北海道大学
HOKKAIDO UNIVERSITY

Language Media Laboratory
Research Group of Information Media Science and Technology
Division of Media and Network Technologies
Graduate School of Information Science and Technology
Hokkaido University

Joanna Eleonora Szwoch

Study on propaganda detection in Polish online news

A doctoral dissertation
supervised by
Rafal Rzepka, PhD

Sapporo, 2025

Thesis Abstract in English

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% 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% for Binary Propaganda Detection task and 69% 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.

Thesis Abstract in Japanese

近年、オンラインニュースソースへの迅速かつ容易なアクセスは、公的意見に大きな影響を与えており、特に偏向報道や誤情報の拡散を通じて顕著である。この傾向は、情報リソースの限られた言語圏において特に問題視されており、信頼性の高い情報へのアクセスが制限されているため、プロパガンダの検出がより困難となっている。本研究は、これらの問題に対処するため、異なる手法で感情分析を行い、大規模言語モデル（LLM）を活用してポーランド語のオンラインニュース記事における政治的および体制的バイアス、およびプロパガンダの検出を試みた段階的研究である。

本研究の第一段階では、「Polish Online News Corpus (PONC)」と呼ばれるデータセットの作成に焦点を当てた。このデータセットは、2019年から2021年にかけてのTVP InfoおよびTVN24の2つの主要ポーランドニュースプロバイダーから取得した20万件以上の記事で構成されており、その後のすべての実験で利用された。PONCデータセットは、BeautifulSoupライブラリを用いた多段階のデータ収集・前処理プロセスを経て作成され、その品質と再現性を確保した。

次に、ニュース記事の出典をテキスト特徴に基づいて分類する実験を実施し、TVP InfoとTVN24の文体および内容の違いを間接的に評価した。使用した機械学習モデルは、ロジスティック回帰、ランダムフォレスト、サポートベクターマシン（SVM）、ナイーブベイズ分類器の4つである。また、特徴抽出手法として、Bag of Words (BoW)とTF-IDFの2種類を適用した。最高の精度を達成したのは、TF-IDFを用いたランダムフォレスト分類器であり、正確率87.45%およびF1スコア88.29%を記録した。この結果は、テキスト特徴のみで両ニュースプロバイダーの違いを識別できることを示し、言語使用やコンテンツ構成に根本的な差異が存在することを示唆している。

次の分析では、これらの記事の感情的負荷（エモーショナルロード）の分析に焦点を当て、語彙ベースおよび統計的手法の両方を用いて感情分析を行った。本研究の目的は、表面的には客観的であると見なされるオンラインニュースに感情的な要素が含まれているかを評価し、異なる分析手法がバイアス検出にどのような影響を与えるかを明らかにすることであった。

初期実験では、ポーランド語の感情辞書であるNencki Affective Word List (NAWL)を使用し、感情を怒り、嫌悪、恐怖、幸福、悲しみ、中立の6つのカテゴリーに分類した語彙ベースの感情分析を行った。また、XLM-RoBERTaモデル（XNLIデータセットでファインチューニング済み）によるゼロショット分類を使用した。結果として、語彙ベースの分析では両ニュースプロバイダー間に統計的に有意な差は認められず、文脈的感情の検出能力が限定的である

ことが示された。一方で、XLM-RoBERTaモデルはより繊細な感情の差異を捉えることができ、TVP Infoの記事では、NAWL辞書のベースラインと比較して、「嫌悪」に分類される単語の出現回数が8倍多く、「恐怖」と「幸福」の出現回数はそれぞれ11倍、5倍少ないことが確認された。 さらに、特定のテーマに関して感情的パターンの違いが明確に示された。たとえば、中絶に関する記事では、TVN24の記事がTVP Infoより3倍多く「幸福」の表現を含んでいた。一方、教会に関する記事では、TVP Infoの記事がTVN24の記事より2倍多く「悲しみ」を表す単語を含み、「幸福」の語数は半分であった。これらの結果は、語彙ベース手法では感情の多様性を過小評価する一方、統計モデル（XLM-RoBERTa）はより微細な感情の差異を捉えられることを示している。

また、本研究は、LLMsを用いた政治的および体制的バイアスの検出に関する課題と制限も明らかにした。特に、比喩表現や皮肉的表現などの非直接的な言語の理解が不十分であることが課題として挙げられる。GPT-4およびGPT-3.5-turboは、政治的傾向や体制的バイアスの自動注釈付けにおいて時間とリソースの削減に貢献する可能性を示したものの、体制的バイアスの検出精度は、政治的傾向の検出よりも10-12%低下した。これにより、明示的な政治的偏向（例：左右の政治的立場）の方が、政府や制度への支持といった間接的なバイアスよりも識別が容易であることが示された。

最後に、プロパガンダ技術の検出に関しても、LLMの有用性を検証した。GPT-4は、2つのタスク（二値分類および技術分類）で評価され、74%の精度で二値分類、69%の精度で技術分類を達成した。しかし、繰り返しや論点すり替えなどの微妙なプロパガンダ手法の検出精度は低く、複雑な表現の解釈には限界があった。 総括として、本研究は、ポーランド語メディアの偏向検出において、LLMsが補助的ツールとして有望であることを示すと同時に、言語リソースの制限やモデルの解釈力の限界が存在することを強調した。さらなるデータ拡張とモデルのファインチューニングの必要性が示唆されており、本研究の成果は、偏向報道の分析とメディアの公正性の促進に貢献するものである。

