学 位 論 文 内 容 の 要 旨

博士の専攻分野の名称　博士（情報科学）　氏名　SZWOCH Joanna Eleonora

学 位 論 文 題 名

Study on propaganda detection in Polish online news
（ポーランドのオンラインニュースにおけるプロパガンダ検出に関する研究）

In recent years, the quick and easy access to online news sources has significantly influenced public opinion, especially through biased reporting and the spread of misinformation. This trend has been particularly problematic in under-resourced languages, where access to reliable information is limited, and detecting propaganda is more challenging. This study is a step-by-step attempt to address these issues by analyzing sentiment with different approaches and leveraging Large Language Models (LLMs) for the detection of political and establishment biases, as well as propaganda in Polish online news articles.

First phase of this research focused on creation of a dataset called Polish Online News Corpus (PONC), which includes more than 200,000 articles from two major Polish news providers, TVP Info and TVN24, published between 2019 and 2021 and was later utilized in all experiments. The PONC dataset was constructed through a multi-stage process with the use of BeautifulSoup library to ensure quality and reproducibility. The classification experiment aimed to determine whether the source of a news article could be predicted based on textual features, indirectly assessing stylistic and content differences between TVP Info and TVN24. The following machine learning models were applied: Logistic Regression, Random Forest, Support Vector Machine (SVM), Naive Bayes Classifier. Each model was tested using two feature extraction techniques: Bag of Words (BoW) and TF-IDF. The best performing model was the Random Forest classifier with TF-IDF vectorization, achieving an accuracy of 87.45 and an F1-score of 88.29. These results confirmed that textual features alone could distinguish between the two news outlets, indicating underlying differences in language use and content framing.

The following task was an analysis of the emotional load in these articles using both lexicon-based and statistical approaches to sentiment analysis. The study aimed to determine whether supposedly objective online news contains emotive content and how different methodologies affect the detection of such biases. Initial experiments involved lexicon-based sentiment analysis using the Nencki Affective Word List (NAWL), a Polish affective lexicon assigning scores across six emotional categories: anger, disgust, fear, happiness, sadness, and neutrality, and zero-shot classification with the XLM-RoBERTa model fine-tuned on the XNLI dataset. The lexicon-based analysis revealed no statistically significant differences in emotional content between the two news outlets, suggesting a limited sensitivity to contextual sentiment. However, the XLM-RoBERTa model provided a more nuanced detection of emotive content, highlighting notable disparities between the sources. Specifically, TVP Info articles exhibited eight times more words classified as disgust compared to NAWL's baseline, while words associated with fear and happiness were reduced by eleven and five times, respectively. Topic-specific emotional patterns were also observed. For Abortion, articles from TVN24 contained three times more words classified as happy compared to TVP Info. For Church, TVP Info articles displayed twice as many words classified as sad and half as many words classified as happy as TVN24. These results demonstrate that while lexicon-based methods may underrepresent emotional variation, statistical models like XLM-RoBERTa can capture more subtle differences in emotive content, particularly in sensitive topics.

The study also highlighted the challenges and limitations associated with using LLMs for political and establishment bias detection, including the models' inability to fully understand non-literal language. Despite the challenges, the research demonstrated the potential of LLMs to automate the annotation process, thus saving time and resources compared to manual annotation. Subsequent analyses showed

that LLMs, such as gpt-3.5-turbo and gpt-4, could annotate political leanings and establishment biases, albeit with varying degrees of accuracy. Establishment bias proved more challenging to detect, with accuracy rates 10-12 percent lower compared to political bias detection tasks. This result indicates that political alignment (e.g., left- or right-wing tendencies) was easier for the models to identify than stance towards the ruling government or institutions. The models struggled with non-literal language and subtle political messaging, especially when bias was embedded in metaphors, irony, or indirect references. For instance, phrases with implicit sarcasm or historical references led to inconsistent classifications.

In addition to detecting biases, the study explored the potential of LLMs to identify propaganda techniques used in Polish news articles. The experiments on propaganda detection using Large Language Models (LLMs), specifically gpt-3.5-turbo and gpt-4, focused on analyzing articles from the Polish Online News Corpus (PONC) to identify rhetorical strategies aimed at influencing public opinion. The analysis was conducted on high-emotional charge articles, selected for their controversial topics where propaganda techniques are more likely to appear. Gpt-4 achieved a classification accuracy of 74 percent for Binary Propaganda Detection task and 69 percent for Technique Classification on PONC subset. Most commonly detected techniques included Appeal to Fear/Prejudice and Loaded Language, whereas Name-calling, Slogans, and Bandwagon were less frequently. More nuanced propaganda techniques such as repetition and whataboutism were rarely detected, suggesting a gap in the models' capacity to identify subtle manipulative patterns. Additionally, despite controlled prompts and temperature settings, variation in results was noted across trials. While gpt-4 showed promising results in detecting propaganda techniques like loaded language, its ability to identify subtle rhetorical techniques remained limited. These findings emphasize the importance of using LLMs as assistive tools for propaganda detection, but currently not able to fully replace human evaluators.

Finally, it is clarified that utilizing LLMs is a promising approach for detecting propaganda or political and establishment biases in online news, but their effectiveness varies depending on the specific task and model used. It is also noted that further research and improvements in these models are necessary to enhance their accuracy and reliability, particularly in under-resourced languages like Polish. This study contributes to the ongoing efforts to ensure media fairness and balance by providing insights into the capabilities and limitations of LLMs in bias detection.