Title: Natural Language Processing in Dutch Free Text Radiology Reports: Challenges in a Small Language Area Staging Pulmonary Oncology

Authors: J. Martijn Nobel, Sander Puts, Frans C. H. Bakers, Simon G. F. Robben, André L. A. J. Dekker

Abstract: To categorize the T-stage of pulmonary oncology from free-text radiological reports in Dutch, this study examines the difficulties and potential solutions in deploying a Natural Language Processing (NLP) tool. The goal of the project is to develop a rule-based system that correctly interprets and categorises data by utilising PyContextNLP, spaCy, and regular expressions. It shows the particular difficulties encountered in a limited language region and the ways in which natural language processing (NLP) might improve radiological reports by providing structure and enabling additional analysis. The algorithm showed promise in the method used in this pilot study, with an overall accuracy of 83% in the training set and 87% in the validation set. Although the results are encouraging, the report suggests that more research is necessary to introduce more machine learning algorithms and maybe lessen the work required to obtain domain-specific knowledge. This research should involve larger datasets and external validation.

Background: The main means of communication between radiologists and referring doctors is through radiological reports. The complexity of the present medical data flow is discussed in the paper, with a focus on radiology since there is a lot of image and textual data collected there. It presents the idea of tumour staging in pulmonary oncology and talks about the difficulties in using radiological report free text unstructured data for data mining.

Methodology: In order to establish a baseline for the upcoming, the study proposes a rule-based natural language processing algorithm with machine learning pre-processing steps. It describes the pre and post-processing actions required to prepare data for analysis. In order to eliminate artifacts and standardise abbreviations, a sectionizer was created to pick pertinent portions of the report. With an emphasis on tumor size, existence, and involvement, the algorithm's structure is intended to categorise the T-stage in accordance with the eighth TNM classification system.

Results: The algorithm attained 83% accuracy in the training set and 87% accuracy in the validation set, indicating positive results from the study. The findings demonstrate that a rule-based natural language processing (NLP) technique can successfully identify radiological data, even in a limited language domain such as Dutch. This establishes a solid basis for future study and development.

Discussion: The conversation emphasises how crucial domain-specific expertise is for developing rule-based algorithms in situations where there is a shortage of training data. It highlights the difficulties in extracting context, the diversity of reporters' lexicon, and how crucial context is to precise descriptions. The study's shortcomings are also covered in the publication, such as the small sample size and the algorithm's exclusive training on a single dataset from a single radiological department.

Conclusion: The study concludes that natural language processing (NLP) holds great promise for the mining of free-text radiological reports written in languages other than English, such as Dutch. However, the context of the ideas discussed in the study has a major role in how well a free-text mining technique is implemented. The study highlights the requirement of using NLP in a balanced way while maintaining standards; next research should concentrate on utilizing additional machine or deep learning techniques as well as external validation.