best system performances peaked in the high 90s (F_1 score) using classical machine learning methods, Conditional Random Fields (CRFs) in particular, hand-written rules, or a hybrid mixture of both. As a successor to I2B2, the CEGS-NGRID Shared Tasks and Workshop on Challenges in NLP for Clinical Data created a corpus of 1,000 manually de-identified psychiatric evaluation records (Stubbs et al., 2017). Interestingly, for the automatic de-identification task performance values dropped significantly down to 79.85 F_1 for the best-performing system indicating an only modest potential for domain and text genre portability (moving from discharge summaries to psychiatric evaluation records).

Recently, the deep learning wave has also hit the (clinical) de-identification community. For this task, bidirectional Long Short-Term Memory Networks (Bi-LSTMs) became quite popular as evidenced by the work of Dernoncourt et al. (2017b) who achieve an F_1 score of 97.85 on the I2B2 2014 dataset, or Liu et al. (2017) who report performance figures ranging from 95.11% over 96.98% up to 98.28% micro F_1 score under increasingly sloppier matching criteria on the same dataset. Another direction to prevent privacy leakage from clinical documents has recently been proposed by Friedrich et al. (2019). They introduce an adversarially learned representation of medical text that allows privacy-preserving sharing of training data for a de-identification classifier by transforming text non-reversibly into a non-interpretable vector space representation as training data. Employing a simple LSTM-CRF de-identification model they achieved an F_1 score of 97.4% on the I2B2 2014 reference dataset.

Yet, the focus of these studies lies on the *recognition* of privacy-relevant text stretches not on *pseudonymization*, a considerably more complex task (Stubbs et al., 2015b).

[4] One of the rare exceptions is described by Jung et al. (2015) for privacy-sensitive services on personal mobile devices.

[5] For medical applications, Bui et al. (2018) recently found no advantage for pseudonymization over anonymization in terms of system accuracy and impact on document readability.

[6] The distinction we make between de-identification and anonymization differs from the one proposed by Meystre et al. (2010) who equate de-identification and our understanding of anonymization, while their understanding of the term 'anonymization' is focused on medical data security concerns and implies that the data cannot be linked to *identify* the patient. Our change of terminology is motivated by the intention to strictly decouple the linguistic layer of different forms of *de*-identification (masking of or substituting privacy-relevant items) from data security concerns (the potential of *re*-identification of individuals), the latter being out of scope of NLP.

[7] Note that we have to distinguish between privacy protection for *structured* tabular data housed in (clinical) information systems (for which k-anonymity (Sweeney, 2002) is a well-known model to minimize a person's re-identification risk) and de-identification in *unstructured* verbal data we here focus on.

[8] `https://www.i2b2.org/`

[9] `https://www.hhs.gov/hipaa/for-professionals/privacy/index.html`

Carrell et al. (2013) examined this problem with the 'Hiding In Plain Sight' approach: detected privacy-bearing identifiers are replaced with realistic synthetic surrogates in order to collectively render the few 'leaked' identifiers difficult to distinguish from the synthetic surrogates—a major advantage for pseudonymization over anonymization. Targeting 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 have been proposed for Swedish (Alfalahi et al., 2012) and Danish (Pantazos et al., 2011) clinical corpora, yet not for German ones, up until now.

The most radical departure to escape from the data privacy problem is to generate fully synthetic, i.e., artificial textual 'fake' documents. Methodologically, this work is rooted in generative adversarial networks (GAN) (Goodfellow et al., 2014). In GANs, two networks, a discriminator model jointly learned with a generator model, play a game-theoretical game in which the generator attempts to generate realistic, but fake, data and the discriminator aims to distinguish between the generated fake data and the real data. This approach has already been applied to camouflage person-identifying demographic data (Li et al., 2018; Elazar and Goldberg, 2018).

For the medical domain, Guan et al. (2019) and Choi et al. (2017), e.g., take clinical features (diagnoses, treatment, medication codes, etc.) as input and automatically generate synthetic textual data incorporating these features as output. Lee (2018) instead uses an encoder–decoder model, as employed in many machine translation systems, for generating chief complaints from discrete variables in Electronic Health Records (EHR), like age group, gender, and diagnosis. After being trained end-to-end on authentic records, the model can generate realistic, privacy-neutral chief complaint text. Most important from our perspective, such synthetic texts include *none* of the PHI elements that was in the training data, suggesting that such models can effectively solve the de-identification problem by introducing comparable, yet artificial *documents* as substitutes for authentic clinical documents, rather than artificial *mentions of PHI-relevant entities* (as pseudonymization does).

De-identification work outside the clinical domain is rare and limited in scope. Minkov et al. (2005) aim at identifying personal names, a subclass of PHI items, within emails as a prerequisite for anonymization in informal texts using a CRF classifier. Recently, Megyesi et al. (2018) report on an anonymization study for a corpus of essays written by second language (L2) learners of Swedish. While we found no work dealing with the comprehensive anonymization or even pseudonymization of emails and Twitter-style social media data,[10] anonymizing SMSes is a topic of active research. Patel et al. (2013) introduce a system capable of anonymizing SMS (Short Message Service) communication. Their study builds on 90,000 authentic French text messages and uses dictionaries as well as decision trees as machine learning technique. Their evaluation task is, however, very coarse-grained—select those SMSes from a test corpus that either have to be anonymized or not. There is no breakdown to PHI-like categories known from the medical domain. Treurniet et al. (2012) consider privacy-relevant data for a Dutch SMS corpus (52,913 messages, in total) in greater detail. They automatically anonymized all occurrences of dates, times, decimal amounts, and numbers with more than one digit (telephone numbers, bank accounts, etc.), email addresses, URLs, and IP addresses. All sensitive information was replaced with corresponding semantic placeholder codes of the encountered semantic *type* (e.g., each email address was replaced by the type symbol `EMAIL`), not by an alternative semantic *token*, i.e., a pseudonym. The same strategy was also chosen by Chen and Kan (2013) for their SMS corpus that contains more than 71,000 messages, focusing on English and Mandarin. However, neither are the methods of automatic anonymization described in detail, nor are performance figures of this process reported in both papers (Chen and Kan (2013) only mention the use of regular expressions for the anonymization process).

In conclusion, pseudonymization has, to the best of our knowledge, only been seriously applied to medical documents, up until now. Hence, our investigation opens this study field for the first time ever to non-medical applications of pseudonymization.

