A novel approach on de-identification of data preserving semantic links

Fabian Singhofer
fabian.singhofer@uni-ulm.de
Ulm University

1 MOTIVATION

Data collected by organizations (e.g. hospitals, retailers, telecommunication providers, etc.) can be valuable for researchers since recent advances in machine learning allow to gain new insights from those datasets. This data can appear in different forms, where relational tables and unstructured texts are popular. However, in many cases data to be shared contains Personally Identifiable Information (PII) which do require to be anonymized in order to comply with privacy regulations like the Health Insurance Portability and Accountability Act (HIPAA) for medical records in the United States or the General Data Protection Regulation (GDPR) in the European Union.

Structured datasets of organisations do not only contain of numerical, categorial, or string based fields, but in many cases also include free text fields. In order for such datasets to be shared, it is important to consider all fields with regard to their risk of identity disclosure. The Safe Harbor policy according to HIPAA defines 18 specific types of fields which have to be removed from health data in order to comply with this policy (see Appendix A) [35]. In contrast, other institutions such as the National Institute of Standards and Technology (NIST) provide a more general view on fields to be de-identified [24]. Common fields which need to be considered for de-identification and have been investigated in prior work are name, age, email address, gender, sex, street address, ZIP, or any other identifying numbers among others [6, 9, 20, 22, 33].

Moreover, keeping the actual semantic context within the dataset is important and increases the utility of the dataset for further analysis. Therefore, de-identification of such data remains a difficult challenge where reducing the risk of re-identification contradicts the utility of data.

1.1 Terminology

In general, attributes of a dataset can be distinguished in multiple categories with regard to their influence for de-identification [10]. The following categories can be applied to fields in structured data as well as to tokens within text. *Direct identifiers* are one or more attributes which can directly be used to identify an individual. Examples are names, email addresses, credit card numbers, or social security numbers. In contrast, *quasi-identifiers* are attributes which can not directly be used to identify an individual, but pose potential to help identifying a person if the adversary can use other public datasets or has background knowledge. Examples are age, gender, ethnic origin, or profession. *Sensitive attributes* are attributes which do not reveal an individual's identity, but depict critical information which should not leak to the adversary since it could harm data subjects. Examples are salary numbers or medical conditions. Finally, *insensitive attributes* are attributes which do not pose a risk to neither identity disclosure nor attribute disclosure. However, there also exist other definitions for categorizing fields. The GDPR introduces the term *personal data* and defines it as any information relating to an identified or identifiable natural person [3]. Furthermore, the NIST defines PIIs as any information that can be used to distinguish or trace an individual's identity (direct identifiers) and any information that is linked or linkable to an individual (quasi-identifiers) [24]. For the remainder of this work, we will use the terms direct identifier, quasi-identifier, sensitive attribute, and insensitive attribute.

Even though the concepts of anonymization, de-identification, and pseudonymization are often used interchangeably, there are some major differences [10]. De-identification is the process of masking or removing identifiers of a dataset in general. Anonymization describes the process of encrypting or deleting direct identifiers and quasi-identifiers within a dataset so that identities cannot be retrieved. Pseudonymization deals with replacing PII fields by artificial identifiers, which can later be used to retrieve a person's identity. However, for the ease of reading, we use de-identification and anonymization interchangeably.

1.2 Existing Anonymization Approaches

Recent work focussed on the task of de-identification of structured datasets by applying privacy concepts like k-anonymity [33], ℓ-diversity [23], and t-closeness [21]. The goal of k-anonymity is to prevent identity disclosure by grouping k data entries together such that their quasi-identifiers are indistinguishable. Higher values for k lead to better privacy since values in groups tend to become more general. However, the de-identified dataset might have decreased utility due to information loss. ℓ-diversity and t-closeness extend k-anonymity to prevent attribute disclosure. ℓ-diversity focuses on distributing sensitive attributes within a group such that an adversary can not gain information about sensitive attributes of individuals in the same group. t-closeness advances the concept of ℓ-diversity by guaranteeing that the distribution within groups does not differ more than a threshold t compared to the global distribution of sensitive values. Moreover, high effort has been conducted to develop systems which can automatically recognize PII within free texts using rule based approaches [25, 28, 30] or machine learning methods [4, 9, 15, 22] to allow for surrogate generation in the next step.

1.3 Our Contribution

Tasks of de-identifying structured data and unstructured texts have been investigated separately. Therefore, it would be interesting to develop a de-identification system which can incorporate methodologies from both fields in order to de-identify multimodal datasets.

