Applying and Sharing pre-trained BERT-models for Named Entity Recognition and Classification in Swedish Electronic Patient Records

Mila Grancharova
Department of Computer
and Systems Sciences
Stockholm University
Kista, Sweden
mila.grant@gmail.com

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

Abstract

To be able to share the valuable information in electronic patient records (EPR) they first need to be de-identified in order to protect the privacy of their subjects. Named entity recognition and classification (NERC) is an important part of this process. In recent years, general-purpose language models pre-trained on large amounts of data, in particular BERT, have achieved state of the art results in NERC, among other NLP tasks. So far, however, no attempts have been made at applying BERT for NERC on Swedish EPR data.

This study attempts to fine-tune one Swedish BERT-model and one multilingual BERT-model for NERC on a Swedish EPR corpus. The aim is to assess the applicability of BERT-models for this task as well as to compare the two models in a domain-specific Swedish language task. With the Swedish model, recall of 0.9220 and precision of 0.9226 is achieved. This is an improvement to previous results on the same corpus since the high recall does not sacrifice precision. As the models also perform relatively well when fine-tuned with pseudonymised data, it is concluded that there is good potential in using this method in a shareable de-identification system for Swedish clinical text.

1 Introduction

Electronic patient records (EPR), also called clinical text, contain valuable information about patients' symptoms, physicians' assessments, diagnoses, treatments and treatment outcomes. Advancements in natural language processing (NLP) and machine learning have made it possible to use large amounts of clinical text to assist physicians and medical researchers in detecting early symptoms of disorders, predicting adverse effects of treatments, etc, see Chapter 10 in (Dalianis, 2018). However, clinical text contains information that can reveal the identity of patients and other mentioned individuals, so called Protected Health Information (PHI). Methods have been developed to detect this information and obscure it in order to protect people's identities (Meystre et al., 2010; Stubbs et al., 2015). One important note to make is that de-identified text cannot be guaranteed to be safe to release and must still be handled with great care. A good de-identification system can, however, help facilitate an efficient anonymisation process.

In this study PHI refers only to the named entities which may reveal a person's identity, such as name, age and location. In this sense, detecting and identifying the PHI before obscuring it is a Named Entity Recognition and Classification (NERC) problem. When it comes to data-driven NERC, models based on recurrent neural networks (RNNs) and long short-term memory (LSTM) networks have been successfully used for several languages (Lê et al., 2020; Lange et al., 2019). In the last two years, however, transformer-based language models such as BERT have achieved state-of-the-art results in several NLP task on commonly used data sets (Devlin et al., 2019).

BERT is a general-purpose language model developed by Devlin et al. (2019). In essence, BERT is a neural network based on transformers. Transformers are a type of deep learning model designed to handle sequential data, such as natural language text. Since their introduction in 2017 (Vaswani et al., 2017), transformers have been widely used across a variety of NLP tasks, not least on clinical text (Lewis et al., 2020). The benefit of transformer-based models over previous architectures is that they do not require the sequential data to be processed in order, allowing for parallelization of the training process. This has made it possible to develop large pre-trained models such as BERT, which have been fitted on larger amounts of data than was previously feasible.

Since the first BERT-model was released in 2018, several models with modified architecture and different data used in pre-training have been released, including the multilingual M-BERT[1]. M-BERT is pre-trained on texts in 104 languages, including Swedish. In 2019, the National Library of Sweden released a Swedish BERT model, KB-BERT[2], pre-trained exclusively on Swedish texts.

To use a pre-trained BERT-model for a downstream task, it needs to be fine-tuned for that task. Both KB-BERT and M-BERT have shown success in the NERC task for Swedish when fine-tuned with the publicly available Stockholm-Umeå Corpus consisting of Swedish texts from the 1990's (Malmsten et al., 2020). To our knowledge, however, no previous attempt has been made at using these models for NERC in Swedish EPR data.

