Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 410–416
Marseille, 20-25 June 2022

Evaluating Pretraining Strategies for Clinical BERT Models

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

Abstract

Research suggests that using generic language models in specialized domains may be sub-optimal due to significant domain differences. As a result, various strategies for developing domain-specific language models have been proposed, including techniques for adapting an existing generic language model to the target domain, e.g. through various forms of vocabulary modifications and continued domain-adaptive pretraining with in-domain data. Here, an empirical investigation is carried out in which various strategies for adapting a generic language model to the clinical domain are compared to pretraining a pure clinical language model. Three clinical language models for Swedish, pretrained for up to ten epochs, are fine-tuned and evaluated on several downstream tasks in the clinical domain. A comparison of the language models' downstream performance over the training epochs is conducted. The results show that the domain-specific language models outperform a general-domain language model, although there is little difference in performance between the various clinical language models. However, compared to pretraining a pure clinical language model with only in-domain data, leveraging and adapting an existing general-domain language model requires fewer epochs of pretraining with in-domain data.

Keywords: language models, domain-adaptive pretraining, Swedish clinical text

1. Introduction

Language models, pretrained in a self-supervised fashion on large unlabeled corpora, and subsequently fine-tuned on downstream tasks using labeled datasets, have led to performance gains across many NLP tasks. More recently Transformer models often outperform recurrent neural networks and LSTMs and as a result, they have become the main focus of recent NLP research. The paradigm of pretraining and fine-tuning language models comes with the advantage of being able to make effective use of large, unlabeled corpora, and subsequently specialize the model to perform a specific task in a process that is relatively resource efficient in terms of the amount of labeled data that is required.

Language models are often pretrained using corpora in the general domain, e.g. Wikipedia. However, the use of generic language models in specialized domains may be sub-optimal due to significant domain differences (Lewis et al., 2020; Gururangan et al., 2020). As a result, there have been many efforts to develop domain-specific language models, e.g. SciBERT (Beltagy et al., 2019) and BioBERT (Lee et al., 2020). There are different approaches to developing domain-specific language models, including pretraining a language model with in-domain data from scratch and continuing to pretrain an existing, generic language model with in-domain data in a process known as domain-adaptive pretraining. However, it is not clear which approach is more effective for creating clinical language models.

This study aims to evaluate and compare various strategies for pretraining clinical BERT models for Swedish. One option is to leverage an existing generic language model and adapt it to the clinical domain through domain-adaptive pretraining, i.e. continued pretraining with in-domain data. Here, we evaluate two such strategies, based on using a general vocabulary (Lamproudis et al., 2021) vs. a clinical vocabulary (Lamproudis et al., 2022). An alternative is to develop a new, purely clinical language model that is pretrained using only in-domain data. We evaluate and compare these three pretraining strategies for creating clinical BERT models and also include a baseline in the form of a generic language model for Swedish, namely KB-BERT (Malmsten et al., 2020). The clinical BERT models are pretrained for up to ten epochs; at the end of each epoch, they are fine-tuned and evaluated on six downstream tasks in the clinical domain. In summary, the main contributions of this study are as follows:

- Three strategies for pretraining clinical BERT models are evaluated on six downstream clinical NLP tasks, including classification and NER. It is shown that all three pretraining strategies result in clinical language models that clearly outperform a generic language model and do so already after a single epoch of pretraining with in-domain data.
- While the best results are obtained when adapting an existing generic language model to the clinical domain, in particular when using a clinical vocabulary, the differences are small when comparing the best models from the pretraining session. However, adapting a generic language model to the clinical domain significantly requires fewer epochs of pretraining compared to pretraining a entirely new model from scratch.
- The two previously evaluated domain-adaptive pretraining strategies are here compared to pretraining a clinical language model from scratch, i.e. using only in-domain data. Also, compared

to the previous evaluations, in which the models were pretrained for only one epoch, the clinical language models are here pretrained for up to ten epochs and evaluated on several more downstream tasks, including a newly introduced task and associated dataset for detecting adverse drug events.

2. Related Research

From the decoder of the transformer, Generative Pre-Training (GPT) (Radford et al., 2018), and later from the encoder of the transformer (Devlin et al., 2019), BERT was introduced. Since then, transformer-based language models have been widespread and the focal point of recent research in language modeling and representation learning. Later research, along with modified architectures of the transformer, yielded improved techniques in the development of these language models with respect to pretraining, RoBERTA being a prominent example of research aiming to optimize training of language models (Liu et al., 2019).