3. Named Entity Types for De-Identification

The most relevant source and starting point for determining types of personally identifying information pieces in written informal documents is a catalog of *Personal Health Information* (PHI) items that has been derived from the *Health Information Privacy Act* (HIPAA). PHI enumerates altogether 18 privacy-sensitive information classes organized into eight main categories (Stubbs and Uzuner, 2015):

- *Name* includes the names of patients, doctors and user names,
- *Profession* practiced by a person,
- *Location* includes rooms, clinical departments, hospital names, names of organizations, street names, city names, state names, names of countries, ZIPs, etc.,
- *Age* of persons,
- *Date* expressions,
- *Communication* codes, e.g., phone or fax numbers, email addresses, URLs, IP addresses,
- all sorts of *IDs*, such as Social Security number, medical record number, account number, license number, vehicle ID, device ID, biometric ID, etc.,
- any *Other* form of personally sensitive data.

While some of the above categories are useful for non-medical anonymization procedures as well, others are merely domain-specific, because they are intrinsically attached to the clinical domain (such as the names of patients, doctors or nurses, hospitals and their departments). Hence, we adapted this list to email documents while, at the same time, we tried to avoid over-fitting to this text genre.

[10] Lüngen et al. (2017) report on *manual* anonymization efforts for German chat data. Boufaden et al. (2005) describe a privacy compliance engine that monitors emails generated in an organization for violation of the privacy policy of this organization.

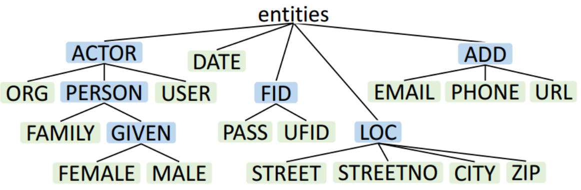


Figure 1: Hierarchy of privacy-bearing information (pi) entity types relevant for emails (leaves in green)

We, finally, came up with the set of privacy-bearing information (henceforth, pi) entity types depicted in Figure 1. It was designed to universally account for all types of emails, irrespective of any particular natural language and email encoding. These entities are organized in a concise hierarchy whose top level categories are *SocialActor (ACTOR)*, *Date (DATE)*, *FormalIdentifier (FID)*, *Location (LOC)*, and *Address (ADD)*. We anticipate that this hierarchy can also be further refined and accommodated other privacy-sensitive text genres (blogs, tweets, SMS, etc.).

The category of *SocialActor* can further be divided into *Organization (ORG)*, which includes all types of legal actors such as companies, brands, institutions and agencies, etc., *Persons (PERSON)*, with subtypes *FamilyName (FAMILY)* and *GivenName (GIVEN)*, with another split into two subcategories, namely *FemaleName (FEMALE)* and *MaleName (MALE)*, both including nicknames and initials. Finally, *UserName (USER)* covers all kinds of invented user names for IT systems and platforms.

Date (DATE) covers all sorts of date descriptions, e.g., date of birth, year of death, starting and ending dates of contracts, etc..

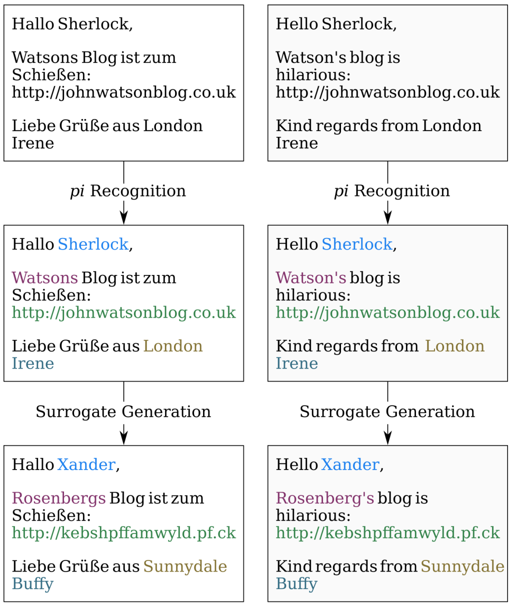
The category of *FormalIdentifier (FID)* includes *Password (PASS)* as user-provided artificial character string for all kinds of technical appliances, and *UniqueFormalIdentifier (UFID)* to capture persons (students, customers, employees, members of social security systems (SSN), authors (ORCHID), etc.), computer systems (IP addresses), or other artifacts (e.g., IBAN, DOI).

The *Location* (LOC) category subsumes *StreetName (STREET)*, *StreetNumber (STREETNO)*, *ZipCode (ZIP)*, and *CityName (CITY)* which stands for villages, towns, cities, metropolitan areas (e.g., *'Larger London'*) and regions smaller than a state (e.g., *'Bay Area'*); it also includes derivations of these names (e.g., *'Roman'*).

Address (ADD) encompasses *EmailAddress (EMAIL)*, *PhoneNumber (PHONE)*, including fax numbers, and *URL (URL)*, as well as other forms of domain names.

Unlike studies in clinical NLP (Stubbs et al., 2015b), we did not take mentions of age or profession into account, because these information units are very rare in emails (yet they often occur, with high relevance, in clinical reports). Furthermore, unspecific dates like *'Christmas'*[11] or *'next week'* and geographical information such as landmarks, rivers or lakes were not tagged for de-identification since their contribution to possible re-identification is fairly limited due to their generality.

In Figure 2, we illustrate our workflow for pseudonymization with an email excerpt highlighting different mentions of pi entity types by different colors.


Hallo Sherlock,

Watsons Blog ist zum Schießen:
http://johnwatsonblog.co.uk

Liebe Grüße aus London
Irene

pi Recognition

Hallo Sherlock,

Watsons Blog ist zum Schießen:
http://johnwatsonblog.co.uk

Liebe Grüße aus London
Irene

Surrogate Generation

Hallo Xander,

Rosenbergs Blog ist zum Schießen:
http://kebshpffamwyld.pf.ck

Liebe Grüße aus Sunnydale
Buffy

Hello Sherlock,

Watson's blog is hilarious:
http://johnwatsonblog.co.uk

Kind regards from London
Irene

pi Recognition

Hello Sherlock,

Watson's blog is hilarious:
http://johnwatsonblog.co.uk

Kind regards from London
Irene

Surrogate Generation

Hello Xander,

Rosenberg's blog is hilarious:
http://kebshpffamwyld.pf.ck

Kind regards from Sunnydale
Buffy

Figure 2: Workflow from an original email via the recognition of privacy-sensitive information units (pi) to its pseudonymized form containing synthetic substitutes for pi entities (German original (left), English translation (right))

4. Named Entity Annotation in Emails

Our work is based on CODE ALLTAG (Krieg-Holz et al., 2016), a resource composed of two non-overlapping collections of emails. The larger portion, CODE ALLTAG$_{XL}$, was extracted from various archived *Internet Usenet Newsgroups* and consists of 1,469,469 German-language emails that merely underwent some rudimentary data cleansing.[12] This huge data set is complemented by a much smaller number of 1,390 German-language emails, CODE ALLTAG$_{S+d}$, collected on the basis of voluntary email donation. The donors have provided their explicit consent that, *after de-identification*, their emails may be made publicly available. Sharing the entire corpus would create lots of opportunities for NLP research, since public access to private emails is generally forbidden.[13]

For our experiments, we exploited both data resources differently. We manually annotated CODE ALLTAG$_{S+d}$ with the privacy-bearing categories from Section 3 for training and testing. Furthermore, we annotated 1,000 randomly picked emails from CODE ALLTAG$_{XL}$ by hand to evaluate our classifier trained on CODE ALLTAG$_{S+d}$ on the much noisier XL corpus; the latter contains less personal information than the former (see Table 1).