In this study, we attempt to improve NERC performance on Swedish electronic patient records by fine-tuning KB-BERT and M-BERT with domain-specific data. More specifically, our aim is to achieve high recall, which is a priority in the de-identification task, without sacrificing precision. A risk with de-identification methods based on machine learning is that a model trained on sensitive data could be re-engineered, revealing the data. In a BERT-model, there are no links between words in the vocabulary, making it infeasible to retrieve the patient records used for fine-tuning. However, due to names and other personal identifiers appearing in the model's vocabulary, there may be legal issues with releasing a model fine-tuned on patient records. Therefore, in an additional experiment, the models are fine-tuned using pseudonymised patient records to see how NERC performance on authentic records is affected.

The outline of this paper is as follows. First, Section 2 presents some previous studies on NERC in clinical text and specifically previous results on the data set at hand. Then, Section 3 describes the data used in this study, gives some more detail on the two BERT models and goes through how the fine-tuning and evaluation are performed. Section 4 presents the results for both models. Finally, the results are discussed in Section 5.

2 Related Research

There are several publicly available BERT-models pre-trained specifically for the biomedical and clinical domains. In 2019, Lewis et al. (2020) released BioBERT[3], a BERT-model pre-trained on PubMed articles as well as Wikipedia articles and books. The authors present an F_1-score of approximately 0.87 on the commonly used i2b2 2010 data set for clinical text NERC. In a different 2019 project, (Peng et al., 2019) continued to pre-train the pre-trained BERT-model released by (Devlin et al., 2019) on PubMed abstracts and clinical notes. This model, named BlueBERT[4], reaches an F_1-score of approximately 0.77 on the i2b2 data set. The same year, (Alsentzer et al., 2019) released clinicalBERT[5] pre-trained on clinical texts but also specifically on discharge summaries. The combined Bio+Discharge Summary model reaches an F_1-score of 0.88 on the i2b2 2010 data set. All of these models are only pre-trained on English texts.

For non-English clinical text NERC, some advancements were made in connection to the 2019 shared task MEDDOCAN which consisted of performing NERC on Spanish electronic patient records with annotated PHI. In a submission to the contest Mao and Liu (2019) used M-BERT, which is also pre-trained on Spanish text (Mao and Liu, 2019) with a decoding CRF layer for token classification. They also applied some post-processing techniques, achieving F_1-score and recall of approximately 0.93.

When it comes to Swedish, several attempts have been made at performing NERC on the annotated data set of electronic patient records Stockholm EPR PHI Corpus. In one study by Berg and Dalianis (2020) the authors extended the annotated data set with data generated using a semi-supervised learning method with the aim of increasing recall without sacrificing precision. The highest recall reported was 0.8920, at which point the precision was 0.9420. These results were achieved using a Conditional Random Field (CRF) model. Grancharova et al. (2020) managed to increase the recall to 0.9209 using the same model by under-sampling negative tokens, thus tokens not belonging to a PHI. However, this came at the cost of significant

[1] M-BERT, https://github.com/google-research/bert

[2] KB-BERT, https://github.com/Kungbib/swedish-bert-models

[3] BioBERT, https://github.com/dmis-lab/biobert

[4] BlueBERT, https://github.com/ncbi-nlp/bluebert

[5] clinicalBERT, https://github.com/EmilyAlsentzer/clinicalBERT

decrease in precision to 0.8819. Regarding the application of models trained on pseudonymised clinical data for NERC on authentic data, there is a study by Berg et al. (2019) where the authors achieved at highest recall of 0.5510 using a LSTM network. The experiment was repeated with a classic CRF and the recall decreased to 0.4983.

3 Data and Methods

This section describes the data, tools and methods used in this study. First, the EPR data set is described in Section 3.1. Then, Section 3.2 describes the BERT-models used and how they were fine-tuned. Lastly, Section 3.3 describes how the models were evaluated in a number of experiments.

3.1 Data

The data used in this study is Stockholm EPR PHI Corpus[6] Stockholm EPR PHI Corpus is part of the research infrastructure Health Bank - The Swedish Health Records Research Bank[7]. Stockholm EPR PHI Corpus consists of 200,000 tokens with nine annotated PHI classes. See Table 1 for the classes and their distribution.

The annotation of Stockholm EPR PHI Corpus is described in more detail in (Velupillai et al., 2009). The data was refined in the first de-identification experiment described in (Dalianis and Velupillai, 2010) and has since been used in several studies. Figure 1 shows an example of an pseudonymised annotated record from the data set, followed by an English translation of the same record.

