Digital Personalized Health and Medicine
L.B. Pape-Haugaard et al. (Eds.)

doi:10.3233/SHTI200140

De-Identifying Swedish EHR Text Using Public Resources in the General Domain

Taridzo CHOMUTARE [a,1], Kassaye Yitbarek YIGZAW [a] Andrius BUDRIONIS [a]
Alexandra MAKHLYSHEVA [a] Fred GODTLIEBSEN [a,c] and Hercules DALIANIS [a,b]

[a] *Norwegian Centre for E-health Research, Tromsø, Norway*
[b] *Department of Computer and Systems Sciences, Stockholm University, Sweden*
[c] *Faculty of Science & Technology, UiT - The Arctic University of Norway*

Abstract. Sensitive data is normally required to develop rule-based or train machine learning-based models for de-identifying electronic health record (EHR) clinical notes; and this presents important problems for patient privacy. In this study, we add non-sensitive public datasets to EHR training data; (i) scientific medical text and (ii) Wikipedia word vectors. The data, all in Swedish, is used to train a deep learning model using recurrent neural networks. Tests on pseudonymized Swedish EHR clinical notes showed improved precision and recall from 55.62% and 80.02% with the base EHR embedding layer, to 85.01% and 87.15% when Wikipedia word vectors are added. These results suggest that non-sensitive text from the general domain can be used to train robust models for de-identifying Swedish clinical text; and this could be useful in cases where the data is both sensitive and in low-resource languages.

Keywords. EHR, clinical text, de-identification, deep learning, wiki word vectors

1. Introduction

De-identifying health data is an important problem for health data reuse, and the topic has generated significant scholarly interest because of increased use of electronic health records (EHR). Re-use of the data in research could give us unique insights into disease etiology and progression, as well as a greater understanding of patient care processes and pathways. Current de-identification methods rely on sensitive health data for training. This presents a number of data-sensitivity problems, such as when there is need to transfer or adapt the models to new target data. In this study, we investigate the usefulness of non-sensitive training data from the general domain.

Two main approaches have so far been used for de-identification namely, rule-based and machine learning-based methods [1]. Studies show that more successful de-identification systems use a hybrid of both these approaches [2]. On the one hand, rule-based methods can go as far as using name lists from the economy/administration software to match against the clinical text [3]. While this can be an effective solution, it is not robust enough for simple variations or for use outside the specified datasets or organizations, and could entail serious risks to patient privacy. On the other hand, machine

[1] Corresponding Author: Taridzo Chomutare; E-mail: firstname.lastname@ehealthresearch.no

learning approaches, while more robust since they learn patterns, instead of matching specific instances, still require a large amount of sensitive data.

Machine learning approaches require a lot of training data or examples to learn from. Creating examples by annotating the data is an expensive proposition because it requires specialist knowledge, and the amounts of data are enormous. Unsupervised methods which can be used to discover discriminating features in new target datasets are emerging. These emerging deep learning architectures do not require any feature engineering to produce state of the art results [4]. So far however, these architectures have only used embeddings from sensitive data or scientific medical publications like PubMed [5]. Out-of-domain sources such as Wikipedia or newer general language models like BERT [6] have not been extensively explored for this task on medical text.

Exploring use of non-sensitive data, the validity of using pseudonymised clinical text for de-identification is studied in [7] where the Stockholm EPR PHI Pseudo Corpus [8] is used and compared with Stockholm EPR PHI Corpus, the non-pseudonymised corpora. It is shown that the results using pseudonymised corpora as training data are slightly decreased, suggesting limited potential.

In another approach, McMurray et al. [3] used both EHR text and text from publicly published medical journals for training purposes. The authors argued medical publications will generally not contain enough protected health information (PHI) information, and this could be a discriminating factor. In contrast, a recent study by Berg et al. [9] found no additional benefits of using out-of-domain training material for de-identification using deep learning approaches.

Whether non-sensitive medical text such as scientific medical publications or even text from the general domain is useful for de-identification, is still a matter without fully resolved clarity. In this study we test both these non-sensitive sources and contribute evidence to help answer the question.

2. Method

Experiments will compare the effect of adding medical scientific text versus text from the general domain to the training set for a de-identification deep learning model. The comparisons are with (i) the base embedding layer from the EHR text, (ii) EHR text plus medical scientific text, and (iii) EHR text plus Wikipedia word vectors. These data sources are detailed in the succeeding subsections.