This work aims to develop an automatic de-identification system which can be used to anonymize relational datasets containing free text fields. Our system achieves privacy by applying privacy models

such as k-anonymity, ℓ-diversity, and t-closeness to primitive attributes as well as to tokens within free text fields. Finally, we propose a de-identification system which tries to persist semantic meaning by keeping links within the dataset. In particular, this work tries to answer the following research questions:

RQ1. Given a structured dataset with free text fields, how can this dataset be de-identified?

RQ2. How can privacy concepts developed for structured data be applied on unstructured texts?

RQ3. How can a dataset be de-identified such that semantic links within this dataset are not broken?

2 RELATED WORK

Automatic de-identifications has been subject of research in the past. Research areas are distinguished into de-identification of structured data (i.e. data within tables) and unstructured data like free texts, images, or audio. Moreover, several toolings already exist which can be used to de-identify structured datasets as well as unstructured free text.

2.1 Structured Data

Early work focussed on de-identification of structured datasets by developing new privacy models. Sweeney [33] introduced the concept of k-anonymity. Based on this model Machanavajjhala et al. [23] introduced the model of ℓ-diversity which then has been extended by Li et al. [21] to build the model of t-closeness.

Several algorithms have been developed to efficiently apply those privacy concepts. Sweeney [32] proposed a greedy approach to achieve k-anonymity with tuple suppression. Moreover, LeFevre et al. [20] suggested a top-down greedy algorithm called Mondrian for implementing multidimensional k-anonymity. Additionally, Ghinita et al. [11] showed how optimal multidimensional k-anonymity can be achieved by reducing the problem to a one-dimensional problem which improves performance while reducing the information loss.

Multiple tools have been developed to apply these privacy models. *ARX*[1] is an open source software for anonymizing personal data within structured datasets [26, 27]. This tool supports multiple privacy and risk models, methods for transforming data, and concepts for analyzing the output data. Moreover, *Amnesia*[2] is a flexible data anonymization tool which allows to ensure privacy on relational data. *μ-ARGUS*[3] is a tool designed to create safe micro-data files and is based on the programming language R, which is specifically built to support statistical analyses. Commercial solutions include *CN-Protect*[4], which is a plugin for data science platforms that allows to protect privacy and sensitive datasets. Finally, *Privacy Analytics*[5] offers a commercial Eclipse plugin which can be used to anonymize structured data.

Besides toolings for de-identification of structured data, there also exist frameworks or modules to achieve this task. *python-datafly*[6] is a Python implementation of the well-established Datafly algorithm introduced by Sweeney [32] as one of the first algorithms to transfer structured data to match k-anonymity. Additionally, *Crowds*[7] is an open-source python module developed to de-identify a dataframe using the Optimal Lattice Anonymization (OLA) algorithm as proposed by El Emam et al. [6] to achieve k-anonymity. Finally, an example for an implementation of the Mondrian algorithm [20] is available for Python[8] to show how k-anonymity, ℓ-diversity, and t-closeness can be used as privacy models.

2.2 Unstructured Texts

For the de-identification of textual data, recent work used rule-based approaches as well as methods from machine learning. Sweeney [30] suggests a rule-based approach using dictionaries with specialized knowledge of the medical domain to detect Protected Health Informations (PHIs). Moreover, Ruch et al. [28] introduced a system for locating and removing PHIs within patient records using a semantic lexicon specialized in medicine. Additionally, Neamatullah et al. [25] introduced an automatic approach to de-identify medical records using lexical look-up tables, regular expressions and heuristics. They tested their systems using random samples of nursing notes from the MIMIC II database. All approaches achieved good results within their domain. However, those systems are hard to be transferred to other domains and are customized towards specific categories of texts (e.g. medical records or nursing reports).

Advances in machine learning lead to new approaches on de-identification of textual data. Gardner and Xiong [9] introduced an integrated system which uses Conditional Random Fields (CRF) to identify PII. Furthermore, Dernoncourt et al. [4] implemented a de-identification system with Recurrent Neural Networks (RNNs) achieving high scores in the 2014 Informatics for Integrating Biology and the Bedside (i2b2) challenge. Liu et al. [22] proposed a hybrid automatic de-identification system which incorporates subsystems which use concepts of rule-based systems and machine learning. For a full comparison of de-identification systems with regard to i2b2 challenges, please refer to the work of Yogarajan et al. [38]. Recent advances in transformer neural networks appear to also be applicable for de-identification. Yan et al. [37] suggested to use transformers for Named Entity Recognition (NER) tasks as an improvement to Bidirectional Long Short-Term Memory (BiLSTM) networks. In addition, Khan et al. [18] showed that transformers can be used for NER within the biomedical domain. Finally, Johnson et al. [15] were first to propose a de-identification system using transformers. Their results indicate that transformers are competitive to modern baseline models in case of de-identification of free text.