When formatting the data for fine-tuning, tagged entities consisting of multiple words were split into separate tokens and tagged according to the BIOES-standard. This means marking whether a positive token is in the beginning ('B'), ending ('E') or inside ('I') a named entity, or if the token itself makes up a named entity ('S') (Reimers and Gurevych, 2017). Negative tokens, thus tokens which are not part of a named entity, were marked 'O'.

3.2 Methods

This section describes the methods used in this study. First, Section 3.2.1 gives more details on the two pre-trained BERT models used. Then, Section 3.2.2 describes how the models were fine-tuned.

PHI Class	Instances
First Name	923
Last Name	931
Phone Number	137
Age	55
Full Date	457
Date Part	709
Health Care Unit	1,414
Location	95
Organisation	43
Total	4,764

Table 1: The class distribution of Stockholm EPR PHI Corpus.

3.2.1 BERT models

The BERT-models used in this study are the Swedish KB-BERT and the multilingual M-BERT. Both models implement the BERT-Base architecture consisting of twelve layers with a hidden size of 768 and $11 \cdot 10^7$ parameters.

KB-BERT was released by the National Library of Sweden in 2019 (Malmsten et al., 2020). It is pre-trained on approximately 20 GB of digitized Swedish texts written between the years 1940 and 2019. The resources include news articles, legal text, social media posts and Swedish Wikipedia articles. This results in a vocabulary size of around 50,000 tokens. The model is cased, meaning that there are separate entries for tokens beginning with an upper case letter and tokens beginning with a lower case letter.

Devlin et al. (2019) released a multilingual BERT model alongside the original English BERT model. The multilingual model used in this study, M-BERT, is the cased version of this model. It has been pre-trained on 104 languages, including Swedish. For each language, the training data consisted of Wikipedia articles written in that language. To balance the data, high-resource languages were under-sampled while low-resource languages were over-sampled using exponentially smoothed weighting of the data. M-BERT has a vocabulary size of around 120,000 tokens.

3.2.2 Fine-tuning

The pre-trained BERT models provide a general representation, or encoding, of input data. To use the models for prediction or inference they need to be fine-tuned for a specific down-stream task. This involves adding an additional output layer and fit-

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

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

Planeringsansvarig: SSK Tjänstgörande
Patientansvarig läkare: <First_Name>Mohamed</First_Name>
<Last_Name>Åström</Last_Name>
Kontaktorsak: Ramlat i hemmet <Full_Date>10/5-2006</Full_Date> och krampat
<Date_Part>12/5</Date_Part>.
Hade inte ätit eller druckit på 4 dygn.
Hälsohistoria: vårderf. Se läkare anteckningar.
Närstående: Dotter <First_Name>Jessica</First_Name><Last_Name>Fredriksson</Last_Name> tel: <Phone_Number>0715-463920</Phone_Number>,
tel hem <Phone_Number>92 35 45</Phone_Number> <Last_Name>Fredriksson</Last_Name>
tel. <Phone_Number>0392-857461</Phone_Number>
Social bakgrund: Bor på gruppboende, <Health_Care_Unit>Lärkan</Health_Care_Unit>
på <Location>Ladugårdsgärdet</Location>.

Planning manager: Nurse on duty
Attending physician: <First_Name>Mohamed</First_Name>
<Last_Name>Åström</Last_Name>
Reason of contact: Fallen at home <Full_Date>10/5-2006</Full_Date> and felt cramps
<Date_Part>12/5</Date_Part>.
Had not eaten or drunk in 4 days.
Health background: See physician's notes.
Family: Daughter <First_Name>Jessica</First_Name><Last_Name>Fredriksson</Last_Name> ph: <Phone_Number>0715-463920</Phone_Number>,
ph. home <Phone_Number>92 35 45</Phone_Number>
<Last_Name>Fredriksson</Last_Name>
ph. <Phone_Number>0392-857461</Phone_Number>
Social background: Lives at nursing home,
<Health_Care_Unit>Lärkan</Health_Care_Unit> at <Location>Ladugårdsgärdet</Location>.