In an effort to create better language models for specific domains, BioBERT was one of the first attempts in trying to adapt BERT to the biomedical domain (Lee et al., 2020). BioBERT was initialized with BERT's original parameters along with the same vocabulary and was pretrained using biomedical text. Similarly, BioMegatron (Shin et al., 2020) was a larger model, trained with more data, and used a domain-specific vocabulary to achieve improved performance in the biomedical domain. Building on previous work, by inheriting the model parameters from BioBERT and pretraining using clinical text, Clinical BERT (Alsentzer et al., 2019) achieved improved results with tasks in the clinical text domain.

In a more extensive study (Gururangan et al., 2020), the benefits of domain-adaptive pretraining were investigated and a number of proposed modifications were evaluated and compared. More specifically, pretraining on unlabeled domain-specific data was compared to pretraining on unlabeled task-specific data. Inspired by this work, Clinical KB-BERT was developed for Swedish and achieved substantial improvements on several downstream tasks in the clinical domain (Lamproudis et al., 2021). Clinical KB-BERT inherited model parameters and the vocabulary from KB-BERT (Malmsten et al., 2020), after which a session of domain-adaptive pretraining was carried out.

Further efforts in adapting existing language models to a particular domain have focused on the vocabulary of the language models. With exBERT, domain-specific terms were included in the model's vocabulary along with extensions in each self-attention layer (Tai et al., 2020). This resulted in a model with slightly more parameters than the original BERT, which was then further pretrained with in-domain data and yielded improved performance on the domain-specific downstream tasks. Similarly, the complete replacement of the model's vocabulary with a domain-specific vocabulary has been explored, yielding promising results (Koto et al., 2021). In this approach, the model is initialized using parameters for whole words and subtokens and then further pretrained with in-domain data. In a similar vain, a new version of Clinical KB-BERT was developed for Swedish, using a clinical vocabulary, inheriting parameters, followed by a session of domain-adaptive pretraining. This model yielded further improvements over the first version of Clinical KB-BERT – which used a general-domain vocabulary – on two downstream tasks (Lamproudis et al., 2022).

A different approach is *not* to leverage an existing, general-domain language model; instead, pretraining is carried out from scratch with in-domain data. In one study (Gu et al., 2021), this approach was shown to outperform domain-adaptive pretraining.

3. Methods and Data

In this study, three clinical language models for Swedish based on the *BERT-base* architecture are developed using different pretraining strategies and compared: (i) domain-adaptive pretraining of a generic Swedish language model with a general-domain vocabulary, (ii) domain-adaptive pretraining of a generic Swedish language model with a clinical vocabulary, and (iii) a pure clinical language model pretrained from scratch. The clinical language models are fine-tuned on six downstream NLP tasks in the clinical domain, covering a variety of important named entity recognition (NER) and classification tasks. The clinical language models are pretrained for up to ten epochs; checkpoints at the end of each epoch are fine-tuned and evaluated in order to produce learning curves. The clinical language models are also compared to a general-domain Swedish language model, namely *KB-BERT* (Malmsten et al., 2020), which is pretrained on government documents, Swedish Wikipedia and newspapers.

3.1. Pretraining Strategies

In this section, we describe the various strategies for creating the clinical language models, which are all based on BERT (Devlin et al., 2019). The models are also pretrained using the same hyperparameters as BERT with some notable exceptions, namely that (i) they are only trained with maximum length sequences (512) instead of 128 length sequences, and (ii) to achieve the maximum sequence length, shorter sequences are concatenated by inserting [SEP] tokens to denote the end and beginning of two original sequences, i.e. text from two distinct clinical notes. See Table 1 for details regarding which hyperparameters were used during pretraining.

