De-identifying Norwegian Clinical Text using Resources from Swedish and Danish

Anastasios Lamproudis, MSc[1], Sara Mora, PhD[2], Therese Olsen Svenning, MSc[1], Torbjørn Torsvik, MD, MPH[1], Taridzo Chomutare, PhD[1,4], Phuong Dinh Ngo, PhD[1,5], Hercules Dalianis, PhD[1,3]
[1]Norwegian Centre for E-health Research, Tromsø, Norway;
[2] Department of Informatics, Bioengineering, Robotics and System engineering (DIBRIS), University of Genoa, Genoa, Italy;
[3]Department of Computer and Systems Science (DSV), Stockholm University, Kista, Sweden.
[4]Department of Computer Science, UiT - The Arctic University of Norway, Tromsø, Norway.
[5]Department of Physics and Technology, UiT - The Arctic University of Norway, Tromsø, Norway.

Abstract

The lack of relevant annotated datasets represents one key limitation in the application of Natural Language Processing techniques in a broad number of tasks, among them Protected Health Information (PHI) identification in Norwegian clinical text. In this work, the possibility of exploiting resources from Swedish, a very closely related language, to Norwegian is explored. The Swedish dataset is annotated with PHI information. Different processing and text augmentation techniques are evaluated, along with their impact in the final performance of the model. The augmentation techniques, such as injection and generation of both Norwegian and Scandinavian Named Entities into the Swedish training corpus, showed to increase the performance in the de-identification task for both Danish and Norwegian text. This trend was also confirmed by the evaluation of model performance on a sample Norwegian gastro surgical clinical text.

Introduction

Electronic Medical Records (EMRs) contain large amounts of free text whose secondary use can further research and benefit patient care. However, patient privacy concerns arose [1] in parallel with the increasing effort in devising systems able to automatically extract information from this kind of data source. EMRs contain information that can unequivocally identify a person and therefore, in order to use them in clinical research, texts written in natural language must be stripped of all Protected Health Information (PHI). In the US, PHIs are defined by the Health Insurance Portability and Accountability Act (HIPAA). PHI can be "direct identifiers", such as first name, last name, and social security numbers, or "quasi-identifiers" that can be used to identify people if combined. As opposed to more structured data sets, there is no easy way of removing all PHIs from unstructured natural language text [2]. Removal of PHIs manually involves a great amount of manual labor, is error prone, and often not feasible if the dataset is very large [3].

Automatic de-identification (DEID) [4] is a specific area of research in the field of clinical text mining, also involved in a vast variety of real-world applications. DEID aims at removing a list of identifiers from free text. De-identifiers have often relied on both ruled-based and machine learning (ML) based methods [5]. More recently, the use of deep learning has improved performance, exploiting text embeddings and neural networks [6, 7]. These methods require a corpus where PHIs are annotated by expert coders. Such resources are labour intensive to create, and few open-source tools are available for languages other than English [8, 9], and they yield the best results.

Considering the Norwegian language, an early approach was presented in [10]. The authors described the architecture of a de-identification system for general practitioner data, both structured and unstructured. The work involved a rule-based approach to identify exact and approximate matches between the clinical text and the vocabulary created by combining several Norwegian sources of entities (names of health institutions, names of geographical areas and locations, etc). In another later research project, [11], 225 Norwegian patient records were de-identified using a rule-based approach where the system obtained 0.68 precision. A more recent approach was a hybrid de-identification system for Norwegian clinical text, [12]. It was based on both Conditional Random Fields (CRF), and regular expressions. It could identify 8 classes, 4 of them with machine learning approach and the rest with regular expressions. *First Name, Last Name, Health Care Unit* and *Location* were identified using machine learning, *Age, Date, Social Security* and *Phone Number* were identified using regex. The de-identifier was trained on the NorNe non-clinical Norwegian corpus and evaluated on the synthetic Norwegian clinical corpus NorSynthClinical PHI. The system using the hybrid approach obtained an F_1-score of 0.73 and a recall of 0.62.