Contents

1 Introduction **13**

1.1 Background and Motivation 13

1.2 Research Objectives 14

1.3 Contributions 14

1.4 Thesis Structure 15

2 Creation of Polish Online News Corpus for Political Polarization Studies **17**

2.1 Introduction 17

2.2 Related Works in Online News Corpus Creation 19

2.3 Corpus Creation 20

2.3.1 Data Collection 20

2.3.2 Dataset Cleansing 22

2.4 Methodology 22

2.4.1 Dataset Preparation 22

2.4.2 Feature Extraction 23

2.4.3 Algorithms 23

2.4.4 Model Quality Measurements 24

2.5 Results of Online News Classification Task 24

2.6 Discussion 24

2.7 Conclusion 25

3 Sentiment Analysis of Polish Online News Covering Controversial Topics **27**

3.1 Introduction 27

3.2 Related works in Sentiment Analysiss of Online News 28
3.3 Datasets 29
3.3.1 NAWL – The Nencki Affective Word List 29
3.3.2 Polish online news articles covering controversial topics 30
3.4 Methods 32
3.4.1 Lexicon-based classification with NAWL dictionary 32
3.4.2 Zero-shot classification on NAWL dictionary 32
3.4.3 Zero-shot classification on articles covering controversial topics 32
3.5 Results 32
3.5.1 Lexicon-based classification with NAWL dictionary on articles covering controversial topics 33
3.5.2 Zero-shot classification on NAWL dictionary 33
3.5.3 Zero-shot classification on articles covering controversial topics 33
3.6 Discussion 35
3.7 Conclusion 36

4 Determining Political Leaning of Polish News Articles with LLMs 39
4.1 Introduction 39
4.2 Related Works in Political Leaning and Media Bias 40
4.3 Dataset 42
4.3.1 Matching news pairs from PONC 42
4.4 Methods for Political Bias Detection 43
4.4.1 Utilized LLMs 44
4.4.2 Human Annotation 44
4.4.3 Evaluation Metrics 45
4.5 Results of LLM-based annotations 45
4.6 Discussion 48
4.6.1 Comparison of human and LLM-generated responses 48
4.6.2 Incorrect output form 49
4.6.3 Ambiguous responses 50
4.6.4 Limitations of LLM-based annotation 50

4.7 Conclusion 50

5 Experiments with Large Language Models in Propaganda Detection Task 53

5.1 Introduction 54

5.1.1 Motivation 54

5.1.2 Objective 55

5.1.3 Contributions 56

5.2 Literature Review 57

5.2.1 Short Introduction to the History of Propaganda 57

5.2.2 Media Bias, Fact Checking and Propaganda in Online News . 59

5.2.3 Propaganda, Media Bias and Fact Checking in Poland 62

5.3 Datasets 65

5.3.1 Propaganda Techniques Corpus 65

5.3.2 Polish Online News Corpus (PONC) 66

5.4 Methods 66

5.4.1 LLM on SemEval2020 Task 3 – English Data, CoT approach for TC 67

5.4.2 LLM on SemEval2020 – English Data 67

5.4.3 Propaganda Technique Detection on the PONC subset with the Use of an LLM 68

5.5 Results of Propaganda Detection Experiments 68

5.5.1 LLM on SemEval2020 Task 3 – English Data, experiments with Sprenkamp et al. approach 68

5.5.2 LLM on SemEval2020 – English Data, results of gpt-3.5-turbo-0125 and gpt-4-0125-preview for SI and TC 69

5.5.3 Propaganda Technique Detection in PONC Subset with the Use of LLM 72

5.6 Error Analysis 72

5.7 Discussion 73

5.8 Conclusion 74

6 Conclusion of the Thesis 75

6.1 Summary of Contributions 75

6.2 Future Work 76

Bibliography **92**

Appendix A **93**

A.1 Appendix A.1. Prompts 93

A.1.1 Appendix A.1.1. Prompt for Task 2 (LLM on SemEval2020 – English Data) and Task 3 (Propaganda Technique Detection in the PONC Subset with the Use of LLM, First Try) – In English 93

A.1.2 Appendix A.1.2. Prompt for Task 2 (LLM on SemEval2020 – English Data – Technique Classification) 94

A.1.3 Appendix A.1.3. Prompt for Task 3, Second Prompt – In Polish, Techniques in English 95

A.1.4 Appendix A.1.4. Prompts for Task 3, Third Prompt – In Polish, Techniques In English; Few-Shot – Examples of Propaganda Techniques in Polish 96

Acknowledgments **100**

List of Figures

2.1 Data cleaning process steps 21

3.1 Number of articles classified by RoBERTa-XNLI for ABORTION category 35

3.2 Number of articles classified by RoBERTa-XNLI for CHURCH category 35

4.1 Inter-coder agreement metric (Krippendorff's alpha) for left- and right-wing leaning per each controversial topic 45

4.2 Inter-coder agreement metric (Krippendorff's alpha) for anti- and pro-government stances per each controversial topic 46

4.3 Accuracy of LLM generated responses to human annotations for anti- and pro-government stances per each controversial topic 47

4.4 Accuracy of LLM-generated responses to human annotations for left- and right-wing leaning per each controversial topic 48

List of Tables

2.1 Number of articles within each category 21

2.2 Extracted data 21

2.3 Performance measurements of models 24

3.1 Keywords used for selecting articles concerning each controversial topic 31

3.2 Number of articles in each topic per news outlet 31

3.3 Percentage of words in each emotion category, based on NAWL dictionary and RoBERTa-XNLI classification 33

3.4 Percentage of articles with dominant emotion per news outlet – based on NAWL classification. Bolded values show significant differences in percentages between news outlets, depending on the topic 34

3.5 Percentage of articles with dominant emotion per news outlet – based on XLM-RoBERTa Large fine-tuned on XNLI classification. Bolded values show significant differences in percentages between news outlets, depending on the topic 34

3.6 Examples of words misclassified by RoBERTa-XNLI 36

4.1 Number of matching pairs of article descriptions per controversial topic after filtering 43

4.2 Accuracy scores of *text-davinci-002*, *text-davinci-003*, *gpt-35-turbo*, *gpt-35-turbo-instruct* and *gpt-4* models 47

4.3 Number of non-numeric answers generated by each model, depending on the temperature parameter 49

5.1 Main organizations dealing with media bias and fact checking 60

5.2 Comparison of news outlets in Poland based on Media Bias/Fact Check (MBFC) data 63

5.3 List of Polish organizations that investigate media bias and perform fact checking of news in Polish 64

5.4 PTC-SemEval2020 corpus distribution 65

5.5 Training dataset columns 66
5.6 Results from Sprenkamp et al. experiment [1] 69
5.7 F1 scores for each of the propaganda techniques 70
5.8 Task span identification (SI) – results 70
5.9 Task technique classification (TC) – F1 scores 71

Chapter 1

Introduction

1.1 Background and Motivation

With the increasing prevalence of digital media, free online news platforms have become one of the most important source of information for much of the population in European Union [2]. This shift has facilitated the rapid dissemination of news but has also introduced new challenges in managing the accuracy and integrity of information. One of the major concerns in this context is the spread of propaganda, which has a profound impact on public opinion, elections, and societal beliefs. In Poland, as in many other countries, the potential for propaganda to distort information and manipulate public sentiment is a growing issue, making the study of these dynamics particularly relevant.