All pretraining sessions use masked language modeling (MLM) as the training task, where a percentage of the words in a sequence – usually 15% – is masked and the model is required to predict the masked words. The maximum training session is set to last for 10 epochs, roughly corresponding to the training length in (De-

hyperparameters	values
learning rate	10^{-4}
batch size	256
Adam optimizer	**yes**
β_1	0.9
β_2	0.999
L2 weight decay	0.01
warm up steps	10 000
linear learning rate decay	**yes**
dropout probability	10%
update steps	$\approx$ 400 000
training sequence length	512
MLM probability	15%

Table 1: Pretraining hyperparameters

vlin et al., 2019), saving checkpoints at intervals corresponding to each epoch. Rather than evaluating the pretraining task itself, each of the saved checkpoints is evaluated in terms of the performance on downstream clinical NLP tasks according to common practice when evaluating the pretraining of language models. Each model is trained on one NVIDIA RTX A5000 with 24 GB of memory for approximately 20 days on the clinical data described in Section 3.3.

Clinical KB-BERT v1 uses a general-domain language model for Swedish, *KB-BERT*, which in turn is based on the *BERT-base* architecture, for model initialization and then carries out domain-adaptive pretraining with in-domain data. The vocabulary is inherited from *KB-BERT*. See (Lamproudis et al., 2021) for more details.

Clinical KB-BERT v2 is also initialized using *KB-BERT*. In contrast to v1, this model however constructs and uses a clinical vocabulary. Existing representations for whole words and subtokens are inherited from *KB-BERT* and then updated during the domain-adaptive pretraining session. See (Lamproudis et al., 2022) for more details.

Pure Clinical BERT is also based on the *BERT-base* architecture and is pretrained from scratch using only in-domain data, i.e. it does not rely on an existing general-domain language model. This model has not been evaluated previously.

Compared to previous evaluations of *Clinical KB-BERT v1* and *Clinical KB-BERT v2*, in which the models were pretrained for only one epoch, in this study all three clinical language models are pretrained for up to ten epochs. This allows us to evaluate and compare how the language models benefit from further pretraining. The evaluation of the clinical language models is also more extensive with six downstream tasks compared to only two and three, respectively, in the previous studies. All language models included in this study are summarized in Table 2.

Model	Vocabulary	In-domain pretraining	Domain adaptation: general $\rightarrow$ clinical
KB-BERT	General	No	No
Clinical KB-BERT v1	General	Yes	Yes
Clinical KB-BERT v2	Clinical	Yes	Yes
Pure Clinical BERT	Clinical	Yes	No

Table 2: Characteristics of the evaluated models

3.2. Fine-Tuning & Evaluation

All language models are fine-tuned on six important clinical NLP tasks. These tasks fall into one of two categories: classification (both binary and multi-label classification) and NER. The datasets and downstream tasks are presented in Section 3.3. The models, except for *KB-BERT*, are evaluated at the end of each epoch, allowing us to track the progress – in terms of performance on downstream tasks – throughout the pretraining session and produce learning curves.

During fine-tuning, no extensive hyperparameter search is performed since the aim is not to outperform the state of the art on the downstream tasks, but rather to compare and evaluate pretraining strategies for creating clinical language models. To that end, the best parameters of a narrow hyperparameter search are used, with the hypothesis being that the evaluation and comparison of the models will be fair regardless of the possible existence of more optimal hyperparameters. The hyperparameters for the various tasks are presented in Table 3. In these fine-tuning sessions, the models are trained until convergence in terms of the validation set loss. Ten experiments for each model are conducted with different, non-overlapping test sets, after which the results are averaged in order to produce a more robust estimation of model performance.

3.3. Data & Downstream Tasks

For pretraining with in-domain data, clinical text from the research infrastructure Health Bank[1] – Swedish Health Record Research Bank at DSV/Stockholm University (Dalianis et al., 2015) is used. The clinical text originates from Karolinska University Hospital and encompasses electronic health records for over 2 million patients from 500 clinical units during 2007-2014. The data has a size of 17.9 GB[2], which is comparable to the size of the training data that was used for developing *KB-BERT*.

For fine-tuning and evaluating the models on downstream tasks, the following five manually annotated clinical datasets, also from Health Bank, are used:

Stockholm EPR Gastro ICD-10 Corpus This corpus consists of 6,062 gastro-related discharge summaries and their assigned ICD-10 diagnosis codes, encompassing 4,985 unique patients and 795,839 tokens. The data is divided into 10

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

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

Model	ICD-10	PHI	Clinical Entity	ADE	Factuality	Factuality
	Classification	NER	NER	Classification	Classification	NER
learning rate	$2 \cdot 10^{-5}$	$3 \cdot 10^{-5}$	$3 \cdot 10^{-5}$	$3 \cdot 10^{-5}$	$2 \cdot 10^{-5}$	$3 \cdot 10^{-5}$
batch size	32	64	64	64	32	64

Table 3: Hyperparameters for the downstream tasks

groups that correspond to different body parts; the ICD-10 codes range from K00 to K99. Each group contains several codes. See (Remmer et al., 2021) for more details.