[11] Yet, in *'Christmas 2019'*, e.g., the year *'2019'* will be tagged as *DATE*.

[12] A similar and simultaneously conducted corpus-building initiative is described by Schröck and Lüngen (2015).

[13] One of the rare exceptions is the ENRON corpus (Klimt and Yang, 2004) whose non-anonymized content was released for open inspection by order of US judges in the course of the destruction of the Enron company. For yet another example, cf. the AVOCADO RESEARCH EMAIL COLLECTION, available from LDC2015T03. Both corpora contain English language only.

	CODE ALLTAG$_{S+d}$	CODE ALLTAG$_{XL}$
# emails	1,390	1,000
# tokens	151,229	94,664
# *pi* entities	8,866	3,231
# *pi* tokens	12,649	3,549

Table 1: Privacy-bearing information in CODE ALLTAG$_{S+d}$ and the 1,000 email sample from CODE ALLTAG$_{XL}$; '#' stands for 'number of'

4.1. CODE ALLTAG$_{S+d}$

For the manual annotation campaign,[14] we set up a team of three annotators who tagged equally sized parts of the corpus, according to the privacy-bearing (pi) categories described in Section 3 (see Figure 1). Annotation was performed on the entity level. Therefore, we did not have to care about token boundaries in the surrogate generation step and, thus, no special handling for compounds and multi-token entities was required.

In order to measure inter-annotator agreement (IAA), the annotators worked on 50 identical emails randomly selected from the corpus within the same annotation phase as the entire corpus. Table 2 shows Cohen's κ (Cohen, 1960) as a measure for IAA for the pairs of annotators calculated on the entities represented by the BIO annotation scheme.[15] Hence, not only the token label itself but also matching starting and ending points of an entity are taken into account. The agreement is quite high, especially between annotator 1 and 3.

A1 - A2	A2 - A3	A3 - A1
0.925	0.933	0.958

Table 2: Cohen's κ for BIO tags on CODE ALLTAG$_{S+d}$

Based on these 50 emails the annotators also examined and discussed differences of their annotations and decided on the gold standard by majority vote, which we applied for further evaluation to measure precision, recall and F$_1$ score. Table 3 shows the outcomes regarding BIO tags per annotator and the overall result calculated over the joint annotations of the annotators. We took the averages from the outcomes of the single categories weighted by the number of true instances for each label.

An error analysis revealed that, besides mostly accidental errors, a higher disagreement on tagging *ORGs* (due to overlap or confusion with city or product names and rather generic organizations)[16] and an uncertainty regarding *DATEs* could be observed. The latter problem was solved by the decision to treat all dates as pi regardless of their specificity. As a consequence, one of the three annotators worked through the entire corpus to re-tag each *DATE* and, if necessary, also re-tag *ORGs* according to the findings from the error analysis. The outcome of this revision constituted the final gold standard annotations of CODE ALLTAG$_{S+d}$ for the de-identification task.

	Prec	Rec	F$_1$
A1	98.06	95.63	96.72
A2	87.67	77.92	80.36
A3	94.14	84.79	86.82
A1+A2+A3	93.37	86.11	88.66

Table 3: Weighted average of Prec(ision), Rec(all) and F$_1$ score in respect to the gold standard for BIO tags of CODE ALLTAG$_{S+d}$

4.2. CODE ALLTAG$_{XL}$

The manual annotation process for the 1,000 emails from CODE ALLTAG$_{XL}$ corresponds to the one described for CODE ALLTAG$_{S+d}$, except for the updated annotation guidelines according to the results of the error analysis reported above. This annotation campaign involved five annotators (Cohen's κ for this set-up is displayed in Table 4).

A1-A2	A1-A3	A1-A4	A1-A5	A2-A3
0.898	0.902	0.917	0.858	0.933
A2-A4	A2-A5	A3-A4	A3-A5	A4-A5
0.952	0.848	0.942	0.900	0.881

Table 4: Cohen's κ for BIO tags on CodE Alltag$_{XL}$

They also settled for a gold standard in a consolidation meeting afterwards. Again, nearly all annotation differences are due to overlooked personal information. We also found several disagreements regarding the gender of person names, because CODE ALLTAG$_{XL}$ contains quite a few initials or even names that are gender-neutral. The outcome of this examination constituted the final gold standard annotations for CODE ALLTAG$_{XL}$. Table 5 shows the scores per annotator and the overall result calculated over the joint annotations of the five annotators. Again, we took the averages from the outcomes of the single categories weighted by the number of true instances for each label.

CODE ALLTAG$_{XL}$	Prec	Rec	F$_1$
A1	95.53	82.46	87.98
A2	94.89	94.46	94.43
A3	93.98	91.38	92.50
A4	98.20	95.69	96.80
A5	90.36	87.69	88.26
A1+A2+A3+A4+A5	94.58	90.34	92.28

Table 5: Weighted average of Prec(ision), Rec(all) and F$_1$ score in respect to gold standard of CODE ALLTAG$_{XL}$

5. Recognition of Privacy-Sensitive Entities

With respect to the structure of emails, we first experimented with different ways of segmenting the emails into

[14] We used the BRAT tool (`http://brat.nlplab.org/`) for annotation (Stenetorp et al., 2012).

[15] 'B' preceding a token's tag stands for the *Beginning* of an entity, 'I' for its continuation (*Inside*), and 'O' for any stretch of text that does not belong to an entity (*Outside*).

[16] Stubbs and Uzuner (2015) also report confusions of organizations with other subcategories from their location class (department, hospital) and Stubbs et al. (2017) witness an uncertainty for tagging quasi-generic organizations.

sequences and tokenization.[17] For that, we compared SoMaJo (Proisl and Uhrig, 2016), a tokenizer and sentence splitter for German and English Web and social media texts, with spaCy[18] and found that the latter yielded better results for our task than the former. Also, taking the lines in the emails as sequences rather than the segmented sentences and keeping the tokenization as is improved performance.

5.1. Recognition Models for *pi* Entity Mentions

For automatically recognizing privacy-bearing (*pi*) entity mentions, we experimented with several systems.[19]

GermaNER and German NER. We ran out of the box GermaNER (Benikova et al., 2015), a CRF-based tagger primarily developed for the task of named entity recognition (NER), and German NER (Riedl and Padó, 2018), a combination of BiLSTM and CRF that utilizes character embeddings as well.