There are multiple tools and frameworks for de-identification of free text. *NLM-Scrubber*[9] is a freely available tool for de-identification of clinical texts according to Safe Harbor principles introduced in the HIPAA Privacy Rule. Moreover, the *MITRE Identification Scrubber Toolkit (MIST)*[10] is a suite of tools for identifying and redacting PII in free-text medical records [17]. *deid*[11] is a tool which allows

[1] https://arx.deidentifier.org/
[2] https://amnesia.openaire.eu/
[3] http://neon.vb.cbs.nl/casc/mu.htm
[4] https://cryptonumerics.com/cn-protect-for-data-science/
[5] https://privacy-analytics.com/health-data-privacy/
[6] https://github.com/alessiovierti/python-datafly
[7] https://github.com/leo-mazz/crowds
[8] https://github.com/Nuclearstar/K-Anonymity
[9] https://scrubber.nlm.nih.gov/
[10] http://mist-deid.sourceforge.net/
[11] https://www.physionet.org/content/deid/1.1/

anonymization of free texts within the medical domain. *deidentify*[12] is a Python library developed especially for de-identification of medical records and comparison of rule-, feature-, and deep-learning-based approaches for de-identification of free texts [34]. Finally, Friedrich et al. [7] introduced a way to transform medical records into a non-reversible vector representation and make their code available on GitHub[13].

2.3 Additional Data Formats

Even though this work only focuses on structured data and free text, recent work on anonymization of other forms of data is worth mentioning. For de-identification of images showing faces, Gross et al. [12] highlighted that pixelation and blurring offers poor privacy and suggested a model-based approach to protect privacy while preserving data utility. In contrast, recent work by Hukkelås et al. [13] applied methods from machine learning by implementing a simple Generative Adversarial Network (GAN) to generate new faces to preserve privacy while retaining original data distribution.

Regarding audio data, recent work focussed either on anonymization of the speaker's identity or the speech content. Justin et al. [16] suggested a framework which automatically transfers speech into an de-identified version using different acoustical models for recognition and synthesis. Moreover, Cohn et al. [2] investigated the task of de-identifying spoken text by first using Automatic Speech Recognition (ASR) to transcribe text, then extracting entities using NER, and finally aligning text elements to the audio and suppressing audio segments which should be de-identified.

Additionally, recent work by Agrawal and Narayanan [1] showed that de-identification of people can also be applied to whole bodies within videos whereas Gafni et al. [8] focussed on live de-identification of faces in video streams.

3 MATERIAL AND METHODS

3.1 Problem Statement

Given a structured dataset, previous work investigated how privacy can be obtained while preserving high utility. Moreover, recent advances in machine learning lead to multiple de-identification systems which can be used for free texts. Our work tries to bring both techniques together by building a de-identification system which can deal with free text fields within structured data.

Scenario: We illustrate the problem of de-identifying multimodal datasets containing structured data and free text fields using an imaginary dataset shown in Table 1. Our example depicts a dataset of an organization which stores customer data together with customer's feedback and contains seven typical fields. ID, name, age, gender, ZIP, and status are primitive fields whereas the feedback field stores free text. A naive approach of de-identifying this dataset would be to remove the feedback field and apply privacy concepts of structured data for the remaining fields. However, this does lead to losing valuable information. Keeping the feedback field and using anonymization techniques known from structured data poses the risk of identity and attribute disclosure since identifiers and sensitive attributes can still remain in the feedback field (e.g the age of Peter Smith). Moreover, the feedback field can also reveal identities or sensitive information on persons which do not actually appear within the structured data. An example would be Julia mentioning her sister Maria in her feedback. Finally, when anonymizing the free text, it is important to keep the dataset consistent and sound with regard to context and links. Therefore, if surrogates are used as placeholders, they need to be replaced in a sound way. For example the name Georg Ries appears in the name column as well as in the corresponding feedback text and needs to be treated equally in order for the data to remain consistent.

