Doctoral School of Information and Biomedical Technologies
Polish Academy of Sciences (TIB PAN)

Topic:
Natural Language Processing for Plagiarism Detection

Supervisor:
Przemyslaw Korytkowski, professor at National Information Processing Institute – National Research Institute (e-mail: pkorytkowski@opi.org.pl.pl)

Description:
Plagiarism is a challenge in science because it violates ethics and intellectual integrity. Improving the efficiency of plagiarism detection contributes to protecting intellectual property rights, strengthens the quality of research, and promotes integrity among researchers.

The scientific goal of the project is to develop plagiarism detection algorithms based on language models. Using language models using deep neural networks, specifically the Transformer architecture, is expected to improve the efficiency and accuracy of identifying text originality compared to algorithms that do not consider semantics.

The essence of the Transformer architecture is to process sequences using the so-called attention mechanism. This allows the interpretation of sequence elements in multiple contexts composed of other elements occurring in different, including distant, parts of the input sequence, making it possible to recognize relationships and patterns in long data sequences.

Recommended reading:

1. El-Rashidy, M.A., et al. (2022) Reliable plagiarism detection system based on deep learning approaches. *Neural Comput & Applic* 34, 18837–18858.
2. Kozłowski, M. (2021) Systemy informatyczne wspierające naukę i szkolnictwo wyższe. JSA: Jednolity System Antyplagiatowy, OPI-PIB, Warszawa.
3. Sanchez-Perez, M.A., Gelbukh, A., Sidorov, G. (2015). Adaptive Algorithm for Plagiarism Detection: The Best-Performing Approach at PAN 2014 Text Alignment Competition. CLEF 2015.
4. Sun, Xiaofei, et al. (2022) Sentence similarity based on contexts. *Transactions of the Association for Computational Linguistics* 10: 573-588.