Stockholm EPR PHI Corpus This corpus consists 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*. See (Dalianis and Velupillai, 2010) for the creation of the gold standard.

Stockholm EPR Clinical Entity Corpus This corpus consists of 70,852 tokens and 7,946 annotated entities corresponding to four clinical entity classes: *Diagnosis*, *Findings*, *Body parts* and *Drugs*. See (Skeppstedt et al., 2014) for more details.

Stockholm EPR ADE ICD-10 Corpus This corpus is new and introduced here for the first time. It contains 21,642 samples and 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.

Stockholm EPR Diagnosis Factuality Corpus This corpus encompasses six levels of annotations regarding the factuality of a diagnosis. It consists of 3,710 samples with 7,066 annotated entities *Certainly Positive*, *Probably Positive*, *Possibly Positive*, *Possibly Negative*, *Probably Negative*, and *Certainly Negative*, in total encompassing 240,000 tokens. See (Velupillai et al., 2011; Velupillai, 2011) for more details. This corpus is used for creating two downstream tasks: one as a multi-label document classification task (**Factuality Classification**) and one as a NER task (**Factuality NER**).

Note that several of the corpora – *Stockholm EPR Gastro ICD-10*, *Stockholm EPR PHI*, *Stockholm EPR Clinical Entity*, and *Stockholm EPR Diagnosis Factuality* – are also included in the pretraining data. Although this may raise concerns regarding possible performance improvements compared to the *KB-BERT* baseline, the impact is likely to be insignificant for two reasons: (i) the size of the corpora used in the downstream tasks is very small ($\sim$ 2 MB) compared to the size of the pretraining data ($\sim$ 18 GB), and (ii) the nature of the two tasks – self-supervised pretraining vs. supervised classification or NER – is different and helps to avoid overfitting. However, as a control task, we have created and here introduce the *Stockholm EPR ADE ICD-10 corpus*, which is completely excluded from the pretraining data and will also act as a validation of the results for the rest of the downstream tasks. Again, it is worth noting that the main purpose of the experiments in this study is not to produce state-of-the-art predictive performance on the downstream tasks, but to compare various pretraining strategies for creating clinical BERT models.

4. Results

In Table 4, we present the best results of the evaluated models in each of the six downstream tasks. These do not always correspond to the model produced in the last epoch of pretraining but can be the results of earlier checkpoints of each model. They are presented in terms of F_1-score, with the best result on each task highlighted in bold.

The best results are obtained by *Clinical KB-BERT v1* and *Clinical KB-BERT v2*. The difference between these versions is small across all six downstream tasks, even if *Clinical KB-BERT v2* obtains the best results on four out of six downstream tasks. The differences between the two versions of *Clinical KB-BERT* and *Pure Clinical BERT* is also small across tasks and, when considering the best results by each model over ten epochs, there seems to be very little difference between the strategies for creating clinical BERT models. However, all clinical BERT models clearly outperform the generic language model, *KB-BERT*, on all downstream tasks.

The best result on the ICD-10 task is obtained by *Clinical KB-BERT v2*, yielding an F_1-score of 0.848. The same model also performs best on the clinical entity task (F_1: 0.862, tied with *Clinical KB-BERT v1*), as well as the two factuality tasks (F_1: 0.734 and 0.696, respectively). On the PHI task, *Clinical KB-BERT v1* outperforms the other models, obtaining an F_1-score of 0.948. This model also performs best on the newly introduced ADE classification task, obtaining an F_1-score of 0.199.