NeuroNER+token(+seq). We adapted NeuroNER, a system particularly designed for clinical de-identification (Dernoncourt et al., 2017b; Dernoncourt et al., 2017a). It is based on BiLSTMs (Hochreiter and Schmidhuber, 1997) and shares many similarities with the model proposed by Lample et al. (2016). Lee et al. (2016) enhanced NeuroNER with manually engineered features by concatenating the output of a feed-forward neural network run on a binary feature vector comprising the token's features to the character embedding and the pre-trained token embedding of a token.[20] We adopted this approach (NeuroNER+token) and decided for similar token features (see Table 6) utilizing spaCy and various lexicon look-ups.[21]

Additionally, we take the structure of emails into account (NeuroNER+token+seq). Lampert et al. (2009) define nine different zones, such as greeting, signoff, signature, new text from the sender of the email or advertising etc., and present an algorithm to automatically classify these parts. Since we presume that there is a higher likelihood to find *pi* information in some zones than in others, we apply basics of their approach on the sequence level, roughly adopting some of their features (Table 6). Also Liu et al. (2017), among others, use sentence information (number of words, unmatched brackets, end punctuation) and section information especially by adding a hidden layer concatenating these features with the output of the token BiLSTM. In contrast, our feature vectors of a document's sequences are fed to a BiLSTM and the output is concatenated directly to a corresponding token embedding that has been constructed the same way as in NeuroNER+token.

Token Features
typographic
is punctuation character, is left punctuation mark, is right punctuation mark (e.g. ')'), is bracket, is quote, is currency, is digit, contains digit, is alphabetic character, contains alphabetic character, is special character (no punctuation or alphanumerical character), contains special character, is all upper case, is all lower case, is title case, is mixed case
lexical
URL, email address, stop word, surname, female given name, male given name, city name, street name, organization name
POS tag
Universal Dependencies v2 tags[1]
Sequence Features
number of tokens, *number of characters*, *number of preceding newlines*, *number of following newlines*, *position* (number of the sequence in the document normalized by the document's length)

[1] https://universaldependencies.org/u/pos/index.html

Table 6: Token features used for the binary token feature vector (NeuroNER+token); sequence features for the sequence feature vector in NeuroNER+token+seq

BPEmb(+char). We considered BPEmb subword embeddings (Heinzerling and Strube, 2018) based on Byte Pair Encoding (BPE) (Sennrich et al., 2016), building on the results from Heinzerling and Strube (2019) who obtained higher scores for NER on the German part of the WikiAnn dataset (Pan et al., 2017) using BPEmb standalone rather than (in combination) with Bert's contextual embeddings (Devlin et al., 2019). We applied BPEmb embeddings solely (BPEmb) and in combination with character embeddings (BPEmb+char).[22] For that, we used the Flair (Akbik et al., 2018) library and its sequence tagger.

5.2. Performance on CodE Alltag$_{S+d}$

Table 7 shows the results of the classifiers from Section 5.1 for a 10-fold cross-validation on CodE Alltag$_{S+d}$. GermaNER, German NER and BPEmb without character embeddings performed worse (in terms of F_1 score) than the best performing model BPEmb+char. Yet, there is no statistically significant difference between BPEmb+char and NeuroNER+token or NeuroNER+token+seq. The

[17] We revised tokenization (without messing up sentence segmentation and parsing) originating from the tools to account for entities which only span part of the token. We split tokens on '-/&@' except for URLs, email addresses and punctuation marks.

[18] https://spacy.io/

[19] Experiments with Bert's contextual embeddings (Devlin et al., 2019) are in progress.

[20] We used fastText word embeddings (Grave et al., 2018) based on Common Crawl and Wikipedia.

[21] Female and male given names are taken from ftp://ftp.heise.de/pub/ct/listings/0717-182.zip, German surnames come from http://www.namenforschung.net/fileadmin/user_upload/dfa/Inhaltsverzeichnisse_etc/Index_Band_I-V_Gesamt_Stand_September_2016.pdf, locations from http://download.geonames.org/export/dump/allCountries.zip, German company names are imported from https://www.datendieter.de/item/Liste_von_deutschen_Firmennamen_.txt (obtained from OpenStreetMap http://www.openstreetmap.org); finally, street names originate from http://www.datendieter.de/item/Liste_von_deutschen_Strassennamen_.csv and http://www.statistik.at/strasse/suchmaske.jsp.

[22] We took the 100-dimensional BPEmb with vocabulary size 100,000.

pi Recognition Model	Prec	Rec	F$_1$
GermaNER	**93.68**	83.62	88.04*
German NER	88.66	80.63	83.99**
NeuroNER+token	89.46	86.99	88.17
NeuroNER+token+seq	90.87	**87.27**	88.99
BPEmb	90.52	86.65	88.31*
BPEmb+char	91.37	87.18	**89.03**

Table 7: Weighted average of Prec(ision), Rec(all) and F$_1$ score of the selected pi recognition models on CodE AlltaG$_{S+d}$ (10-fold cross-validation); statistically significant differences (using the two-sided Wilcoxon signed-rank test on F$_1$) are marked with '*' and '**' for $p < 0.05$ and 0.01, respectively relative to BPEmb+char

latter achieved a slightly lower F$_1$ score, with a slightly higher recall. In conclusion, we utilized BPEmb+char for the task of de-identifying CodE AlltaG$_{XL}$.

5.3. Performance on CodE AlltaG$_{XL}$

As Table 8 reveals, performance plummeted when applying the BPEmb+char model trained on CodE AlltaG$_{S+d}$ to the 1,000 emails drawn from CodE AlltaG$_{XL}$ (row 1). This is due to the fact that CodE AlltaG$_{XL}$ is much noisier and far less structured than CodE AlltaG$_{S+d}$. Training and testing on CodE AlltaG$_{XL}$ (row 3) yielded better results, but they are still far behind the ones we achieved on CodE AlltaG$_{S+d}$ (row 4, repeated from Table 7). As CodE AlltaG$_{XL}$ contains less entities and some categories appear only rarely and are thus more difficult to learn, we joined both corpora to benefit from the higher amount and diversity of entities in CodE AlltaG$_{S+d}$. We merged both corpora for training while testing solely on the quarter of CodE AlltaG$_{XL}$ left out from training in a 4-fold cross-validation setting. Omitting CodE AlltaG$_{S+d}$ from testing prevents getting too high scores due to the larger proportion of tokens and pi entities in CodE AlltaG$_{S+d}$ (those are obviously easier to classify as the outcome reveals). Hence, this setting allows more realistic results in respect to the recognition of pi entities on the entire CodE AlltaG$_{XL}$ afterwards. Compared to training exclusively on CodE AlltaG$_{XL}$ including CodE AlltaG$_{S+d}$ improved performance, reaching an overall F$_1$ score of 70.96 (row 2).