[12]https://github.com/nedap/deidentify
[13]https://github.com/maxfriedrich/deid-training-data

3.2 De-Identification Process

The typical process of de-identifying a dataset involves the following steps [14]:

(1) Assess the intended audience to decide on the release model. There exist public and non-public release models which have an impact on the risk of re-identification.
(2) Specify and name direct-identifiers, quasi-identifiers (also called semi-identifiers), and sensitive attributes. For structured data, each column of a dataset can directly be labeled as either one of the three previously mentioned types, or as an insensitive attribute. For unstructured free text, manual or automatic recognition of attribute types has to be done.
(3) Assess the risk of re-identification by analyzing the adversary and re-identification attacks possible.
(4) Calculate the amount of de-identification needed. This also includes the evaluation of de-identification techniques which should be applied (refer to Section 2).
(5) De-identify the dataset.
(6) Check the de-identified dataset on privacy, but also data utility. Metrics like precision, discernability metric, and non-uniform entropy [6] as well as Global Certainty Penalty (GCP) [11] can be used to measure the information loss.
(7) Document the process.

3.3 De-Identification Algorithms

Regarding the de-identification process above, we describe three well-known algorithms which can be used for Step 5 of the de-identification process. It is important to note that all algorithms can be configured such that types can be assigned to fields (Step 2) and results from Steps 3 and 4 can be used to parameterize the level of privacy within the algorithms.

To guarantee that re-identification is not possible given a dataset, it is not enough to just remove identifiers since linking of publicly available data has shown to make re-identification possible [31]. Therefore, a technique called k-anonymity was introduced by Sweeney [33]. Within a dataset, k-anonymity is met if each value of quasi-identifiers appears at least k times. Higher values for k guarantee better privacy. To achieve this, the concepts of generalization and suppression have been introduced [32]. Generalization describes the process of grouping specific values together to a more general group. In contrast, suppression focuses on reducing information by replacing certain values with asterisks. Grouping ages of persons is a popular example for generalization (e.g. $15 < age \leq 25$) whereas suppression is often used in ZIP codes (e.g. codes *02138, 02139* can be reduced to *0213**) [32]. Even though k-anonymity can be implemented rather easily, it does not prevent attacks such

Table 1: Imaginary dataset which contains numerical (age), categorial (gender), string (name), and free text fields (feedback).

ID	Name	Age	Gender	ZIP	Status	Feedback
12	Julia Brooks	35	Female	02432	Premium	I had a really bad phone connection the last time I spoke with Maria on the 23rd of August. She's living close by in Bridgeton. What is the reason?
23	Peter Smith	44	Male	02433	Basic	I'm 44 years old and with you since I moved to Greenberg. Could you finally upgrade me to premium.
56	Joe Small	25	Male	02452	Basic	Everything worked out nicely. Thanks!
77	Erica Lopez	52	Female	02312	Basic	I recently asked my sister to join you. We are looking forward to receive the 25 € reward.
89	Georg Ries	22	Male	42312	Basic	Best service available! Our company BestPrint appreciates your efforts. Best regards, Georg Ries.

as the *Homogeneity Attack* or the *Background Knowledge Attack* [23]. The Homogeneity Attack tries to gain knowledge about an individual residing in a homogenous equivalence class. Homogenous refers to the lack of diversity within sensitive attributes. The Background Knowledge Attack enables the adversary to deduce information about an individual by leveraging background knowledge using a combination of quasi-identifiers and sensitive attributes. To cope with those attacks, a modification to k-anonymity called ℓ-diversity has been introduced [23]. This principle states that if each equivalence class has at least ℓ "well-represented" values for each sensitive attribute, the table is said to be ℓ-diverse. The meaning of "well-represented" depends on which type of ℓ-diversity is implemented. Machanavajjhala et al. [23] defined three submodels for ℓ-diversity, namely distinct ℓ-diversity, entropy ℓ-diversity, and recursive (c,ℓ)-diversity. Distinct ℓ-diversity states that at least ℓ distinct values must exist for each equivalence class. Entropy ℓ-diversity guarantees that for each equivalence class, the entropy distribution is evenly enough. Recursive (c,ℓ)-diversity ensures that most common values do not appear too frequently. A drawback of ℓ-diversity is that similarity attacks are still possible and information can be leaked. Finally, t-closeness was introduced by Li et al. [21] to cope with limitations of ℓ-diversity. In order for a dataset to fulfill the t-closeness property, the distribution of sensitive attributes within an equivalence class must be close to the distribution of the whole dataset. If those distributions are close to each other, an adversary cannot gain any additional information from any equivalence class with respect to the whole dataset.

3.4 Methodology

This work aims to develop a de-identification system which combines methodologies of anonymizing structured data as well as free texts. We try to achieve this by answering three research questions which have been introduced in Section 1. In particular, developing a solution to answer each research question defines a milestone for this work.