One major challenge involving the development of these models for the Norwegian language is the lack of annotated data. The manual annotation of a clinical training corpus in Norwegian requires many resources both from a time use and from a human resources point of view, and therefore may not be possible in a short-term time period. Consequently, it might be beneficial to investigate the use of datasets from other similar languages for such tasks until openly available datasets for Norwegian are available.

These datasets can be evaluated with a multitude of model choices from the ones openly available. Specifically, after the advent of self-attention architectures and the breakthrough they introduced in the field of NLP, with the most recent breakthroughs demonstrated from models such as chatGPT[1], LLaMa [13], and BARD[2], more accessible, older models of the same technology, can still be relevant. Models such as the Bidirectional Encoder Representations from Transformers (BERT) are still quite popular as they can achieve good performance while being very efficient in terms of computational resources. Furthermore, there is strong support for these models from previous literature in the field of cross-lingual transfer learning dealing with low resources languages [9, 14]. In our case, and because Large Language Models (LLMs) are usually provided as services and are not run locally due to the heavy need for computational resources to do so, it is not possible to use them for the task due to the strict privacy framework which is applied for clinical data.

To summarise the aim of this study, the goal is to develop a model that can adequately de-identify 10 million gastro surgery notes in Norwegian encompassing 30 000 patients from the University Hospital of North Norway, UNN in Tromsø that are going to be used for predicting ICD-10 diagnosis codes. This model is part of the effort of developing a deidentification application for the clinical Norwegian language. Since a proper clinical learning corpus in Norwegian is not yet available the usability of a BERT-based model learned on Swedish equivalent data is investigated. Specifically, the aim is to explore the benefit of using and extending the training dataset using a sequence of different interventions .

Table 1: The table presents the Stockholm EPR PHI Pseudo Corpus annotated entities in comparison with the oversampled balanced corpora created in this work. (see Figure 1 in Experimental Design of Methods section).

	Pseudonymized Stockholm EPR PHI Corpus	**Balanced corpora (Dataset II)**
First Name	928	5,999
Last Name	923	5,170
Phone Number	135	4,514
Age	56	2,102
Full Date	500	4,373
Date Part	710	4,381
Health Care Unit	1,021	7,653
Location	95	4,355
Organisation	53	3,727
Total	4,421	42,274

[1] https://openai.com/blog/chatgpt

[2] https://en.wikipedia.org/wiki/Bard_(chatbot)

Data

This section presents the complete list of data resources used in this project.

Training dataset. The learning sample or **training dataset** includes Swedish PHI annotated data used to train the model to perform the Named Entity Recognition (NER) task. Furthermore, there is access to a list of Norwegian entities used for text permutation.

- **Pseudonymized version of the Stockholm EPR PHI Corpus.** The original Stockholm EPR PHI Corpus is first described in [15] and later processed and updated to a gold standard in [16]. The dataset, after an initial preprocessing, contains 100 patient records written in Swedish from four different medical specialities. The corpus has 4,421 PHI annotations distributed in 9 PHI categories. These are *First Name, Last Name, Phone Number, Age, Location, Health Care Unit, Date Part, Full Date* and *Organisation*, see Table 1. These entities is a subset of the HIPAA standard. The dataset encompasses in total approximetely 174 000 tokens and has been pseudonymised, where the real entities have been replaced with surrogates [17].
- **Lexicon of Norwegian and Sami identifiers.** It includes toponyms and locations in Norway, both in Norwegian and Sami, names extracted from Norwegian text and from the website of the Statistics Norway, and Norwegian Health Care unit names. These resources have been extracted from the Statistisk sentralbyrå (Statistics Norway)[3], some resources are from [17] and others from [11].

Test dataset. The test samples or (test datasets) are composed of a number of open-source datasets belonging to the Norwegian and Danish languages, a test partition of the **Stockholm EPR PHI Pseudo Corpus**, and a test dataset belonging to the clinical Norwegian language, specifically annotated for this work. The goal is to evaluate the general model performance over all the different datasets in an effort to approximate more accurately model performance in languages other than Swedish. The datasets used as test samples in this work are summarised bellow:

Pseudonymized version of the Stockholm EPR PHI Corpus The test sample of this corpus, excluded from the learning session is used in the evaluation of the model. It consists of 38,287 tokens, containing 1,168 entities.

