

# Multi-label text classification of German language medical documents

Stephan Spat<sup>a</sup>, Bruno Cadonna<sup>a</sup>, Ivo Rakovac<sup>a</sup>, Christian Gütl<sup>b</sup>, Huber Leitner<sup>c</sup>, Günther Stark<sup>c</sup>, Peter Beck<sup>a</sup>

<sup>a</sup> Institute of Medical Technologies and Health Management, JOANNEUM Research Forschungsgesellschaft mbH, Austria

<sup>b</sup> Institute for Information Systems and Computer Media, Graz University of Technology, Austria

<sup>c</sup> Steiermärkische Krankenanstaltenges. m.b.H., Graz, Austria

## Abstract and Objective

*Nearly at every patient visit medical documents are produced and stored in a medical record, often in unstructured form as free text. Growing amount of stored documents increases the need for effective and timely retrieval of information. We developed a multi-label classification system to categorize German language free text medical documents (e.g. discharge letters, clinical findings, reports) into predefined classes. A random sample of 1,500 free text medical documents was retrieved from a general hospital information system, and was assigned manually to 1 to 8 categories by a domain expert. This sample was used to train and evaluate the performance of 4 classification schemes: Naïve Bayes, kNN, SVM and J48. Additional tests of the effect of text preprocessing were done. In our study preprocessing improved the performance, and best results were obtained by J48 classification.*

## Keywords:

Machine learning, Classification, Medical Records, multi-label