This research focuses on Polish online news platforms and the content they produce, specifically news articles written in Polish. A key challenge in this domain is the detection of propaganda, which often takes a subtle and sophisticated form, embedded within otherwise factual reporting. This complexity extends to related tasks such as detecting political leanings and establishment bias, where identifying subtle biases poses a similarly intricate problem.

Recent advances in artificial intelligence (AI) and natural language processing (NLP) have provided new tools and methodologies to address these detection challenges. These innovations are critical in improving the precision and efficiency of detecting propaganda and related biases. Additionally, the increasing focus on combating misinformation, fake news, and propaganda within the research community underscores the importance and timeliness of this study.

The motivation for this research is rooted in the need to maintain public trust in the media. Ensuring that news consumers can differentiate between objective reporting and biased or propagandistic content is essential for upholding media credibility and informed public discourse. This study aims to contribute to these efforts by leveraging AI and NLP advancements to detect and analyze propaganda and bias in Polish online news.

1.2 Research Objectives

The primary objective of this research is to explore and evaluate various methods for detecting emotionally-charged content in Polish online news. This approach aims to identify politically biased articles and, ultimately, those containing propaganda. By developing efficient techniques for this purpose, the study seeks to provide valuable tools that can be utilized by journalists, researchers, and policymakers to detect and counteract propaganda in a more time-effective and cost-efficient way.

1.3 Contributions

This research has made several significant contributions to the study of media bias and propaganda detection in Polish online news. A key outcome is the creation of a specialized corpus of Polish news articles from two major TV broadcasters, which serves as a foundation for investigating media bias and propaganda. Additionally, the research conducted sentiment analysis on controversial topics in these news articles using two distinct methods—a lexicon-based approach and a statistical classification method. While word-level analysis revealed no major emotional differences between the news outlets, the zero-shot classification approach highlighted emotional divergences between them.

Another important contribution lies in the experimentation with Large Language Models (LLMs) for annotating political and establishment biases in the opening paragraphs of Polish news articles. These experiments demonstrated that while this task is challenging for both LLMs and human annotators, the models achieved higher agreement rates than the human annotators across all categories. The study also identified limitations of LLMs when dealing with under-resourced languages like Polish, illustrating that while LLMs may not fully replace human annotation, they offer a promising starting point in automating this process.

The research also explored LLMs in propaganda detection, revealing that while the outputs of generative models can be inconsistent, they offer potential in preliminary screening of news articles that may contain propaganda. This can reduce the time and labor required for more detailed human analysis.

It is also worth noting that the data collected in the form of the Polish Online News Corpus (PONC) are available for use upon request and has been shared with various researchers interested in topics such as media propaganda and the representation of the LGBT community in online media.

Overall, this work represents one of the first attempts to create a corpus of online news in Polish, an under-resourced language, and contributes to the field of quantitative sentiment analysis of Polish news. Unlike previous research, which has largely focused on short texts such as tweets or reviews, this study applies these methods to full-length online news articles. Additionally, it is among the pioneering efforts to automate the detection of political leanings and establishment stances in Polish news, employing LLMs to raise awareness of potential media bias, particularly in the context of elections. The application of LLMs in propaganda detection also

demonstrates their potential in streamlining the identification of propagandistic content, helping to reduce human effort in this labor-intensive task.

1.4 Thesis Structure

The structure of this dissertation is organized as follows – after the Introduction, Chapter 2 describes the Creation of the Polish Online News Corpus (PONC), including data collection and cleaning from two major news outlets, TVP Info and TVN24, providing the foundation for the analyses. Chapter 3 focuses on sentiment analysis of controversial topics, comparing lexicon-based and statistical methods to assess emotive content in the news articles. Chapter 4 investigates political bias detection using Large Language Models (LLMs), examining the challenges and potential of LLMs in identifying political leanings and establishment stances. Chapter 5 explores the use of LLMs for propaganda detection, discussing the complexities of the task and potential applications for initial screening of biased content.

Finally, Chapter 6 is the conclusion which summarizes the findings and outlines future research directions, including expanding the dataset and refining detection models for greater accuracy in bias and propaganda analysis.

Chapter 2

Creation of Polish Online News Corpus for Political Polarization Studies

In this chapter I describe a Polish news corpus as an attempt to create a filtered, organized and representative set of texts coming from contemporary online press articles from two major Polish TV news providers: commercial TVN24 and state-owned TVP Info. The process consists of web scraping, data cleansing and formatting. A random sample was selected from prepared data to perform a classification task. The random forest achieved the best prediction results out of all considered models. I believe that this dataset is a valuable contribution to existing Polish language corpora as online news are considered to be formal and relatively mistake-free, therefore, a reliable source of correct written language, unlike other online platforms such as blogs or social media. Furthermore, such corpus has not been created before. In the future I would like to expand this dataset with articles coming from other online news providers, repeat the classification task on a bigger scale, utilizing other algorithms. My data analysis outcomes might be a relevant basis to improve research on a political polarization and propaganda techniques in media.

This chapter was created with the use of my previous work – a conference paper presented at the LREC conference in 2022 [3].

2.1 Introduction

Nowadays, a piece of information is the most valuable asset. There is a growing problem of distinguishing valuable information from noise and fake news. In Poland it is significantly noticeable during the Russia-Ukraine conflict. One could see inaccuracies regarding the influx of Ukrainian refugees into the European Union, the course of the fighting or the actions of Western countries toward Russia [4].

In the 21st century we can observe a sociological phenomenon called the filter bubble [5, 6]. Sometimes the same topic is presented in extremely different ways, depending on the news provider. Users tend to visit news sources matching

their political attitudes and spend more time on biased content [7]. Manipulation is performed with a variety of language techniques and each language requires a tailored approach to detect them.