RQ1 poses the question how structured datasets with free text fields can be de-identified in general. A naive approach to solve this problem is to see this de-identification task as two separate problems, where in the first step all free text fields are de-identified and afterwards k-anonymity is applied to the remaining fields. For de-identification of free text, first privacy violating tokens need to be identified. This can be achieved using rule- and dictionary based approaches, or methods from machine learning like CRFs, Long-Short Term Memory (LSTM) cells, or transformer networks. A combined approach is preferable [22]. Those tokens can then either be removed, blanked out, or replaced by surrogates. In all cases, privacy is preserved. However, the method of choice can have an impact on the utility of the text. After the free text is anonymized, privacy concepts for structured data can be applied on the remaining fields. The resulting dataset is then de-identified, but might lack consistency. An example of how a de-identified dataset could look like is shown in Table 2. There, the imaginary dataset introduced in Table 1 has been de-identified by removing direct identifiers, implementing 2-anonymity for quasi-identifiers, and replacing privacy violating tokens within the feedback field by their types. Since the last entry of the dataset cannot be grouped with other entries to match 2-anonymity, this entry is omitted.

RQ2 addresses the question how privacy concepts known to work for structured data (k-anonymity, ℓ-diversity, t-closeness) can be used on free texts. Given that quasi-identifiers can be identified in free texts, there might be no need to remove them if the anonymized dataset fulfills k-anonymity, ℓ-diversity, or t-closeness. In order to apply those concepts, tokens within free text can be treated like fields in a structured dataset. However, techniques like generalization or suppression might not be appropriate for free texts and therefore alternative approaches need to be investigated. Moreover, both de-identification problems need to be solved together since attributes like age or gender can appear in primitive fields as well as in free texts.

RQ3 raises the question whether it is possible to de-identify datasets while preserving semantic links between fields of the dataset and tokens within the text. To achieve this, a straightforward approach is to create a map where direct identifiers (e.g. names) are mapped to appropriate surrogates and all occurrences within columns or texts are replaced according to this mapping. However, links might even exist if tokens do not match exactly. Therefore, we will use concepts of Natural Language Processing (NLP) to determine links within a dataset.

4 GOALS AND LIMITATIONS

4.1 Mandatory Goals

- Develop a de-identification system which can cope with structured datasets containing free text fields.
- Evaluate the developed system regarding precision, recall, and F1-measure.

Table 2: Anonymized version of the imaginary dataset implementing 2-anonymity and de-identified free texts using token types as placeholders.

ID	Name	Age	Gender	ZIP	Status	Feedback
—	—	30 - 60	Female	02***	Premium	I had a really bad phone connection the last time I spoke with **name** on the **date**. She's living close by in **location**. What is the reason?
—	—	20 - 50	Male	024**	Basic	I'm **age** years old and with you since I moved to **location**. Could you finally upgrade me to **status**.
—	—	20 - 50	Male	024**	Basic	Everything worked out nicely. Thanks!
—	—	30 - 60	Female	02***	Basic	I recently asked my sister to join you. We are looking forward to receive the **money** reward.

- Propose methods how privacy concepts (k-anonymity) can be applied to different kinds of data fields including texts and their tokens.
- Show generalization capabilities by de-identifying at least two different publicly available datasets.
- Keep logs on changes and actions taken for de-identification to document the process.

4.2 Optional Goals

- Incorporate ℓ-diversity and t-closeness within the system to prevent attribute disclosure.
- Evaluate the utility of the anonymized dataset.
- Design system to be able to cope with multilingual datasets.

4.3 Limitations

- Size and complexity of the de-identification system are not taken into consideration.
- Performance on large datasets is not evaluated.
- Image, audio, or video data is not considered to be de-identified.

5 DATASETS

Datasets for the de-identification task can be grouped into two categories, namely structured and unstructured datasets. For anonymizing structured data any arbitrary dataset can be used which incorporates direct identifiers and quasi-identifiers. One real and popular dataset used to evaluate k-anonymity algorithms is the Adult dataset from the UCI Machine Learning Repository[14]. This dataset contains direct identifiers like names, quasi-identifiers like age and gender, and sensitive attributes like workclass and education, but lacks free text fields.

Moreover, for unstructured data, several datasets exist from the medical domain. Two commonly used datasets were introduced by i2b2. Those datasets are the 2006 de-identification challenge of clinical records [36] and the 2014 i2b2/UTHealth corpus [29]. However, Eder et al. [5] criticizes that de-identification systems are only evaluated using medical texts. They generated an anonymized German email corpus and transferred the problem of de-identification out of the medical domain, but unfortunately the corpus with labels is not available.

