Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 4245–4252
Marseille, 20-25 June 2022

Downstream Task Performance of BERT Models Pre-Trained Using Automatically De-Identified Clinical Data

Thomas Vakili, Anastasios Lamproudis, Aron Henriksson, Hercules Dalianis
Department of Computer and Systems Sciences (DSV), Stockholm University, Kista, Sweden
{thomas.vakili, anastasios, aronhen, hercules}@dsv.su.se

Abstract

Automatic de-identification is a cost-effective and straightforward way of removing large amounts of personally identifiable information from large and sensitive corpora. However, these systems also introduce errors into datasets due to their imperfect precision. These corruptions of the data may negatively impact the utility of the de-identified dataset. This paper de-identifies a very large clinical corpus in Swedish either by removing entire sentences containing sensitive data or by replacing sensitive words with realistic surrogates. These two datasets are used to perform domain adaptation of a general Swedish BERT model. The impact of the de-identification techniques is assessed by training and evaluating the models using six clinical downstream tasks. The results are then compared to a similar BERT model domain-adapted using an unaltered version of the clinical corpus. The results show that using an automatically de-identified corpus for domain adaptation does not negatively impact downstream performance. We argue that automatic de-identification is an efficient way of reducing the privacy risks of domain-adapted models and that the models created in this paper should be safe to distribute to other academic researchers.

Keywords: Privacy-preserving machine learning, pseudonymization, de-identification, Swedish clinical text, pre-trained language models, BERT, downstream tasks, NER, multi-label classification, domain adaptation

1. Introduction

Natural Language Processing (NLP) research is currently dominated by so-called pre-trained language models based on transformers (Vaswani et al., 2017), which were popularized by the introduction of the BERT model by Devlin et al. (2019). These language models typically consist of millions – even billions – of parameters that are learned from enormous corpora. The success of pre-trained language models in general-domain tasks has prompted research into whether these models also succeed in medical-domain tasks.

Language models are taught to model language by learning the statistical distributions of the words in their training data. However, words often have different meanings depending on in which domain they are used. The word *chest* has a dual meaning in everyday language – something used for storage or a region of the body – but only one of these is relevant in a medical context. A language model which has learned the word *chest* from a general-domain corpus may have a representation of the word that is sub-optimal in the medical domain.

Indeed, many researchers have found that performance on domain-specific tasks is helped by adapting existing language models or pre-training new models using in-domain data (Lee et al., 2019; Beltagy et al., 2019; Lamproudis et al., 2021; Lamproudis et al., 2022b; Lamproudis et al., 2022a). Better performance means that the models will be more useful in helping medical professionals improve patient outcomes.

However, the scale of the data used to train these models means that researchers cannot know what sensitive information the corpora contain. In the medical domain, we can be certain that the texts contain sensitive information. This is cause for concern since pre-trained language models are susceptible to privacy attacks (Bender et al., 2021).

This paper examines one way of reducing the privacy risks: automatic de-identification. Two different approaches are studied: pseudonymization (Sweeney, 1996; Dalianis, 2019) and removal of sensitive data. Two different clinical BERT models are created by applying these techniques to the pre-training data. The impact of automatic de-identification on the performance of the models is then evaluated on downstream tasks.

2. Related Research

The two main topics of this paper are automatic de-identification and the privacy risks of large language models. This section introduces these concepts by providing a brief summary of results related to the topic of this paper.

2.1. Privacy Attacks on Language Models

Large pre-trained language models are susceptible to a wide range of attacks on privacy. One reason for this is due to their size, which gives them a tendency to unintentionally memorize parts of their training data. The attacks can generally be separated into two main categories:

Training data extraction An attacker that successfully mounts a model inversion attack is able to extract details about the training data. One example of a training data extraction attack was mounted by Carlini et al. (2020). They managed to extract entire passages from IRC logs from the model GPT-2 (Radford et al., 2019).

Membership inference If an attacker is able to discern whether or not a datapoint was part of the training data, they have successfully mounted a membership inference attack (Shokri et al., 2017). Although these attacks are typically less severe than training data extraction attacks, they can also expose sensitive data.