Training	Testing	Prec	Rec	F$_1$
S+d	XL	63.53	53.26	56.28
S+d + $\frac{3}{4}$ XL	$\frac{1}{4}$ XL	76.16	68.07	70.96
XL		78.17	62.74	68.62
S+d		91.37	87.18	89.03

Table 8: Weighted average of Prec(ision), Rec(all) and F$_1$ score for training on CodE AlltaG$_{S+d}$ (denoted as 'S+d') and evaluation on CodE AlltaG$_{XL}$ (denoted as 'XL', 3 runs) and for CodE AlltaG$_{S+d}$ merged with CodE AlltaG$_{XL}$, with 4-fold cross-validation solely on the latter, as well as for 10-fold cross-validation exclusively on CodE AlltaG$_{XL}$ and CodE AlltaG$_{S+d}$, respectively

6. Pseudonymized CodE AlltaG 2.0

After applying a BPEmb+char model trained on CodE AlltaG$_{S+d}$ and the 1,000 emails from CodE AlltaG$_{XL}$ (as specified in the second row of Table 8) to the remaining part of the latter corpus we substituted the recognized pi entities with type-preserving surrogates as described in Eder et al. (2019) (and illustrated in Figure 2 above).[23]
The two portions of the *p*seudonymized version of CodE AlltaG, CodE AlltaG$_{pS+d}$ and CodE AlltaG$_{pXL}$ constitute CodE AlltaG 2.0. Table 9 shows quantitative properties of CodE AlltaG$_{pS+d}$ as well as of CodE AlltaG$_{pXL}$ (based on processing with spaCy), with splits into seven categories (*Finance*, *German*, *Movies*, *Philosophy*, *Teens*, *Travels* and *Events*) (Krieg-Holz et al., 2016). Note that the number of emails in CodE AlltaG$_{pS+d}$ is smaller than the amount of emails used for training in Section 5, because we are not allowed to distribute emails that were not written by the donors themselves. CodE AlltaG$_{pS+d}$ comprises 4,104 pi entities, while CodE AlltaG$_{pXL}$ contains over 8,3M pis altogether (Table 10).

6.1. Demographic Data for CodE AlltaG$_{pS+d}$

As detailed in Krieg-Holz et al. (2016) the donors of the emails from CodE AlltaG$_{S+d}$ were asked to complete a questionnaire after submitting an email. We thus collected demographic data including gender, age, language and regional provenance, educational and professional background, frequency of writing texts and proficiency of emailing. For privacy reasons, we provide this information in an *aggregated* manner only.
Roughly 73% of the emails from CodE AlltaG$_{pS+d}$ were written by females; more than two thirds were donated by persons between 18 and 34 years old (see Figure 3 (a)). Regarding language provenance most of the donated emails were written by native speakers of German (close to 92%) and two thirds also speak German dialects. With 39% daily and 45% weekly writing frequency, the majority of emails were sent by donors with strong writing experience as part of their daily routines (cf. Figure 3 (b)). Further, most of them have a long-standing experience using email as a communication medium (Figure 3 (c)).

7. Conclusion

The pseudonymization of privacy-sensitive information has almost exclusively been a topic of research in the medical NLP domain. However, privacy concerns also become increasingly relevant for user-generated content spread over social media. Our investigation opens this study field for the first time ever to a non-medical application, the de-identification of email corpora.
We distinguish two steps in this process. First, privacy-relevant information pieces have to be identified. We treat this task primarily as a named entity recognition problem and devise an entity hierarchy that captures the relevant named entity types for the de-identification of emails.
We evaluated several system architectures for the task of automatically recognizing these entities on CodE

[23] The code for the generation of surrogates is available at `https://github.com/ee-2/SurrogateGeneration`.

	# emails	# sentences	# tokens	# tokens*	# types*	# lemmas*
Finance	174,271	4,668,397	32,448,628	16,993,163	1,336,645	1,292,513
German	240,703	1,987,861	15,835,884	6,781,721	784,992	737,933
Movies	205,856	3,153,523	26,386,837	12,527,035	1,001,364	952,693
Philosophy	209,322	4,077,644	43,496,863	16,770,793	1,160,975	1,097,913
Teens	238,977	2,401,965	12,532,029	5,716,143	580,081	551,368
Travels	154,040	1,813,318	15,404,633	7,462,448	744,206	711,323
Events	246,300	3,232,851	28,260,682	11,763,229	961,800	910,021
$\sum$ CODE ALLTAG$_{pXL}$	1,469,469	21,335,559	174,365,556	78,014,532	4,506,550	4,411,765
CODE ALLTAG$_{pS+d}$	800	7,975	87,384	33,483	13,925	12,146

Table 9: Quantitative breakdown of CODE ALLTAG$_{pXL}$ and CODE ALLTAG$_{pS+d}$; '#' stands for 'number of'; '*' denotes punctuation marks and stop words excluded

Entity	*Finance*	*German*	*Movies*	*Philosophy*	*Teens*	*Travels*	*Events*	$\sum$ $_{pXL}$	$_{pS+d}$
ORG	210,617	26,988	89,987	34,974	17,652	69,288	93,508	543,014	336
FAMILY	225,006	264,255	344,185	299,713	296,528	121,282	298,906	1,849,875	954
FEMALE	64,913	56,763	85,914	55,005	79,145	48,889	48,849	439,478	831
MALE	296,171	383,742	510,450	367,763	400,244	234,425	367,092	2,559,887	507
USER	351	67	226	213	44	1,872	3,003	5,776	2
DATE	369,014	103,556	104,375	68,248	65,520	51,203	114,816	876,732	282
PASS	14	3	8	22	3	22	31	103	1
UFID	89,915	4,111	9,956	4,678	5,444	4,668	5,665	124,437	91
STREET	15,351	5,706	4,970	4,986	4,914	6,292	12,827	55,046	96
STREETNO	40,258	1,378	4,239	4,174	1,527	3,850	9,626	65,052	108
CITY	197,078	57,373	66,175	29,628	20,012	152,302	82,482	605,050	486
ZIP	16,307	610	3,939	2,634	1,132	3,355	6,728	34,705	74
EMAIL	67,333	21,235	63,321	70,533	121,479	42,248	40,765	426,914	56
PHONE	51,858	6,837	11,245	5,515	4,830	9,256	6,777	96,318	139
URL	103,994	137,082	95,305	69,295	45,438	89,082	134,610	674,806	141
$\sum$	1,748,180	1,069,706	1,394,295	1,017,381	1,063,912	838,034	1,225,685	8,357,193	4,104

