The Legal and Ethical issues Workshop @LREC2022, pages 49–52
Marseille, 20 June 2022

Cross-Clinic De-Identification of Swedish Electronic Health Records: Nuances and Caveats

Olle Bridal*, Thomas Vakili, Marina Santini*****
*Linköping University,**DSV, Stockholm University, ***RISE Research Institutes of Sweden
ollbr616@student.liu.se, thomas.vakili@dsv.su.se, marina.santini@ri.se

Abstract

Privacy preservation of sensitive information is one of the main concerns in clinical text mining. Due to the inherent privacy risks of handling clinical data, the clinical corpora used to create the clinical Named Entity Recognition (NER) models underlying clinical de-identification systems cannot be shared. This situation implies that clinical NER models are trained and tested on data originating from the same institution since it is rarely possible to evaluate them on data belonging to a different organization. These restrictions on sharing make it very difficult to assess whether a clinical NER model has overfitted the data or if it has learned any undetected biases. This paper presents the results of the first-ever cross-institution evaluation of a Swedish de-identification system on Swedish clinical data. Alongside the encouraging results, we discuss differences and similarities across EHR naming conventions and NER tagsets.

Keywords: de-identification, clinical NLP, NER, electronic health records, cross-clinic evaluation

1. Introduction

Clinical text mining is a subfield of Natural Language Processing (NLP). Current NLP state of the art is based on pre-trained language models, which are typically trained on gigabytes – or even terabytes – of data (Devlin et al., 2019; Smith et al., 2022). Since any manual inspection or fine-grained annotation of sensitive data of this size would be unthinkable, there is a risk of leaking sensitive information about persons mentioned in the datasets (Carlini et al., 2020). The privacy-breaching risks of models trained on sensitive data are especially problematic in the clinical domain, where training corpora often consist of sensitive electronic health records (EHR). While general-purpose datasets *can* contain sensitive documents, nearly *all* EHRs contain sensitive data to some degree. One source of concern is the prevalence of *Protected Health Information* (PHI) in the data, such as names and other identifiers.

De-identification of PHI can be addressed using Named Entity Recognition (NER), a prolific subfield of NLP. Removing a PHI or replacing it with a surrogate value is called *automatic de-identification*. Due to the privacy regulations of the GDPR[1], the datasets containing PHI used to train clinical NER systems cannot be shared. Typically only researchers who have signed a confidentiality agreement have access to the source EHRs. Because of this restriction, clinical NER systems are trained and tested on data from the same institution. Furthermore, it is rarely possible to evaluate a de-identification system on data from outside the institution that trained the model. In these cases, it is impossible to assess whether a NER system has overfitted to the particular ways that the sources of the electronic health records have been written.

Since we have the rare opportunity to test a clinical NER model trained on EHRs from one hospital on EHRs from a different hospital, we present the results of such an evaluation together with a discussion about differences and similarities across EHR naming conventions and NER tagsets. Specifically, we evaluate a de-identification system pre-trained on a dataset based on EHRs from Karolinska University Hospital (Region Stockholm)[2] on a test set built on the EHRs from Linköping University Hospital (Region Östergötland). All of the EHRs are written in Swedish. The EHRs used for pre-training the de-identification model belong to the *Health Bank* (Dalianis et al., 2015), while the EHRs used for testing[3] come from the *LIU-Hospital-EMRs-collection* (Jerdhaf et al., 2021).

Our results are encouraging. However, they also show an urgent need to harmonize annotation standards, since many institutions and regions in Sweden follow different naming conventions and thus require different NER tagsets.

2. Related Work

Certain tasks, such as de-identifying EHRs, have significant ethical implications. Thus, it is extra important that benchmark results for such problems are not only *internally* valid but also generalize to the problem more broadly. However, internal (or intrinsic) evaluation is the norm in machine learning or deep learning. Normally, intrinsic evaluation is ”self-asserted”, as pointed out by Liao et al. (2021), who examine the reliance on benchmarking as the primary evaluation

[1]The General Data Protection Regulation (GDPR) is a regulation of data protection and privacy in the European Union (EU) and the European Economic Area (EEA).

