MeDa-BERT: A medical Danish pretrained transformer model

Abstract

This paper introduces a medical Danish BERT-based language model (MeDa-BERT) and medical Danish word embeddings. The word embeddings and MeDa-BERT were pretrained on a new medical Danish corpus consisting of 133M tokens from medical Danish books and text from the internet. The models showed improved performance over general-domain models on medical Danish classification tasks. The medical word embeddings and MeDa-BERT are publicly available at [1].

1 Introductions

Large language models (LLM) are powerful representation learners and have become the backbone structure of many modern natural language processing (NLP) systems. To learn text representations, LLM are first pretrained on a large-scale text corpus using self-supervised learning, e.g., masked language modelling. After pretraining, LLM are fine-tuned on specific downstream tasks where they have achieved state-of-the-art results on NLP benchmarks such as GLUE (Wang et al., 2018).

However, directly applying these general pretrained models to specialized domains such as the medical have led to unsatisfactory results (Peng et al., 2019). As a solution to this, a second round of in-domain pretraining (domain-adaptive pretraining) has shown to improve the performance of LLMs that were first trained on a general domain corpus (Gururangan et al., 2020). Domain-adaptive pretraining adjusts the weights of the LLM to better capture the terminology, style, and nuances that are relevant to the target domain.

Resource-rich languages such as English have large domain-specific corpuses available that have been used to develop e.g., biomedical (Lee et al., 2020), clinical (Alsentzer et al., 2019), scientific (Beltagy et al., 2019), and financial (Peng et al., 2021) LLMs that perform better than models trained on general corpuses. These models could potentially be used to improve human decision making, save time, and reduce costs, e.g., by extracting information from scientific articles, identifying potential drug interactions, and helping with NLP tasks such as text classification, named entity recognition, and question answering for each of their specialized domains.

For the Danish language, only LLMs trained on a general domain have been published. This paper presents a medical Danish BERT model (MeDa-BERT) — a LLM trained on a new medical Danish text corpus. We also used the medical corpus to train medical word embeddings. To evaluate the medical word embeddings and MeDa-BERT, we used existing medical Danish classification datasets. We found that an LSTM model using the medical word embeddings outperformed a similar model using general-domain word embeddings, and that MeDa-BERT performed slightly better than a general-domain BERT model.

2 Method

This section first describes how the medical corpus was collected and used to pretrain the medical Danish word embeddings and MeDa-BERT. Next, the datasets used to compare model performances and the fine-tuning procedure is described.

2.1 Danish medical corpus

We collected data from the internet and from medical books. The owners of the data resources approved that we used their data in this study. We describe the data collection for each text contributor below. An overview of the text corpuses and their size can be seen in Table 1.

[1] ANONYMIZED

Corpus	Type	Date retrieved	Tokens
Clinical guidelines	Guidelines	October - November 2022	80,567,576
Medicin.dk	Information portal	June 2021	28,878,335
FADL	Books	January 2022	12,531,373
Sundhed.dk	Information portal	May 2022	6,767,409
Netdoktor.dk	Information portal	October 2022	3,227,051
Wikipedia	Encyclopedia	October 2022	1,992,796
Total			**133,964,540**

Table 1: Number of tokens and date retrieved for each data source

2.1.1 Clinical guidelines

We collected text from the document management systems of the five Danish regions. The documents describe guidelines and instructions for diagnostics and treatment of patients and all workflows that support this. The document systems also include non-medical documents from purchasing, logistics, and service departments which were removed. All departments that were excluded and the number of tokens retrieved from each region can be seen in Appendix A.

2.1.2 Medical information portals

We collected text from webpages that provide information to medical doctors and patients. The text was collected from Medicin.dk, Netdoktor.dk, and Sundhed.dk. The resources provide information about diseases, symptoms, and medical treatments. Moreover, the resources contain information specifically for health care professionals, e.g., medication guidelines and information about best practices in the field. Text not related to the medical domain and text written by non-professionals were removed from the corpus. A description of this process can be seen in appendix A.

2.1.3 Books

This part of the corpus consisted of 107 medical books from publisher FADLs Forlag that publishes books for medicine and nursing school.

2.1.4 Wikipedia

We used PetScan[2] to search for medical Wikipedia documents within predefined categories and its subcategories. We used a maximum depth of 5 for searching for subcategories. The following categories were used: anatomi, physiology, diseases, medication, epidemiology, diagnostics, medical procedures, medical specialities, medical physics, and medical equipment. We excluded documents with the categories: persons and companies. This process resulted in 5,391 documents. Next, we manually removed non-medical articles from that list which resulted in 5,266 documents.

[2] https://petscan.wmflabs.org/

2.2 Preprocessing of data

For all text corpusses, we defined a sample as one paragraph, i.e., a continuous stream of text without line breaks. We inserted spaces between alphanumeric and non-alphanumeric characters. Samples were further preprocessed to fit the pretraining procedure for either word embeddings or the transformer model, as detailed below.