Table 10: pi entities in numbers for each category of CODE ALLTAG$_{pXL}$ as well as for the entire CODE ALLTAG$_{pXL}$ ($_{pXL}$) and CODE ALLTAG$_{pS+d}$ ($_{pS+d}$)

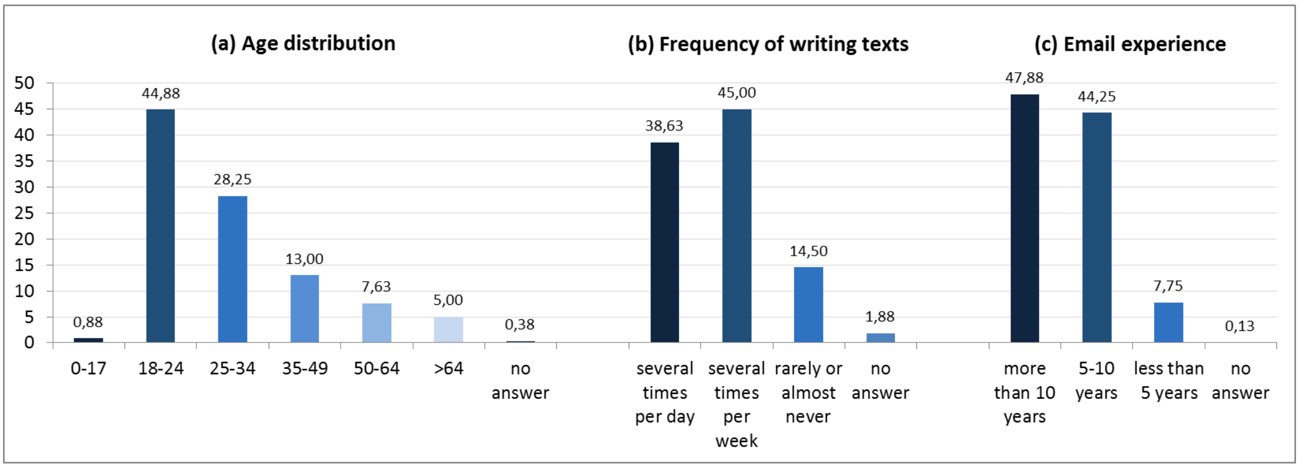


Figure 3: Distribution of email donors (CODE ALLTAG$_{pS+d}$) by age (a), frequency of writing texts (b) and email experience in years (c)

ALLTAG$_{S+d}$, a small, yet well-curated German email corpus, with roughly 1k emails and applied the best performing model to CODE ALLTAG$_{XL}$, another German email corpus, with roughly 1,5M emails. The recognition results for the latter corpus are degraded due to a high noise level and OOV problems. Hence, determining a more robust recognition model will require further work.

In a second step, we transform privacy-bearing entities from their original form into an entity type-preserving variant and thus create pseudonym forms. We finally come up with the pseudonymized German email corpus CODE ALLTAG 2.0 which is composed of CODE ALLTAG$_{pS+d}$ and CODE ALLTAG$_{pXL}$. This corpus is available at `https://github.com/codealltag`.

Acknowledgments

We thank the donors of the emails, and especially the annotators, for their efforts and support. This work was funded for the first two authors by the Austrian HRSM project *Kompetenznetzwerk Digitale Edition* (KONDE).

8. Bibliographical References

Adar, E. (2007). User 4xxxxx9: anonymizing query logs. In Einat Amitay, et al., editors, *Proceedings of the Workshop on Query Log Analysis: Social and Technological Challenges @ WWW 2007. Banff, Alberta, Canada, May 8, 2017.*

Akbik, A., Blythe, D., and Vollgraf, R. (2018). Contextual string embeddings for sequence labeling. In Pierre Isabelle, et al., editors, *COLING 2018 — Proceedings of the 27th International Conference on Computational Linguistics: Main Conference. Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 1638–1649.

Alfalahi, A., Brissman, S., and Dalianis, H. (2012). Pseudonymisation of personal names and other PHIs in an annotated clinical Swedish corpus. In Sophia Ananiadou, et al., editors, *BioTxtM 2012 — Proceedings of the 3rd Workshop on Building and Evaluating Resources for Biomedical Text Mining @ LREC 2012. Istanbul, Turkey, May 26, 2012*, pages 49–54.

Benikova, D., Yimam, S. M., Santhanam, P., and Biemann, C. (2015). GERMANER: free open German Named Entity Recognition tool. In Bernhard Fisseni, et al., editors, *GSCL 2015 — Proceedings of the International Conference of the German Society for Computational Linguistics and Language Technology. Duisburg-Essen, Germany, September 30 - October 2, 2015*, pages 31–38.

Boufaden, N., Elazmeh, W., Ma, Y., Matwin, S., El-Kadri, N., and Japkowicz, N. (2005). PEEP: an information extraction base approach for privacy protection in email. In *CEAS 2005 — Proceedings of the 2nd Conference on Email and Anti-Spam. Stanford, California, USA, July 21-22, 2005.*

Bui, D., Redden, D. T., and Cimino, J. J. (2018). Is multiclass automatic text de-identification worth the effort? *Methods of Information in Medicine*, 57(4):177–184.

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, March.

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 Lucila Ohno-Machado et al., editors, *MEDINFO 2019 — Proceedings of the 17th World Congress on Medical and Health Informatics: Health and Wellbeing e-Networks for All. Lyon, France, 25-30 August 2019*, number 264 in Studies in Health Technology and Informatics, pages 70–73, Amsterdam, Berlin, Tokyo, Washington, D.C. IOS Press.

Choi, E., Biswal, S., Malin, B. A., Duke, J., Stewart, W. F., and Sun, J. (2017). Generating multi-label discrete patient records using generative adversarial networks. In Finale Doshi-Velez, et al., editors, *MLHC 2017 — Proceedings of the 2nd Machine Learning for Health Care Conference. Boston, Massachusetts, USA, 18-19 August 2017*, pages 286–305.

Coavoux, M., Narayan, S., and Cohen, S. B. (2018). Privacy-preserving neural representations of text. In Ellen Riloff, et al., editors, *EMNLP 2018 — Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium, October 31 - November 4, 2018*, pages 1–10, Stroudsburg/PA. Association for Computational Linguistics (ACL).

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

De Choudhury, M. and De, S. (2014). Mental health discourse on Reddit: self-disclosure, social support, and anonymity. In Eytan Adar, et al., editors, *ICWSM 2014 — Proceedings of the 8th International AAAI Conference on Weblogs and Social Media. Ann Arbor, Michigan, USA, June 1-4, 2014*, pages 71–80, Palo Alto/CA.

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., and Szolovits, P. (2017a). NEURONER: an easy-to-use program for named-entity recognition based on neural networks. In Martha Palmer, et al., editors, *EMNLP 2017 — Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. Copenhagen, Denmark, September 9-11, 2017*, pages 97–102. Association for Computational Linguistics (ACL).