To the best of our knowledge, there are no examples of successful training data extraction attacks on BERT models. Lehman et al. (2021) and Vakili and Dalianis (2021) found that BERT models are at least less susceptible to such attacks than GPT-2. Both studies attempted to extract training data from a BERT model trained on a version of MIMIC-III (Johnson et al., 2016) which had its masked entities populated with realistic but fake values.

Nakamura et al. (2020) performed a related attack that attempted to re-predict pseudonymized information. They trained a BERT model on a version of the MIMIC-III (with inserted surrogate values) and then re-masked the surrogate entities in this dataset. They then attempted to reconstruct the surrogate names but did not succeed, concluding that this does not seem to be a viable attack.

Lehman et al. (2021) also performed membership inference attacks on their BERT model. Their results indicated a small risk of memorizing patients' names. At the same time, they were not able to link a patient's name to any of their conditions. Jagannatha et al. (2021) also performed membership inference attacks on BERT and found that there is a risk of privacy leakage from BERT models. However, this risk is significantly smaller than for models like GPT-2.

2.2. De-Identification of Clinical Text Data

The electronic health records (EHRs) used in clinical NLP are inherently sensitive. For example, the data used in this study was found to have an estimated protected health information (PHI) density[1] of 1.57% (Henriksson et al., 2017). However, the PHI density varied considerably across medical specialties and classes of clinical notes. For example, almost 20% of the sentences in discharge summaries contained at least one PHI. The prevalence of PHI has caused many researchers to explore ways of reducing the risks to patient privacy that comes with using their health data. One active area of research is automatic de-identification.

Automatic de-identifiers typically rely on named entity recognition (NER) models to detect sensitive data in datasets. Thus, the recall of the model needs to be balanced against its precision. In this context, the classic precision-recall trade-off translates to one between utility and privacy. Low recall means that a lot of sensitive data will be undetected, but a low precision results in a dataset where a lot of non-sensitive data is corrupted.

[1]PHI density was defined as the number of PHI mentions divided by the number of tokens.

Berg et al. (2020) used various high recall models to de-identify several Swedish clinical datasets. This did not seem to lower the utility of the datasets, as training with the datasets did not significantly decrease downstream performance. The authors tried out four strategies for the de-identification: pseudonymization (replacing sensitive data with surrogates), masking the sensitive data, replacing a sensitive word with its class name (e.g., replacing "John" with "First Name"), and removing the sensitive data along with the sentence in which it appeared. All of the downstream tasks were NER tasks and were approached using a machine learning algorithm based on conditional random fields (CRFs). The tasks were clinical entity identification, adverse drug effect identification, and cervical cancer symptom detection. Pseudonymization resulted in the smallest negative impact on the downstream tasks, while the sentence removal strategy resulted in a greater deterioration of the performance.

Vakili and Dalianis (2022) automatically de-identified three Swedish clinical datasets using pseudonymization. Each dataset was associated with a task: two sequence classification tasks (ICD-10 classification and factuality classification) and one NER task (clinical entity recognition). Different BERT models were trained using unaltered and pseudonymized data, and the performances on all tasks were compared. There was no significant difference in the performance of the models trained on unaltered data and the models trained on pseudonymized data.

Obeid et al. (2019) de-identified clinical data and evaluated the impact of this by building detectors of altered mental status (AMS) using a variety of machine learning models. These included Naïve Bayes Classifiers, Single Decision Trees, Random Forests, and Multilayer Perceptrons. The deep learning models performed the strongest, but no model showed any significant deterioration in performance when trained using de-identified text instead of the original text.

No automatic de-identification system has perfect recall, and some sensitive data will remain in a processed corpus. However, pseudonymizing the data makes it difficult to determine which data are real and which data are pseudonymized. Carrell et al. (2019) explored the concept of *hiding in plain sight* (HIPS). They were able to train a tagger to distinguish between pseudonymized data and data that were HIPS in a pseudonymized dataset. The tagger performed significantly better than random guessing but had a high rate of false positives and false negatives. Thus, the authors concluded that HIPS is still helpful for protecting privacy.