2.1. Stockholm EPR PHI Psuedo Corpus

Stockholm EPR PHI Pseudo Corpus[2] is a Swedish EHR corpus, which has been de-identified and pseudonymized [8], and where the tokens are annotated with PHI information. Stockholm EPR PHI Psuedo Corpus is part of the Health Bank [10], the Swedish Health Record Research Bank[3]. The Health Bank encompasses structured and unstructured patient records data from 512 clinical units from Karolinska University Hospital collected from the years 2007 to 2014 encompassing over 2 million patients. The dataset uses a less fine-grained annotation scheme (IOB), indicating [I=inside token], [B=begin token], and [O = not PHI token].

[2]Research approved by the Regional Ethical Review Board in Stockholm; permission no. 2014/1607-32.
[3]Health Bank, http://www.dsv.su.se/healthbank

2.2. Scientific medical journal and Swedish Wiki word vectors

Scientific medical text is based on the Läkartidningen corpus (The Swedish scientific medical journal from 1996 to 2005). Läkartidningen has publicly available articles at Språkbanken[4]. Wiki word vectors are pre-trained word vectors created with fastText from Swedish Wikipedia text [11], and are publicly available at fastText[5]. They are designed with no specific downstream task in mind, but what makes them interesting is their use of character-level n-grams, where a single word can be represented by several character n-grams.

2.3. Deep recurrent neural networks

A state of the art deep learning algorithm previously used on health data [5], the Bidirectional Long Short-Term Memory algorithm with conditional random fields (BI-LSTM-CRF), was used in the experiments, as implemented in TensorFlow/Keras[6]. For the scientific medical text, we used another state of the art method, Word2Vec, to create the word embeddings. Wikipedia word vectors are made available to the public pre-trained and ready for downstream tasks. Both sources have 300 dimensional vector representation.

3. Results

The results in Table 1 show a clear improvement in results, from adding Wiki word vectors to the base embedding layer with EHR data only. We also observe that adding scientific medical text improves performance, but falls short of Wiki word vectors.

PHI	EHR			EHR + Scientific medical text			EHR + Wikipedia		
	P %	R %	F1	P %	R %	F1	P %	R %	F1
Age	66.67	40.00	50.00	100.00	80.00	88.89	100.00	80.00	88.89
Date Part	62.87	83.24	71.63	92.09	91.06	91.57	87.76	96.09	91.73
First Name	72.22	87.39	79.09	89.83	66.81	76.63	95.78	95.38	95.58
Full Date	50.00	85.54	63.11	67.23	96.39	79.21	80.41	93.98	86.67
Health Care U.	40.39	77.15	53.02	67.10	77.15	71.78	71.43	82.4	76.52
Last Name	91.61	97.26	94.35	77.01	98.63	86.49	92.95	99.32	96.03
Location	21.15	18.64	19.82	87.50	11.86	20.90	100.00	15.25	26.47
Phone Number	17.39	42.11	24.62	66.67	31.58	42.86	92.86	68.42	78.79
Avg	55.62	80.02	65.62	77.83	77.21	77.52	85.01	87.15	86.07

Table 1. De-identification results based on the three comparisons, P=Precision, R= recall, both percentage

There are a number of reasons that could explain why the Wikipedia text performed better than medical text. First, Wikipedia is a rich source of information which contains both general text and medicine-related text as well. In addition, a number of PHI information such as first and last names, ages, year, and location are present in the text. Also, the scientific medical journal corpus in Swedish (Läkartidningen) produced 118,683 vectors while Wikipedia, on the other hand, produced 1,143,274 vectors.

Further, we observed that the scientific medical text start's out with a relatively high error loss in each epoch, while initial error loss is much lower for Wikipedia. In terms of

[4]Läkartidningen, https://spraakbanken.gu.se/swe/resurs/lakartidn-vof
[5]fastText, https://fasttext.cc
[6]TensorFlow, http://www.tensorflow.org

the improvement in F1 measures (see Figure 1), there was significant performance gain for *Age* and *Phone Number*. For scientific medical text, we noted poorer performance for some PHI information like first names and last names, compared to the EHR baseline.

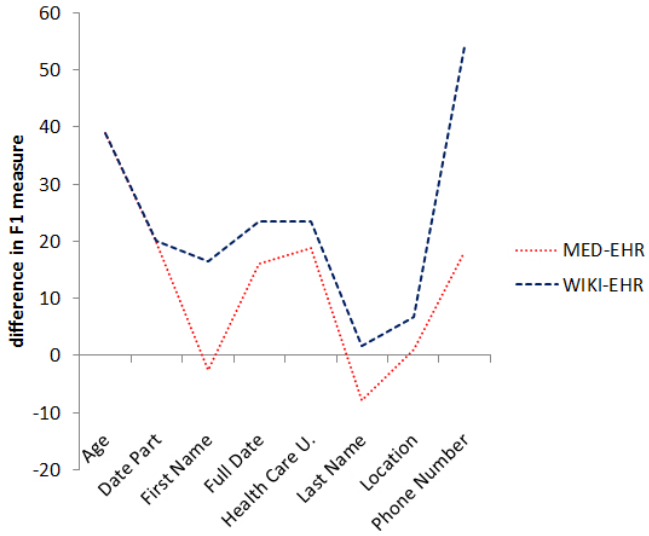

Figure 1. The graph shows the PHI differences in F1 measures between scientific medical text and the EHR baseline (MED-EHR) and between Wikipedia and the EHR baseline (WIKI-EHR) respectively.