Figure 1: Example of a pseudonymised electronic patient record in Swedish from Stockholm EPR PHI Corpus and its translation to English.

ting the model with task-specific data. In this case, the down-stream task is NERC and the data used for fine-tuning is that described in Section 3.1. The pre-trained models were loaded and fine-tuned using the *HuggingFace's Transformers library* (Wolf et al., 2020). Both models were loaded with the library's *BertForTokenClassification structure* which providers a linear output layer on top of the hidden-states output.

A challenge with fine-tuning BERT is hyper-parameter optimization. The model is sensitive to several parameters such as number of epochs, batch size and learning rate. Devlin et al. (2019) found that for large data sets the hyper-parameters do not have great impact on performance. On smaller data sets, the authors recommend performing some hyper-parameter optimization for the task at hand. Due to the size of the models, the time it takes to fine-tune them presents a limit on how much resources can be delegated to hyper-parameter optimization. In this study, the optimization is limited to a simple parameter search with starting point at the values recommended by Devlin et al. (2019).

3.3 Application of methods: Experiments

This section presents the different experiments performed to generate the results presented in this paper. First, 20% of the original data, selected at random, was held out for testing. Out of the remaining data, 20% was reserved for development. The purpose of the development set was to evaluate different hyper-parameter settings. The remaining data, which we call the training set, was used for fine-tuning.

In addition to the original training set, Stockholm EPR PHI Corpus, we created a version of the training set, Stockholm EPR PHI Pseudo Corpus, where the PHIs have been replaced by surrogates. We call this the pseudonymised training set[8], or *pseudo* for short.

The surrogate generation is lexical, based on the collection of Swedish named entity lists used in (Dalianis, 2019). In this study, however, the variation of surrogate names is much larger, containing 123,000 female first names, 121,000 male first names and 35,000 last names, rather than only the 100 most common first- and last names used in (Dalianis, 2019).

After fine-tuning on the pseudonymised training set, the models were evaluated on the original test set. The motivation behind these tests is that models trained on pseudonymised data are safer to release for further development by other researchers, without risking that the PHI is revealed. Therefore, it is of interest to see how well such models perform on authentic, not de-identified, patient records.

For both KB-BERT and M-BERT, a search over hyper-parameters was performed. The batch size was set to 16 and the learning rate to $5 \cdot 10^{-5}$. When it comes to the number of epochs, the results differed slightly between the models. Figures 2 and 3 show the precision and recall for different number of epochs over the training set when fine-tuning KB-BERT and M-BERT, respectively. When choosing the number of epochs, most attention was paid to recall as that is of highest priority in a de-identification system. For all models except one, recall either decreased or did not improve significantly after three epochs. Thus, the models were fine-tuned for three epochs. The exception was M-BERT fitted with the pseudonymised data which was fine-tuned for four epochs. The precision was also monitored and it was observed, as the figures show, that precision continued to increase longer than recall. Since recall was prioritized and resources were limited, no experiments were made with training the models further.

After the models were fine-tuned, they were evaluated on the original test set, namely the held out data set, 20% of Stockholm EPR PHI Corpus. We call this set *test set A*. In order to test how well the models perform on a broader range of EPR data, they were also evaluated on other medical specialities of Swedish EPR Corpora from Health Bank. For the purpose of this report, this second test set is called *test set B*. For the most part, *test set B* is annotated according to the same standard as Stockholm EPR PHI Corpus but is lacking the *Organisation* class which is thus excluded from the evaluation on this test set. Further, *test set B* contains ages and dates but their annotation differs from those in Stockholm EPR PHI Corpus. In order to minimize the error caused by different annotation standards, the classes *Age*, *Date Part* and *Full Date* are also excluded from evaluation.

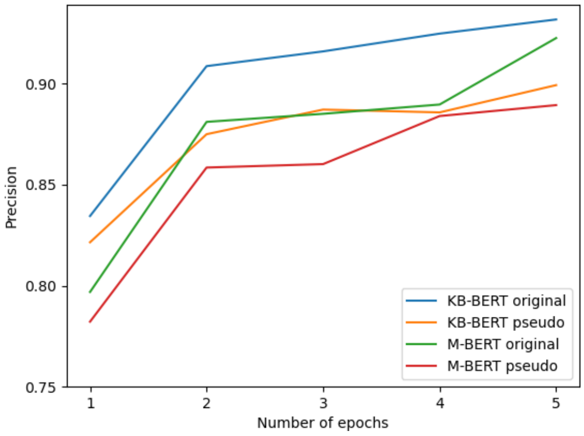

Figure 2: Precision on the development set after different number of epochs for all four models.