This study applies two of the de-identification approaches outlined in Berg et al. (2021) to a clinical corpus data. However, the data used in this paper is much larger in scale and is used to pre-train language models rather than to build task-specific classifiers.

3. Data

The clinical data used to train and evaluate the BERT models originate from the Karolinska University Hospital. The data are stored in the research infrastructure Health Bank – The Swedish Health Record Research Bank[2] (Dalianis et al., 2015) at DSV/Stockholm University.

3.1. EHRs from the Health Bank

The BERT models in this paper were pre-trained using a 17.9 GB subset of the Health Bank. The clinical texts come from a large number of clinical units and encompass over 2 million EHRs[3]. This dataset is comparable in size to the general domain Swedish corpus of newspapers, Swedish Wikipedia, and government documents that was used to pre-train *KB-BERT* (Malmsten et al., 2020).

These EHRs were de-identified according to the process outlined in Section 4.1, and the resulting dataset was used to train two BERT models, as will be described in Section 4.2. Lamproudis et al. (2021) also use this dataset in its unaltered form to train the baseline model used for evaluating the impact of de-identifying the pre-training data.

3.2. Datasets for Downstream Tasks

Five manually annotated datasets, all created from the Health Bank, were used to evaluate the downstream performance of the models. All of the downstream tasks concern clinical NLP tasks and make it possible to compare the BERT models to each other.

Stockholm EPR Gastro ICD-10 Corpus A Gastro ICD-10 data set consisting of 6,062 gastro-related discharge summaries and their assigned ICD-10 diagnosis codes. The data set encompasses 4,985 unique patients and 795,839 tokens. The data are divided into 10 groups that correspond to different body parts; the ICD-10 codes range from K00 to K99. Each group contains several codes (Remmer et al., 2021).

Stockholm EPR PHI Corpus A PHI data set of 4,480 annotated entities and 380,000 tokens. The PHIs correspond to nine PHI classes: *First Name*, *Last Name*, *Age*, *Phone Number*, *Location*, *Health Care Unit*, *Organization*, *Full Date*, and *Date Part* (Dalianis and Velupillai, 2010).

Stockholm EPR Clinical Entity Corpus A clinical entity data set comprising 70,852 tokens and 7,946 annotated entities corresponding to four clinical entity classes *Diagnosis*, *Findings*, *Body parts*, and *Drugs* (Skeppstedt et al., 2014).

Stockholm EPR Diagnosis Factuality Corpus A factuality diagnosis data set encompassing six levels of annotations regarding the factuality of a diagnosis. The data set consists of 3,710 samples with 7,066 annotated entities *Certainly Positive*, *Probably Positive*, *Possibly Positive*, *Possibly Negative*, *Probably Negative*, and *Certainly Negative* encompassing 240,000 tokens (Velupillai et al., 2011; Velupillai, 2011). The dataset is used for two tasks. One is a NER task where the goal is to identify tokens specifying diagnoses and assigning them a factuality label. The second task treats the sample as a single datapoint and performs a multi-label classification of the entire sample to predict its factuality.

Stockholm EPR ADE ICD-10 Corpus A newly introduced ADE corpus containing 16,858 samples encompassing 634,000 tokens. The samples are distributed over 12 different ICD-10 codes describing adverse drug events. The task is treated as a binary classification task where positive samples have been assigned a specific ICD-10 code that denotes an adverse drug event. Negative samples in each group have been assigned a code describing a similar condition that was not drug-induced. The goal of the task is to determine whether or not the condition defined by the ICD-10 code was induced by an ADE.

4. Experiments

The study encompasses three steps. First, the Health Bank corpus is processed to detect and deal with sensitive data. This leads to two different clinical corpora that are then used for domain-adaptive pre-training. The resulting models are evaluated on downstream tasks, and the results are compared to other models trained on the Health Bank data. This section gives a detailed account of the experiments and their results.

