PRETRAINING AND VALIDATING T5 MODEL ON BULGARIAN DATA

Simeon Monov, Nikolay Pavlov, Detelinka Trifonova

Abstract. *There is insufficient resource availability for the Bulgarian language in the field of Natural Language Processing (NLP), since most of the data used in the majority of the research is in other languages. In this paper, we pretrain a T5 model on collected data from different Bulgarian text corpora and evaluate its performance against other multilingual models on different tasks.*

Key words: NLP, Natural Language Processing, Bulgarian, Transformer model, T5, Pretraining, Fine-tunning.

Acknowledgments

This work is supported by the project MUPD23-FMI-009 of the Scientific Fund of the Paisii Hilendarski University of Plovdiv, Bulgaria.

Simeon Monov[1], Nikolay Pavlov[2], Detelinka Trifonova[3]
[1, 2, 3] Paisii Hilendarski University of Plovdiv,
Faculty of Mathematics and Informatics,
236 Bulgaria Blvd., 4003 Plovdiv, Bulgaria
Corresponding author: smonov@uni-plovdiv.bg