4. Discussion

It appears the general consensus in scholarship is that training on general-domain text is not appropriate for tasks on clinical text, since clinical text is so different that it represents a unique linguistic genre. The language in clinical notes is meant for other healthcare professionals. Clinicians and nurses write these notes under time pressure, therefore the text has abbreviations, misspellings, unusual grammatical constructs and other errors and ambiguities.

Our results support a counter-argument that PHI information is distinct from clinical text since PHI information is general, as opposed to clinical procedures, medication or medical concepts that are present in clinical text. Therefore, it could be appropriate to use non-sensitive text in the general domain as training data for detecting PHI information. Also, deep learning architectures have been reported to show good performance under different domains and languages.

The poor results obtained with scientific medical text is consistent with previous assertions made in the literature, that is, scientific text is not likely to contain names and surnames in meaningful contexts [3]. However, the significant improvement in *Age* and *Phone Number* suggest that scientific medical text could still be useful for detecting specific PHI information. Therefore, combining this medical text with other sources could be a viable option.

5. Conclusion

Current results suggest that non-sensitive resources in the general domain can be useful for de-identification tasks on clinical notes. Even though deep learning models are generally thought of as data-hungry, current results raise the prospect of creating robust models; where the primary training data is sensitive and low resourced. In the future, we will test non-sensitive resources and language models to adapt and transfer deep learning models for de-identifying clinical notes between closely similar Nordic languages; such as between Swedish and Norwegian clinical notes.

Acknowledgments

This work is partially supported by the Northern Norway Regional Health Authority, Helse Nord; research grant HNF1395-18.

References

[1] O. Ferrández, B.R. South, S. Shen, F.J. Friedlin, M.H. Samore and S.M. Meystre, Evaluating current automatic de-identification methods with Veteran's health administration clinical documents, *BMC medical research methodology* **12**(1) (2012), 109.

[2] A. Dehghan, A. Kovacevic, G. Karystianis, J.A. Keane and G. Nenadic, Combining knowledge-and data-driven methods for de-identification of clinical narratives, *Journal of biomedical informatics* **58** (2015), S53–S59.

[3] A.J. McMurry, B. Fitch, G. Savova, I.S. Kohane and B.Y. Reis, Improved de-identification of physician notes through integrative modeling of both public and private medical text, *BMC medical informatics and decision making* **13**(1) (2013), 112.

[4] F. Dernoncourt, J.Y. Lee, O. Uzuner and P. Szolovits, De-identification of Patient Notes with Recurrent Neural Networks, 2016.

[5] Z. Liu, B. Tang, X. Wang and Q. Chen, De-identification of clinical notes via recurrent neural network and conditional random field, *Journal of Biomedical Informatics* **75** (2017), S34–S42.

[6] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, *arXiv preprint arXiv:1810.04805* (2018).

[7] H. Berg, T. Chomutare and H. Dalianis, Building a De-identification System for Real Swedish Clinical Text Using Pseudonymised Clinical Text, in: *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019),in conjuction with Conference on Empirical Methods in Natural Language Processing, (EMNLP) November 2019, Hongkong, ACL.*, 2019, pp. 118–125.

[8] H. Dalianis, Pseudonymisation of Swedish Electronic Patient Records using a rule-based approach, in: *Proceedings of the Workshop on NLP and Pseudonymisation, NoDaLiDa, Turku, Finland September 30, 2019*, 2019.

[9] H. Berg and H. Dalianis, Augmenting a De-identification System for Swedish Clinical Text Using Open Resources (and Deep learning), in: *Proceedings of the Workshop on NLP and Pseudonymisation, NoDaLiDa, Turku, Finland September 30*, 2019.

[10] H. Dalianis, A. Henriksson, M. Kvist, S. Velupillai and R. Weegar, HEALTH BANK-A Workbench for Data Science Applications in Healthcare., in: *CAiSE Industry Track*, 2015, pp. 1–18.

[11] P. Bojanowski, E. Grave, A. Joulin and T. Mikolov, Enriching word vectors with subword information, *Transactions of the Association for Computational Linguistics* **5** (2017), 135–146.