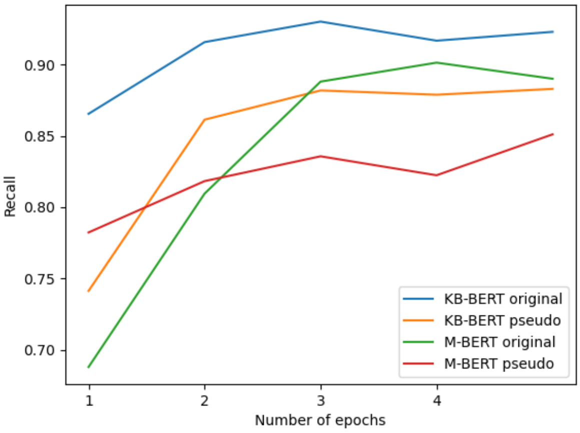

Figure 3: Recall on the development set after different number of epochs for all four models.

4 Results

The results presented in this section were achieved with the best hyper-parameter values found, see Section 3.3. Note that the hyper-parameter optimization was not exhaustive and this may have significant effects on the results.

Table 2 shows the precision (P), recall (R) and F_1-score for the two models fine-tuned with the

[8] Generally, most research on clinical text is carried out on pseudonymised data while most studies on Health Bank data have used real data.

original training set, as well as those fine-tuned on the pseudonymised training set, when evaluated on *test set A*. Table 3 shows the corresponding scores for *test set B*.

Model	**Data**	**P**	**R**	**F$_1$**
KB	Original	0.9226	0.9220	0.9223
	Pseudo	0.8827	0.8822	0.8824
M	Original	0.9051	0.8899	0.8974
	Pseudo	0.8602	0.8357	0.8478

Table 2: Precision, recall and F$_1$-score of KB-BERT and M-BERT fine-tuned with the original training set and the pseudonymised training set respectively, and evaluated on *test set A*.

Model	**Data**	**P**	**R**	**F$_1$**
KB	Original	0.6923	0.7272	0.7093
	Pseudo	0.6427	0.7439	0.6896
M	Original	0.6494	0.6847	0.6666
	Pseudo	0.6398	0.6963	0.6669

Table 3: Precision, recall and F$_1$-score of KB-BERT and M-BERT fine-tuned with the original training set and the pseudonymised training set respectively, and evaluated on *test set B*.

Table 4 shows the recall per class for the models fine-tuned with the original training set and evaluated on *test set A*. Table 5 shows the corresponding results for *test set B*. In the same manner, Tables 6 and 7 shows the recall per class for all the models fine-tuned with the pseudonymised training set and evaluated on *test set A* and *test set B*, respectively. Note that the averages in all tables are weighted based on the number of instances from each class present in the test set at hand. The number of instances per class are given by the figures within the parentheses in the tables' first column.

5 Discussion and Conclusions

The results show that the fine-tuned KB-BERT achieves recall on the same level as that reported in (Grancharova et al., 2020) on the same data set, see Table 2. In this study, however, the relatively high recall does not come at the price of low precision. The precision achieved using KB-BERT is on par with the highest recorded precision on Stockholm EPR PHI Corpus which was documented in (Berg and Dalianis, 2020). There, again, recall was below 0.9. Thus, the BERT-model seems to offer a good balance between precision and recall. From a pure de-identification perspective, high precision is not a priority. However, for the de-identified data to be of use to physicians and researchers, precision remains important. In this sense, the results presented in this paper can be considered an overall improvement of NERC on this data.

Class (instances)	**KB-BERT**	**M-BERT**
First Name (195)	0.9385	0.9077
Last Name (213)	0.9531	0.9296
Phone Number (21)	0.9048	0.8571
Age (9)	1.0000	0.7778
Full Date (83)	0.9518	0.9518
Date Part (131)	0.9847	0.9824
Health Care Unit (293)	0.8737	0.8396
Location (19)	0.7895	0.4221
Organisation (10)	0.5000	0.5000
Weighted average	0.9220	0.8899

Table 4: Recall per class of the models fine-tuned with the original training set and evaluated on *test set A*.

