HB Deid - HB De-identification tool demonstrator

Hercules Dalianis
Department of Computer
and Systems Sciences
Stockholm University
Kista, Sweden
hercules@dsv.su.se

Hanna Berg
Department of Computer
and Systems Sciences
Stockholm University
Kista, Sweden
hanna@hidetext.se

Abstract

This paper describes a freely available web-based demonstrator called HB Deid. HB Deid identifies so-called protected health information, PHI, in a text written in Swedish and removes, masks, or replaces them with surrogates or pseudonyms. PHIs are named entities such as personal names, locations, ages, phone numbers, dates. HB Deid uses a CRF model trained on non-sensitive annotated text in Swedish, as well as a rule-based post-processing step for finding PHI. The final step in obscuring the PHI is then to either mask it, show only the class name or use a rule-based pseudonymisation system to replace it.

1 Introduction

Electronic patient records and other health data contain information that can identify a patient. These data need to be *washed*, for both legal and ethical reasons before they can be re-used for various purposes as research or machine learning.

This paper presents a machine learning-based demonstrator called HB Deid, Health Bank De-Identification tool. It is a tool for automatic de-identification and pseudonymisation of protected health information, PHI. PHIs are information that can identify a patient. PHIs are similar to named entities and encompass for example *personal names, locations, phone numbers, dates, ages*, etc. PHIs can be present both in structured and unstructured clinical data such as free text. In structured data PHIs are easily identifiable, the table information informs of the category of the data the whole table can be removed or obscured. In the free clinical text there is a greater effort to identify a PHI since, for example, the PHI entity *Parkinson* is difficult to classify whether it is a disease name or a personal name. Similarly, *Sjögren's* in *Sjögren's syndrome* may be mistaken for a patient's name. In this paper, we focus on PHIs[1] in free text in electronic patient records.

A demonstrator is a pedagogical instrument to show a system or an idea and let a broader audience explore - in our case - a research system or pilot system.

The process of de-identifying text typically uses two steps. Firstly, a PHI is identified through named entity recognition. Secondly, the PHI is obscured. Strategies for hiding information may be masking the information or replacing it with a surrogate through a process called pseudonymisation. Pseudonymisation makes the text more fluent to read and when also removing the annotation tags it inconceives the identification of potentially remaining sensitive data in plain sight and protects the identification of PHIs. This method is called Hiding in Plain Sight (HIPS), (Carrell et al., 2012).

A substantial amount of studies have been published on the de-identification of text (Meystre et al., 2010; Stubbs et al., 2015). While most of these studies have focused on English, research have been carried out for French (Grouin and Névéol, 2014), Spanish (Marimon et al., 2019), Danish (Pantazos et al., 2016) and Swedish (Berg and Dalianis, 2019) as well as Japanese (Kajiyama et al., 2020).

Generally, high recall is preferred over high precision in de-identification research as the privacy of the individuals describe is of paramount importance. It is therefore important to not miss any sensitive information.

With regards to pseudonymisation, there are fewer studies. One of the first was a study by Sweeney (1996), which described a system for

HideText, http://www.hidetext.se is the platform where HB Deid is commercialised.

[1] PII, Personally identifying information, is a more general term which includes other domains.

identifying PHI and then replacing them with surrogates, but not how this process was carried out. In another study by Douglass et al. (2004) this pseudonymisation process is described elaborated such as that dates were shifted, personal names were shifted to other personal names in the Boston area. Locations were shifted randomly, and hospitals were given fictitious names.

For de-identification and pseudonymisation for English there is yet another study (Deleger et al., 2014). For pseudonymisation for Swedish, there is a system described in (Dalianis, 2019).

While there are plenty of demonstrators of Named Entity Recognition systems the only available one for de-identification, to our knowledge, is the HitzalMed demonstrator (Lopez et al., 2020). The HitzalMed demonstrator is constructed for de-identification and pseudonymisation of Spanish electronic patient records. To try it out a registration must be carried out here[2]. The system identifies and categorizes entities into different categories, as: *first name, last name, location, phone number, age, date* and *health care unit*. While the system is intended for Spanish, this part works relatively well for English and Swedish. The system then either masks or replaces the sensitive information with surrogates - which are in Spanish.

2 The HB Deid demonstrator

The HB Deid demonstrator[3], see Figure 1, is an attempt to show the possibilities of de-identification and pseudonymisation techniques for electronic patient records written in Swedish. The system is based on the work carried out by Berg and Dalianis (2019) for de-identification and by Dalianis (2019) for the pseudonymisation.

