Gregor Engels, Regina Hebig, Matthias Tichy (Hrsg.): SE 2023,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2023 93

On the validity of pre-trained transformers for natural language processing in the software engineering domain

Julian von der Mosel[1], Alexander Trautsch[2], Steffen Herbold[3]

Abstract: We summarize the article *On the validity of pre-trained transformers for natural language processing in the software engineering domain* [VTH22], which was published in the IEEE Transactions on Software Engineering in 2022.

Keywords: Defect Prediction; Costs; Return On Investment

1 Overview

The article "On the validity of pre-trained transformers for natural language processing in the software engineering domain" was published in the IEEE Transactions on Software Engineering in 2022. Transformers are the current state-of-the-art of natural language processing in many domains and are using traction within software engineering research as well. Such models are pre-trained on large amounts of data, usually from the general domain. However, we only have a limited understanding regarding the validity of transformers within the software engineering domain, i.e., how good such models are at understanding words and sentences within a software engineering context and how this improves the state-of-the-art. Within this article, we shed light on this complex, but crucial issue. We compare BERT transformer models trained with software engineering data with transformers based on general domain data in multiple dimensions: their vocabulary, their ability to understand which words are missing, and their performance in classification tasks.

2 Results

Our results show that for tasks that require understanding of the software engineering context, pre-training with software engineering data is valuable, while general domain models are sufficient for general language understanding, also within the software engineering domain.

[1] Georg-August-Universität Göttingen, Fakultät für Mathematik und Informatik, Goldschmidtstr. 7, 37077 Göttingen, Deutschland

[2] Universität Passau, Fakultät für Informatik und Mathematik, Dr.-Hans-Kapfinger-Straße 30, 94032 Passau, Deutschland alexander.trautsch@uni-passau.de

[3] Universität Passau, Fakultät für Informatik und Mathematik, Dr.-Hans-Kapfinger-Straße 30, 94032 Passau, Deutschland steffen.herbold@uni-passau.de

3 Data Availability

The seBERT model we pre-trained for this work, as well as all code to reproduce our experiments, is available online.[4]

Literatur

[VTH22] Von der Mosel, J.; Trautsch, A.; Herbold, S.: On the validity of pre-trained transformers for natural language processing in the software engineering domain. IEEE Transactions on Software Engineering/, S. 1–1, 2022.

[4] https://github.com/smartshark/seBERT