Polish language cannot be called a typical lesser-resourced language, but compared to others, such as English, German or Russian, it has a significantly smaller base of available corpora. Additionally, vast majority of text resources are paid and not disclosed to the public. The National Corpus of Polish[1] is one of few initiatives which provides a reference corpus containing roughly fifteen hundred million words for free. Sources include literature, newspapers, specialist magazines, transcripts of conversations and Internet texts. However, none of them are online news websites. Moreover, the project was finished in 2012 and has not been updated since. This means that all resources come from year 2011 or older, sometimes even reaching back to 1920s.

Modern languages are very flexible and their use changes constantly. That is why I think that the aforementioned corpora should be expanded by newer sources. Websites providing online news in Polish should therefore be perfect for that as they contain contemporary version of the language in everyday use. Another premise is that they are written in a correct manner, prepared by professional journalists, unlike other online sources such as blogs or social media.

My dataset was created with the use of two online news websites, run by Polish major TV news providers, namely state-owned TVP Info[2] and commercial TVN24[3]. These are examples of the most watched TV news programs in Poland[4]. Aforementioned websites were scraped and news from years 2019-2021 from different categories were stored in CSV format.

This chapter focuses on two tasks – data collection aspect and news outlet classification, which can be treated as a baseline for further experiments. It explains step by step how my corpus was created – I describe web scraping method which I find the most convenient, readable and therefore reproducible. Having created the dataset with scraped articles, I explain how to perform data cleansing to prepare the corpus for modeling. After lemmatization and tokenization, cleaned text is vectorized and classification task is performed with the use of several machine learning models. Unlike most of experiments concerning news articles processing, I trained models to predict which news provider wrote certain article, instead of predicting the article category.

People often state that media which do not share their political views or opinions are biased and on the contrary, the ones that they follow are not. In the era of news flowing constantly from different sources it would be beneficial to be able to measure media bias objectively. It is claimed to be possible via data-driven analyses whose results should be free of subjectivity [8]. This topic may be of particular interest in connection with the on-going war in Ukraine as reports about it are

[1] http://nkjp.pl/index.php

[2] https://www.tvp.info/

[3] https://tvn24.pl/

[4] https://www.wirtualnemedia.pl/artykul/fakty-lider-ogladalnosci-luty-programy-informacyjne

presented differently, sometimes to an extreme extend, depending on the source.

I think that this corpus is a good starting point for further analyses of contemporary Polish language. Although, the professional journalists should focus on conveying a clear message, holding subjective information only, I believe that news outlets could be examined whether they show any political bias and what kind of propaganda techniques are being used, if any. However, I want to direct attention to that matter in future works, with the use of this data set, possibly extended with more articles from other online news sources[5].

2.2 Related Works in Online News Corpus Creation

Before online news, there were first attempts to create corpora based on other media. For example, the News-on-Demand application, developed as part of the Informedia™ Digital Video Library project, demonstrates the use of speech recognition technology for creating transcripts from video, aligning them with closed-captioned text, segmenting audio into paragraphs, and enabling spoken query interfaces. While speech recognition accuracy varies significantly depending on data quality, informal tests indicate that even moderate accuracy can achieve reasonable recall and precision in information retrieval tasks. This highlights the potential of speech recognition to make spoken content in video and audio media accessible for indexing, search, and retrieval [9].

Another notable example is the Boston University Radio Speech Corpus [10]. This corpus comprises over seven hours of professionally read radio news data from FM news announcers associated with a public radio station called WBUR. It was primarily developed to support research in text-to-speech synthesis. The corpus includes both the speech recordings and accompanying annotations, making it suitable for various speech and language research applications.

In the past few years there was plenty of studies which tried to tackle the problem of collecting online text resources in an efficient manner. Researchers from MIT managed to collect over three million articles from 2019 and 2020 from about 100 online media outlets with the use of the open-source Newspaper3k[6] software [8]. Another method to achieve this goal is to perform web scraping. This process consists of a few steps: desired websites identification, URLs collection, HTML retrieval, text parsing and finally, persistence (long-lasting storage and retention for long-term use) [11]. One of the most popular and widely used Python libraries for imitating human alike behavior of entering the URL and retrieving necessary data is Beautiful Soup which automatically obtains data from HTML and XML files [12].

Slovak Categorized News Corpus was created in a similar way. It contains words, automatic morphological as well as named entity annotations. It consists of almost five thousand articles, with over one hundred thousand sentences and million and a half of tokens [13].

[5]Upon request, my dataset can be provided for research purposes.

[6]`https://newspaper.readthedocs.io/en/latest/`

For text classification task itself, firstly, input data needs to be converted to a computer-readable form. Bag-of-Words and TF-IDF word vectors are two possibilities to handle this task [14], [15]. Then, logistic regression can be trained on such data. Except for logit models, other methods include neural networks, support vector machines, random forests, or naive Bayes classifiers as it was experimented on Vox Media dataset or smaller samples of online news from ABC news, Fox news, New York Times, and BBC news [16, 17].

2.3 Corpus Creation

I decided to collect online articles from two Polish major TV news providers, TVP Info and TVN24. Firstly, I tried using Newspaper3k library for this task. Although library documentation states that it handles Polish language, it failed to parse chosen websites correctly and therefore I had to give up on this method. However, one feature that worked properly was listing the subcategories of the main website. In the end I decided to prepare a tailored solution for scraping these resources with Beautiful Soup library in Python, as it was utilized in other works [12], [18] [19].

2.3.1 Data Collection

Aforementioned websites were scraped and news from different categories were collected between 1st January 2019 and 31st December 2021.

Data collection and cleansing process consisted of the following steps, as presente in Figure 2.1:

- Identifying main pages of news providers
- Listing all contexts (website subpages)
- Web crawling to gather all URLs from designated time span
- Parsing websites to retrieve data from HTML files; first text cleaning with the use of regular expressions to filter markups from retrieved HTMLs
- Saving data to CSV file

The aforementioned process resulted in the collection of articles from two sources in the following amounts and categories presented in Table 2.1.