Class (instances)	**KB-BERT**	**M-BERT**
First Name(208)	0.7212	0.7596
Last Name (282)	0.7270	0.6915
Phone Number (22)	0.8636	0.7727
Health Care Unit (208)	0.7163	0.6394
Location (57)	0.7368	0.5088
Weighted average	0.7272	0.6847

Table 5: Recall per class of the models fine-tuned with the original training set and evaluated on *test set B*.

Class (instances)	**KB-BERT**	**M-BERT**
First Name (195)	0.9128	0.8564
Last Name (213)	0.9202	0.8638
Phone Number (21)	0.8095	0.9048
Age (9)	1.0000	0.8889
Full Date (83)	0.9398	0.8554
Date Part (131)	0.9695	0.9847
Health Care Unit (293)	0.8029	0.7577
Location (19)	0.6842	0.4737
Organisation (10)	0.6000	0.5000
Weighted average	0.8822	0.8357

Table 6: Recall per class of the models fine-tuned with the pseudonymised version of the training set and evaluated on *test set A*.

Class (instances)	KB-BERT	M-BERT
First Name(208)	0.8077	0.7548
Last Name (282)	0.7447	0.7092
Phone Number (22)	0.7273	0.7273
Health Care Unit (208)	0.6490	0.6731
Location (57)	0.7368	0.4912
Weighted average	0.7349	0.6963

Table 7: Recall per class of the models fine-tuned with the pseudonymised version of the training set and evaluated on *test set B*.

Regarding the comparison between KB-BERT and M-BERT, the first achieves higher precision and recall on both test sets, see Tables 2 and 3. The difference is more prevalent in some PHI classes than in others. For instance, the recall on *Location* drops significantly when using M-BERT compared to using KB-BERT. This suggests that pre-training specialized toward one language is more beneficial than broader pre-training. This is only a speculation since there are other differences between the two models that could affect performance on the task at hand, such as the nature and amount of Swedish texts used in pre-training.

It is also worth mentioning that the difference in recall between the two models is small, averaging at approximately 0.5 percentage points when fine-tuning on the original data and 1 percentage point when fine-tuning on the pseudonymised data. Since only a limited amount of time was spent on optimisation, it is possible that M-BERT could achieve results similar to KB-BERT if fine-tuned with better settings or more data.

Tables 2 and 3 also show that the models fine-tuned on the original records perform better than those fine-tuned on the pseudonymised records. This is not surprising, as the surrogates have limited range compared to the authentic named entities. Tables 4 - 7 show that, for instance, the recall on *Age* is more negatively affected by fine-tuning on pseudonymised records than the recall on *First name* and *Last name*. An explanation could be that the formats in which surrogate ages are given do not cover all formats present in the authentic records, resulting in greater discrepancies between the training set and the test set when fine-tuning with the pseudonymised records. The formats of names, on the other hand, are less varied in this domain.

Although the models fine-tuned with pseudonymised data perform worse overall, the differences between them and the same models fine-tuned with the original data are not huge. In some cases, such as *Phone number* in M-BERT, the pseudonymised model actually performs better, see Tables 4 and 6. It is clear that the BERT-models are less sensitive to the discrepancies between the original and pseudonymised data than the CRF and LSTM models used on this data set previously, see Section 2 Related research and (Berg et al., 2019). This suggests that this method should be explored further for the purpose of being able to share models trained on electronic patient records while reducing the risks of breaching the privacy of patients or other individuals mentioned in the text.

A comparison between Table 2 and Table 3 demonstrates that there is a loss in recall and an even greater loss in precision when applying the models to data in a slightly broader domain. Differences in the annotation of the two test sets make a direct comparison difficult, but it is clear that the models have learned enough to generalize relatively well to a broader range of electronic patient records. Future work includes creating more annotated data for evaluation as well as training on a broader range of records in order to improve generalization.