4.1. De-Identifying the Health Bank

A NER model was built based on a clinical BERT model trained by Lamproudis et al. (2021) using the *Stockholm EPR PHI Corpus*. The model was used to detect the nine PHI classes described in Section 3.2 and by Dalianis and Velupillai (2010). This model was then applied to the 17.9 GBs of EHRs extracted from the Health Bank. This processing uncovered a large amount of possibly sensitive data. The number of detected instances for each PHI type is listed in Table 1.

Two approaches to de-identification were taken, as illustrated in Figure 1. In the first approach, which we refer to as *pseudonymization*, each detected entity was replaced by a realistic surrogate value of the same class. For example, a detected name will be replaced with another (generated but realistic) name. Pseudonymization preserves the semantics of the text as long as the entity has been correctly classified and allows the model to

[2]Health Bank: `http://dsv.su.se/healthbank`

[3]This research has been approved by the Swedish Ethical Review Authority under permission no. 2019-05679.

PHI Type	# Predicted Instances	NER Recall	NER Precision
Health Care Unit	19,659,127	80%	87%
Partial Date	19,374,711	83%	94%
Last Name	14,332,309	97%	96%
First Name	12,525,688	97%	98%
Full Date	10,459,935	55%	77%
Location	3,158,031	89%	85%
Age	2,064,111	35%	47%
Organisation	1,078,115	36%	71%
Phone Number	1,262,313	40%	63%

Table 1: The PHI types in order of frequency as classified by the de-identification system. The per-class recall and precision for the NER model are also displayed and were calculated on the test data from Dalianis and Velupillai (2010). In total, 83,914,340 sensitive entities are found in 49,715,558 sentences.

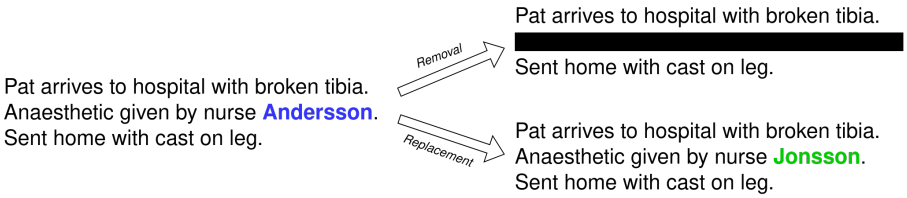

Figure 1: This hypothetical example illustrates the two approaches taken to de-identify the data. One approach *replaces* the sensitive data with realistic surrogates and is used to train the model *KB-BERT + Pseudo*. The other approach instead *removes* the entire sentence from the dataset and this filtered dataset is used to train the model *KB-BERT + Filtered*.

learn essentially the same information without exposing any sensitive information.

The second and more aggressive approach is to remove all sentences that contain sensitive entities. This approach removes 49,715,558 out of 364,385,114 sentences in the original dataset. In other words, 13.65% of all sentences were identified as containing sensitive entities. The removal of these sentences reduced the size of the dataset by approximately 19%.

Combined with the total number of entities shown in Table 1, these statistics indicate a slight tendency for sensitive entities to cluster in the same sentences, with around 1.69 entities per sensitive sentence. If this tendency holds for the entire dataset, then removing entire sentences should help remove some additional sensitive entities that the de-identifier has missed.

4.2. Training the BERT Models

The models in this paper are trained using a setup similar to Lamproudis et al. (2021), whose model is used for comparison in this study. Their model was trained using unaltered sensitive EHR data and is referred to as *KB-BERT + Real* in this paper. The two new models are built using the datasets described in Section 4.1:

KB-BERT + Pseudo The data used to train this model has had all sensitive entities (as listed in Table 1) replaced with realistic surrogates of the same entity class.

KB-BERT + Filtered This model is built using the dataset where all sentences found to contain sensitive data have been *removed*. This filtered version of the dataset is 19% smaller than the version used to train *KB-BERT + Pseudo*.