2.2.1 Danish medical transformer model

MeDa-BERT was initialized with weights from a pretrained Danish BERT model[3] trained on 10.7 GB Danish text from Common Crawl (9.5 GB), Danish Wikipedia (221 MB), debate forums (168 MB), and Danish OpenSubtitles (881 MB).

For domain-adaptive pretraining, samples from the collected medical corpus were appended a [CLS] and [SEP] token in the start and end of each sample, respectively. Samples were concatenated to fit the maximum sequence length of 512 tokens and document boundaries were indicated by adding an extra [SEP] token in between samples. After this process, we removed duplicates corresponding to 0.2% of the total corpus. The model was trained using Adam (Kingma and Ba, 2015) with a weight decay of 0.01 as described in (Loshchilov and Hutter). Using gradient accumulation, the model was trained with a batch size of 4,032, a learning rate of 1e-4, and a linear learning rate decay warmed up over 1 epoch. The model was trained for a total of 48 epochs and evaluated after 16, 32, and 48 epochs. We used 5% of the samples as a validation set to evaluate the model during training and trained the model on the remaining data using dynamic masked language modeling. The model was optimized using four Tesla v100 GPUs using the Huggingface (Wolf et al., 2020) library. All model parameters and pretraining losses are shown in Appendix B.

2.3 Danish medical word embeddings

We trained 300-dimensional FastText (Bojanowski et al., 2017) word embeddings. The embeddings were trained for 10 epochs using a window size of 5 and 10 negative samples. The

[3] https://github.com/certainlyio/nordic_bert

Dataset	Label	Train	Validation	Test
Bleeding	Positive	10,331	1,300	1,300
	Negative	10,331	1,300	1,300
Bleeding site	Airways	1,000	125	125
	Cerebral	1,000	125	125
	Ear-nose-throat	1,000	125	125
	Eyes	1,000	125	125
	Gastrointestinal	1,000	125	125
	Gynecological	1,000	125	125
	Internal	1,000	125	125
	Skin	1,000	125	125
	Urogenital	1,000	125	125
	Unknown	1,000	125	125
VTE	Positive	9,064	1,100	1,100
	Negative	9,064	1,100	1,100
VTE site	Airways	1,600	200	200
	Lungs	1,600	200	200
	Unknown	1,600	200	200

Table 2: Dataset distributions

hyperparameters were chosen to be able to compare the produced embeddings with the Danish FastText word embeddings from Grave et al. (2018) that were trained on a general domain.

2.4 Datasets

We compared performances between models using four medical datasets: bleeding classification, bleeding site classification, venous thromboembolism (VTE) classification, and VTE site classification. All samples were annotated with a consensus label from three medical doctors. The dataset distributions can be seen in Table 2.

2.4.1 Bleeding classification

The bleeding dataset (Pedersen et al., 2021) is a binary classification problem with 25,862 samples. The dataset was constructed from 900 Danish electronic health records (EHR) from Odense University Hospital. The samples had an average token length of 13.3.

2.4.2 Bleeding site classification

The bleeding site dataset (Pedersen et al., 2022b) is a 10-class classification problem with 11,250 unique bleeding-positive samples annotated for the bleeding site. The bleeding site labels were: airways, cerebral, ear-nose-throat, eyes, gastrointestinal, gynecological, internal, skin, urogenital, and unknown. The dataset was constructed from 149,523 Danish EHR notes from Odense University Hospital. The samples had an average token length of 14.4.

2.4.3 VTE classification

The VTE dataset (Pedersen et al., 2022a) is a binary classification problem with 22,528 samples. The dataset was constructed from 94,520 Danish EHR notes from Odense University hospital. The samples had an average token length of 13.8.

2.4.4 VTE site classification

The VTE site dataset (Pedersen et al., 2022a) is a 3-class classification problem with 6,000 VTE-positive samples annotated for the VTE site. The VTE site labels were: airways, lungs, and unknown. The dataset was constructed from 94,520 Danish EHR notes from Odense University Hospital. The samples had an average token length of 14.5.

2.5 Fine-tuning

2.5.1 MeDa-BERT and BERT

We used the [CLS] token followed by a classification layer to classify samples of the datasets. We searched for the best models five times using Adam with learning rates [5e-5, 3e-5, 1e-5], i.e., we fine-tuned each model 15 times. The models were trained for a maximum of 10 epochs. For MeDa-BERT, we evaluated the model after 16, 32, and 48 epochs.

2.5.2 LSTM

We used the medical word embeddings as input to a bidirectional LSTM layer with a hidden layer size of 512. The last hidden state of the LSTM was followed by a dropout layer with probability 0.2, a dense layer of size 256, a ReLU activation function, a dropout layer of probability 0.2, and a dense classification layer. This model is referred to as LSTM+MeDa-WE.

