Caring is Sharing – Exploiting the Value in Data for Health and Innovation
M. Hägglund et al. (Eds.)

doi:10.3233/SHTI230273

Diagnosis Classification in the Emergency Room Using Natural Language Processing

Marieke M. VAN BUCHEM[a,1], Hanna H. 'T HART[a,b], Pablo J. MOSTEIRO[b], Ilse M.J. KANT[a] and Martijn P. BAUER[a]
[a] *Leiden University Medical Center, Leiden, The Netherlands*
[b] *University of Utrecht, Utrecht, The Netherlands*
ORCiD ID: Marieke van Buchem https://orcid.org/0000-0002-2917-0842

Abstract. Diagnosis classification in the emergency room (ER) is a complex task. We developed several natural language processing classification models, looking both at the full classification task of 132 diagnostic categories and at several clinically applicable samples consisting of two diagnoses that are hard to distinguish.

Keywords. Natural language processing, diagnosis classification, emergency medicine

1. Introduction

In the emergency room (ER), action by clinicians must be taken rapidly. Patients come in with various symptoms and illnesses, making it difficult for clinicians to diagnose a patient quickly and accurately. Errors in diagnosis occur more often at the ER than in the rest of the hospital [1], which can result in serious harm to the patient [2]. Natural Language Processing (NLP) has the potential to assist ER clinicians with diagnosing patients based on the clinical notes that they write while examining a patient. In the current study, we investigate the feasibility of developing a diagnosis classification model to assist ER clinicians in correctly and timely diagnosing a patient.

2. Methods

2.1. Data

We used the description of the history of present illness from the letter that is sent to the general practitioner after a patient visits the ER. The dataset consisted of 72.990 letters from unique encounters that took place between 2011 and 2021 at the Leiden University Medical Center. The number of unique diagnoses was 1997, which we categorized using the Clinical Classifications Software Refined tool. After removing encounters that could not be linked to an ICD-10 code, encounters that did not have a valid diagnosis (such as 'fever'), and categories that occurred less than 10 times, the number of encounters was 29871, with 132 unique diagnostic categories.

[1] Corresponding Author: Marieke M. van Buchem, E-mail: m.m.van_buchem@lumc.nl

2.2. Models

We trained a baseline model using TF-IDF and SVM and finetuned two pretrained BERT models (the Dutch BERTje and the Dutch medical MedRoBERTa.nl)[3].

2.3. Experiments

We conducted two experiments. In experiment 1, we trained the models on the full classification task, including all 132 diagnostic categories. In experiment 2, a clinician defined three sets of diagnoses that are difficult to distinguish from each other: heart failure versus Covid-19 (sample 1); biliary tract disease versus aortic peripheral and visceral artery aneurysms (sample 2); and acute hemorrhagic cerebrovascular disease versus meningitis (sample 3). Then, using only the encounters that included one of the diagnoses in a set, we trained the model to choose the correct diagnosis per encounter.

3. Results and Discussion

All models in experiment 2 outperformed the models in experiment 1. The BERT models outperformed the baseline model in all experiments, although the difference was small for experiment 2, sample 3. Within experiment 1, BERTje performed best, while within experiment 2, MedRoBERTa.nl performed best in all samples (see Table 1).

Table 1. Micro F1-score of the three different models in experiment 1 and experiment 2, samples 1-3.

Experiment number	Baseline	BERTje	MedRoBERTa.nl
Experiment 1	0.28	**0.35**	0.32
Experiment 2, sample 1	0.72	0.90	**0.91**
Experiment 2, sample 2	0.54	0.71	**0.79**
Experiment 2, sample 3	0.8	0.83	**0.86**

The current experiments show that looking at samples of diagnoses might be more feasible to develop for clinical practice than trying to create a classification model for all diagnostic categories. Within the next months, we will refine the samples we created with a larger group of clinicians. Furthermore, we will use LIME, an explanation algorithm, to explain the differences in output between the different models [4]. Lastly, we will develop multimodal models that also include structured data to optimize performance for both tasks.

References

[1] Hussain F, Cooper A, et alDiagnostic error in the emergency department: learning from national patient safety incident report analysis. BMC Emergency Medicine. 2019;19:1-9.

[2] Balogh EP, Miller BT. Committee on diagnostic error in health care. National Academies Press. 2015.

[3] Devlin J, Chang MW, Lee K, Toutanova K. Bert: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2019;1:4171–4186.

[4] Ribeiro MT, Singh S, Guestrin C. “Why should I trust you?” Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowlegde discovery and data mining. 2016;1135-1144.