Since both de-identification tasks have been treated separately in the past, to the best of our knowledge, no dataset incorporating both features we need exists. However, by merging textual datasets with structured data, we can generate an appropriate dataset to test our system. An example would be to take the structured Adult dataset and match it with medical records from the i2b2 challenges. Additionally, another dataset could be generated by incorporating emails from Enron Email corpus [19] to evaluate our system outside of the medical domain.

[14]https://archive.ics.uci.edu/ml/datasets/adult

6 ADDITIONAL FRAMEWORKS

Since the task of identifying PII in unstructured text can be compared to the task of tagging individual words, widely used frameworks for NER could possibly be applied. Popular frameworks to fulfill this task and which might be of interest for our implementation are NLTK[15] and spaCy[16]. Additionally, Faker[17] is a Python package which can be used to create unreal data for names, addresses, companies, among others and can be used for replacing identifiers with appropriate surrogates.

7 SCHEDULE

Given the timeframe of six months, Table 3 provides a rough overview on the schedule. During the implementation phase, methods for de-identification are implemented. Starting with two separate tasks for structured and unstructured data, the ultimate goal is to have a single de-identification system which combines both implementations. Using this implementation, experiments are conducted and our implementation is evaluated. During the course of this work but especially at the end of the timeframe, details on methodology, implementation, experiments, and results are documented.

Table 3: Estimated time schedule for implementation, experimental evaluation, and documentation of the to-be-developed de-identification system.

Task	Time Span
Implementation	2020-09-01 - 2020-12-31 (4 months)
Experiments (and Evaluation)	2020-12-01 - 2021-01-31 (2 month)
Finishing Paper and Supplementary Materials	2021-02-01 - 2021-02-28 (1 month)

[15]https://www.nltk.org/
[16]https://spacy.io/
[17]https://github.com/joke2k/faker

8 PRELIMINARY OUTLINE

8.1 Paper

1. Introduction
2. Related Work
3. Methods on Data Anonymization
   3.1. Structured Data
   3.2. Unstructured Data
   3.3. Combined Approach
4. Experimental Apparatus
   4.1. Procedure
   4.2. Datasets
   4.3. Measures and Metrics
5. Results
6. Discussion
7. Conclusion

8.2 Supplementary Material

1. Extended Related Work
2. Implementation
   2.1. Architecture Overview
   2.2. De-identification of Structured Data
   2.3. De-identification of Unstructured Data
   2.4. Combined approach
3. Experiments and Results
4. Discussion
5. Conclusion and Future Work

REFERENCES

[1] Prachi Agrawal and P. J. Narayanan. 2011. Person De-Identification in Videos. *IEEE Trans. Circuits Syst. Video Techn.* 21, 3 (2011), 299–310.

[2] Ido Cohn, Itay Laish, Genady Beryozkin, Gang Li, Izhak Shafran, Idan Szpektor, Tzvika Hartman, Avinatan Hassidim, and Yossi Matias. 2019. Audio De-identification - a New Entity Recognition Task. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 2 (Industry Papers)*, Anastassia Loukina, Michelle Morales, and Rohit Kumar (Eds.). Association for Computational Linguistics, 197–204.

[3] Council of European Union. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance).

[4] Franck Dernoncourt, Ji Young Lee, Ozlem Uzuner, and Peter Szolovits. 2016. De-identification of patient notes with recurrent neural networks. *Journal of the American Medical Informatics Association* (Dec. 2016), ocw156.

[5] Elisabeth Eder, Ulrike Krieg-Holz, and Udo Hahn. 2019. De-Identification of Emails: Pseudonymizing Privacy-Sensitive Data in a German Email Corpus. In *Proceedings - Natural Language Processing in a Deep Learning World*. Incoma Ltd., Shoumen, Bulgaria, 259–269.

[6] Khaled El Emam, Fida Kamal Dankar, Romeo Issa, Elizabeth Jonker, Daniel Amyot, Elise Cogo, Jean-Pierre Corriveau, Mark Walker, Sadrul Chowdhury, Regis Vaillancourt, Tyson Roffey, and Jim Bottomley. 2009. A Globally Optimal k-Anonymity Method for the De-Identification of Health Data. *Journal of the American Medical Informatics Association* 16, 5 (Sept. 2009), 670–682.

[7] Max Friedrich, Arne Köhn, Gregor Wiedemann, and Chris Biemann. 2019. Adversarial Learning of Privacy-Preserving Text Representations for De-Identification of Medical Records. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 5829–5839.

[8] Oran Gafni, Lior Wolf, and Yaniv Taigman. 2019. Live Face De-Identification in Video. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. IEEE, 9377–9386.

