# Adversarial Learning of Privacy-Preserving Text Representations for De-Identification of Medical Records

**Max Friedrich[1], Arne Köhn[2], Gregor Wiedemann[1], Chris Biemann[1]**

[1]Language Technology Group, Universität Hamburg
{2mfriedr,gwiedemann,biemann}@informatik.uni-hamburg.de

[2]Department of Language Science and Technology, Saarland University
koehn@coli.uni-saarland.de

## Abstract

De-identification is the task of detecting protected health information (PHI) in medical text. It is a critical step in sanitizing electronic health records (EHRs) to be shared for research. Automatic de-identification classifiers can significantly speed up the sanitization process. However, obtaining a large and diverse dataset to train such a classifier that works well across many types of medical text poses a challenge as privacy laws prohibit the sharing of raw medical records. We introduce a method to create privacy-preserving shareable representations of medical text (i.e. they contain no PHI) that does not require expensive manual pseudonymization. These representations can be shared between organizations to create unified datasets for training de-identification models. Our representation allows training a simple LSTM-CRF de-identification model to an $F_1$ score of $97.4\%$, which is comparable to a strong baseline that exposes private information in its representation. A robust, widely available de-identification classifier based on our representation could potentially enable studies for which de-identification would otherwise be too costly.

## 1 Introduction

Electronic health records (EHRs) are are valuable resource that could potentially be used in large-scale medical research (Botsis et al., 2010; Birkhead et al., 2015; Cowie et al., 2017). In addition to structured medical data, EHRs contain free-text patient notes that are a rich source of information (Jensen et al., 2012). However, due to privacy and data protection laws, medical records can only be shared and used for research if they are sanitized to not include information potentially identifying patients. The PHI that may not be shared includes potentially identifying information such as names, geographic identifiers, dates, and account

numbers; the American Health Insurance Portability Accountability Act[1] (HIPAA, 1996) defines 18 categories of PHI. De-identification is the task of finding and labeling PHI in medical text as a step toward sanitization. As the information to be removed is very sensitive, sanitization always requires final human verification. Automatic de-identification labeling can however significantly speed up the process, as shown for other annotation tasks in e.g. Yimam (2015).

Trying to create an automatic classifier for de-identification leads to a "chicken and egg problem" (Uzuner et al., 2007): without a comprehensive training set, an automatic de-identification classifier cannot be developed, but without access to automatic de-identification, it is difficult to share large corpora of medical text in a privacy-preserving way for research (including for training the classifier itself). The standard method of data protection compliant sharing of training data for a de-identification classifier requires humans to pseudonymize protected information with substitutes in a document-coherent way. This includes replacing e.g. every person or place name with a different name, offsetting dates by a random amount while retaining date intervals, and replacing misspellings with similar misspellings of the pseudonym (Uzuner et al., 2007).

In 2019, a pseudonymized dataset for de-identification from a single source, the *i2b2 2014* dataset, is publicly available (Stubbs and Uzuner, 2015). However, de-identification classifiers trained on this dataset do not generalize well to data from other sources (Stubbs et al., 2017). To obtain a universal de-identification classifier, many medical institutions would have to pool their data. But, preparing this data for sharing using the document-coherent pseudonymization ap-
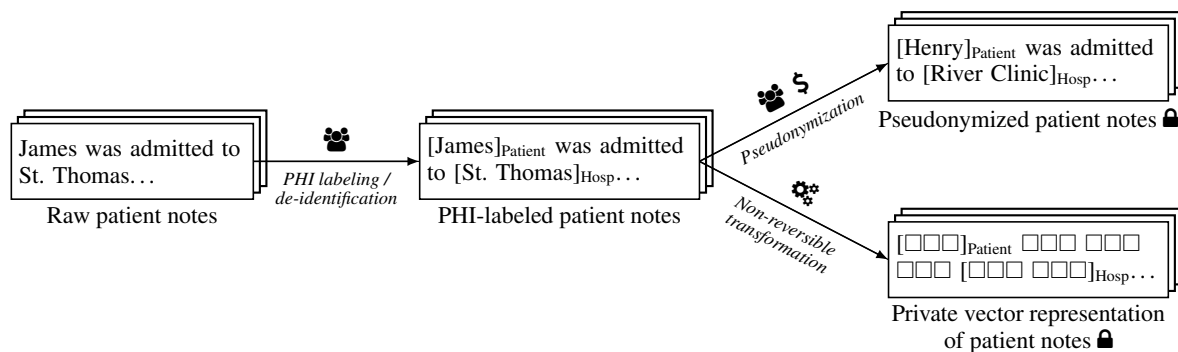
---

[1]https://legislink.org/us/pl-104-191

Figure 1: Sharing training data for de-identification. PHI annotations are marked with [brackets]. Upper alternative: traditional process using manual pseudonymization. Lower alternative: our approach of sharing private vector representations. The people icon represents tasks done by humans; the gears icon represents tasks done by machines; the lock icon represents privacy-preserving artifacts. Manual pseudonymization is marked with a dollar icon to emphasize its high costs.

proach requires large human effort (Dernoncourt et al., 2017).