Danish Dependency Treebank (DaNE) dataset. This dataset, introduced in [18], is a named entity annotation produced using the CoNLL-2003 schema for the Danish Universal Dependencies treebank. It is a publicly available resource, which follows the Universal Dependencies standards [19], and consists of a collection of 474 annotated texts.

DaNED The dataset derived from DaCoref [20], is annotated with Silver-standard Name Annotations from Wikipedia Markups [21].

Norwegian Named Entities (NorNE) [4] It is a manually annotated corpus written in Norwegian language containing a rich set of entities that belong to several identifiers categories, such as Norwegian names, organisations, locations, etc [22].

Norwegian synthetic data The only known open available clinical dataset in Norwegian is the NorSynthClinical PHI Corpus. It was manually annotated for PHIs by two native Norwegian speakers with an inter annotator agreement of 0.94 F_1-score, [12].

Norwegian clinical data from the University Hospital of North Norway The Norwegian gastrosurgical dataset contains 30,000 adult patient records encompassing approximately 10 million clinical notes. Patients who were in contact with the gastrosurgical department at the University Hospital of North Norway between 2017 and 2022. A sample of 200 patient records[5] from the Norwegian gastro dataset was used for the evaluation. Among them,

[3] Statistisk sentralbyrå, https://www.ssb.no/befolkning/

[4] https://huggingface.co/datasets/norne

[5] This research was approved by The Norwegian Regional Committees for Medical and Health Research Ethics (REK) North, decision number 260972

a subset containing 12,535 tokens was selected, as well representative for each class of entity. This subset was annotated by one Norwegian native speaker. She used the Visual Studio Code editor[6] to annotate the text, which was in CoNLL format[7]. In total 610 annotations were carried out for *First_Name, Last_Name, Phone_Number, Age, Location, Health_Care_Unit, Date_Part, Full_Date* and *Organisation* see Table 2.

Table 2: Distribution of manually annotated named entities in the Norwegian gastro corpus.

Manually annotated entities	
Age	31
Date_Part	54
Full_Date	71
First_Name	88
Last_Name	97
Health_Care_Unit	161
Location	80
Organisation	15
Phone_Number	13
Sum	610

Methods

This work was configured as a Named Entity Recognition task involving a BERT based model which was not previously used for de-identification of Norwegian clinical text. Different iterations of the **Stockholm EPR PHI pseudo corpus** were created with the aim to evaluate a set of potential steps needed in developing a model trained on text from a different but related language source (Swedish), and if it could yield a good performance in Norwegian tasks.

This work is part of a broader project which aims at building a de-identification application, called NorDeid, for Norwegian clinical text. In order to complete the final application, more identification steps will be added, i.e. using regular expressions to match phone numbers, and social security numbers.

Classifier

The common practise was followed, starting with a pretrained transformer (BERT) model, and further fine-tuned for the task of interest. Three different pretrained models were considered and evaluated:

SweDeClin-BERT SweDeClin-BERT [23] is based on KB-BERT [24], further adapted to the clinical domain by a pretraining process, using pseudonymized Swedish clinical text.

NorBERT NorBERT[8] [25] is a BERT model, developed for the Norwegian language. It is available in the **Huggingface**[9] online repository for language models.

ScandiBERT ScandiBERT[10] is a transformer model based on the RoBERTA architecture [26]. The model is pretrained with text from all Scandinavian languages including Swedish, Norwegian, Danish, Icelandic, and Faroese. It is also available in the **Huggingface** repository.

