Pseudonymisation of Patient Records

Graduate

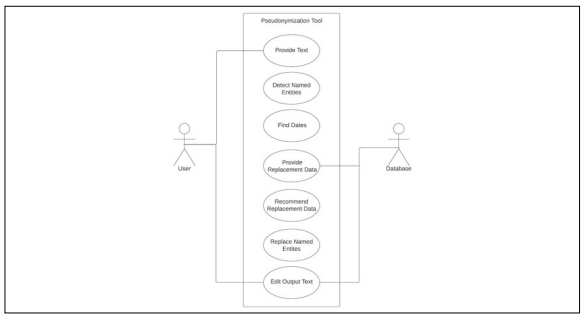
Conradin Kleinstein

Introduction: The automatic evaluation of patient records is well suited for artificial intelligence using natural language processing. However, patient records are full of sensitive and personal data and cannot be used to build datasets. This makes research difficult up to impossible. In order to make the patient records usable for future research, they need to be pseudonymised first. The goal of this Master Thesis is to provide an algorithm that pseudonymises a text so that there is no way to trace it back to the original persona. However, pseudonymisation has to be consistent and preserve the scientific value of the data.

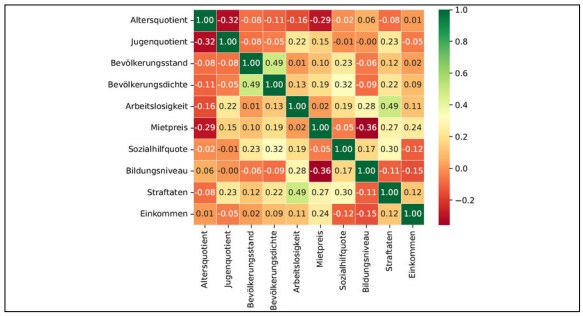
Approach / Technology: State-of-the-art named entity recognition models were analysed and implemented to detect sensitive data. In the second step, a database was built that stores names and locations used to replace the found sensitive data. Furthermore, to increase the consistency of the pseudonymised text, a nearest neighbour classifier was fit on ten different criteria and provided similar locations for replacement. Finally, the algorithm was wrapped in a graphical user interface.

Result: The result of this Master Thesis is a support software which automatically pseudonymises a given text. It can detect named entities with a performance of 92.32 \% and uses the determined pseudonymisation strategies to suggest a suitable replacement from the database. The database provides over 150'000 names and holds all Swiss cantons, districts and municipalities. The graphical user interface visualises the found and replaced entities and allows the user to customise the results.However, the developed system cannot cover all possible cases and needs human supervision.

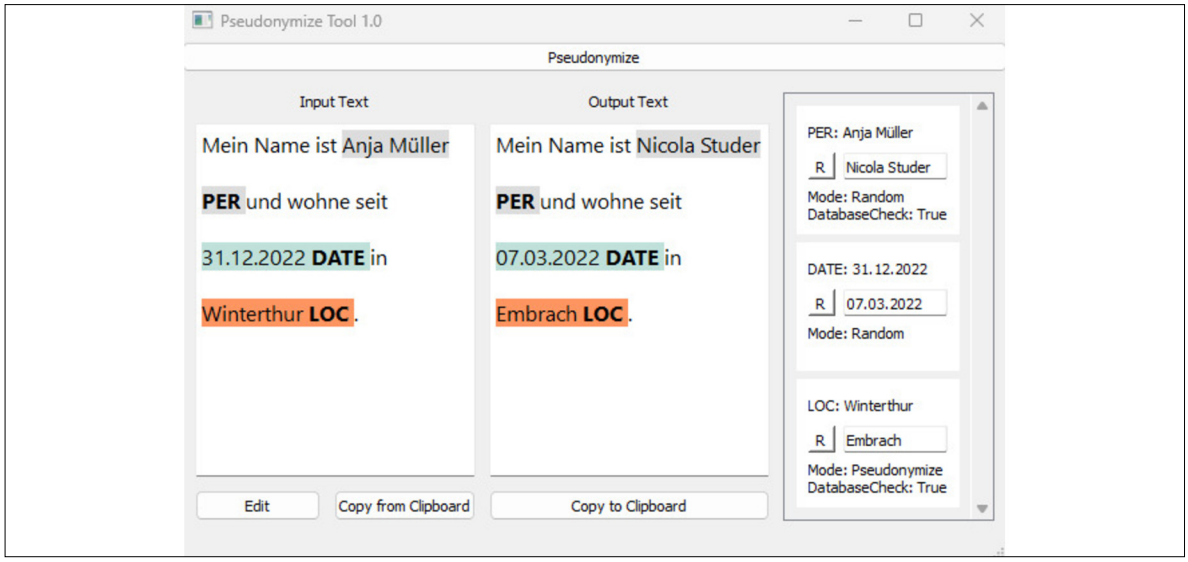
Use Case Diagramm of the Developed Software
Own presentment



Correlation Matrix of Features Used in the Recommender System
Own presentment



Pseudonymize Tool 1.0 Graphical User Interface
Own presentment



Advisor
Hannes Badertscher

Co-Examiner
Gabriel Sidler, Teamup Solutions AG, Zürich, ZH

Subject Area
Data Science