To address this problem, we 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 vector space and only sharing this representation. Our approach still requires humans to annotate PHI (as this is the training data for the actual de-identification task) but the pseudonymization step (replacing PHI with coherent substitutes) is replaced by the automatic transformation to the vector representation instead. A classifier then trained on our representation cannot contain any protected data, as it is never trained on raw text (as long as the representation does not allow for the reconstruction of sensitive information). The traditional approach to sharing training data is conceptually compared to our approach in Fig. 1.

## 2 Related Work

Our work builds upon two lines of research: firstly de-identification, as the system has to provide good de-identification performance, and secondly adversarial representation learning, to remove all identifying information from the representations to be distributed.

### 2.1 Automatic De-Identification

Analogously to many natural language processing tasks, the state of the art in de-identification changed in recent years from rule-based systems and shallow machine learning approaches like conditional random fields (CRFs) (Uzuner et al.,

2007; Meystre et al., 2010) to deep learning methods (Stubbs et al., 2017; Dernoncourt et al., 2017; Liu et al., 2017).

Three i2b2 shared tasks on de-identification were run in 2006 (Uzuner et al., 2007), 2014 (Stubbs et al., 2015), and 2016 (Stubbs et al., 2017). The organizers performed manual pseudonymization on clinical records from a single source to create the datasets for each of the tasks. An $F_1$ score of $95\%$ has been suggested as a target for reasonable de-identification systems (Stubbs et al., 2015).

Dernoncourt et al. (2017) first applied a long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997) model with a CRF output component to de-identification. Transfer learning from a larger dataset slightly improves performance on the i2b2 2014 dataset (Lee et al., 2018). Liu et al. (2017) achieve state-of-the-art performance in de-identification by combining a deep learning ensemble with a rule component.

Up to and including the 2014 shared task, the organizers emphasized that it is unclear if a system trained on the provided datasets will generalize to medical records from other sources (Uzuner et al., 2007; Stubbs et al., 2015). The 2016 shared task featured a sight-unseen track in which de-identification systems were evaluated on records from a new data source. The best system achieved an $F_1$ score of $79\%$, suggesting that systems at the time were not able to deliver sufficient performance on completely new data (Stubbs et al., 2017).

## 2.2 Adversarial Representation Learning

Fair representations (Zemel et al., 2013; Hamm, 2015) aim to encode features of raw data that allows it to be used in e.g. machine learning algorithms while obfuscating membership in a protected group or other sensitive attributes. The domain-adversarial neural network (DANN) architecture (Ganin et al., 2016) is a deep learning implementation of a three-party game between a representer, a classifier, and an adversary component. The classifier and the adversary are deep learning models with shared initial layers. A gradient reversal layer is used to worsen the representation for the adversary during back-propagation: when training the adversary, the adversary-specific part of the network is optimized for the adversarial task but the shared part is updated against the gradient to make the shared representation less suitable for the adversary.

Although initially conceived for use in domain adaptation, DANNs and similar adversarial deep learning models have recently been used to obfuscate demographic attributes from text (Elazar and Goldberg, 2018; Li et al., 2018) and subject identity (Feutry et al., 2018) from images. Elazar and Goldberg (2018) warn that when a representation is learned using gradient reversal methods, continued adversary training on the frozen representation may allow adversaries to break representation privacy. To test whether the unwanted information is not extractable from the generated information anymore, adversary training needs to continue on the frozen representation after finishing training the system. Only if after continued adversary training the information cannot be recovered, we have evidence that it really is not contained in the representation anymore.

