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

doi:10.3233/SHTI230275

Few-Shot and Prompt Training for Text Classification in German Doctor's Letters

Phillip RICHTER-PECHANSKI[a,b,c,d,e], Philipp WIESENBACH[a,b,d], Dominic M. SCHWAB[b], Christina KIRIAKOU[b], Mingyang HE[a,b,e], Nicolas A. GEIS[b,d], Anette FRANK[e*] and Christoph DIETERICH[a,b,c,d*1]

[a] *Klaus Tschira Institute for Computational Cardiology, Heidelberg, Germany*
[b] *Department of Internal Medicine III, University Hospital Heidelberg, Germany*
[c] *German Center for Cardiovascular Research (DZHK) - Partner site Heidelberg/Mannheim, Germany*
[d] *Informatics for Life, Heidelberg, Germany*
[e] *Department of Computational Linguistics, Heidelberg University, Germany*

Abstract. To classify sentences in cardiovascular German doctor's letters into eleven section categories, we used pattern-exploiting training, a prompt-based method for text classification in few-shot learning scenarios (20, 50 and 100 instances per class) using language models with various pre-training approaches evaluated on CARDIO:DE, a freely available German clinical routine corpus. Prompting improves results by 5-28% accuracy compared to traditional methods, reducing manual annotation efforts and computational costs in a clinical setting.

Keywords. deep learning, prompting, language models, cardiology

1. Introduction and Methods

By using methods of natural language processing (NLP) and machine learning (ML) we aim to extract clinical information from unstructured doctor's letters. While most supervised ML approaches rely on large amounts of manually annotated training data, recent developments in NLP showed promising results in text classification tasks using pre-trained language models (PLM) and prompts [1]. Prompting exploits the ability of PLMs to infer knowledge from context, in combination with supervised methods they achieve state-of-the-art results on various text classification tasks. Doctor's letters are separated into sections, e.g. anamnesis, diagnosis and risk factors, which contain semantically related sentences. Here we present our initial results using pattern-exploiting training (PET) with various domain and task-adapted PLMs [2,3] on the task of section classification in German doctor's letters from the cardiology domain.
Our data is based on the CARDIO:DE corpus, a freely available and distributable large German clinical corpus from the cardiovascular domain encompassing 500 clinical routine German doctor's letters from Heidelberg University Hospital

[1] *Corresponding authors: Phillip RICHTER-PECHANSKI, Christoph DIETERICH E-mails: phillip.richter-pechanski@med.uni-heidelberg.de , christoph.dieterich@med.uni-heidelberg.de, frank@cl.uni-heidelberg.de. CD and AF share the last authorship.

(https://doi.org/10.11588/data/AFYQDY). For evaluation we used eleven CDA-compliant section categories included in CARDIO:DE such as medication and anamnesis.

We evaluated four medium-sized PLMs based on bidirectional transformer-based BERT encoder models [4]: (1) based on a publicly available German BERT model (*gbert-base*), (2) further task-adapted on our CARDIO:DE data set (*gbert-base-cardiode*), (3) further domain-adapted on an internal medical data set (200,000 German cardiology doctor's letters; *gbert-fine*), and (4) the two latter approaches combined (*gbert-fine-cardiode*). To evaluate prompting in a few-shot learning scenario we used PET, a state-of-the-art semi-supervised few-shot learning method using prompts, to classify sequences of text, achieving promising results in various domains [5]. We evaluated PET on three different training set sizes |T| = 20, 50, 100 and compared PET results with a baseline based on BERT with a sequence classification head.

2. Results

Table 1. Mean accuracy scores for |T| = 20, 50, 100 for four PLMs using a traditional sequence classification (SC) baseline model and PET. Trained on two different random seeds and training sets randomly extracted from CARDIO:DE400. We evaluated the models on 13,563 separate section annotations of CARDIO:DE100.

Model	SC |T|=20	PET |T|=20	SC |T|=50	PET |T|=50	SC |T|=100	PET |T|=100
gbert-base	28,2	54,7	45,3	67,3	62,9	72,3
gbert-base-cardiode	32,6	57,7	58,7	70,4	70	75,1
gbert-fine	37,1	57,6	60,2	70	71,7	76,4
gbert-fine-cardiode	28,4	**64,2**	48,1	**76**	67,5	**79,4**

3. Discussion and Conclusion

PET outperforms all SC baselines by large margins using any type of PLMs. The domain- and task-adapted PLM *gbert-fine-cardiode* outperforms the baseline and all other PET results. Already *gbert-cardiode* outperforms *gbert-base*, for both the baseline and PET.

PET can significantly improve classification results in a clinical setup on low-resource languages like German and can both accelerate and improve the development of accurate section classification models to e.g. support automatic medication extraction.

References

[1] P. Liu, W. Yuan, Z. Jiang, H. Hayashi, G. Neubig, J. Fu, W. Yuan, Z. Jiang, H. Hayashi, G. Neubig, and J. Fu, Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing, ACM Comput Surv. 55 (2023) 1–35. doi:10.1145/3560815.

[2] S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, and N.A. Smith, Don't Stop Pretraining: Adapt Language Models to Domains and Tasks, in: ACL, 2020: pp. 8342–8360. doi:10.18653/v1/2020.acl-main.740.

[3] T. Schick, and H. Schütze, Exploiting Cloze-Questions for Few-Shot Text Classification and Natural Language Inference, in: EACL 2021 , 2021: pp. 255–269. doi:10.18653/v1/2021.eacl-main.20.

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, (2018). http://arxiv.org/abs/1810.04805 (Mar. 7, 2019).

[5] T. Schick, and H. Schütze, True Few-Shot Learning with Prompts -- A Real-World Perspective, (2021). https://arxiv.org/abs/2111.13440v1 (Dec. 14, 2021).