Both models were trained using *KB-BERT* (Malmsten et al., 2020) as the starting point and are the same size as $BERT_{BASE}$ (Devlin et al., 2019). As in Lamproudis et al. (2021), the vocabularies of both models are identical to that of *KB-BERT*. Pre-training was resumed for three epochs of the datasets using hyperparameters shown in Table 2.

One way in which the training of these two models differs from *KB-BERT + Real* is that our training data does not contain any document boundaries. This means that some datapoints in the training data contain two sentences from different clinical notes. In theory, this can harm the training process. As will be shown in Section 4.3, it does not seem to matter very much in practice.

4.3. Evaluating on Downstream Tasks

After training each model for three epochs, the resulting models were fine-tuned and evaluated on each of the six downstream tasks described in Section 3.2.

Table 3 displays the results of the downstream evaluation. Each model, except for *KB-BERT*, is evaluated on all three epochs, and we report the best result of the three evaluations. The best result is selected as the

Hyperparameter	Value
Max epochs	3
Batch size	256
Training sequence length	512
Mask probability	15%
Optimizer	Adam
Learning decay rate	Linear
Learning rate	1e-4
Dropout	0.1
Warm-up steps	10,000

Table 2: The hyperparameters used for continuing the pre-training with *KB-BERT* as a starting point. These hyper-parameters were used to train *KB-BERT + Real* (Lamproudis et al., 2021), *KB-BERT + Filtered*, and *KB-BERT + Pseudo*.

aim of this study is not to determine the optimal number of epochs which could vary depending on the de-identification approach.

All three models outperform the non-clinical baseline *KB-BERT* on every clinical downstream task. This is expected and indicates that the models have adapted to the language of the domain. More surprisingly, de-identification does not lead to any discernible drop in performance. In fact, *KB-BERT + Pseudo* even outperforms *KB-BERT + Real* on some tasks.

5. Discussion & Conclusions

The results in Section 3.2 show that performance on downstream tasks is not harmed by de-identifying the data used for domain adaptation of language models. This section contextualizes these findings and provides suggestions for future research.

5.1. Absence of Performance Drops

Automatic de-identification leads to a certain degree of corruption of the training data. The models used in this paper have a strong level of precision for many entity classes, as shown in Table 1. On the other hand, the evaluation indicates that around 15% of all detected locations are actually something else. The de-identification system will then either erroneously replace the word with a location name – corrupting the data – or unnecessarily discard the sentence.

Surprisingly, Table 3 indicates that this does not adversely affect the usefulness of the resulting models on the downstream tasks. *KB-BERT + Pseudo* is trained on data that is possibly corrupted due to precision issues but still performs similarly to *KB-BERT + Real*.

KB-BERT + Filtered also performs comparably to *KB-BERT + Real* even though the data is reduced to a non-trivial degree. It does, however, perform noticeably worse on the PHI NER task. This is expected since the de-identification approach aims to remove all such entities from the continued pre-training.

5.2. Reliability of the De-Identification

The NER model used in this paper is evaluated on in-domain clinical NER data. This strongly suggests that the recall and precision estimates are accurate. Nevertheless, the efficacy of the de-identification can only be assessed using the testing data. Due to the very nature of the problem, this means that the amount of sensitive information remaining in the training data can only be estimated.

However, not all entity classes are equally sensitive. Table 3 shows that our system detects and de-identifies 97% of all first and last names which are arguably the most sensitive classes. Furthermore, an attacker cannot target a specific person as they do not know if their names are among the 3% retained in the dataset.

5.3. Releasing the Models?

As explained in Section 2.1, there has been a growing interest in evaluating how susceptible pre-trained language models are to privacy attacks. While GPT-2 has been found to be very susceptible to attacks, BERT seems to be more resilient.

The performance of the de-identification system suggests that the overwhelming majority of sensitive data are removed from the training data of our models. If only 3% of all names in the data used for domain adaptation are sensitive, and the risk of exposing *any* name is less than 10% (Jagannatha et al., 2021), then the risk of exposing a *real* name is very small.

