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

SUBJECT: Spotting and Analysing Manipulative Discourse: Propaganda, Disinformation and Other Manipulative Techniques in Polish Data

SUPERVISOR: dr hab. Maciej Ogrodniczuk (Institute of Computer Science, Polish Academy of Sciences)

DESCRIPTION:

In the age of information overload, the ability to critically evaluate textual data is more crucial than ever. Propaganda, disinformation and other related forms of data manipulation appearing in news articles, social media posts, opinion pieces, or even official statements can distort truth, manipulate emotions, and influence public opinion.

One of the primary challenges in spotting forms of manipulation is the limited availability of annotated data. The annotation process requires expertise in linguistics, psychology and social science and is highly subjective. The procedure is also highly time-consuming as it requires the annotator to read large quantities of text. This challenge is especially pronounced for Polish, where high-quality propaganda or disinformation detection datasets remain scarce despite increasing manipulation in online discourse.

This research will create a comprehensive annotated dataset of manipulation techniques in Polish media and develop detection methods tailored to Polish language characteristics. The study will establish a data manipulation taxonomy adapted to Polish contexts, including techniques such as appeal to authority, loaded language, and flag-waving [1,2].

A key innovation will be implementing an LLM-assisted annotation framework where multiple LLMs provide initial annotations of manipulation techniques, with human experts reviewing difficult cases [4]. This human-in-the-loop approach creates an efficient workflow that continuously improves through feedback while reducing annotation time.

For automated detection, the research will evaluate methods from traditional ML approaches [3] to advanced transformer-based models for Polish. The study will explore parameter-efficient techniques like LoRA [5] to fine-tune LLMs for manipulation detection and investigate cross-lingual transfer learning to leverage existing English datasets [1,2].

The research will distinguish between rhetorical techniques used with propagandistic and disinformative intent versus those employed in legitimate informative contexts, addressing the critical challenge that similar linguistic devices can serve different communicative purposes depending on context and intent. Where possible, multiple modalities may be considered.

The research will deliver a Polish spotting manipulation benchmark for evaluating both specialized models and general-purpose LLMs (and possibly LMMs), including diverse examples across techniques and domains. Implementation will utilize modern deep learning frameworks like PyTorch and optimization tools such as Unsloth for efficient fine-tuning of large language models.

The candidate must hold a Master's degree in Data Science or a closely related field. Proficiency in both Polish and English (minimum C1 level) is required. Applicants should have practical experience with deep learning frameworks (e.g. PyTorch), along with demonstrated expertise in hosting and fine-tuning local large language models (LLMs), web scraping, annotation of data, and classical natural language processing (NLP) techniques.

Familiarity with multimodal approaches – especially involving vision or audio – is highly desirable. Prior work engaging with linguistic or sociological themes is strongly appreciated and will be considered a significant asset.

The candidate should contact the author of the proposal before a formal submission of documents.

BIBLIOGRAPHY:

[1] Da San Martino, G., et al. (2020). SemEval-2020 Task 11: Detection of Propaganda Techniques in News Articles.

[2] Maarouf, A., et al. (2023). HQP: A Human-Annotated Dataset for Detecting Online Propaganda.

[3] Oliinyk, V., et al. (2020). Propaganda Detection in Text Data Based on NLP and Machine Learning.

[4] Zheng, L., et al. (2023). Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena.

[5] Hu, E. J., et al. (2021). LoRA: Low-Rank Adaptation of Large Language Models.