In summary, this paper presents an improvement on previous results on the Stockholm EPR PHI Corpus in the sense that the same high recall is achieved without sacrificing precision. It is also demonstrated that performance is somewhat negatively affected by fine-tuning on pseudonymised electronic patient records but the models still achieve relatively high recall. Due to the benefit of being able to share non-sensitive models in compliance with preserving the privacy of patients, this approach should be studied and developed further. The results also show that KB-BERT outperforms M-BERT overall but both models perform relatively well. We can not make any concrete conclusions on the limitations of the models due to the limited resources delegated to optimisation and the limited data used for fine-tuning. Future work includes optimising the models further and fine-tuning on a larger data set.

On a final note, even with a de-identification system with high recall, the de-identified data could be re-identified using external sources. Therefore, the de-identified data must be be handled with care. To improve the privacy where there could be some false negatives, thus missed PHI, one could remove the tags of the true positive so the false negatives are not distinguishable, performing what is known as HIPS (Hide In Plain Sight) (Carrell et al., 2013).

Acknowledgments

We are grateful to the DataLEASH project for funding this research work. Great thanks also to John Valik Karlsson, M.D., for the assistance with the translation of the patient record to English.

References

Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, WA Redmond, and Matthew BA McDermott. 2019. Publicly Available Clinical BERT Embeddings. *NAACL HLT 2019*, page 72.

Hanna Berg, Taridzo Chomutare, and Hercules Dalianis. 2019. 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)*, pages 118–125.

Hanna Berg and Hercules Dalianis. 2020. A Semi-supervised Approach for De-identification of Swedish Clinical Text. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille*, pages 4444–4450.

David Carrell, Bradley Malin, John Aberdeen, Samuel Bayer, Cheryl Clark, Ben Wellner, and Lynette Hirschman. 2013. 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. 2018. *Clinical text mining: Secondary use of electronic patient records*. Springer.

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.

Hercules Dalianis and Sumithra Velupillai. 2010. De-identifying Swedish Clinical Text - Refinement of a Gold Standard and Experiments with Conditional Random Fields. *Journal of Biomedical Semantics*, 1:6.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. http://arxiv.org/abs/1810.04805 BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *arXiv*, `https://arxiv.org/abs/1810.04805`.

Mila Grancharova, Hanna Berg, and Hercules Dalianis. 2020. Improving Named Entity Recognition and Classification in Class Imbalanced Swedish Electronic Patient Records through Resampling. In *Proceedings of Eighth Swedish Language Technology Conference (SLTC) 2020, Göteborg, Sweden.*

Lukas Lange, Heike Adel, and Jannik Strötgen. 2019. NLNDE: The Neither-Language-Nor-Domain-Experts' Way of Spanish Medical Document De-Identification. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019))*, pages 671–678.

Patrick Lewis, Myle Ott, Jingfei Du, and Veselin Stoyanov. 2020. Pretrained Language Models for Biomedical and Clinical Tasks: Understanding and Extending the State-of-the-Art. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 146–157.

Ngoc C. Lê, Ngoc-Ye Nguyen, Anh-Duong Trinh, and Hue Vu. 2020. On the Vietnamese Name Entity Recognition: A Deep Learning Method Approach. In *IEEE Access*.

Martin Malmsten, Love Börjeson, and Chris Haffenden. 2020. http://arxiv.org/abs/2007.01658 Playing with Words at the National Library of Sweden – Making a Swedish BERT. In *arXiv*, `https://arxiv.org/abs/2007.01658`.

Jihang Mao and Wanli Liu. 2019. Hadoken: a BERT-CRF Model for Medical Document Anonymization. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019))*, pages 720–726.

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.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. http://arxiv.org/abs/1906.05474 Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets. In *arXiv*, `https://arxiv.org/abs/1906.05474`.

Nils Reimers and Iryna Gurevych. 2017. Optimal Hyperparameters for Deep LSTM-Networks for Sequence Labeling Tasks. *arXiv preprint arXiv:1707.06799*.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Sumithra Velupillai, Hercules Dalianis, Martin Hassel, and Gunnar H Nilsson. 2009. Developing a standard for de-identifying electronic patient records written in Swedish: precision, recall and F-measure in a manual and computerized annotation trial. *International Journal of Medical Informatics*, 78(12):e19–e26.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. http://arxiv.org/abs/2007.01658 HuggingFace's Transformers: State-of-the-art Natural Language Processing. In *arXiv,* *`https://arxiv.org/abs/1910.03771`*.