## 3 Dataset and De-Identification Model

We evaluate our approaches using the i2b2 2014 dataset (Stubbs and Uzuner, 2015), which was released as part of the 2014 i2b2/UTHealth shared task track 1 and is the largest publicly available dataset for de-identification today. It contains 1304 free-text documents with PHI annotations. The i2b2 dataset uses the 18 categories of PHI defined by HIPAA as a starting point for its own set of PHI categories. In addition to the HIPAA set of categories, it includes (sub-)categories such as doctor names, professions, states, countries, and ages under 90.

| Hyperparameter | Value |
| --- | --- |
| Pre-trained embeddings | FastText, GloVe |
| Casing feature | Yes |
| Batch size | 32 |
| Number of LSTM layers | 2 |
| LSTM units per layer/dir. | 128 |
| Input embedding dropout | 0.1 |
| Variational dropout | 0.25 |
| Dropout after LSTM | 0.5 |
| Optimizer | Nadam |
| Gradient norm clipping | 1.0 |

Table 1: Hyperparameter configuration of our de-identification model.

We compare three different approaches: a non-private de-identification classifier and two privacy-enabled extensions, automatic pseudonymization (Section 4) and adversarially learned representations (Section 5).

Our non-private system as well as the privacy-enabled extensions are based on a bidirectional LSTM-CRF architecture that has been proven to work well in sequence tagging (Huang et al., 2015; Lample et al., 2016) and de-identification (Dernoncourt et al., 2017; Liu et al., 2017). We only use pre-trained FastText (Bojanowski et al., 2017) or GloVe (Pennington et al., 2014) word embeddings, not explicit character embeddings, as we suspect that these may allow easy re-identification of private information if used in shared representations. In place of learned character features, we provide the casing feature from Reimers and Gurevych (2017) as an additional input. The feature maps words to a one-hot representation of their casing (*numeric*, *mainly numeric*, *all lower*, *all upper*, *initial upper*, *contains digit*, or *other*).

Table 1 shows our raw de-identification model's hyperparameter configuration that was determined through a random hyperparameter search.

## 4 Automatic Pseudonymization

To provide a baseline to compare our primary approach against, we introduce a naïve word-level automatic pseudonymization approach that exploits the fact that state-of-the-art de-identification models (Liu et al., 2017; Dernoncourt et al., 2017) as well as our non-private de-identification model work on the sentence level and do not rely on document coherency. Before training, we shuffle the
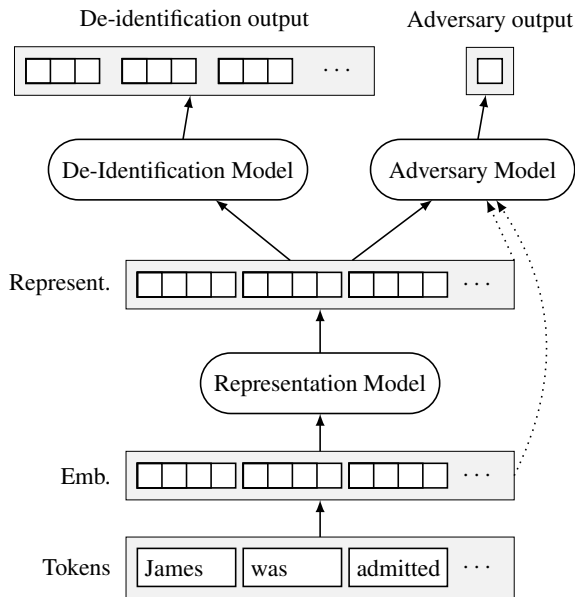
Figure 2: Simplified visualization of the adversarial model architecture. Sequences of squares denote real-valued vectors, dotted arrows represent possible additional real or fake inputs to the adversary. The casing feature that is provided as a second input to the de-identification model is omitted for legibility.

training sentences and replace all PHI tokens with a random choice of a fixed number $N$ of their closest neighbors in an embedding space (including the token itself), as determined by cosine distance in a pre-computed embedding matrix.

Using this approach, the sentence

[James] was admitted to [St. Thomas]

may be replaced by

[Henry] was admitted to [Croix Scott].

While the resulting sentences do not necessarily make sense to a reader (e.g. "Croix Scott" is not a realistic hospital name), its embedding representation is similar to the original. We train our de-identification model on the transformed data and test it on the raw data. The number of neighbors $N$ controls the privacy properties of the approach: $N = 1$ means no pseudonymization; setting $N$ to the number of rows in a precomputed embedding matrix delivers perfect anonymization but the resulting data may be worthless for training a de-identification model.

# 5 Adversarial Representation

We introduce a new data sharing approach that is based on an adversarially learned private repre-

sentation and improves on the pseudonymization from Section 4. After training the representation on an initial publicly available dataset, e.g. the i2b2 2014 data, a central model provider shares the frozen representation model with participating medical institutions. They transform their PHI-labeled raw data into the pseudonymized representation, which is then pooled into a new public dataset for de-identification. Periodically, the pipeline consisting of the representation model and a trained de-identification model can be published to be used by medical institutions on their unlabeled data.