Another feature of the approach taken in this paper is that the models use a pre-trained model as their starting point. This means that any memorized names can come both from the Health Bank or the data used to train *KB-BERT*. This can be viewed as a form of *hiding in plain sight* (HIPS). Thus, an attacker who has extracted a name not only needs to determine whether or not it is a surrogate but also whether it came from a sensitive or non-sensitive data source.

BERT models have been shown to be quite resistant to training data extraction attacks (Nakamura et al., 2020; Lehman et al., 2021; Vakili and Dalianis, 2021). Furthermore, the limited susceptibility to membership inference attacks (Lehman et al., 2021; Jagannatha et al., 2021) is likely negligible when most of the data memorized by the model has been made non-sensitive through de-identification. Based on this, as well as other points made in this paper, we believe that the models can safely be shared among academic researchers. The model *KB-BERT + Pseudo* will be distributed under the name *SweDeClin-BERT*[4] once we have obtained the necessary permissions from the Swedish Ethical Review Authority.

5.4. Future Research

As noted in Section 2.2, previous research has shown that training on pseudonymized data can adversely impact model performance. In this paper, we show that

[4]This is short for **Swe**dish **De**-identified **Clin**ical **BERT**.

Model	ICD-10	PHI	Clinical Entity	Factuality	Factuality	ADE
	Classification	NER	NER	Classification	NER	Classification
KB-BERT	0.799	0.91	0.803	0.635	0.630	0.183
KB-BERT + Real	**0.833**	**0.941**	0.858	0.732	0.682	**0.199**
KB-BERT + Filtered	**0.833**	0.929	0.854	0.731	0.672	**0.199**
KB-BERT + Pseudo	0.832	**0.941**	**0.861**	**0.736**	**0.684**	0.191

Table 3: The table compares the downstream performances of each BERT model. *KB-BERT* and *KB-BERT + Real* are used as baselines. *KB-BERT* is also the starting point for the continued pre-training of all three models, as described in Section 4.2. All values are F_1-scores and the best results are bolded.

this does not seem to be a problem when pre-training for domain adaptation. However, the data used for the downstream tasks is unaltered sensitive data, and further research into the impacts of pseudonymization on task-specific training data is needed.

It could also be interesting to perform a similar experiment on English data. A natural candidate would be to use the freely available and anonymized MIMIC-III dataset (Johnson et al., 2016), though this would require replacing all the masked PHIs with realistic surrogates. This has been done by Lehman et al. (2021). On the other hand, using a non-anonymized dataset – as done in this paper – helps ensure that the results are realistic and not contingent on the quality of the surrogate selection.

Another way to avoid leaking private information is to use synthetic data. This can be generated using generative models. Generative models such as GANs[5] (Goodfellow et al., 2014) have successfully been applied to generate very realistic image data, targeting many different domains (Jetchev and Bergmann, 2017; Han et al., 2018; Brock et al., 2018).

Choi et al. (2017; Guan et al. (2018) use GANs to generate EHR data, and a more recent paper by Al Aziz et al. (2021) use generative transformer-based models to generate synthetic EHRs. None of these papers use the synthetic data to pre-train a new language model. Performance limitations are likely a barrier to generating a dataset of the scale needed for domain adaptation of a pre-trained language model.

5.5. Conclusions

This paper compares the impact of automatically de-identifying a large corpus which is used to domain-adapt Swedish BERT models. The consequences for the utility of the de-identified corpus are determined by comparing the downstream performance of the resulting BERT models with a model domain-adapted using an unaltered version of the corpus.

The results from six clinical downstream tasks show that there is no negative impact from using an automatically de-identified clinical corpus. Indeed, the results show a slight increase in performance for some tasks.

We suggest that practitioners who use clinical data for domain adaptation incorporate automatic de-identification into their workflow to decrease the risk of privacy leaks. Automatic de-identification is an easily implemented measure that reduces the risks of unintentionally memorizing sensitive information without harming utility.