The data used to train HB Deid is not the original set of annotated sensitive electronic patient records in Swedish - the Stockholm EPR PHI Corpus, but the corpus is indirectly used since its trained model is used to improve the public available Swedish news corpora called Webbnyheter 2012 that are semi-manually annotated for the NER classes PER and LOC and ORG and MISC[4]. Webbnyheter 2012 was machine annotated using the original sensitive trained model from Stockholm EPR PHI Corpus and then it was corrected manually, using both the manual annotations and the machine annotations to decide on the correct annotation. This effort was carried out to avoid using the sensitive Stockholm EPR PHI Corpus directly.

The bootstrapped model in the HB Deid demonstrator has not yet been evaluated, but the original sensitive model has been evaluated and the results reported in (Berg and Dalianis, 2019).

2.1 The HB Deid classes

The HB Deid demonstrator identifies the following PHI classes: *First name, last name, location, phone number, age, date, health care unit, organisation* and *personal number (social security number)*.

2.2 Programming languages and machine learning environment

HB Deid is developed in the programming language Python and is machine learning-based with a rule-based post-processing step, (Berg and Dalianis, 2019). It uses the CRF Conditional Random Fields algorithm (Lafferty et al., 2001) as implemented in CRFSuite (Okazaki, 2007) with a sklearn-crfsuite wrapper[5].

The pseudonymiser is completely rule-based and uses dictionaries to generate surrogates in a fashion similar to (Dalianis, 2019). Compared to the pseudonymiser in (Dalianis, 2019) personal names are replaced with greater variation. While common names are replaced with other common names, uncommon names are replaced with uncommon names. The uncommon names in the dictionaries are 123,000 female first names, 121,000 male first names and 35,000 last names.

The web interface for the HB-demo is written in Flask[6] that in turn is coded in Python.

2.3 Interface

The Flask web interface for HB Deid uses encryption according to the HTTPS protocol. Nothing processed is saved on the web-server.

See Figure 1 for all the possible menu choices and below for a detailed description:

- *Ersättare - Replacer*. Decides the shapes of the processed text, see the following choices.