In Figure 1, we present the evolution of the average performance across tasks of each model during the pre-

Model	ICD-10	PHI	Clinical Entity	ADE	Factuality	Factuality
	Classification	NER	NER	Classification	Classification	NER
KB-BERT	0.799	0.920	0.803	0.183	0.635	0.630
Clinical KB-BERT v1	0.841	**0.948**	**0.862**	**0.199**	0.732	0.690
Clinical KB-BERT v2	**0.848**	0.946	**0.862**	0.196	**0.734**	**0.696**
Pure Clinical BERT	0.844	0.939	0.857	0.193	0.726	0.694

Table 4: The predictive performance, in terms of F_1-score, of all models on the six downstream tasks.

training session. As can be seen, for all models, there is a general improvement in performance with more pretraining epochs, although the improvement is not monotonic. The two versions of *Clinical KB-BERT* – which both inherit model parameters from an existing pretrained language model – benefit less from further pretraining epochs compared to *Pure Clinical BERT*. The difference between the two versions of *Clinical KB-BERT*, on the one hand, and *Pure Clinical BERT*, on the other hand, is clear up until around seven epochs of pretraining.

5. Discussion

The results demonstrate that the domain-specific clinical language models clearly outperform the general-domain language model. These results corroborate the findings of previously published research, see e.g. (Lee et al., 2020; Gururangan et al., 2020; Alsentzer et al., 2019). It is evident that in-domain pretraining leads to better performance on downstream tasks in the target domain. When selecting the best models over ten pretraining epochs, there was little observed difference between the different pretraining strategies for developing clinical BERT models. However, the best results were consistently obtained by the two versions that leveraged an existing pretrained language model, followed by domain-adaptive pretraining. Between these two versions, the one utilizing a domain-specific vocabulary performed slightly better, which is also in line with previous work (Koto et al., 2021; Lamproudis et al., 2022).

From Figure 1, we see that *Clinical KB-BERT v1* and *Clinical KB-BERT v2* have an improved performance from a very early stage of their domain-adaptive pretraining session. In contrast, *Pure Clinical BERT*, although also outperforming *KB-BERT* very early in its pretraining session, reaches the performance of *Clinical KB-BERT v1* and *Clinical KB-BERT v2* only after around seven pretraining epochs. This is to be expected as *Clinical KB-BERT v1* and its vocabulary-adapted counterpart, *Clinical KB-BERT v2*, essentially inherit the parameters of an already trained language model, in this case *KB-BERT*. This can be seen as a form of warm start to their domain-adaptive pretraining session. In contrast, *Pure Clinical BERT* is initialized with random parameters and therefore needs more training time to reach a comparable level of performance.

This is further illustrated in Figure 2, where the regression lines of the performances of each of the models across all six downstream tasks during the pretraining session are presented. *Clinical KB-BERT v1* and *Clinical KB-BERT v2* reach almost optimal performance from very early on in their domain-adaptive pretraining session. In contrast, *Pure Clinical BERT* generally keeps improving with more pretraining epochs, which is to be expected since the model is newly initialized and pretrained from scratch. In relation to this result, a warm start – to inherit parameters from an already trained model – seems to be advantageous compared to random initialization and pretraining from scratch. In essence, both *Clinical KB-BERT v1* and *Clinical KB-BERT v2* are warm-start versions of *Pure Clinical BERT* as they are initialized with non-random parameters, but rather with the parameters of a generic Swedish language model. Furthermore, *Clinical BERT v2* illustrates how beneficial a joint warm start approach of both the model's parameters and vocabulary can be for the development of a language model as it is overall the best-performing model of the three.

6. Conclusions

The results confirm previous studies that demonstrate the benefits of domain-specific language models, all outperforming a general-domain language model on clinical downstream tasks. Furthermore, the best model is the domain-adapted clinical language model with a clinical vocabulary, *Clinical KB-BERT v2*, further agreeing with the literature (Koto et al., 2021; Lamproudis et al., 2022). However, there is little difference in performance between the various clinical language models, i.e. whether a general-domain language model is adapted to the clinical domain or a new language model is pretrained from scratch with in-domain data. Compared to pretraining a pure clinical language model with only in-domain data, leveraging and adapting an existing general-domain language model, however, requires fewer epochs of pretraining with in-domain data. This can possibly be extended to broader domains, e.g. developing a Norwegian language model with the use of an already existing Swedish language model, which also might prove more beneficial than starting from scratch.