Dernoncourt, F., Lee, J. Y., Uzuner, O., and Szolovits, P. (2017b). 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 Jill Burstein, et al., editors, *NAACL-HLT 2019 — Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis, Minnesota, USA, June 2-7, 2019*, volume 1: Long and Short Papers, pages 4171–4186, Stroudsburg/PA. 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 Galia Angelova, et al., editors, *RANLP 2019 — Proceedings of the 12th International Conference on "Recent Advances in Natural Language Processing:" Natural Language Processing in a Deep Learning World. Varna, Bulgaria, 2-4 September, 2019*, pages 259–269, Shoumen, Bulgaria. Incoma Ltd.

Elazar, Y. and Goldberg, Y. (2018). Adversarial removal of demographic attributes from text data. In Ellen Riloff, et al., editors, *EMNLP 2018 — Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium, October 31 - November 4, 2018*, pages 11–21, Stroudsburg/PA. Association for Computational Linguistics (ACL).

Flick, C. (2016). Informed consent and the Facebook emotional manipulation study. *Research Ethics*, 12(1):14–28.

Friedrich, M., Köhn, A., Wiedemann, G., and Biemann, C. (2019). Adversarial learning of privacy-preserving text representations for de-identification of medical records. In Lluís Màrquez, et al., editors, *ACL 2019 — Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy, July 28 - August 2, 2019*, pages 5829–5839, Stroudsburg/PA. Association for Computational Linguistics (ACL).

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. C., and Bengio, Y. (2014). Generative adversarial nets. In Zoubin Ghahramani, et al., editors, *Advances in Neural Information Processing Systems 27 — NIPS 2014. Proceedings of the 28th Annual Conference on Neural Information Processing Systems 2014. Montréal, Québec, Canada, December 8-13, 2014*, pages 2672–2680.

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., and Mikolov, T. (2018). Learning word vectors for 157 languages. In Nicoletta Calzolari, et al., editors, *LREC 2018 — Proceedings of the 11th International Conference on Language Resources and Evaluation. Miyazaki, Japan, May 7-12, 2018*, pages 3483–3487, Paris. European Language Resources Association (ELRA).

Grodzinsky, F. and Tavani, H. T. (2010). Applying the "contextual integrity" model of privacy to personal blogs in the blogoshere. *International Journal of Internet Research Ethics*, 3(1):38–47, December.

Guan, J., Li, R., Yu, S., and Zhang, X. (2019). A method for generating synthetic electronic medical record text. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, page [Epub ahead of print], October. https://arxiv.org/pdf/1812.02793.

Heinzerling, B. and Strube, M. (2018). BPEMB: tokenization-free pre-trained subword embeddings in 275 languages. In Nicoletta Calzolari, et al., editors, *LREC 2018 — Proceedings of the 11th International Conference on Language Resources and Evaluation. Miyazaki, Japan, May 7-12, 2018*, pages 2989–2993, Paris. European Language Resources Association (ELRA).

Heinzerling, B. and Strube, M. (2019). Sequence tagging with contextual and non-contextual subword representations: a multilingual evaluation. In Lluís Màrquez, et al., editors, *ACL 2019 — Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy, July 28 - August 2, 2019*, pages 273–291, Stroudsburg/PA. Association for Computational Linguistics (ACL).

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.

Jones, R., Kumar, R., Pang, B., and Tomkins, A. D. (2007). "I know what you did last summer": query logs and user privacy. In *CIKM 2007 — Proceedings of the 16th ACM Conference on Information and Knowledge Management. Lisbon, Portugal, 6-10 November 2007*, pages 909–914, New York/NY. Association for Computing Machinery (ACM).

Jung, Y., Stratos, K., and Carloni, L. P. (2015). LN-ANNOTE: an alternative approach to information extraction from emails using locally-customized named-entity recognition. In Aldo Gangemi, et al., editors, *WWW 2015 — Proceedings of the 24th International Conference on World Wide Web. Florence, Italy, May 18–22, 2015*, pages 538–548, New York/NY. 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 Nicoletta Calzolari, et al., editors, *LREC 2018 — Proceedings of the 11th International Conference on Language Resources and Evaluation. Miyazaki, Japan, May 7-12, 2018*, pages 135–139, Paris. European Language Resources Association (ELRA).

Klimt, B. and Yang, Y. (2004). The ENRON corpus: a new dataset for email classification research. In Jean-François Boulicaut, et al., editors, *Machine Learning. ECML 2004 – Proceedings of the 15th European Conference on Machine Learning, Pisa, Italy, September 20-24, 2004*, number 3201 in Lecture Notes in Computer Science, pages 217–226, Berlin, Heidelberg. Springer.

Krieg-Holz, U., Schuschnig, C., Matthies, F., Redling, B., and Hahn, U. (2016). CODE ALLTAG: a German-language e-mail corpus. In Nicoletta Calzolari, et al., editors, *LREC 2016 — Proceedings of the 10th International Conference on Language Resources and Evaluation. Portorož, Slovenia, 23-28 May 2016*, pages 2543–2550, Paris. European Language Resources Association (ELRA-ELDA).

Lampert, A., Dale, R., and Paris, C. L. (2009). Segmenting email message text into zones. In *EMNLP 2009 — Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing. A Meeting of SIGDAT, a Special Interest Group of ACL @ ACL-IJCNLP 2009. Singapore, 6-7 August 2009*, pages 919–928, Stroudsburg/PA. Association for Computational Linguistics & Asian Federation of Natural Language Processing.

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., and Dyer, C. (2016). Neural architectures for named entity recognition. In Kevin C. Knight, et al., editors, *NAACL-HLT 2016 — Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego, California, USA, June 12-17, 2016*, pages 260–270. Association for Computational Linguistics (ACL).

Lee, J. Y., Dernoncourt, F., Uzuner, O., and Szolovits, P. (2016). Feature-augmented neural networks for patient note de-identification. In Anna Rumshisky, et al., editors, *ClinicalNLP 2016 — Proceedings of the Clinical Natural Language Processing Workshop @ COLING 2016. Osaka, Japan, December 11, 2016*, pages 17–22.

Lee, S. H. (2018). Natural language generation for electronic health records. *npj Digital Medicine*, 1(1):#63.

Li, Y., Baldwin, T., and Cohn, T. (2018). Towards robust and privacy-preserving text representations. In Claire Cardie, et al., editors, *ACL 2018 — Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne, Victoria, Australia, July 15-20, 2018*, volume 2: Short Papers, pages 25–30, Stroudsburg/PA. 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, November.