Since both the representation model and the resulting representations are shared in this scenario, our representation procedure is required to prevent two attacks:

A1. Learning an inverse representation model that transforms representations back to original sentences containing PHI.

A2. Building a lookup table of inputs and their exact representations that can be used in known plaintext attacks.

## 5.1 Architecture

Our approach uses a model that is composed of three components: a representation model, the de-identification model from Section 3, and an adversary. An overview of the architecture is shown in Fig. 2.

The representation model maps a sequence of word embeddings to an intermediate vector representation sequence. The de-identification model receives this representation sequence as an input instead of the original embedding sequence. It retains the casing feature as an auxiliary input. The adversary has two inputs, the representation sequence and an additional embedding or representation sequence, and a single output unit.

## 5.2 Representation

To protect against A1, our representation must be invariant to small input changes, like a single PHI token being replaced with a neighbor in the embedding space. Again, the number of neighbors $N$ controls the privacy level of the representation.

To protect against A2, we add a random element to the representation that makes repeated transformations of one sentence indistinguishable from representations of similar input sentences.

5832

We use a bidirectional LSTM model to implement the representation. It applies Gaussian noise $N$ with zero mean and trainable standard deviations to the input embeddings $E$ and the output sequence. The model learns a standard deviation for each of the input and output dimensions.

$$R = N_{\text{out}} + \text{LSTM}(E + N_{\text{in}}) \qquad (1)$$

In a preliminary experiment, we confirmed that adding noise with a single, fixed standard deviation is not a viable approach for privacy-preserving representations. To change the cosine similarity neighborhoods of embeddings at all, we need to add high amounts of noise (more than double of the respective embedding matrix's standard deviation), which in turn results in *unrealistic* embeddings that do not allow training a de-identification model of sufficient quality.

In contrast to the automatic pseudonymization approach from Section 4 that only perturbs PHI tokens, the representation models in this approach processes all tokens to represent them in a new embedding space. We evaluate the representation sizes $d \in \{50, 100, 300\}$.

## 5.3 Adversaries

We use two adversaries that are trained on tasks that directly follow from A1 and A2:

T1. Given a representation sequence and an embedding sequence, decide if they were obtained from the same sentence.

T2. Given two representation sequences (and their cosine similarities), decide if they were obtained from the same sentence.

We generate the representation sequences for the second adversary from a copy of the representation model with shared weights. We generate real and fake pairs for adversarial training using the automatic pseudonymization approach presented in Section 4, limiting the number of replaced PHI tokens to one per sentence.

The adversaries are implemented as bidirectional LSTM models with single output units. We confirmed that this type of model is able to learn the adversarial tasks on random data and raw word embeddings in preliminary experiments. To use the two adversaries in our architecture, we average their outputs.
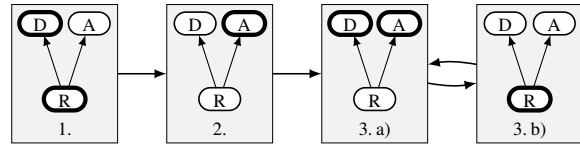


Figure 3: Visualization of Feutry et al.'s three-part training procedure. The adversarial model layout follows Fig. 2: the representation model is at the bottom, the left branch is the de-identification model and the right branch is the adversary. In each step, the thick components are trained while the thin components are frozen.

## 5.4 Training

We evaluate two training procedures: DANN training (Ganin et al., 2016) and the three-part procedure from Feutry et al. (2018).

In DANN training, the three components are trained conjointly, optimizing the sum of losses. Training the de-identification model modifies the representation model weights to generate a more meaningful representation for de-identification. The adversary gradient is reversed with a gradient reversal layer between the adversary and the representation model in the backward pass, causing the representation to become less meaningful for the adversary.

The training procedure by Feutry et al. (2018) is shown in Fig. 3. It is composed of three phases:

P1. The de-identification and representation models are pre-trained together, optimizing the de-identification loss $l_{\text{deid}}$.

P2. The representation model is frozen and the adversary is pre-trained, optimizing the adversarial loss $l_{\text{adv}}$.

P3. In alternation, for one epoch each:

    (a) The representation is frozen and both de-identification model and adversary are trained, optimizing their respective losses $l_{\text{deid}}$ and $l_{\text{adv}}$.