[5]GAN stands for **G**enerative **A**dversarial **N**etwork.

Acknowledgements

We would like to thank Sonja Remmer for creating the *Stockholm EPR ADE ICD-10 Corpus*.

This work was partially funded by the *DataLEASH* project and by Region Stockholm through the project *Improving Prediction Models for Diagnosis and Prognosis of COVID-19 and Sepsis with NLP*.

References

Al Aziz, M. M., Ahmed, T., Faequa, T., Jiang, X., Yao, Y., and Mohammed, N. (2021). Differentially Private Medical Texts Generation Using Generative Neural Networks. *ACM Transactions on Computing for Healthcare*, 3(1):5:1–5:27, October.

Beltagy, I., Lo, K., and Cohan, A. (2019). SciBERT: A Pretrained Language Model for Scientific Text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China, November. Association for Computational Linguistics.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623.

Berg, H., Henriksson, A., and Dalianis, H. (2020). The Impact of De-identification on Downstream Named Entity Recognition in Clinical Text. In *Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis, Louhi 2020, in conjunction with EMNLP 2020*, pages 1–11.

Berg, H., Henriksson, A., Fors, U., and Dalianis, H. (2021). De-identification of Clinical Text for Secondary Use : Research Issues. pages 592–599. SciTePress.

Brock, A., Donahue, J., and Simonyan, K. (2018). Large Scale GAN Training for High Fidelity Natural Image Synthesis. In *International Conference on Learning Representations*.

Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, U., Oprea, A., and Raffel, C. (2020). Extracting Training Data from Large Language Models. *arXiv:2012.07805 [cs]*, December. arXiv: 2012.07805.

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

Choi, E., Biswal, S., Malin, B., Duke, J., Stewart, W. F., and Sun, J. (2017). Generating Multi-label Discrete Patient Records using Generative Adversarial Networks. In *Proceedings of the 2nd Machine Learning for Healthcare Conference*, pages 286–305. PMLR, November. ISSN: 2640-3498.

Dalianis, H. and Velupillai, S. (2010). De-identifying Swedish clinical text - refinement of a gold standard and experiments with Conditional random fields. *Journal of Biomedical Semantics*, 1(1):6, April.

Dalianis, H., Henriksson, A., Kvist, M., Velupillai, S., and Weegar, R. (2015). HEALTH BANK- A Workbench for Data Science Applications in Healthcare. *CEUR Workshop Proceedings Industry Track Workshop*, pages 1–18, 1.

Dalianis, H. (2019). Pseudonymisation of Swedish electronic patient records using a rule-based approach. In *Proceedings of the Workshop on NLP and Pseudonymisation, September 30, 2019, Turku, Finland*, number 166, pages 16–23. Linköping University Electronic Press.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, pages 2672–2680, Cambridge, MA, USA, December. MIT Press.

Guan, J., Li, R., Yu, S., and Zhang, X. (2018). Generation of Synthetic Electronic Medical Record Text. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 374–380, December.

Han, C., Hayashi, H., Rundo, L., Araki, R., Shimoda, W., Muramatsu, S., Furukawa, Y., Mauri, G., and Nakayama, H. (2018). GAN-based synthetic brain MR image generation. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 734–738, April. ISSN: 1945-8452.

Henriksson, A., Kvist, M., and Dalianis, H. (2017). Prevalence estimation of protected health information in swedish clinical text. In *Informatics for Health: Connected Citizen-Led Wellness and Population Health*, pages 216–220. IOS Press.

Jagannatha, A., Rawat, B. P. S., and Yu, H. (2021). Membership Inference Attack Susceptibility of Clinical Language Models. *arXiv:2104.08305 [cs]*, April. arXiv: 2104.08305.

Jetchev, N. and Bergmann, U. (2017). The Conditional Analogy GAN: Swapping Fashion Articles on People Images. pages 2287–2292.

Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L.-w. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L., and Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(1):160035, May. Number: 1 Publisher: Nature Publishing Group.