[9] James Gardner and Li Xiong. 2008. HIDE: An Integrated System for Health Information DE-identification. In *2008 21st IEEE International Symposium on Computer-Based Medical Systems*. IEEE, Jyvaskyla, Finland, 254–259.

[10] Simson L. Garfinkel. 2015. *De-identification of personal information.* Technical Report NIST IR 8053. National Institute of Standards and Technology. NIST IR 8053 pages.

[11] Gabriel Ghinita, Panagiotis Karras, Panos Kalnis, and Nikos Mamoulis. 2007. Fast Data Anonymization with Low Information Loss. In *Proceedings of the 33rd International Conference on Very Large Data Bases (VLDB '07)*. VLDB Endowment, 758–769. event-place: Vienna, Austria.

[12] Ralph Gross, Latanya Sweeney, Fernando De la Torre, and Simon Baker. 2006. Model-Based Face De-Identification. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2006, New York, NY, USA, 17-22 June, 2006*. IEEE Computer Society, 161.

[13] Håkon Hukkelås, Rudolf Mester, and Frank Lindseth. 2019. DeepPrivacy: A Generative Adversarial Network for Face Anonymization. In *Advances in Visual Computing - 14th International Symposium on Visual Computing, ISVC 2019, Lake Tahoe, NV, USA, October 7-9, 2019, Proceedings, Part I (Lecture Notes in Computer Science)*, George Bebis, Richard Boyle, Bahram Parvin, Darko Koracin, Daniela Ushizima, Sek Chai, Shinjiro Sueda, Xin Lin, Aidong Lu, Daniel Thalmann, Chaoli Wang, and Panpan Xu (Eds.), Vol. 11844. Springer, 565–578.

[14] Information and Privacy Commissioner of Ontario. 2016. *De-identification Guidelines for Structured Data.* Technical Report. https://www.ipc.on.ca/wp-content/uploads/2016/08/Deidentification-Guidelines-for-Structured-Data.pdf

[15] Alistair E. W. Johnson, Lucas Bulgarelli, and Tom J. Pollard. 2020. Deidentification of free-text medical records using pre-trained bidirectional transformers. In *Proceedings of the ACM Conference on Health, Inference, and Learning*. ACM, Toronto Ontario Canada, 214–221.

[16] Tadej Justin, Vitomir Struc, Simon Dobrisek, Bostjan Vesnicer, Ivo Ipsic, and France Mihelic. 2015. Speaker de-identification using diphone recognition and speech synthesis. In *11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, FG 2015, Ljubljana, Slovenia, May 4-8, 2015*. IEEE Computer Society, 1–7.

[17] Mehmet Kayaalp, Allen C. Browne, Zeyno A. Dodd, Pamela Sagan, and Clement J. McDonald. 2014. De-identification of Address, Date, and Alphanumeric Identifiers in Narrative Clinical Reports. *AMIA ... Annual Symposium proceedings. AMIA Symposium* 2014 (2014), 767–776.

[18] Muhammad Raza Khan, Morteza Ziyadi, and Mohamed AbdelHady. 2020. MT-BioNER: Multi-task Learning for Biomedical Named Entity Recognition using Deep Bidirectional Transformers. *arXiv:2001.08904 [cs, stat]* (Jan. 2020). arXiv: 2001.08904.

[19] Bryan Klimt and Yiming Yang. 2004. The Enron Corpus: A New Dataset for Email Classification Research. In *Machine Learning: ECML 2004*, David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Madhu Sudan, Demetri Terzopoulos, Dough Tygar, Moshe Y. Vardi, Gerhard Weikum, Jean-François Boulicaut, Floriana Esposito, Fosca Giannotti, and Dino Pedreschi (Eds.). Vol. 3201. Springer Berlin Heidelberg, Berlin, Heidelberg, 217–226. Series Title: Lecture Notes in Computer Science.

[20] K. LeFevre, D.J. DeWitt, and R. Ramakrishnan. 2006. Mondrian Multidimensional K-Anonymity. In *22nd International Conference on Data Engineering (ICDE'06)*. IEEE, Atlanta, GA, USA, 25–25.

[21] Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. 2007. t-Closeness: Privacy Beyond k-Anonymity and l-Diversity. In *2007 IEEE 23rd International Conference on Data Engineering*. IEEE, Istanbul, 106–115.

[22] Zengjian Liu, Buzhou Tang, Xiaolong Wang, and Qingcai Chen. 2017. De-identification of clinical notes via recurrent neural network and conditional random field. *Journal of Biomedical Informatics* 75 (Nov. 2017), S34–S42.