    (b) The de-identification model and adversary are frozen and the representation is trained, optimizing the combined loss

$$l_{\text{repr}} = l_{\text{deid}} + \lambda |l_{\text{adv}} - l_{\text{random}}| \qquad (2)$$

In each of the first two phases, the respective validation loss is monitored to decide at which point

the training should move on to the next phase. The alternating steps in the third phase each last one training epoch; the early stopping time for the third phase is determined using only the combined validation loss from Phase P3b.

Gradient reversal is achieved by optimizing the combined representation loss while the adversary weights are frozen. The combined loss is motivated by the fact that the adversary performance should be the same as a random guessing model, which is a lower bound for anonymization (Feutry et al., 2018). The term $|l_{\mathrm{adv}} - l_{\mathrm{random}}|$ approaches 0 when the adversary performance approaches random guessing[2]. $\lambda$ is a weighting factor for the two losses; we select $\lambda = 1$.

## 6 Experiments

To evaluate our approaches, we perform experiments using the i2b2 2014 dataset.

**Preprocessing:** We apply aggressive tokenization similarly to Liu et al. (2017), including splitting at all punctuation marks and mid-word e.g. if a number is followed by a word ("25yo" is split into "25", "yo") in order to minimize the amount of GloVe out-of-vocabulary tokens. We extend spaCy's[3] sentence splitting heuristics with additional rules for splitting at multiple blank lines as well as bulleted and numbered list items.

**Deep Learning Models:** We use the Keras framework[4] (Chollet et al., 2015) with the TensorFlow backend (Abadi et al., 2015) to implement our deep learning models.

**Evaluation:** In order to compare our results to the state of the art, we use the token-based binary HIPAA $F_1$ score as our main metric for de-identification performance. Dernoncourt et al. (2017) deem it the most important metric: deciding if an entity is PHI or not is generally more important than assigning the correct category of PHI, and only HIPAA categories of PHI are required to be removed by American law. Non-PHI tokens are not incorporated in the $F_1$ score. We perform the evaluation with the official i2b2 evaluation script[5].

---

[2]In the case of binary classification: $L_{\mathrm{random}} = -\log \frac{1}{2}$.
[3]https://spacy.io
[4]https://keras.io
[5]https://github.com/kotfic/i2b2_evaluation_scripts

| Model | $F_1$ (%) |
|---|---|
| Our non-private FastText | 97.67 |
| Our non-private GloVe | 97.24 |
| Our non-private GloVe + casing | 97.62 |
| Dernoncourt et al. (LSTM-CRF) | 97.85 |
| Liu et al. (ensemble + rules) | **98.27** |
| Our autom. pseudon. FastText | 96.75 |
| Our autom. pseudon. GloVe | 96.42 |
| Our adv. repr. FastText | **97.40** |
| Our adv. repr. GloVe | 96.89 |

Table 2: Binary HIPAA $F_1$ scores of our non-private (top) and private (bottom) de-identification approaches on the i2b2 2014 test set in comparison to non-private the state of the art. Our private approaches use $N = 100$ neighbors as a privacy criterion.

## 7 Results

Table 2 shows de-identification performance results for the non-private de-identification classifier (upper part, in comparison to the state of the art) as well as the two privacy-enabled extensions (lower part). The results are average values out of five experiment runs.

### 7.1 Non-private De-Identification Model

When trained on the raw i2b2 2014 data, our models achieve $F_1$ scores that are comparable to Dernoncourt et al.'s results. The casing feature improves GloVe by $0.4$ percentage points.

### 7.2 Automatic Pseudonymization

For both FastText and GloVe, moving training PHI tokens to random tokens from up to their $N = 200$ closest neighbors does not significantly reduce de-identification performance (see Fig. 4). $F_1$ scores for both models drop to around $95\%$ when selecting from $N = 500$ neighbors and to around $90\%$ when using $N = 1\,000$ neighbors. With $N = 100$, the FastText model achieves an $F_1$ score of $96.75\%$ and the GloVe model achieves an $F_1$ score of $96.42\%$.