Data is not evenly distributed and some categories existed only in one website, but not in the other one.

Datasets have the following structure as presented in Table 2.2.

TVP Info dataset has two additional columns when compared to TVN24, namely modification time and hash tags. I decided to include them as they might

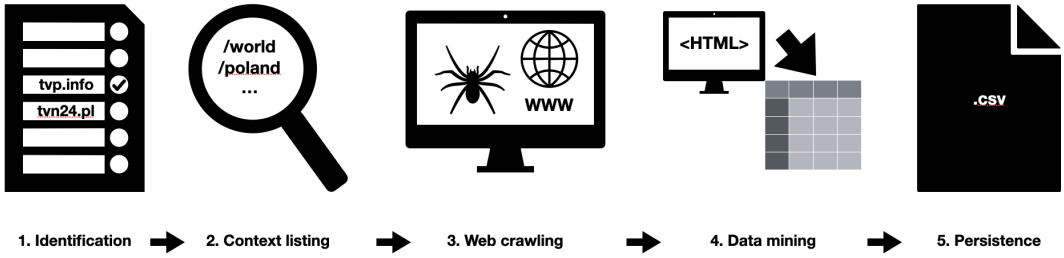

Figure 2.1: Data cleaning process steps

Category	**TVP Info**	**TVN24**
Poland	36,223	37,511
World	19,982	28,318
Society	11,484	-
Business	4,488	12,629
Warsaw	-	12,532
Sport	3,628	26,289
Science	2,297	108
Culture	1,698	-
Miscellaneous	1,399	-
Weather	495	10,558
Politics	-	513
Entertainment	-	69
Total	81,694	128,527

Table 2.1: Number of articles within each category

Variable	**TVP Info**	**TVN24**
`url`	✓	✓
`magazine_title`	✓	✓
`website_category`	✓	✓
`title`	✓	✓
`description`	✓	✓
`authors`	✓	✓
`article`	✓	✓
`pub_time`	✓	✓
`mod_time`	✓	-
`hash_tags`	✓	-

Table 2.2: Extracted data

be interesting to be examined in the future for other purposes. TVN24 dataset did not have any information regarding the modification time of the article and tags appeared only recently in their articles, therefore these columns were not added to this dataset.

2.3.2 Dataset Cleansing

Dataset cleaning process consisted of 5 steps in the following order:

- **Duplicates removal** - some articles were repeated during the web crawl.
- **Removing repetitions from retrieved text** - in some cases, the opening paragraph of an article, appearing in bold, right after the title, which I will refer to as *description* of the article, appeared also in the article text – such duplicate pieces of text were not desired.
- **Filtering ad words** - most of the articles consisted of phrases which encouraged the reader to watch a related video or read an article whose topic is connected.
- **Deleting special characters such as punctuation, double spaces or tabulation as well as numbers** - regular expressions were used to eliminate all unnecessary elements from this group.
- **Stop words removal** - I used *stop-words* Python library[7] and a set listed by user *bieli* on GitHub[8] to create an extended collection of Polish stop words, as part of them were not included in *SpaCy* library.

As a result, I obtained fairly clean 197,606 articles from both TVP and TVN, consisting of 4,042,638 sentences and 44,528,641 tokens.

2.4 Methodology

In this section all methods which were used to perform text classification task are explained.

2.4.1 Dataset Preparation

Firstly, cleaned dataset had to undergo a few more processes before it is eventually used as an input to train classification models, namely:

- **Lemmatization** - *SpacyPL* handled this for Polish language, using a lemma dictionary imported from Morfeusz morphological analyzer[9].
- **Tokenization** - *nltk* Python library was used for this task.

As the website category groups were unevenly distributed, I decided to take 2,000 randomly selected articles both from TVN24 and TVP Info from the following

[7] `https://github.com/Alir3z4/python-stop-words`
[8] `https://github.com/bieli/stopwords`
[9] `http://morfeusz.sgjp.pl/`

four most numerous categories: World, Poland, Business and Sports. This selection aimed to ensure a representative dataset that balances coverage of diverse yet popular topics from both news outlets, allowing for meaningful analysis of their content. Eventually, I trained the models on a reduced subset of 16,000 online news. Dataset was then divided into two sets - training set which consists of 12,000 records and test set that has remaining 4,000 tuples. This approach was intended to facilitate a robust exploration of potential patterns and biases across the selected categories.

2.4.2 Feature Extraction

Although classifying text is not an easy task to be performed by computers, it can be done if input data is converted into a numerical representation.

- **Bag of Words (BoW)** - this method is considered to be simpler both computationally and conceptually than other methods. It is assumed that it could record higher performance scores on common used benchmarks of text [14].
- **TF-IDF** - word vectors are also able to help with converting characters into a format that is processable by a computer [15].

Both methods were implemented with the use of *Scikit-learn* library.

2.4.3 Algorithms

Based on the suggestions from previous works, I trained following four machine learning models:

- **Logistic regression** - commonly used for NLP classification tasks also for fake news detection [20]
- **Random forest** - recommended for news articles classification task, along with N-gram textual features [21]
- **Support Vector Machine** - better results with high dimension data like large volumes of text, comparing with Neural Networks or Naive Bayes methods [22]
- **Naive Bayes** - high accuracy in online news category classification task [23]

Scikit-learn Python library allows us to use already built-in functions which perform all the calculations of the aforementioned supervised learning methods[10].

[10] `https://scikit-learn.org/stable/`

2.4.4 Model Quality Measurements

In order to check the performance of trained models, I calculated two of the standard metrics, namely **Accuracy** and **F1 Score** [24]. I also used *Scikit-learn* library for retrieval of these statistics.

2.5 Results of Online News Classification Task

Having trained four different models with two possible types of feature extraction, they yielded the following results shown in Table 2.3.

	Feature Representation			
	BoW		TF-IDF	
Model	Accuracy	F-score	Accuracy	F-score
SVM	0.8668	0.8713	0.8598	0.8629
Random forest	0.8703	0.8798	**0.8745**	**0.8829**
Logistic Regression	0.8700	0.8729	0.7228	0.7790
Naive Bayes	0.8048	0.8013	0.7640	0.7627

Table 2.3: Performance measurements of models

The best results were achieved by random forest with TF-IDF vectorization method. Accuracy was 87.45% and F-score was equal to 88.29%. The worst results were obtained by logistic regression which also had features vectorized with TF-IDF method. Accuracy and F-score were equal 72.28% and 77.90% accordingly and these were the lowest scores among other models.