[2]This research has been approved by the Swedish Ethical Review Authority under permission nr. 2019-05679.

[3]The research has been approved by the Swedish Ethical Review Authority (Etikprövningsmyndigheten), authorization nr.: 2021-00890.

method for machine learning research. They argue that benchmarking, in which a model is trained on a subset of the available data and evaluated on a held-out dataset (Gareth et al., 2013), mainly focuses on confirming the *internal validity* of a model. The validity of the results relies on the assumption that the held-out dataset is representative of the problem that the benchmark aims to model. However, this assumption is rarely stated explicitly, nor is the problem that the benchmark is meant to represent always clearly defined.

When building a NER system to detect PHIs, the training data typically originates from a small set of related clinics located in a limited geographical area. The commonly used MIMIC and i2b2 datasets (Johnson et al., 2016; Johnson et al., 2020; Stubbs and Uzuner, 2015) share this trait. A de-identification system, however, should also be useful to users in other locations and settings than the creators of the system. The sensitive nature of clinical data, however, prohibits the free dissemination of training data which makes it difficult to assess how representative the data are in reality.

Yang et al. (2019) build a de-identifier using LSTM-CRFs trained using i2b2 data and evaluate it new data created by annotating EHRs from other clinics. Their evaluation shows that the performance of their de-identifier drops slightly when evaluating on data from other clinics. They suggest that de-identification systems be customized for a target clinic and their results highlight the importance of evaluating the cross-clinic validity of systems.

Since we have the rare opportunity to test a clinical NER model trained on EHRs from one hospital on EHRs from a different hospital, we start filling this gap with the findings presented in this paper.

3. Data and Datasets

3.1. Stockholm Health Bank EHRs

The NER dataset used for fine-tuning was the *Stockholm EPR PHI Corpus*. This corpus contains 4,480 manually annotated PHI entities spanning nine PHI classes and a total of 380,000 tokens (Dalianis and Velupillai, 2010). The annotated texts are from the aforementioned *Health Bank* and are EHRs from Stockholm hospitals that were written between 2006 and the first half of 2008. The annotators processed 100 EHRs sampled equally from five clinics in the following specializations: neurology, orthopaedics, oral surgery, infectious diseases and clinical nutrition.

3.2. LIU Test Set

The sample of EHRs used for testing come from the *LIU-Hospital-EMRs-collection* (Jerdhaf et al., 2021). This collection contains EHRs from three clinics, i.e. cardiology, neurology and orthopaedics (two locations). The size and the chronology of the collections are shown in Table 1. From each of the clinics, 1,000 sentences were randomly sampled, amounting to a total of 3,000 sentences. This sample set was pre-annotated using the Swedish BERT-NER model (Malmsten et al., 2020) fine-tuned on the SUC 3.0 dataset[4]. The pre-annotated sentences were then presented to an annotator who manually validated the tags and fixed the errors. The distribution of the NER tags in the test set are shown in Table 2, where PER stands for Person Name, LOC for location and ORG for Organization.

4. Experiments

4.1. NER with a Clinical BERT Model

A new clinical Swedish NER model was created using data from the *Health Bank*. This model is based on the SweDeClin-BERT model that is described and evaluated in Vakili et al. (2022). SweDeClin-BERT is based on the Swedish KB-BERT model (Malmsten et al., 2020) that has been adapted to the clinical domain through *continued pre-training* using de-identified data from the Health Bank (Dalianis et al., 2015).

The fine-tuned model – SweDeClin-BERT NER – was trained for three epochs using the *Stockholm EPR PHI Corpus* (described in section 3.1) and evaluated on a held-out test set containing 10% of the dataset. Table 4 shows the fine-tuned model's recall and precision for each of the PHI classes in the held-out test data.