### 7.3 Adversarial Representation

We do not achieve satisfactory results with the conjoint DANN training procedure: in all cases, our models learn representations that are not sufficiently resistant to the adversary. When training the adversary on the frozen representation for an
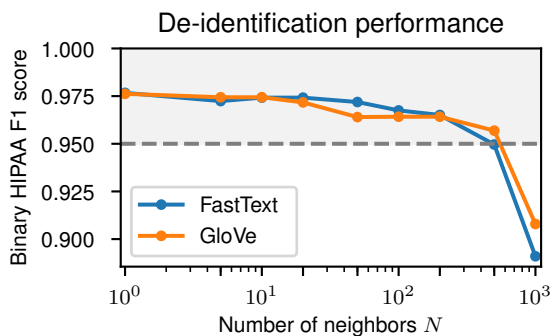
Figure 4: $F_1$ scores of our models when trained on automatically pseudonymized data where PHI tokens are moved to one of different numbers of neighbors $N$. The gray dashed line marks the 95% target $F_1$ score.

additional 20 epochs, it is able to distinguish real from fake input pairs on a test set with accuracies above 80%. This confirms the difficulties of DANN training as described by Elazar and Goldberg (2018) (see Section 2.2).

In contrast, with the three-part training procedure, we are able to learn a representation that allows training a de-identification model while preventing an adversary from learning the adversarial tasks, even with continued training on a frozen representation.

Figure 5 (left) shows our de-identification results when using adversarially learned representations. A higher number of neighbors $N$ means a stronger invariance requirement for the representation. For values of $N$ up to $1\,000$, our FastText and GloVe models are able to learn representations that allow training de-identification models that reach or exceed the target $F_1$ score of 95%. However, training becomes unstable for $N > 500$: at this point, the adversary is able to break the representation privacy when trained for an additional 50 epochs (Fig. 5 right).

Our choice of representation size $d \in \{50, 100, 300\}$ does not influence de-identification or adversary performance, so we select $d = 50$ for further evaluation. For $d = 50$ and $N = 100$, the FastText model reaches an $F_1$ score of 97.4% and the GloVe model reaches an $F_1$ score of 96.89%.

## 8 De-Identification Performance

In the following, we discuss the results of our models with regard to our goal of sharing sensitive training data for automatic de-identification. Overall, privacy-preserving representations come at a cost, as our best privacy-preserving model

scores 0.27 points $F_1$ score lower than our best non-private model; we consider this relative increase of errors of less than 10% as tolerable.

**Raw Text De-Identification:** We find that the choice of GloVe or FastText embeddings does not meaningfully influence de-identification performance. FastText's approach to embedding unknown words (word embeddings are the sum of their subword embeddings) should intuitively prove useful on datasets with misspellings and ungrammatical text. However, when using the additional casing feature, FastText beats GloVe only by 0.05 percentage points on the i2b2 test set. In this task, the casing feature makes up for GloVe's inability to embed unknown words.

Liu et al. (2017) use a deep learning ensemble in combination with hand-crafted rules to achieve state-of-the-art results for de-identification. Our model's scores are similar to the previous state of the art, a bidirectional LSTM-CRF model with character features (Dernoncourt et al., 2017).

**Automatically Pseudonymized Data:** Our naïve automatic word-level pseudonymization approach allows training reasonable de-identification models when selecting from up to $N = 500$ neighbors. There is almost no decrease in $F_1$ score for up to $N = 20$ neighbors for both the FastText and GloVe model.

**Adversarially Learned Representation:** Our adversarially trained vector representation allows training reasonable de-identification models ($F_1$ scores above 95%) when using up to $N = 1\,000$ neighbors as an invariance requirement. The adversarial representation results beat the automatic pseudonymization results because the representation model can act as a task-specific feature extractor. Additionally, the representations are more general as they are invariant to word changes.

## 9 Privacy Properties

In this section, we discuss our models with respect to their privacy-preserving properties.

**Embeddings:** When looking up embedding space neighbors for words, it is notable that many FastText neighbors include the original word or parts of it as a subword. For tokens that occur as PHI in the i2b2 training set, on average 7.37 of their $N = 100$ closest neighbors in the FastText embedding matrix contain the original token
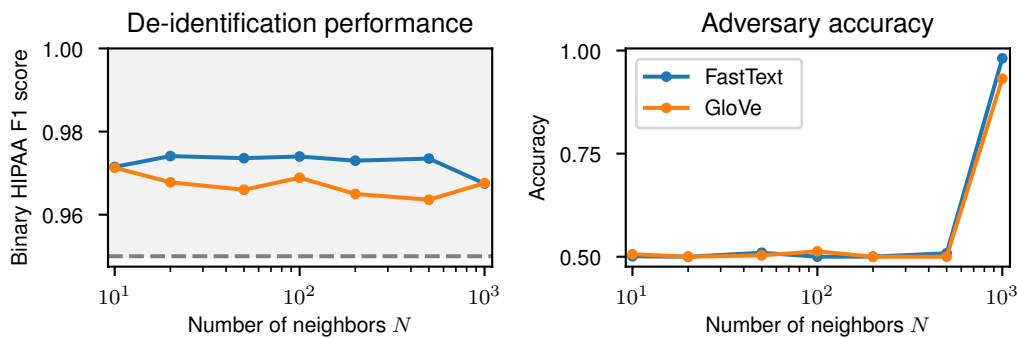
Figure 5: Left: de-identification $F_1$ scores of our models using an adversarially trained representation with different numbers of neighbors $N$ for the representation invariance requirement. Right: mean adversary accuracy when trained on the frozen representation for an additional 50 epochs. The figure shows average results out of five experiment runs.