The performance of the model is compared with another LSTM model (LSTM+General-WE) with the same parameters but using FastText embeddings trained on the general domain as input (Grave et al., 2018). We searched for the best models five times using Adam with learning rates [5e-5, 3e-5, 1e-5], i.e., we fine-tuned each model 15 times.

For all models we report the mean test accuracy and standard deviation for the five best performing models on the validation dataset.

3 Results

Table 3 shows the results of each model on the four classification datasets.

	Bleeding	Bleeding site	VTE	VTE site
LSTM+General-WE	$83.8_{.7}$	$69.3_{.8}$	$88.5_{.2}$	$86.4_{.7}$
LSTM+MeDa-WE	$\mathbf{91.4_{.3}}$	$\mathbf{84.9_{1.1}}$	$\mathbf{94.1_{.3}}$	$\mathbf{93.4_{.4}}$
BERT	$94.3_{.6}$	$86.7_{.8}$	$96.7_{.3}$	$94.7_{.3}$
MeDa-BERT_16	$94.7_{.3}$	$88.4_{.6}$	$97.1_{.4}$	$95.5_{.2}$
MeDa-BERT_32	$95.1_{.5}$	$88.7_{.6}$	$96.9_{.3}$	$95.7_{.3}$
MeDa-BERT_48	$\mathbf{95.3_{.4}}$	$\mathbf{89.1_{.2}}$	$\mathbf{97.0_{.5}}$	$\mathbf{95.8_{.3}}$

Table 3: Mean accuracy and standard deviation (subscript) for each model on four medical classification tasks. Best results for the LSTM and BERT-based models highlighted in bold. MeDa-BERT_16 denotes the MeDa-BERT model pretrained for 16 epochs.

3.1 Word embedding comparison

Using the medical word embeddings as input to an LSTM model resulted in large improvements compared to using general word embeddings. On average, LSTM+MeDa-WE outperformed the LSTM+General-WE model by 8.9 percentage points (PP). The largest improvement was seen on the 10-class bleeding site classification with an improvement of 15.6 PP.

3.2 Language model comparison

Comparing BERT and MeDa-BERT, the performance improvements were smaller. However, MeDa-BERT performed better on all datasets with an average improvement of 1.2 PP. The largest improvement was on the 10-class bleeding site classification with an improvement of 2.4 PP.

4 Discussion and limitations

This paper presented a new Danish medical corpus that was used to train NLP models. The corpus included medical books and text scraped from medical websites that provide information for both citizens and healthcare professionals. We applied different techniques to filter out non-medical data, e.g., by removing documents from non-medical departments or text written by non-healthcare professionals. While these steps did remove a large part of non-medical text, we cannot guarantee that the corpus did not include some of it. However, the results showed that models pretrained on the medical corpus performed better than general-domain models, especially for multiclass classification problems.

For the Danish language, few medical evaluation datasets are available and therefore the models were only evaluated on classification tasks. Moreover, the evaluation datasets were constructed from EHR text which has its own nuances compared to the text of the medical pretraining corpus, e.g., EHR text contains many spelling mistakes whereas the medical corpus contains few grammatical errors. These factors might limit the generalizability of the results. Future work should evaluate the models on other tasks, e.g., named-entity recognition and question answering which will provide a better understanding of the models' capabilities.

We found continuous small performance improvements by pretraining MeDa-BERT for more epochs. The model might improve with further pretraining but because of limited computational resources and the small rate of improvement, we did not explore this further. The model would also benefit from more medical pretraining data. Although this paper presented a large part of the available medical Danish text, more data could be collected, e.g., from other medical book publishers and websites.

The medical datasets used to evaluate the models are not publicly available because of privacy concerns. For future work, we will strive to publish parts of the medical corpus which requires permission from the text owners. We advise interested researchers to contact us for sharing possibilities.

5 Conclusion

This paper presented the first Danish medical corpus consisting of 133M tokens. The corpus was used to pretrain medical word embeddings and language models. The models trained on the medical corpus performed better than similar models trained on a general domain.

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. https://doi.org/10.18653/v1/W19-1909 Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. https://doi.org/10.18653/v1/D19-1371 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 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. https://aclanthology.org/Q17-1010 Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5.

Édouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomáš Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. https://doi.org/10.18653/v1/2020.acl-main.740 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, Online. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. http://arxiv.org/abs/1412.6980 Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Jannik Pedersen, Martin Laursen, Pernille Just, Anne Alnor, and Thiusius Savarimuthu. 2022a. Investigating anatomical bias in clinical machine learning algorithms. In *Findings of the European Chapter of the Association for Computational Linguistics: EACL 2023*, Dubrovnik, Croatia. Association for Computational Linguistics.

