Zgłoszenie tematu pracy dyplomowej **:: STUDIA II STOPNIA ::**

rok akademicki 2025/26

Promotor:	**dr hab. Jozef Kapusta, prof. UKEN**
Temat pracy magisterskiej (j. polski, j.angielski):	Data Augmentation Techniques for Natural Language Processing *Techniki powiększania danych w przetwarzaniu języka naturalnego*
Zakres pracy i oczekiwane rezultaty praktyczne:	Nowadays, the limited scope of text training data is a challenge for creating NLP models. One potential solution to this problem is to apply data augmentation techniques to expand the existing text datasets. Data Augmentation can be defined as any technique for increasing the diversity of training examples without explicitly collecting new data. The goal of data augmentation is to enhance the performance and robustness of machine learning models by exposing them to a wider range of variations and situations. There are a few popular techniques for text data augmentation: Back Translation, Synonym replacement, Paraphrasing, Random Insertion, Random Swap, Random Deletion, etc. The thesis aims to determine the extent to which data augmentation techniques enhance performance metrics in classification tasks. In the theoretical part: The theoretical part of the thesis will scrutinize a subset of commonly employed data augmentation techniques. It will also provide a concise summary of existing research in the field of text data augmentation and other techniques related to word embedding models. In the practical part: The practical part of the thesis will focus on creating a few training data sets based on selected popular techniques for text data augmentation. The created datasets will be used for training NLP models (student selects type of models, for example, fake news classifier, sentiment classifier, etc.) The student will evaluate the results from classifiers due to performance measures of the classifiers (accuracy, precision, recall, f1-score). The aim of the practical part is to determine the importance of the application of the coreference resolution method on text classification. The contribution of the practical part lies in determining the significance of these techniques for classification tasks and selecting the most suitable data augmentation technique for classification tasks.
Aspekt naukowy, problemowy, innowacyjny pracy:	definition and implementation of the data augmentation techniques, implementation methods natural language, implementation machine learning methods.
*Oprogramowanie, język programowania, środowisko systemowe:	

*Środowisko uruchomieniowe	
Dodatkowe wymagania i uwagi:	English language
*Literatura:	• Kapusta, J., et. al. (2024). Text Data Augmentation Techniques for Word Embeddings in Fake News Classification. IEEE ACCESS. 12, 31538-31550 • Feng, S. Y., et al. (2021). A Survey of Data Augmentation Approaches for NLP, CoRR, vol. abs/2105.03075, 2021, [Online]. Available: https://arxiv.org/abs/2105.03075 • Pellicer, L. F. A. O. , Ferreira, T. M., Costa, A. H. R. (2023) Data augmentation techniques in natural language processing. Appl Soft Comput, vol. 132, p. 109803, doi: 10.1016/j.asoc.2022.109803. • Wei, J., Zou, K. (2019). EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, 2019, pp. 6381–6387 • Marivate, V., Sefara, T. (2020). Improving Short Text Classification Through Global Augmentation Methods. pp. 385–399, 2020. doi: 10.1007/978-3-030-57321-8_21. • Salah, I., Jouini, K., Korbaa, O. (2023). On the use of text augmentation for stance and fake news detection. Journal of Information and Telecommunication, vol. 7, no. 3, pp. 359–375 • Haralabopoulos, G., et al. (2021). Text data augmentations: Permutation, antonyms and negation. Expert Syst Appl, vol. 177, p. 114769

*pola opcjonalne