The fine-tuning of the models consisted of a token classification task where a model was trained to predict whether each token in an input sentence belonged to any of the classes of interest or not. It was a multi-class classification

[6] Visual Studio Code editor: https://code.visualstudio.com/

[7] SIGNLL, ACL's Special Interest Group on Natural Language Learning, https://www.signll.org/conll/

[8] https://huggingface.co/ltg/norbert

[9] https://huggingface.co/

[10] https://huggingface.co/vesteinn/ScandiBERT

task executed at the token level of the sentence. Having performed a narrow hyper-parameter search, the finetuning parameters were finalised and presented in Table 3, and The model was trained until convergence in terms of validation loss. The **development** set of the Pseudonymized version of the Stockholm EPR PHI Corpus was used as a **validation** set with the final reported results generated using test samples. As test sets, a multitude of datasets were used, in Norwegian and Danish respectively, as well as the Pseudonymized Stockholm EPR PHI Corpus test partition. These datasets are listed in Method section.

Table 3: Values of the finetuning hyperparameters used for all the experiments.

Hyperparameter	**Value**
Learning rate	$1 \cdot 10^{-5}$
Batch size	64

Experimental Design

The goal of present study is to explore the performance of models trained on augmented versions of the Pseudonymized Stockholm EPR PHI Corpus in a de-identification task involving Norwegian clinical text. The performances of the models are presented in terms of F_1-scores, both **macro** and **micro** averaged.

However, as first step, the performance of the three aforementioned BERT-based pretrained models was evaluated. Each model was trained on the **Pseudonymized Stockholm EPR PHI Corpus**, and evaluated on the **Norwegian synthetic data**, results are reported in Table 4.

The model which showed the highest performance was **ScandiBERT** and was selected for the experiments in the rest of this work which are visualized in Figure 1.

Dataset I - Pseudonymized Stockholm EPR PHI corpus In the first step, the train dataset was composed only by the Pseudonymized Stockholm EPR PHI Corpus, described in the Methods section, without any modifications or alterations.

Dataset II - Balancing by oversampling The first intervention produced an over-sampled version of the corpus. This is expected to benefit the models trained with this dataset as it will alleviate the influence of the most frequent classes during the training process. More specifically, a sampling with replacement was performed on the sentences containing those entities overall less frequent. The aim was to reach at least 400 occurrences for each entity within the total dataset. However, each sample could contain a multitude of entities, therefore **Dataset II** ended up with an inflated count of each entity. Table 1 compares, for each entity class, the original imbalanced dataset and the over-sampled balanced version.

Dataset III a - Permuting the new samples with in-dataset resources As a second intervention, the newly created samples of Dataset II were permuted with the aim of increasing the variability of the new samples. To do that, a vocabulary containing all known entities was created by extracting all distinct occurrences of entities within the original dataset. Each entity was then substituted within the new samples with one of the same class randomly extracted from the vocabulary.

Table 4: Results in terms of F_1-score for the different models, using **Pseudonymized Stockholm EPR PHI Corpus** as a learning set and the **Norwegian synthetic data** as a test set.

	F_1-score macro averaged	**F_1-score micro averaged**
SweDeClin-BERT	0.41	0.39
NorBERT	0.62	**0.57**
ScandiBERT	**0.64**	0.55

Dataset III b - Permuting the new samples with Norwegian entities As an alternative to the previous step, the influence of directly injecting Norwegian and Sami entities in the dataset was explored. Specifically, randomly extracted entities from the lexicon of Norwegian and Sami identifiers, described in the Methods section, were used to replace some of the entities in **Dataset II**.

Dataset IV - Generating entities with ScandiBERT Next, to further increase the variability of new samples within **Dataset III a**, Scandi-BERT was used to generate new entities and use them to replace some of the entities. Specifically, new entities were generated by tasking the model to fill the ”mask”. This resulted in the generation of entities conditioned on each of the sample that was subjected to this process. This was carried out on 50% of the total population of entities present in the over-sampled subset of the dataset, which originally belonged to classes with 100 or less occurrences.