2.6 Discussion

TVN24 and TVP Info are two of the most watched news providers in Poland. Nonetheless, the way the information is conveyed by both is considered to be often much different from each other [25], [26], [27]. I believe it would be interesting to supplement social studies with machine learning techniques for measuring political bias. Another step to be considered is error analysis to check which words were most often confused when predicting the news outlet.

This showcase study illustrated that accuracy scores for classification task in Polish were below 90%. This is lower than results reported for other languages, such as English, which have achieved higher accuracy in tasks like fake news detection [28] or news categorization [29] as well as news topic analysis [30]. Further investigation is needed to explore potential improvements in performance for Polish language tasks.

2.7 Conclusion

In this chapter, I introduced a collection of online news from two TV news broadcasters as the first step of building a full-fledged corpus containing modern, journalist-created and hopefully correct sample of Polish language. I showed that this corpus can be used for such a task like news outlet recognition based on news article contents. Out of four trained models, random forest with TF-IDF vectorization achieved the highest accuracy and F-score metrics.

American news outlets have already been analyzed in terms of political polarization problem [8]. This dataset is unique and I hope my contribution will become a base for other tasks, such as news category prediction or also sentiment analysis in Polish language. With this corpus of distinctly different sources I also look forward to encouraging researchers to use it to study polarization of Polish society and to develop methods (e.g. for automatic summarization) freeing information of political biases and possible manipulation.

Building on the foundation of the Polish Online News Corpus and the methodologies for analyzing its content, the subsequent chapter delves into sentiment analysis of news articles covering controversial topics. By examining emotional content through lexicon-based and statistical approaches, this analysis seeks to uncover patterns of emotive language use in online news. The insights gained from these experiments provide a deeper understanding of how media content can subtly influence public perception, setting the stage for more advanced tasks such as political bias and propaganda detection.

Chapter 3

Sentiment Analysis of Polish Online News Covering Controversial Topics

In this paper I present results of a sentiment analysis performed on a part of the Polish Online News Corpus which consists of articles covering controversial topics in Poland. I perform both lexicon-based and statistical experiments and compare the obtained results for this task. The main goal of this comparative chapter is to discover whether supposedly objective online news contain any emotive content and how different the results are, depending on the approach. This study found no meaningful differences in emotional load at the word level, but the multilingual language model trained with the Natural Language Inference objective has partially confirmed its existence. I believe this work can be used for further research on media polarization in Polish language.

This chapter was created with the use of my previous work – a conference paper presented at the LTC conference in 2023 [31].

3.1 Introduction

The Internet is a place full of content generated by people. To a large extent, readers' views and decisions that follow them are shaped by the content they encounter online [32]. Moreover, in present days, traditional media such as newspapers, magazines, radio or even TV are slowly fading away, in favour of news outlets available on the Internet. News providers, according to the highest journalistic standards, are supposed to deliver pure information without any bias. However, there are examples where this condition is not met [33]. In this work I want to investigate if two popular online news providers in Poland are following the aforementioned standards.

News outlets which are considered biased, tend to present information in favour of a certain side, whether it is a political party or a social movement. Their promotion of a specific perspective can lead to a spread of confusing or inaccurate information, causing mistrust and divisions within a society [34].

Sentiment analysis and emotion detection are ways to determine the emotions hidden in text and to gain insights on how people feel about certain topics [35]. It can help to detect fake and biased news as well as misinformation. In general, the problem of including opinions and emotions in contemporary media spotlights the importance of critical thinking and literacy in current digital age [36].

The goal of sentiment analysis is to identify subjective information within a text, such as opinions or emotions. There are two approaches to this task, namely a rule-based and a statistical one. The former one relies on manually-crafted rules, including dictionaries and corpora to classify text into categories. Three classes, that is positive, negative and neutral are examples of such categorization, but they can also consist of various types of emotions. On the other hand, the latter approach can be divided into three methods – supervised, unsupervised and self-supervised. Both approaches have strong points and limitations depending on the type of input data and the use case.

Polish language is rather neglected when it comes to the number of studies on sentiment analysis. In comparison with other languages, there is a relatively small number of available resources. It is an obstacle for researchers wishing to delve into this topic. Despite this challenge, in this paper I experiment with two different approaches to opinion mining in Polish language. I also discuss the pros and cons of the utilized techniques and potential future directions in this field.

3.2 Related works in Sentiment Analysiss of Online News

Opinion mining became a fairly popular branch of natural language studies, which resulted in a bountiful of works covering the topic [37, 38]. Majority of studies relate to short inputs such as tweets or reviews. In this chapter I will focus on examining emotions in news, which is a fairly uncommon target.

One of the examples of lexicon-based approaches is a study which performed sentiment analysis on English BBC news [39]. Bing sentiment dictionary allowed classification of positive, negative or neutral content. Such approach showed that in different news categories, information tends to be polarized rather than impartial. Another popular lexicon-based technique includes VADER[1]. It is prepared for English vocabulary, but a common practice is to translate this source to any desired language, as in the study of sentiment analysis of health-related online news in Brazilian Portuguese [40]. Other examples of dictionaries include WordStat Sentiment Dictionary [41], SentiWordNet [42], AFINN [43], Loughran & McDonald Financial Sentiment Dictionary (L&M) [44], Lexicoder Sentiment Dictionary [45, 46]. Out of these, only SentiWordNet includes a dedicated Polish version – plWordNet [47].

Sentiment analysis conducted on online financial texts indicates that RoBERTa-based model outperformed results obtained by SVM, Logistic Regression or Naive Bayes [48]. Moreover, opinion mining of Italian reviews indicated that machine

[1] `https://github.com/rafjaa/LeIA`

and deep learning approaches such as BERT perform better in general, however, lexicon-based methods are recommended for small datasets with limited resources [49].