[2]HitzalMed registration, https://snlt.vicomtech.org/hitzalmed/demo/help

[3]HB Deid, http://hbdeid.dsv.su.se

[4]Swedish NER Corpus, https://github.com/klintan/swedish-ner-corpus

[5]sklearn-crfsuite, https://github.com/TeamHG-Memex/sklearn-crfsuite

[6]Flask, https://flask.palletsprojects.com/en/1.1.x/

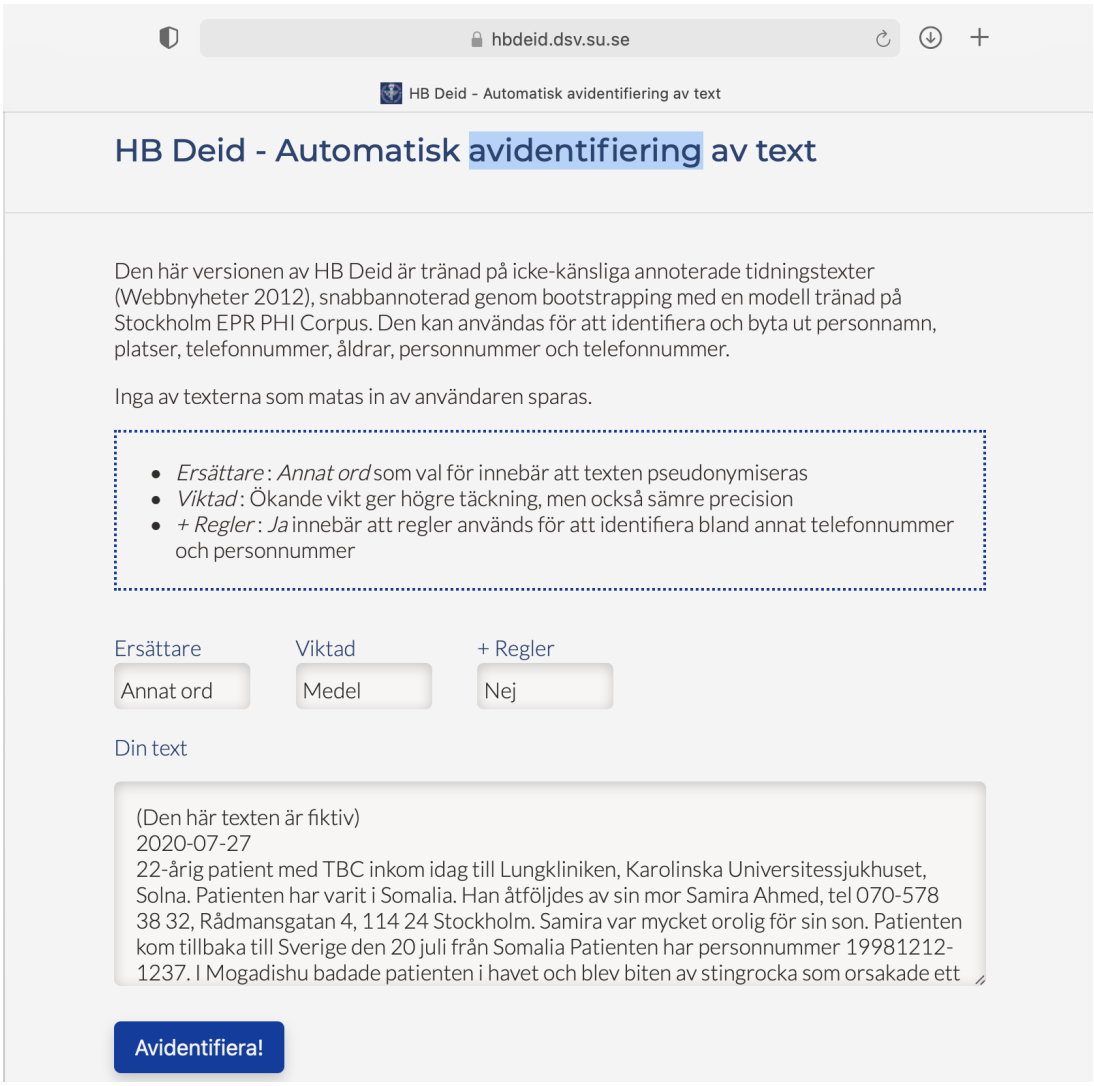

Figure 1: The interface of HB Deid (In Swedish) with the various choices. The text shown is fictitious.

  - *Ersätt inte - Do not replace*. Tag the PHI with the class name.
  - *Annat ord - Other word*. Replace the PHI with a surrogate or pseudonym.
  - *Klass - Class*. Replace the PHI with the class name.
  - *Mask*. Mask the PHI with XXX.

- *Vikter - Weights*. Increases the recall in three steps.

- *+ Regler - Rules* uses a post-processing step utilising rules, mainly controlling the output from the machine learning tool but also uses regular expressions to find personal numbers and phone numbers.

The output from HB Deid can be seen in Figure 2.

The interface and the functionality of the HB Deid demonstrator have not been evaluated yet, since one bottleneck is that the demonstrator must either be installed at the hospital or be set up inside the Health Bank[7] infrastructure laboratory environment at the university to be evaluated by clinicians.

3 Conclusion

We have shown how to train and construct an automatic de-identification and pseudonymisation tool for clinical text in Swedish. We have also used a bootstrapped language model that is privacy protected. Finally, we have made a user friendly and freely available web interface to the demonstrator called HB Deid.

The demonstrator has been presented in the

[7]Health Bank, `https://dsv.su.se/healthbank`

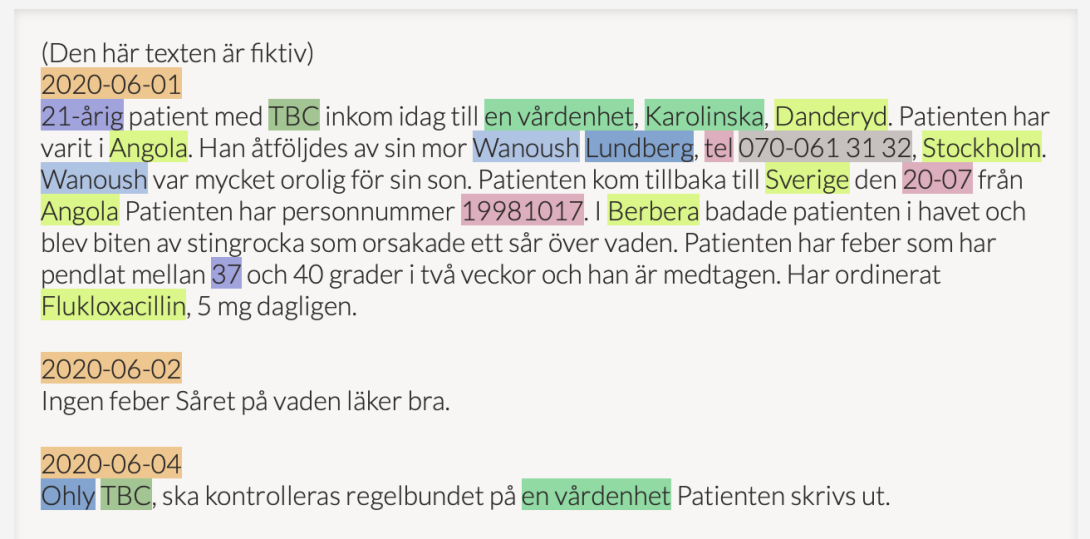

Figure 2: The output from HB Deid in form of a de-identified and pseudonymised text. The various colours represent the different classes. Moving the mouse pointer over an identified coloured entity will display the corresponding class name.