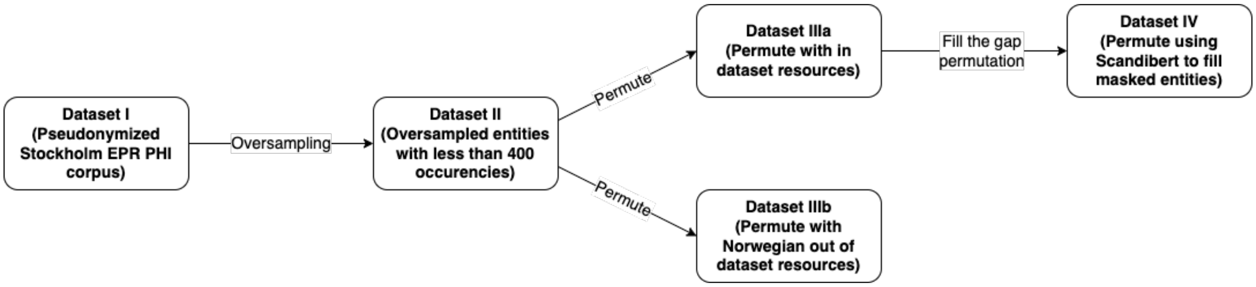

Figure 1: Graphical visualization of **experimental design** steps. Each model was trained using these incrementally modified datasets with the aim to evaluate the impact in the final performance.

Results

The performances obtained considering all the combinations of models and test sets are presented in Table 5 (**macro-averaged** F_1-score) and Table 6 (**micro-averaged** F_1-score).

Overall, as expected, balancing the entities of the dataset is beneficial, even in the form of sampling with replacement. Comparing the performances reported in the first two lines of both tables, that correspond to the model trained on **Dataset I (Original)** and the one trained on **Dataset II**, an increase is evident across all languages.

The following permutation step (**Dataset III a**), also yields models that generally perform better. This can be seen in Tables 5 and 6 respectively, where the models developed outperform both of the previous models trained on **Dataset II** and on the **Original** version.

Excluding the performance on the Pseudonymized Stockholm EPR PHI Corpus, this trend of improved performance continues with the models trained with the permuted dataset obtained using **ScandiBERT** (**Dataset IV**) while we can think the drop in performance on the original dataset as a trade-off in the generalization of the resulting model. The model yields the best performance, in terms of **macro** averaged scores, on all test sets in Norwegian and Danish.

Lastly, the Norwegian surrogate inclusion dataset (**Dataset III b**), yields the lowest performance out of all tested steps except for **NorNE** and the original version of the learning set.

Regarding the de-identification of the Norwegian gastro data set, the de-identifier performed best for *First_Name, Last_Name, Phone_Number, Location* and *Full_Date* and worst on *Age* and *Organisation*.

Discussion

In the present work, we aimed at investigating the possibility of using a Swedish in-domain dataset along with potential data processing steps, to train a model for the process of automatic de-identification for the Norwegian language, given that a proper annotated corpus for Norwegian clinical text is not available. To that end, we first evaluated the

Table 5: Results in **macro** F_1-score. Each experiment is performed 10 times. As a final performance is reported the mean F_1-score for all 10 experiments.

	Pseudonymized Stockholm EPR PHI Corpus	**Norwegian synthetic data**	**NorNE**	**DaNED**	**DaNE**	**Norwegian gastro dataset**
Dataset I (Original)	0.77	0.64	0.43	0.50	0.55	0.46
Dataset II	0.92	**0.71**	0.47	0.54	0.63	0.56
Dataset III a	**0.95**	0.69	0.49	**0.55**	0.65	0.56
Dataset III b	0.89	0.63	**0.52**	0.51	0.60	0.53
Dataset IV	0.91	**0.71**	0.51	**0.55**	**0.68**	**0.57**

Table 6: Results in **micro** F_1-score. Each experiment is performed 10 times. As a final performance the mean F_1-score is reported for all 10 experiments.

