Comorbidity identification in clinical documents with medical terminology-based weak supervision.

Sylvain Brouwer [12], Maurice van Keulen [1], Jeroen Geerdink [2], Johannes H. Hegeman [12]

[1] University of Twente, the Netherlands
sylvainbrouwer@gmail.com, m.vankeulen@utwente.nl
[2] Ziekenhuisgroep Twente, the Netherlands
{j.geerdink,h.hegeman}@zgt.nl

Knowledge of patient comorbidities is crucial for effective healthcare decision-making and predictive modeling, particularly for vulnerable populations such as geriatric hip fracture patients. While electronic health records (EHRs) contain a wealth of data, the nature of communication in healthcare results in much of this data being in the form of unstructured text, posing a challenge for data extraction. This challenge extends to information regarding patient comorbidity. Natural language processing and machine learning may offer solutions for processing this unstructured data for improving healthcare processes [1].

In this work, we apply machine learning to classify emergency department intake notes for elderly hip fracture patients in the groups of diagnoses (e.g. myocardial infarction, kidney diseases) of the Charlson Comorbidity Index (CCI). We first evaluate Naïve Bayes, Gradient Boosting, Random Forest, and a Transformer model in a fully supervised setting based on 3200 hand-labeled documents. While we find promising results for the more common categories of the CCI ($> 5\%$ prevalence), limited availability of labeled data coupled with a large class imbalance between categories led to poor performance for rarer classes.

To mitigate the effects of this class imbalance we augment our dataset with ±20000 documents for patients outside the hip fracture cohort. These documents are labeled programmatically [2], by matching their contents to key terms from established medical terminologies and ontologies like SNOMED CT and the Unified Medical Language System, complemented by pseudo-labels generated by a fully supervised Random Forest.

The best performing model was a Random Forest, trained on the augmented dataset. This model was able to predict the correct CCI-categories for a patient in 75% of test cases, and in 92% of test cases the predicted CCI-score was within 1 point of the correct CCI-score. While there is still room for improvement, particularly in classifying rarer groups of diagnoses, our results offer an encouraging outlook for a more complete overview of comorbidity in hospital information systems, and the inclusion of comorbid conditions as inputs for research and predictive models.

References

[1] Maurice van Keulen et al. ”Exploiting Natural Language Processing for Improving Health Processes”. In: *Proceedings of the 7th International Symposium on Data-Driven Process Discovery and Analysis (SIMPDA 2017).* (2017)

[2] Alexander Ratner et al. ”Data Programming: Creating Large Training Sets, Quickly”. (2017)