as a subword. When looking up neighbors using GloVe embeddings, the value is $0.44$. This may indicate that FastText requires stronger perturbation (i.e. higher $N$) than GloVe to sufficiently obfuscate protected information.

**Automatically Pseudonymized Data:** The word-level pseudonymization does not guarantee a minimum perturbation for every word, e.g. in a set of pseudonymized sentences using $N = 100$ FastText neighbors, we found the phrase

[Florida Hospital],

which was replaced with

[Miami-Florida Hosp].

Additionally, the approach may allow an adversary to piece together documents from the shuffled sentences. If multiple sentences contain similar pseudonymized identifiers, they will likely come from the same original document, undoing the privacy gain from shuffling training sentences across documents. It may be possible to infer the original information using the overlapping neighbor spaces. To counter this, we can re-introduce document-level pseudonymization, i.e. moving all occurrences of a PHI token to the same neighbor. However, we would then also need to detect misspelled names as well as other hints to the actual tokens and transform them similarly to the original, which would add back much of the complexity of manual pseudonymization that we try to avoid.

**Adversarially Learned Representation:** Our adversarial representation empirically satisfies a strong privacy criterion: representations are invariant to *any* protected information token being

replaced with *any* of its $N$ neighbors in an embedding space. When freezing the representation model from an experiment run using up to $N = 500$ neighbors and training the adversary for an additional 50 epochs, it still does not achieve higher-than-chance accuracies on the training data. Due to the additive noise, the adversary does not overfit on its training set but rather fails to identify any structure in the data.

In the case of $N = 1\,000$ neighbors, the representation never becomes stable in the alternating training phase. The adversary is always able to break the representation privacy.

## 10 Conclusions & Future Work

We introduced a new approach to sharing training data for de-identification that requires lower human effort than the existing approach of document-coherent pseudonymization. Our approach is based on adversarial learning, which yields representations that can be distributed since they do not contain private health information. The setup is motivated by the need of de-identification of medical text before sharing; our approach provides a lower-cost alternative than manual pseudonymization and gives rise to the pooling of de-identification datasets from heterogeneous sources in order to train more robust classifiers. Our implementation and experimental data are publicly available[6].

As precursors to our adversarial representation approach, we developed a deep learning model for de-identification that does not rely on explicit character features as well as an automatic

---

[6]https://github.com/maxfriedrich/
deid-training-data

word-level pseudonymization approach. A model trained on our automatically pseudonymized data with $N = 100$ neighbors loses around one percentage point in $F_1$ score when compared to the non-private system, scoring $96.75\%$ on the i2b2 2014 test set.

Further, we presented an adversarial learning based private representation of medical text that is invariant to any PHI word being replaced with any of its embedding space neighbors and contains a random element. The representation allows training a de-identification model while being robust to adversaries trying to re-identify protected information or building a lookup table of representations. We extended existing adversarial representation learning approaches by using two adversaries that discriminate real from fake sequence pairs with an additional sequence input.

The representation acts as a task-specific feature extractor. For an invariance criterion of up to $N = 500$ neighbors, training is stable and adversaries cannot beat the random guessing accuracy of $50\%$. Using the adversarially learned representation, de-identification models reach an $F_1$ score of $97.4\%$, which is close to the non-private system ($97.67\%$). In contrast, the automatic pseudonymization approach only reaches an $F_1$ score of $95.0\%$ at $N = 500$.

Our adversarial representation approach enables cost-effective private sharing of training data for sequence labeling. Pooling of training data for de-identification from multiple institutions would lead to much more robust classifiers. Eventually, improved de-identification classifiers could help enable large-scale medical studies that eventually improve public health.

**Future Work:** The automatic pseudonymization approach could serve as a data augmentation scheme to be used as a regularizer for de-identification models. Training a model on a combination of raw and pseudonymized data may result in better test scores on the i2b2 test set, possibly improving the state of the art.

Private character embeddings that are learned from a perturbed source could be an interesting extension to our models.