In the near future – once we have obtained the necessary permissions from the Swedish Ethical Review Authority – we plan to distribute a de-identified version

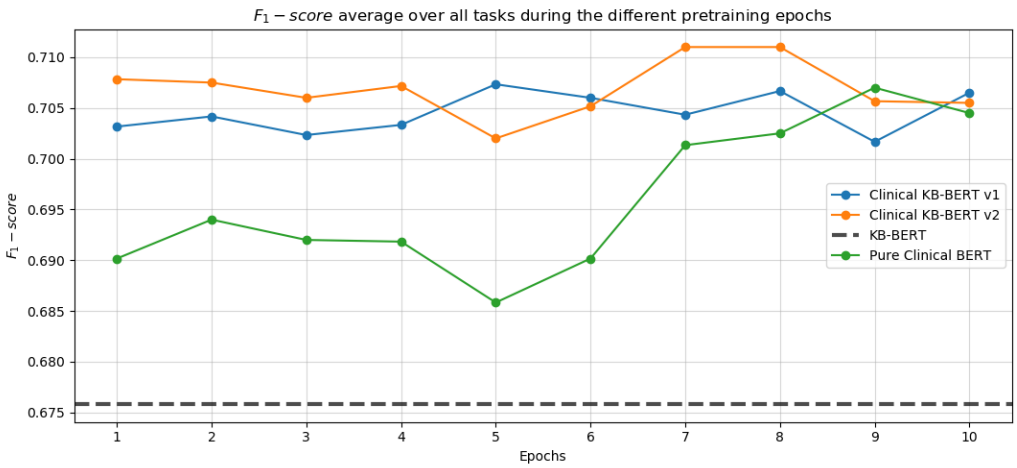

Figure 1: Average performance, in terms of F_1-score, of each model across the six downstream tasks during the pretraining session.

of *Clinical KB-BERT v1* under the name *SweDeClin-BERT*[3]. For details regarding *SweDeClin-BERT*, please see (Vakili et al., 2022).

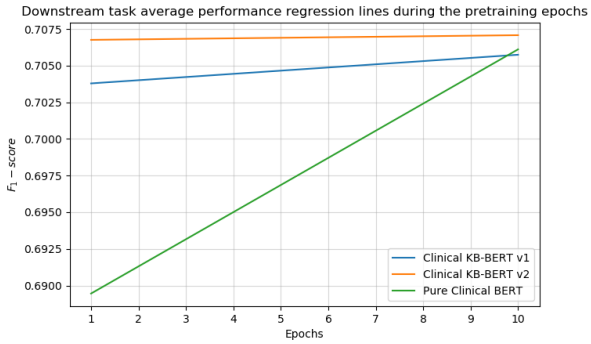

Figure 2: Regression lines for the average performance of each model across all six downstream tasks during the pretraining session. The linear regression lines indicate an approximation of the rate of learning during the pretraining session.

Acknowledgments

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

Alsentzer, E., Murphy, J., Boag, W., Weng, W.-H., Jindi, D., Naumann, T., and McDermott, M. (2019). Publicly Available Clinical BERT Embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78.

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 3606–3611.

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.

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.

Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., and Poon, H. (2021). Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., and Smith, N. A. (2020). Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360.

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

Koto, F., Lau, J. H., and Baldwin, T. (2021). IndoBERTweet: A Pretrained Language Model for Indonesian Twitter with Effective Domain-Specific Vocabulary Initialization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10660–10668.

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 RANLP 2021: Recent Advances in Natural Language Processing*, pages 790–797.

Lamproudis, A., Henriksson, A., and Dalianis, H. (2022). 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. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Lewis, P., Ott, M., Du, J., and Stoyanov, V. (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.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.

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*.

Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. (2018). Improving Language Understanding by Generative Pre-training. OpenAI `https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf/`. [Online; accessed 23 Dec 2021].

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.

Shin, H.-C., Zhang, Y., Bakhturina, E., Puri, R., Patwary, M., Shoeybi, M., and Mani, R. (2020). BioMegatron: Larger Biomedical Domain Language Model. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4700–4706.

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.

Tai, W., Kung, H., Dong, X. L., Comiter, M., and Kuo, C.-F. (2020). exBERT: Extending Pre-trained Models with Domain-specific Vocabulary Under Constrained Training Resources. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 1433–1439.

Vakili, T., Lamproudis, A., Henriksson, A., and Dalianis, H. (2022). Downstream Task Performance of BERT Models Pre-Trained Using Automatically De-Identified Clinical Data. In *Proceedings of the Thirteenth International Conference on Language Resources and Evaluation (LREC 2022)*.

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*.