4.2. Evaluation on LIU Test Set

The *Stockholm EPR PHI Corpus* used to create SweDeClin-BERT NER uses a different and more fine-grained NER tagset than the tagset employed by KB-BERT-NER on which the LIU test set has been based upon. Because of this difference, the output of SweDeClin-BERT NER needed to be mapped to the tags used for the LIU test set. Mapping First names and Last names to Person names was rather straightforward. However, the existence of the Health Care Unit-class in the SweDeClin-BERT NER tagset rendered the evaluation on the entities Person and Location somewhat problematic. Some of the entities tagged as Locations and Organizations in the LIU testset were flagged as health care units by the SweDeClin model, thus being counted as false negatives. It was, however, deemed that the Health Care Unit-class was suitable in some of these cases. The classes Location and Organization were therefore evaluated on a case-by-case.

The KB-BERT-NER model (Malmsten et al., 2020) used to pre-annotate the LIU test set was used as a baseline classifier. Both models were evaluated on the LIU test set. The recall, precision and F_1 scores of both models are shown in Table 4 where we can observe an increase of precision and a drop in recall compared to the baseline model and an overall better F_1 for Person. We can also observe a steep increase of precision and a similarly steep drop of recall for the Location-class resulting in an unchanged F_1. The SweDeClin-NER model did not tag any entities as Organizations, which is why there is no precision score. However, the model

[4]https://spraakbanken.gu.se/en/resources/suc3

Clinics	Size (MB)	Raw Words	EMRs	Time Span
Cardiology	543.278	52 610 553	664 821	2013-2019
Neurology	294.745	29 622 531	314 669	2013-2019
Orthopaedics US	332.414	35 835 451	481 902	2015-2020
Orthopaedics ViN	280.130	29 791 200	361 097	2013-2020
Total	1450.567	147 859 735	1 822 489	5-7 years

Table 1: Clinics, size and chronology of *LIU-Hospital-EMRs-collection*

Clinic	**PER**	**LOC**	**ORG**
Cardiology	99	33	5
Neurology	95	10	7
Orthopedics	89	14	12
Total	283	57	24

Table 2: Distribution of entities in the test set.

PHI Class	**Recall**	**Precision**	**F1**
Age	100%	100%	1.0
First Name	97%	98%	0.97
Last Name	96%	97%	0.96
Partial Date	99%	98%	0.98
Full Date	87%	91%	0.89
Phone Number	93%	89%	0.91
Health Care Unit	89%	88%	0.97
Location	89%	81%	0.85
Organization	29%	80%	0.43

Table 3: SweDeClin-BERT NER's recall and precision for each PHI class are displayed and were calculated on the test data from Dalianis and Velupillai (2010).

tagged 59 % of the Organization entities as Health Care Units.

5. Discussion

The results of our evaluation are informative and highlight a number of issues that are currently uncharted. We focus on two influential factors, namely non-standardized NER tagsets and differing naming conventions across institutions.

5.1. NER Tagsets

There is no consensus on what NER-tags to use for automatic de-identification, and all configurations come with advantages and drawbacks. The *Stockholm EPR PHI Corpus* departs from the standard set of HIPAA categories that frequently serve as a starting point.

For example, while HIPAA only considers *names*, *Stockholm EPR PHI Corpus* and SweDeClin-BERT NER classifies first and last names separately. This finer-grained label has the advantage that it allows for higher-quality surrogate replacement, since a de-identification system can maintain separate word lists for the different types of names. On the other hand, the sets of names overlap and this introduces ambiguity when determining whether a classification was correct or not. In contrast, the LIU test set is closer to the HIPAA definition of PHIs as it considers all names equal by including both in the *Person* label.

Merging the first and last names into a single class is trivial, making the mapping between the labels of the datasets easy. However, we also discovered a discrepancy regarding *titles*. The *Stockholm EPR PHI Corpus* does not consider a persons title as part of the name, but the Linköping dataset includes the title in their definition of the *Person* entity.

Other classes are ambiguous in more subtle ways. The *Stockholm EPR PHI Corpus* treats *Locations*, *Organizations* and *Heath Care Units* as separate classes while the Linköping dataset only distinguishes between *Locations* and *Organizations*. Whether or not a health care unit should be considered as an organization or a location is not obvious. In fact, the entity can fill both functions depending on the context. For example, a patient can be *treated by* a clinic (organization) or be *physically at* a clinic (location).