In adversarial learning with the three-part training procedure, it might be possible to tune the $\lambda$ parameter and define a better stopping condition that avoids the unstable characteristics with high values for $N$ in the invariance criterion. A further possible extension is a dynamic noise level in the representation model that depends on the LSTM output instead of being a trained weight. This might allow using lower amounts of noise for certain inputs while still being robust to the adversary.

When more training data from multiple sources become available in the future, it will be possible to evaluate our adversarially learned representation against unseen data.

## Acknowledgments

## References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. Accessed May 31, 2019.

Guthrie S. Birkhead, Michael Klompas, and Nirav R. Shah. 2015. Uses of electronic health records for public health surveillance to advance public health. *Annual Review of Public Health*, 36(1):345–359. PMID: 25581157.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Taxiarchis Botsis, Gunnar Hartvigsen, Fei Chen, and Chunhua Weng. 2010. Secondary use of EHR: Data quality issues and informatics opportunities. *AMIA Summits on Translational Science Proceedings*, 2010:1–5.

François Chollet et al. 2015. Keras. Accessed May 31, 2019.

Martin R. Cowie, Juuso I. Blomster, Lesley H. Curtis, Sylvie Duclaux, Ian Ford, Fleur Fritz, Samantha Goldman, Salim Janmohamed, Jörg Kreuzer, Mark Leenay, Alexander Michel, Seleen Ong, Jill P. Pell, Mary Ross Southworth, Wendy Gattis Stough, Martin Thoenes, Faiez Zannad, and Andrew Zalewski. 2017. Electronic health records to facilitate clinical research. *Clinical Research in Cardiology*, 106(1):1–9.

Franck Dernoncourt, Ji Young Lee, Özlem Uzuner, and Peter Szolovits. 2017. De-identification of patient notes with recurrent neural networks. *Journal of the American Medical Informatics Association*, 24(3):596–606.

Yanai Elazar and Yoav Goldberg. 2018. Adversarial removal of demographic attributes from text data. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 11–21, Brussels, Belgium. Association for Computational Linguistics.

Clément Feutry, Pablo Piantanida, Yoshua Bengio, and Pierre Duhamel. 2018. Learning anonymized representations with adversarial neural networks. *arXiv preprint arXiv:1802.09386*.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(1):2096–2030.

Jihun Hamm. 2015. Preserving privacy of continuous high-dimensional data with minimax filters. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, volume 38 of *Proceedings of Machine Learning Research*, pages 324–332, San Diego, CA, USA. PMLR.

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

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Peter B. Jensen, Lars J. Jensen, and Søren Brunak. 2012. Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics*, 13:395–405.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, CA, USA. Association for Computational Linguistics.

Ji Young Lee, Franck Dernoncourt, and Peter Szolovits. 2018. Transfer Learning for Named-Entity Recognition with Neural Networks. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, pages 4470–4473, Miyazaki, Japan. European Language Resources Association.

Yitong Li, Timothy Baldwin, and Trevor Cohn. 2018. Towards robust and privacy-preserving text representations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 25–30, Melbourne, Australia. Association for Computational Linguistics.

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(S):S34–S42.

Stephane M. Meystre, F. Jeffrey Friedlin, Brett R. South, Shuying Shen, and Matthew H. Samore. 2010. Automatic de-identification of textual documents in the electronic health record: a review of recent research. *BMC Medical Research Methodology*, 10(1):70.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2017. Optimal hyperparameters for deep LSTM-networks for sequence labeling tasks. *arXiv preprint arXiv:1707.06799*.

Amber Stubbs, Michele Filannino, and Özlem Uzuner. 2017. De-identification of psychiatric intake records: Overview of 2016 CEGS N-GRID shared tasks track 1. *Journal of Biomedical Informatics*, 75(S):S4–S18.

Amber Stubbs, Christopher Kotfila, and Özlem Uzuner. 2015. Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/UTHealth shared task track 1. *Journal of Biomedical Informatics*, 58:11–19.

Amber Stubbs and Özlem Uzuner. 2015. Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/UTHealth corpus. *Journal of Biomedical Informatics*, 58:20–29.

The United States Congress. 1996. Health insurance portability and accountability act of 1996. Accessed May 31, 2019.

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

Seid Muhie Yimam. 2015. Narrowing the loop: Integration of resources and linguistic dataset development with interactive machine learning. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 88–95, Denver, CO, USA. Association for Computational Linguistics.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. Learning fair representations. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 325–333, Atlanta, GA, USA. PMLR.