Lamproudis, A., Henriksson, A., and Dalianis, H. (2021). Developing a Clinical Language Model for Swedish: Continued Pretraining of Generic BERT with In-Domain Data. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 790–797, Held Online, September. INCOMA Ltd.

Lamproudis, A., Henriksson, A., and Dalianis, H. (2022a). Evaluating Pretraining Strategies for Clinical BERT Models. In *Proceedings of the Thirteenth International Conference on Language Resources and Evaluation (LREC 2022)*.

Lamproudis, A., Henriksson, A., and Dalianis, H. (2022b). Vocabulary modifications for domain-adaptive pretraining of clinical language models. In *Proceedings of the 15th International Joint Conference on Biomedical Engineering Systems and Technologies – HEALTHINF*, volume 5, pages 180–188.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. (2019). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, page btz682, September. arXiv: 1901.08746.

Lehman, E., Jain, S., Pichotta, K., Goldberg, Y., and Wallace, B. (2021). Does BERT Pretrained on Clinical Notes Reveal Sensitive Data? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 946–959, Online, June. Association for Computational Linguistics.

Malmsten, M., Börjeson, L., and Haffenden, C. (2020). Playing with Words at the National Library of

Sweden–Making a Swedish BERT. *arXiv preprint arXiv:2007.01658*.

Nakamura, Y., Hanaoka, S., Nomura, Y., Hayashi, N., Abe, O., Yada, S., Wakamiya, S., and Aramaki, E. (2020). KART: Privacy Leakage Framework of Language Models Pre-trained with Clinical Records. *arXiv:2101.00036 [cs]*, December. arXiv: 2101.00036.

Obeid, J. S., Heider, P. M., Weeda, E. R., Matuskowitz, A. J., Carr, C. M., Gagnon, K., Crawford, T., and Meystre, S. M. (2019). Impact of de-identification on clinical text classification using traditional and deep learning classifiers. *Studies in Health technology and Informatics*, 264:283.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Remmer, S., Lamproudis, A., and Dalianis, H. (2021). Multi-label Diagnosis Classification of Swedish Discharge Summaries – ICD-10 Code Assignment Using KB-BERT. In *Proceedings of RANLP 2021: Recent Advances in Natural Language Processing, RANLP 2021, 1-3 Sept 2021, Varna, Bulgaria*, pages 1158–1166.

Shokri, R., Stronati, M., Song, C., and Shmatikov, V. (2017). Membership Inference Attacks Against Machine Learning Models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18, May. ISSN: 2375-1207.

Skeppstedt, M., Kvist, M., Nilsson, G. H., and Dalianis, H. (2014). Automatic recognition of disorders, findings, pharmaceuticals and body structures from clinical text: An annotation and machine learning study. *Journal of biomedical informatics*, 49:148–158.

Sweeney, L. (1996). (replacing personally-identifying information in medical records, the scrub system). In *Proceedings of the AMIA annual fall symposium*, page 333. American Medical Informatics Association.

Vakili, T. and Dalianis, H. (2021). Are Clinical BERT Models Privacy Preserving? The Difficulty of Extracting Patient-Condition Associations. In *Proceedings of the AAAI 2021 Fall Symposium on Human Partnership with Medical AI: Design, Operationalization, and Ethics (AAAI-HUMAN 2021)*, volume 3068. CEUR Workshop Proceedings.

Vakili, T. and Dalianis, H. (2022). Utility Preservation of Clinical Text After De-Identification. In *Proceedings of the 21st Workshop on Biomedical Language Processing, BioNLP@ACL 2022*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention Is All You Need. In *Advances in neural information processing systems*, pages 5998–6008.

Velupillai, S., Dalianis, H., and Kvist, M. (2011). Factuality levels of diagnoses in Swedish clinical text. In *User Centred Networked Health Care*, pages 559–563. IOS Press.

Velupillai, S. (2011). Automatic classification of factuality levels: A case study on Swedish diagnoses and the impact of local context. In *Fourth International Symposium on Languages in Biology and Medicine, LBM 2011*.