Similarly, sometimes a hospital will only be referred to by its geographical location. For example, the Linköping University Hospital may be referred to simply as *Linköping* because the rest is obvious from the context. In such cases, the correct entity might be *Organization* even though the word is only referring to a Location.

5.2. Cross-Institutional Research Challenges

This cross-institutional study is, to the best of our knowledge, the first study measuring the generalizability of a Swedish de-identification system. Considerable efforts were made to lessen the impacts of the legal hurdles that arise from complying with privacy laws.

The restrictions arising from the sensitive nature of the data made it challenging to interpret the results. For example, the co-authors could not look at each others classifications across institutions. This made the error analysis data more challenging than it had otherwise been. Any nuances in annotation standards, such as the lack of titles in the *Stockholm EPR PHI Corpus*, had to be discovered on the results without context.

5.3. Conclusions and Future Work

In this exploratory study, we cross the institutional boundaries and test a BERT-based Swedish clinical

PHI Class	KB-BERT NER			SweDeClin-BERT NER		
	Recall	Precision	F_1	Recall	Precision	F_1
Person	97%	72%	0.83	85%	98%	0.91
Location	94%	68%	0.79	67%	95%	0.79
Organization	50%	54%	0.51	59%	0%	-

Table 4: The recall, precision, and F_1 for the PHI classes labeled in the LIU test set. Metrics for KB-BERT NER are shown on the left while the metrics for SweDeClin-BERT NER are shown on the right.

NER model pre-trained on EHRs from Stockholm on the EHRs from clinics in Linköping. Results are encouraging and highlight nuances and caveats that we had not foreseen, such as the difficulty of mapping different NER tagsets. Future work includes retagging the LIU test set using Stockholm tagset, using SweDeClin-BERT to create pre-annotations. This would yield a more detailed gold-standard for that would be useful for anonymization. Moreover, harmonizing the NER tagset would facilitate the evaluation of NER models across institutions.

Acknowledgements

Research funded by *Vinnova* (Sweden's innovation agency), *Grant*: 2021-0169 and by the *DataLEASH* project.

Bibliographical References

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.

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, January.

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.

Gareth, J., Sohil, F., Sohali, M. U., Shabbir, J., Witten, D., Hastie, T., and Tibshirani, R., (2013). *An introduction to statistical learning with applications in R*, page 198. Springer Science and Business Media, New York.

Jerdhaf, O., Santini, M., Lundberg, P., Karlsson, A., and Jönsson, A. (2021). Implant term extraction from swedish medical records–phase 1: Lessons learned. In *Swedish Language Technology Conference and NLP4CALL*, pages 35–49.

Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L.-W. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., and Mark, R. G. (2016). Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Johnson, A., Bulgarelli, L., Pollard, T., Horng, S., Celi, L., and Mark, R. (2020). Mimic-iv (version 1.0).

Liao, T., Taori, R., Raji, I. D., and Schmidt, L. (2021). Are We Learning Yet? A Meta-Review of Evaluation Failures Across Machine Learning.

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*.

Smith, S., Patwary, M., Norick, B., LeGresley, P., Rajbhandari, S., Casper, J., Liu, Z., Prabhumoye, S., Zerveas, G., Korthikanti, V., Zhang, E., Child, R., Aminabadi, R. Y., Bernauer, J., Song, X., Shoeybi, M., He, Y., Houston, M., Tiwary, S., and Catanzaro, B. (2022). Using DeepSpeed and Megatron to Train Megatron-Turing NLG 530B, A Large-Scale Generative Language Model. *arXiv:2201.11990 [cs]*, February. arXiv: 2201.11990.

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

Vakili, T., Lamproudis, A., Henriksson, A., and Dalianis, H. (2022). Downstream Task Performance of BERT Models Pre-Trained Using Automatically De-Identified Clinical Data. (Accepted to LREC 2022).

Yang, X., Lyu, T., Li, Q., Lee, C.-Y., Bian, J., Hogan, W. R., and Wu, Y. (2019). A study of deep learning methods for de-identification of clinical notes in cross-institute settings. *BMC medical informatics and decision making*, 19(Suppl 5):232, December.