[23] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam. 2006. L-diversity: privacy beyond k-anonymity. In *22nd International Conference on Data Engineering (ICDE'06)*. IEEE, Atlanta, GA, USA, 24–24.

[24] Erika McCallister, Tim Grance, and Karen Scarfone. 2010. *Guide to Protecting the Confidentiality of Personally Identifiable Information (PII).* Technical Report. National Institute of Standards and Technology.

[25] Ishna Neamatullah, Margaret M Douglass, Li-wei H Lehman, Andrew Reisner, Mauricio Villarroel, William J Long, Peter Szolovits, George B Moody, Roger G Mark, and Gari D Clifford. 2008. Automated de-identification of free-text medical records. *BMC Medical Informatics and Decision Making* 8, 1 (Dec. 2008), 32.

[26] Fabian Prasser, Johanna Eicher, Helmut Spengler, Raffael Bild, and Klaus A. Kuhn. 2020. Flexible data anonymization using ARX—Current status and challenges ahead. *Software: Practice and Experience* 50, 7 (July 2020), 1277–1304.

[27] Fabian Prasser, Florian Kohlmayer, Ronald Lautenschläger, and Klaus A. Kuhn. 2014. ARX–A Comprehensive Tool for Anonymizing Biomedical Data. *AMIA ... Annual Symposium proceedings. AMIA Symposium* 2014 (2014), 984–993.

[28] P. Ruch, R. H. Baud, A. M. Rassinoux, P. Bouillon, and G. Robert. 2000. Medical document anonymization with a semantic lexicon. *Proceedings. AMIA Symposium* (2000), 729–733.

[29] Amber Stubbs and Özlem Uzuner. 2015. Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/UTHealth corpus. *Journal of Biomedical Informatics* 58 (Dec. 2015), S20–S29.

[30] L. Sweeney. 1996. Replacing personally-identifying information in medical records, the Scrub system. *Proceedings: a conference of the American Medical Informatics Association. AMIA Fall Symposium* (1996), 333–337.
[31] Latanya Sweeney. 2000. Simple Demographics Often Identify People Uniquely. *Data Privacy Working Paper 3* (2000).
[32] Latanya Sweeney. 2002. ACHIEVING k-ANONYMITY PRIVACY PROTECTION USING GENERALIZATION AND SUPPRESSION. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10, 05 (Oct. 2002), 571–588.
[33] Latanya Sweeney. 2002. k-ANONYMITY: A MODEL FOR PROTECTING PRIVACY. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10, 05 (Oct. 2002), 557–570.
[34] Jan Trienes, Dolf Trieschnigg, Christin Seifert, and Djoerd Hiemstra. 2020. Comparing Rule-based, Feature-based and Deep Neural Methods for De-identification of Dutch Medical Records. *arXiv:2001.05714 [cs]* (Jan. 2020). arXiv: 2001.05714.
[35] U.S. Department of Health & Human Services. 2015. *Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule.* Technical Report.
[36] O. Uzuner, Y. Luo, and P. Szolovits. 2007. Evaluating the State-of-the-Art in Automatic De-identification. *Journal of the American Medical Informatics Association* 14, 5 (Sept. 2007), 550–563.
[37] Hang Yan, Bocao Deng, Xiaonan Li, and Xipeng Qiu. 2019. TENER: Adapting Transformer Encoder for Named Entity Recognition. *arXiv:1911.04474 [cs]* (Dec. 2019). arXiv: 1911.04474.
[38] Vithya Yogarajan, Bernhard Pfahringer, and Michael Mayo. 2020. Automatic end-to-end De-identification: Is high accuracy the only metric? *Applied Artificial Intelligence* 34, 3 (Feb. 2020), 251–269. arXiv: 1901.10583.

A IDENTIFIERS

The following types of attributes are determined as identifiers according to the HIPAA Safe Harbor method and must be removed [35]:

(1) Names
(2) Geographic entities smaller than states (street address, city, county, ZIP, etc.)
(3) Dates (except year)
(4) Phone numbers
(5) Vehicle identifiers and serial numbers
(6) Fax numbers
(7) Device identifiers and serial numbers
(8) Email addresses
(9) URLs
(10) Social security numbers
(11) IP addresses
(12) Medical record numbers
(13) Biometric identifiers, including finger and voice prints
(14) Health plan beneficiary numbers
(15) Full-face photographs
(16) Account numbers
(17) Any other unique identifying number, characteristic, code, etc.
(18) Certificate and license numbers