Jannik S Pedersen, Martin S Laursen, Thiusius Rajeeth Savarimuthu, Rasmus Søgaard Hansen, Anne Bryde Alnor, Kristian Voss Bjerre, Ina Mathilde Kjær, Charlotte Gils, Anne-Sofie Faarvang Thorsen, Eline Sandvig Andersen, et al. 2021. Deep learning detects and visualizes bleeding events in electronic health records. *Research and practice in thrombosis and haemostasis*, 5(4):e12505.

Jannik S Pedersen, Martin S Laursen, Cristina Soguero-Ruiz, Thiusius R Savarimuthu, Rasmus Søgaard Hansen, and Pernille J Vinholt. 2022b. Domain over size: Clinical electra surpasses general bert for bleeding site classification in the free text of electronic health records. In *2022 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, pages 1–4. IEEE.

Bo Peng, Emmanuele Chersoni, Yu-Yin Hsu, and Chu-Ren Huang. 2021. Is domain adaptation worth your investment? comparing bert and finbert on financial tasks. In *Proceedings of the Third Workshop on Economics and Natural Language Processing*, pages 37–44.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. https://doi.org/10.18653/v1/W19-5006 Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 58–65, Florence, Italy. Association for Computational Linguistics.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. https://doi.org/10.18653/v1/W18-5446 GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi 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 Rush. 2020. https://doi.org/10.18653/v1/2020.emnlp-demos.6 Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Appendices

A Preprocessing of text corpuses

A.1 Medical information portals

Netdoktor.dk provides information for citizens about diseases, symptoms, medication, and treatment. Netdoktor.dk contains sections that are not related to the medical domain and discussion forums where users can communicate. Therefore, we removed documents having links containing the following strings: debat, kultur, testdigselv, behandlerguiden, nyhedsbrev, nyheder, privacy-policy, kontaktnetdoktor, cookieinformation, disclaimer, sponsorindhold and discussions.

Region	Categories removed (in Danish)	Date retrieved	Tokens
Capital Region	Den sociale virksomhed Center for ejendomme Center for HR Center for Regional Udvikling Region Hovedstadens Apotek Steno Diabetes Center Copenhagena	October 2022	13,443,269
Northern Region	Logistik afdeling Teknisk Afdeling Himmerland Teknik Logistik Service	October 2022	6,505,559
Southern Region	Administration Service PsykInfo	September - November 2022	29,075,187
Region Zealand	Administration HR organisation og ledelse Indkøb IT PortørCentral Rengøring Økonomi Uddannelse	November 2022	6,387,083
Central Region		November 2022	25,156,478

Table 4: Categories removed and number of tokens from each region.

Moreover, citizens can ask medical questions [4] that are answered by medical professionals. We only included the answers to these questions.

Medicin.dk has three sub-pages: `www.min.medicin.dk` that provides information to citizens, `www.pro.medicin.dk` that provides information to health care professionals, and `www.indlaegssedler.dk` that contains information about medicine. We included all documents from these webpages.

Sundhed.dk provides information for medical professionals [5] and citizens [6] about diseases, symptoms, medication and treatment. We included all documents from these webpages.

A.2 Clinical guidelines

We collected clinical guidelines from the 5 regions of Denmark: The Capital Region of Denmark, The Region of Northern Denmark, The Region of Southern Denmark, The Region of Zealand, and The Central Region of Denmark.

For each region we removed non-medical documents, seen in Table 4.

B Model parameters and pretrainng loss

Table 5 shows the architecture and optimization parameters for MeDa-BERT. Table 6 shows the masked language modelling loss for MeDa-BERT during pretraining.

[4] `www.netdoktor.dk/brevkasser/`

[5] `https://www.sundhed.dk/sundhedsfaglig/`

[6] `https://www.sundhed.dk/borger/`

Parameter	Value
Architecture	
Number of layers	12
Hidden size	768
FFN inner hidden size	3072
Attention heads	12
Attention head size	64
Dropout	0.1
Attention dropout	0.1
Max seq. length	512
Optimization	
Learning rate	1e-4
Optimizer	AdamW
Adam weight decay	0.01
Adam epsilon	1e-6
Adam beta1	0.90
Adam beta2	0.98
Learning rate decay	Linear
Batch size	4032
Warm up	1 epoch
Epochs	16, 32, 48
Gradient clipping	1.0

Table 5: Architecture and optimization parameters for pretraining MeDa-BERT

	Train loss	Validation loss
MeDa-BERT_16	2.122	2.019
MeDa-BERT_32	1.874	1.792
MeDa-BERT_48	1.766	1.673

Table 6: Masked language modelling loss for MeDa-BERT during pretraining. MeDa-BERT_16 denotes the model pretrained for 16 epochs.