Polish language studies include sentiment analysis of conversation transcripts with a social robot [50]. For this type of long text, lexicon-based approach was applied. Two different dictionaries – English NRC [51] and Polish plWordNet – were used for this task. Both of them allow to associate words with 8 different types of emotions. However, due to the limited number of analyzed conversations, the variation between results was not observed. The cross-sectional analysis of the selected long conversations showed that both plNetWord and NRC lexicons produced similar results, indicating consistent emotion classification across the two tools.

On the other hand, multilingual BERT model yielded good results in sentiment analysis of news in another Slavic language, namely Slovenian [52]. Moreover, this zero-shot learning approach was also used on Croatian language for the same task and it outperformed the results of other classifying models.

However, no similar experiments were conducted on Polish online news.

3.3 Datasets

In this section I describe the datasets used for sentiment analysis of news covering controversial topics in Polish language.

3.3.1 NAWL – The Nencki Affective Word List

The Nencki Affective Word List (NAWL) is a database consisting of 2,902 Polish words such as nouns, verbs, and adjectives [53]. All of them are rated for emotional valence, arousal, and imageability[2]. Furthermore, measures of frequency, grammatical class, and number of letters are also included in this set. NAWL is a Polish version of the Berlin Affective Word List - Reloaded [56], which was created for the same purpose, but in German language. This dictionary describes each word with a point value in terms of 5 categories: anger, disgust, fear, happiness and sadness. Additionally, neutral and unknown classes are included for words that do not have a dominant category. The higher the value, the stronger the word expresses the particular emotion.

NAWL is used in the non-statistical approach to sentiment analysis of words which appeared in examined online news. I chose this lexicon due to its accessibility and freeware licence.

[2]According to the authors, imageability reflects how easily words evoke sensory experiences and strongly correlates with concreteness [54, 55].

3.3.2 Polish online news articles covering controversial topics

In this chapter, I used the Polish Online News Corpus to retrieve articles on contemporary controversial topics in Poland [3], focusing on their potential for emotive content and biases. Controversial topics, often eliciting strong emotional reactions and diverse perspectives [57], were selected to analyze how media outlets frame information and influence public opinion. The selection included arbitrarily chosen issues prevalent in Polish media, with topics like EU and ABORTION inspired by prior research on political leaning detection [58].

This study focuses on controversial topics in Polish online news to examine their potential for emotive content and biases. Controversial topics often elicit strong emotional reactions and differing perspectives, making them ideal for analyzing how media outlets frame and present information. By selecting these topics, the study aims to uncover patterns of sentiment and bias that might influence public opinion. Articles were extracted from the Polish Online News Corpus, filtered for relevance to the selected topics, and analyzed to provide insights into media reporting on polarizing issues.

The following subjects were selected as controversial ones in Polish society:

- ABORTION: in Poland, it is only legal when pregnant woman's life is in danger or in case the pregnancy is a result of rape or incest. In 2021, it became illegal in case the fetus is severely malformed[3].
- CHURCH: the Roman Catholic Church is the dominant religion in Poland, playing a significant role in the country's culture, politics, and public life. Its influence often sparks debate, particularly in discussions about secularism, societal values, and the separation of church and state.
- CONSTITUTION: recent constitution changes resulted in financial penalties from the European Union (EU)[4].
- PRESIDENTIAL ELECTIONS: last ones took place in 2020[5].
- EU: Poland is a part of the EU. Citizens' attitude towards Poland's membership in the EU varies in time[6].
- INDEPENDENCE MARCH: it is an annual event that takes place in Warsaw on November 11th to commemorate regaining independence by Poland[7]. While intended as a patriotic celebration, the event has often sparked controversy due to the participation of far-right groups and occasional clashes with counter-protesters. Critics argue that the march sometimes promotes divisive ideologies, whereas supporters view it as a rightful expression of national pride, reflecting the polarized opinions surrounding this event [8].

[3]Dziennik Ustaw (Journal of Laws) Dz.U.2022.1575

[4]Order ECLI:EU:C:2021:878, Court of Justice of EU

[5]https://wybory.gov.pl/index

[6]https://www.cbos.pl/PL/trendy/trendy.php

[7]https://marszniepodleglosci.pl/

[8]https://www.reuters.com/default/far-right-independence-day-march-warsaw-draws-thousands-2024-11-12/?utm_source=chatgpt.com

- LGBT PARADE: the Equality Parade is an annual event that takes place in several cities in Poland to celebrate the LGBT community and promote tolerance and equality[9].

- REFUGEES: the Polish-Belarusian border has become a site of increased activity in recent years, as refugees and migrants try to enter the EU via Poland[10].

To select news articles on controversial topics, I filtered them using keywords that are listed in Table 3.1. I attempted to choose a non-biased wording. The keyword "strike" for abortion was chosen due to its connection to the Polish Women's Strike movement. To avoid repetitions of articles between categories, keywords from one subject were banned in others. Moreover, I limited website category to "Poland" to focus on matters within the country.

Topic	Keywords
ABORTION	abortion, strike
CHURCH	church, catholic, priest
CONSTITUTION	constitution, court, law, judge, tribunal
PRESIDENTIAL ELECTIONS	elections
EU	union, european, cjeu [11]
INDEPENDENCE MARCH	independence, march, day
LGBT PARADE	lgbt, parade
REFUGEES	refugee, border, migrant

Table 3.1: Keywords used for selecting articles concerning each controversial topic

Table 3.2 shows the distribution of articles between topics. Analyses showed that the number of words in articles is resembling normal distribution. I decided to remove the outliers that were outside the three standard deviation distances from the mean length.

Topic	Number of articles	
	TVN	TVP
ABORTION	62	143
CHURCH	166	47
CONSTITUTION	147	92
ELECTIONS	234	205
EU	260	155
INDEPENDENCE MARCH	46	86
LGBT PARADE	29	34
REFUGEES	124	147

Table 3.2: Number of articles in each topic per news outlet

[9] http://www.paradarownosci.eu/

[10] https://www.strazgraniczna.pl/pl/aktualnosci/9689,Nielegalne-przekroczenia-granicy-z-Bialorusia-w-2021-r.html

[11] CJEU – The Court of Justice of the European Union

3.4 Methods

This section describes all methods of performed analyses.