	Pseudonymized Stockholm EPR PHI Corpus	**Norwegian synthetic data**	**NorNE**	**DaNED**	**DaNE**	**Norwegian gastro dataset**
Dataset I (Original)	0.92	0.55	0.50	0.52	0.57	0.53
Dataset II	0.94	**0.65**	0.52	0.55	0.61	0.61
Dataset III a	**0.95**	0.59	**0.54**	**0.57**	0.62	0.61
Dataset III b	0.91	0.60	0.52	0.48	0.56	0.54
Dataset IV	0.91	0.61	0.53	0.56	**0.65**	**0.62**

performance of a model trained with a Swedish pseudonymised clinical corpus on corpora written in two related languages, Danish and Norwegian. Then we evaluated the impact of extending the train set with subsequent interventions. Results showed that the operation of text augmentation benefits the overall model performance on all the open source test sets including, Norwegian synthetic data, NorNE, DaNED, and DaNE. We then selected the best model and evaluated its performance on an annotated subsample of the Norwegian gastro dataset. This further confirmed the previous evaluation on a sample of clinical real world data, collected from the gastrosurgical department of the University Hospital of North Norway. These results confirmed that the operation of text augmentation, in conjunction with dataset permutation using generated and in-dataset resources, induced a better performance in the de-identification task and that the performance gain is comparable with that obtained using the other test sets.

The main advantage of our approach is that, as it is using machine learning (BERT-based), it partially omits the need of a rule-based system. This yields benefits, such as the simplification of a usually very big rule-set that would be otherwise needed to perform the task, and the generalization of the performance to future unseen entities since the model does not search for specific words/tokens but it assigns labels based on the context in which these entities are found. For this reason, it reaches a good performance training the model on **Dataset IV** even though it does not contain many specific examples of Norwegian and Danish entities.

The approach however presents some limitations as well. Even though the performance is good, it is still far from the adequate level which would allow the use of the model in a real scenario. Furthermore, due to the lack of appropriate annotations in the **Stockholm EPR PHI Corpus**, the model is not trained in recognising some of the categories of interest that are found in the Norwegian gastro dataset and that should be addressed, i.e. *Email Address, Social Security Number, User Name*[11] and *Family number*[12]. This kind of information is not always written in a standard format but at least a common schema can be identified. We plan to implement a rule-based system, i.e. regular expressions, to cover the recognition of these patterns.

Regarding the choice of sensitive entities in this work, it is important to consider the current state of regulations in Europe. Currently there is an ongoing debate among physicians, researchers, and lawyers about which entities in patient records are sensitive and which are not. A the time of writing of this work, a consensus that would result in a list

[11] Clinician's domain user name.

[12] Genetic tracking in families.

of widely accepted sensitive entities has not been reached. As such, the EU General Data Protection Regulation does not give a clear guidance, in contrast to the USA and the HIPAA standard. The choice of entities then is dependent on the context and level of sensitivity of each set of patient records that is going to be used in the research. For example it is not always clear if age or date is sensitive or not, or if health care unit is sensitive. Other possible entities to de-identify are for example *ethnicity, religion, profession, relation* such as *mother, father, son, daughter, spouse, wife, husband*, etc.

In conclusion, this first attempt for de-identification reaches encouraging results. We see that we can build adequately trained models using a neighbouring language (Swedish), along with traditional data manipulation techniques, and more recent data augmentation techniques using machine learning to improve and build a model that will perform the task in a different but similar language (Norwegian). However, further steps will be needed before a model will be ready to be used in a real de-identification task.

References

1. Shin SY, Park YR, Shin Y, Choi HJ, Park J, Lyu Y, et al. A de-identification method for bilingual clinical texts of various note types. Journal of Korean medical science. 2015;30(1):7-15.

2. Lison P, Pilán I, Sánchez D, Batet M, Øvrelid L. Anonymisation models for text data: State of the art, challenges and future directions. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers); 2021. p. 4188-203.

3. Bier E, Chow R, Gollé P, King TH, Staddon J. The rules of redaction: Identify, protect, review (and repeat). IEEE Security & Privacy. 2009;7(6):46-53.

4. Uzuner Ö, Luo Y, Szolovits P. Evaluating the state-of-the-art in automatic de-identification. Journal of the American Medical Informatics Association. 2007;14(5):550-63.

5. Meystre SM, Friedlin FJ, South BR, Shen S, Samore MH. Automatic de-identification of textual documents in the electronic health record: a review of recent research. BMC medical research methodology. 2010;10(1):1-16.