Lüngen, H., Beißwenger, M., Herzberg, L., and Pichler, C. (2017). Anonymisation of the Dortmund Chat Corpus 2.1. In Egon W. Stemle et al., editors, *cmccorpora17 — Proceedings of the 5th Conference on CMC and Social Media Corpora for the Humanities. Bolzano, Italy, October 3-4, 2017*, pages 21–24.

Medlock, B. (2006). An introduction to NLP-based textual anonymisation. In Nicoletta Calzolari, et al., editors, *LREC 2006 — Proceedings of the 5th International Conference on Language Resources and Evaluation. Genoa, Italy, 22-28 May, 2006*, 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 Ildikó Pilán, et al., editors, *NLP4CALL 2018 — Proceedings of the 7th Workshop on NLP for Computer Assisted Language Learning. Stockholm, Sweden, 7 November 2018*, number 152 in Linköping Electronic Conference Proceedings, pages 47–56.

Meystre, S. M., Friedlin, F. J., South, B. R., Shen, S., and Samore, M. H. (2010). Automatic de-identification of textual documents in the electronic health record: a review of recent research. *BMC Medical Research Methodology*, 10:#70.

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 International Publishing.

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. Vancouver, British Columbia, Canada, 6-8 October 2005*, pages 443–450. 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, March.

Pan, X., Zhang, B., May, J., Nothman, J., Knight, K. C., and Ji, H. (2017). Cross-lingual name tagging and linking for 282 languages. In Chris Callison-Burch, et al., editors, *ACL 2017 — Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver, British Columbia, Canada, July 30 - August 4, 2017*, volume 1: Long Papers, pages 1946–1958, Stroudsburg/PA. Association for Computational Linguistics (ACL).

Pantazos, K., Lauesen, S., and Lippert, S. (2011). De-identifying an EHR database: anonymity, correctness and readability of the medical record. In Anne Moen, et al., editors, *MIE 2011 — Proceedings of the 23rd Conference of the European Federation of Medical Informatics: User Centred Networked Health Care. Oslo, Norway, August 28-31, 2011*, number 169 in Studies in Health Technology and Informatics, pages 862–866, Amsterdam, Berlin, Tokyo, Washington, D.C. IOS Press.

Patel, N., Accorsi, P., Inkpen, D. Z., Lopez, C., and Roche, M. (2013). Approaches of anonymisation of an SMS corpus. In *CICLing 2013 — Proceedings of the 14th International Conference on Computational Linguistics and Intelligent Text Processing. Karlovasi, Samos, Greece, March 24-30, 2013*, pages 77–88. Springer.

Proisl, T. and Uhrig, P. (2016). SoMaJo: state-of-the-art tokenization for German Web and social media texts. In Paul Cook, et al., editors, *WAC-X — Proceedings of the 10th Web as Corpus Workshop and the EmpiriST Shared Task @ ACL 2016. Berlin, Germany, August 12, 2016*, pages 57–62, Stroudsburg/PA. Association for Computational Linguistics (ACL).

Riedl, M. and Padó, S. (2018). A named entity recognition shootout for German. In Claire Cardie, et al., editors, *ACL 2018 — Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne, Victoria, Australia, July 15-20, 2018*, volume 2: Short Papers, pages 120–125, Stroudsburg/PA. Association for Computational Linguistics (ACL).

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

Scantamburlo, T. and Pelillo, M. (2016). Contextualizing privacy in the context of data science. In Laurence Devillers, et al., editors, *ETHI-CA2 2016 — Proceedings of the Workshop on ETHics In Corpus Collection, Annotation & Application @ LREC 2016. Portorož, Slovenia, 24 May 2016*, pages 1–7, Paris. European Language Resources Association (ELRA).

Schröck, J. and Lüngen, H. (2015). Building and annotating a corpus of German-language newsgroups. In Michael Beißwenger et al., editors, *NLP4CMC 2015 — Proceedings of the 2nd Workshop on Natural Language Processing for Computer-Mediated Communication / Social Media. Essen, Germany, September 29, 2015*, pages 17–22. German Society for Computational Linguistics & Language Technology (GSCL).

Sennrich, R., Haddow, B., and Birch, A. (2016). Neural machine translation of rare words with subword units. In Antal P. J. van den Bosch, et al., editors, *ACL 2016 — Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin, Germany, August 7-12, 2016*, volume 1: Long Papers, pages 1715–1725, Stroudsburg/PA. Association for Computational Linguistics (ACL).

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., and Tsujii, J. (2012). BRAT: a Web-based tool for NLP-assisted text annotation. In Frédérique Segond, editor, *EACL 2012 — Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics: Demonstrations. Avignon, France, April 25-26, 2012*, pages 102–107, Stroudsburg/PA. Association for Computational Linguistics (ACL).

Stubbs, A. and Uzuner, O. (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, O. (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, O., 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 International Publishing.

Stubbs, A., Filannino, M., and Uzuner, O. (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 James J. Cimino, editor, *AMIA '96 — Proceedings of the 1996 AMIA Annual Fall Symposium (formerly SCAMC). Beyond the Superhighway: Exploiting the Internet with Medical Informatics. Washington, D.C., USA, October 26-30, 1996*, pages 333–337. Hanley & Belfus.

Sweeney, L. (2000). Simple demographics often identify people uniquely. Technical Report LIDAP-WP3, Laboratory for International Data Privacy, Carnegie Mellon University, Pittsburgh/PA.

Sweeney, L. (2002). k-anonymity: a model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-based Systems*, 10(5):557–570.

Thomas, D., Pastrana, S., Hutchings, A., Clayton, R., and Beresford, A. (2017). Ethical issues in research using datasets of illicit origin. In Steve Uhlig, et al., editors, *IMC '17 — Proceedings of the 2017 ACM Internet Measurement Conference. London, United Kingdom, November 1-3, 2017*, pages 445–462, New York/NY. Association for Computing Machinery (ACM).

Treurniet, M., De Clercq, O., van den Heuvel, H., and Oostdijk, N. (2012). Collecting a corpus of Dutch SMS. In Nicoletta Calzolari, et al., editors, *LREC 2012 — Proceedings of the 8th International Conference on Language Resources and Evaluation. Istanbul, Turkey, May 21-27, 2012*, pages 2268–2273, Paris. European Language Resources Association (ELRA).

Uzuner, O., 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.

Wilson, S., Schaub, F., Dara, A. A., Liu, F., Cherivirala, S., Leon, P. G., Andersen, M. S., Zimmeck, S., Sathyendra, K. M., Russell, N. C., Norton, T. B., Hovy, E. H., Reidenberg, J., and Sadeh, N. (2016). The creation and analysis of a website privacy policy corpus. In Antal P. J. van den Bosch, et al., editors, *ACL 2016 — Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin, Germany, August 7-12, 2016*, volume 1: Long Papers, pages 1330–1340. Association for Computational Linguistics (ACL).

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.