POPULAR SCIENTIFIC ABSTRACT

Jannik Skyttegaard Pedersen

Unlocking the Potential of Electronic Health Records With Danish Clinical Language Models for Text Mining

This PhD dissertation focuses on the development of language technology that can be used to extract clinical information from Danish electronic health records (EHRs). EHRs contain important health-related information that can be used to guide the treatment of the patient. However, a large part of the information is stored in unstructured narrative text, making it difficult and time-consuming to extract relevant details, especially in acute situations. This can lead to important information being lost, and increase the risk of misdiagnosis and adverse treatment outcomes.

The recent paradigm shift in the field of natural language processing (NLP) has produced automatic textprocessing tools with unprecedented performances. These tools can potentially be used to extract and structure the information from the narrative text of EHRs. This information could be used by healthcare professionals to support the treatment of patients. However, research in modern language technology has mostly been explored for high-resource languages like English, while the development of Danish language technology has received less attention, especially for specialized domains such as the clinical.

This dissertation explores the potential of language technology to automatically extract information from the narrative text of Danish EHRs. Moreover, it emphasizes the importance of developing language resources tailored for the Danish clinical domain, as it could be used to enhance clinical research possibilities, improve patient treatment, and reduce costs in the Danish healthcare sector.

The dissertation includes the development of two Danish clinical language models and clinical evaluation datasets. One of the models, Clin-ELECTRA, was developed using ~300,000 EHRs from Odense University Hospital. The other model, MeDa-BERT, was developed using medical text collected from the internet and medical books. Both models show promising results for extracting clinically important information from the narrative text of Danish EHRs.

The dissertation also explores how dataset curation impacts biases in clinical language models. Specifically, it shows that the performance of clinical text classification tools could vary significantly if the datasets used to train the models are not distributed properly.

Furthermore, the dissertation presents the results of a language model that can be used to extract bleeding events from Danish EHRs and evaluates medical doctors' performance when using the bleeding model as an assistive tool. Finally, the dissertation presents a language model that can be used to extract important medical information such as diseases, symptoms, and treatments in the narrative text of Danish EHRs.