6. Liu Z, Tang B, Wang X, Chen Q. De-identification of clinical notes via recurrent neural network and conditional random field. Journal of biomedical informatics. 2017;75:S34-42.

7. Dernoncourt F, Lee JY, Uzuner O, Szolovits P. De-identification of patient notes with recurrent neural networks. Journal of the American Medical Informatics Association. 2017;24(3):596-606.

8. Becker M, Kasper S, Böckmann B, Jöckel KH, Virchow I. Natural language processing of German clinical colorectal cancer notes for guideline-based treatment evaluation. International journal of medical informatics. 2019;127:141-6.

9. Catelli R, Gargiulo F, Casola V, De Pietro G, Fujita H, Esposito M. Crosslingual named entity recognition for clinical de-identification applied to a COVID-19 Italian data set. Applied soft computing. 2020;97:106779.

10. Tveit A, Edsberg O, Rost T, Faxvaag A, Nytro O, Nordgard T, et al. Anonymization of general practioner medical records. In: Proceedings of the second HelsIT Conference; 2004. .

11. Bjurstrøm R, Singh J. De-identification of Norwegian Health Record Notes. Master Thesis, NTNU; 2013.

12. Bråten S, Wie W, Dalianis H. Creating and evaluating a synthetic Norwegian clinical corpus for de-identification. In: Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa); 2021. p. 222-30.

13. Touvron H, Lavril T, Izacard G, Martinet X, Lachaux MA, Lacroix T, et al. Llama: Open and efficient foundation language models. arXiv preprint arXiv:230213971. 2023.

14. Alyafeai Z, AlShaibani MS, Ahmad I. A survey on transfer learning in natural language processing. arXiv preprint arXiv:200704239. 2020.

15. Velupillai S, Dalianis H, Hassel M, Nilsson GH. 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. 2009;78(12):e19-26.

16. Dalianis H, Velupillai S. De-identifying Swedish clinical text - refinement of a gold standard and experiments with Conditional random fields. Journal of Biomedical Semantics. 2010 Apr;1(1):6. Available from: https://doi.org/10.1186/2041-1480-1-6.

17. Dalianis H. Pseudonymisation of Swedish Electronic Patient Records Using a Rule-Based Approach. In: Proceedings of the Workshop on NLP and Pseudonymisation, In conjunction with Nodalida 2019. Turku, Finland: Linköping Electronic Press; 2019. p. 16-23.

18. Hvingelby R, Pauli AB, Barrett M, Rosted C, Lidegaard LM, Søgaard A. DaNE: A named entity resource for Danish. In: Proceedings of the 12th Language Resources and Evaluation Conference; 2020. p. 4597-604.

19. Johannsen A, Alonso HM, Plank B. Universal dependencies for danish. In: International Workshop on Treebanks and Linguistic Theories (TLT14); 2015. p. 157.

20. Kromann M, Lynge S. The Danish Dependency Treebank v. 1.0; 2004. Linguistic resource containing 100.000 words of linguistically annotated text.

21. Pan X, Zhang B, May J, Nothman J, Knight K, Ji H. Cross-lingual name tagging and linking for 282 languages. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); 2017. p. 1946-58.

22. Jørgensen F, Aasmoe T, Husevåg ASR, Øvrelid L, Velldal E. NorNE: Annotating named entities for Norwegian. arXiv preprint arXiv:191112146. 2019.

23. Vakili T, Lamproudis A, Henriksson A, Dalianis H. Downstream task performance of bert models pre-trained using automatically de-identified clinical data. In: Proceedings of the Thirteenth Language Resources and Evaluation Conference; 2022. p. 4245-52.

24. Malmsten M, Börjeson L, Haffenden C. Playing with Words at the National Library of Sweden–Making a Swedish BERT. arXiv preprint arXiv:200701658. 2020.

25. Kutuzov A, Barnes J, Velldal E, Øvrelid L, Oepen S. Large-scale contextualised language modelling for Norwegian. arXiv preprint arXiv:210406546. 2021.

26. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:190711692. 2019.