Swedish trade magazine Computer Sweden, (Lindström, 2020).

We have also been contacted by many users that had asked us about HB Deid and given us feedback on how to improve the system. They have also asked us if they can use HB Deid for other purposes as de-identification of transcribed interviews.

One more proposal was to have the possibility to correct wrong predictions by re-annotate them directly in the HB Deid interface to re-learn HB Deid.

We plan to let Stockholm Regional Council and Karolinska University Hospital test HB Deid on their electronic patient record texts with the future aim to de-identify them before they are handed out to researchers or for machine learning purposes.

Acknowledgements

We are grateful to the DataLEASH project for funding this research work.

References

Hanna Berg and Hercules Dalianis. 2019. Augmenting a de-identification system for Swedish clinical text using open resources and deep learning. In *Proceedings of the Workshop on NLP and Pseudonymisation*, pages 8–15, Turku, Finland. Linköping Electronic Press.

David Carrell, Bradley Malin, John Aberdeen, Samuel Bayer, Cheryl Clark, Ben Wellner, and Lynette Hirschman. 2012. Hiding in plain sight: use of realistic surrogates to reduce exposure of protected health information in clinical text. *Journal of the American Medical Informatics Association*, 20(2):342–348.

Hercules Dalianis. 2019. Pseudonymisation of Swedish electronic patient records using a rule-based approach. In *Proceedings of the Workshop on NLP and Pseudonymisation*, pages 16–23, Turku, Finland. Linköping Electronic Press.

Louise Deleger, Todd Lingren, Yizhao Ni, Megan Kaiser, Laura Stoutenborough, Keith Marsolo, Michal Kouril, Katalin Molnar, and Imre Solti. 2014. Preparing an annotated gold standard corpus to share with extramural investigators for de-identification research. *Journal of Biomedical Informatics*, 50:173–183.

Margaret Douglass, Gari D. Clifford, Andrew Reisner, George B. Moody, and Roger G. Mark. 2004. Computer-assisted de-identification of free text in the MIMIC II database. In *Computers in Cardiology, 2004*, pages 341–344. IEEE.

Cyril Grouin and Aurélie Névéol. 2014. De-identification of clinical notes in french: towards a protocol for reference corpus development. *Journal of biomedical informatics*, 50:151–161.

Kohei Kajiyama, Hiromasa Horiguchi, Takashi Okumura, Mizuki Morita, and Yoshinobu Kano. 2020. De-identifying free text of japanese electronic health records. *Journal of Biomedical Semantics*, 11(1):1–12.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional Random Fields: Probabilistic models for segmenting and labeling se-

quence data. In *Proceedings 18th International Conference on Machine Learning*, pages 282–289. Morgan Kaufmann.

Karin Lindström. 2020. AI förbereder för AI – tvättar bort känsliga uppgifter ur texter (In Swedish), https://computersweden.idg.se/2.2683/1.744319/ai-tvattar-kansliga-uppgifter. *Computer Sweden*.

Salvador Lima Lopez, Naiara Perez, Laura García-Sardiña, and Montse Cuadros. 2020. HitzalMed: Anonymisation of Clinical Text in Spanish. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, May 13-15, Marseille*, pages 7038–7043.

Montserrat Marimon, Aitor Gonzalez-Agirre, Ander Intxaurrondo, Heidy Rodriguez, Jose Lopez Martin, Marta Villegas, and Martin Krallinger. 2019. Automatic De-identification of Medical Texts in Spanish: the MEDDOCAN Track, Corpus, Guidelines, Methods and Evaluation of Results. In *IberLEF@ SEPLN, Sociedad Española para el Procesamiento del Lenguaje Natural*, pages 618–638.

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.

Naoaki Okazaki. 2007. CRFsuite: a fast implementation of Conditional Random Fields, http://www.chokkan.org/software/crfsuite. Accessed 2021-02-02.

Kostas Pantazos, Søren Lauesen, and Søren Lippert. 2016. Preserving medical correctness, readability and consistency in de-identified health records. *Health Informatics Journal*, page 1460458216647760.

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:S11–S19.

Latanya Sweeney. 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.