3.4.1 Lexicon-based classification with NAWL dictionary

In the first experiment, sentiment analysis is conducted by matching words from selected articles, lemmatized with *SpacyPL*[12] library, with pre-defined NAWL sentiment dictionary consisting of 6 classes. All words within an article are a subject of the analysis. The sentiment scores are aggregated to determine the sentiment of each topic per news provider. Two-tailed t-test is performed to compare the results between news providers depending on the topic and check if the differences are statistically significant.

3.4.2 Zero-shot classification on NAWL dictionary

In the second step I perform zero-shot classification of NAWL dictionary words with XLM-RoBERTa Large model[13], fine-tuned on XNLI multilingual dataset[14]. I decided to use zero-shot approach as it yielded promising results for Slavic languages [52].

3.4.3 Zero-shot classification on articles covering controversial topics

Last experiment is a zero-shot classification on full article texts within each controversial topic. Six labels were provided in English, the same as in NAWL dictionary. Articles have been truncated to BERT-based models' maximum capacity of 512 tokens. Obtained results were used to perform two-tailed t-test for any statistically significant differences between groups.

3.5 Results

In this section I present obtained results of performed experiments.

[12] `https://spacy.io/models/pl`
[13] `https://huggingface.co/xlm-roberta-large`
[14] `https://huggingface.co/joeddav/xlm-roberta-large-xnli`

3.5.1 Lexicon-based classification with NAWL dictionary on articles covering controversial topics

Results of lexicon-based sentiment analysis on the article content are shown in the Table 3.4.

A t-test conducted for all emotions revealed that, at the 5% significance level, the null hypothesis of no difference between means across categories could not be rejected. This indicates that the observed differences between groups were not statistically significant.

3.5.2 Zero-shot classification on NAWL dictionary

Results of RoBERTa classification with NAWL dictionary as predicted labels are shown in Table 3.3.

Emotion	NAWL	RoBERTa-XNLI
Anger	13.26	7.58
Disgust	6.50	39.51
Fear	22.06	2.841
Happiness	19.89	4.33
Neutral	29.63	42.35
Sadness	8.66	3.38

Table 3.3: Percentage of words in each emotion category, based on NAWL dictionary and RoBERTa-XNLI classification

Percentages showed in the Table 3.3 indicate that RoBERTa based model differently classifies most of the groups. It is worth mentioning that there are more words classified by RoBERTa as "Neutral" than in the original dictionary. Moreover, there is also a noticeable difference in a number of words associated with emotions "Disgust", "Fear" and "Happiness". The first one is almost 8 times bigger than in the original dictionary. The other two are respectively almost eleven and five times smaller than the lexicon.

3.5.3 Zero-shot classification on articles covering controversial topics

Results of lexicon-based sentiment analysis on article content are shown in the Table 3.5.

The t-test performed for all emotions indicated that, at the 5% significance level, the null hypothesis of no difference between means was rejected only for Abortion and Church categories. For better visibility of this outcome I display the counts of classified examples with the regard to news outlet for Abortion category on Figures 3.1. and 3.2.

Emotion	Abortion		Church		Constitution		Elections		EU		Independence March		LGBT		Refugees	
	TVN	TVP	TVN	TVP	TVN	TVP	TVN	TVP	TVN	TVP	TVN	TVP	TVN	TVP	TVN	TVP
Anger	13.88	10.40	5.81	5.84	4.68	4.76	4.57	5.12	4.38	4.30	5.65	5.66	6.22	5.87	5.34	5.32
Disgust	1.37	0.63	1.39	1.47	0.15	0.11	0.17	0.51	0.09	0.09	0.23	0.21	2.58	0.26	0.37	1.06
Fear	34.22	34.74	29.55	29.29	64.39	64.97	49.56	45.40	64.38	61.98	27.30	31.08	30.90	25.77	20.21	23.71
Happiness	13.13	15.43	16.67	17.43	6.53	6.06	16.47	16.98	5.45	6.09	31.68	29.22	19.31	22.70	14.46	15.31
Neutral	33.92	36.11	40.54	39.97	23.12	22.93	26.63	29.04	24.06	25.90	29.15	29.48	35.19	41.07	56.17	51.14
Sadness	3.49	2.69	6.04	6.00	1.14	1.17	2.59	2.94	1.65	1.64	**5.99**	4.36	5.79	4.34	3.45	3.46

Table 3.4: Percentage of articles with dominant emotion per news outlet – based on NAWL classification. Bolded values show significant differences in percentages between news outlets, depending on the topic

Emotion	Abortion		Church		Constitution		Elections		EU		Independence March		LGBT		Refugees	
	TVN	TVP	TVN	TVP	TVN	TVP	TVN	TVP	TVN	TVP	TVN	TVP	TVN	TVP	TVN	TVP
Anger	11.29	18.88	15.06	19.15	11.49	7.69	16.87	18.29	11.15	14.84	8.70	12.79	**3.45**	**20.59**	4.84	19.05
Disgust	11.29	16.08	10.84	10.64	7.43	12.09	7.23	10.98	8.46	14.19	6.52	15.12	**10.34**	**26.47**	11.29	14.29
Fear	30.65	30.77	29.52	25.53	**29.05**	**42.86**	34.94	35.37	27.69	29.68	36.96	38.37	**34.48**	**17.65**	33.87	36.05
Happiness	**16.13**	**5.59**	**17.47**	**8.51**	**21.62**	**13.19**	9.64	7.32	**26.54**	**11.61**	19.57	16.28	17.24	5.88	19.35	6.12
Neutral	17.74	14.69	15.66	14.89	15.54	13.19	19.28	17.07	16.92	14.19	**19.57**	**8.14**	20.69	26.47	18.55	11.56
Sadness	12.90	13.99	**11.45**	**21.28**	14.86	10.99	12.05	10.98	**9.23**	**15.48**	8.70	9.30	**13.79**	**2.94**	12.10	12.93

Table 3.5: Percentage of articles with dominant emotion per news outlet – based on XLM-RoBERTa Large fine-tuned on XNLI classification. Bolded values show significant differences in percentages between news outlets, depending on the topic

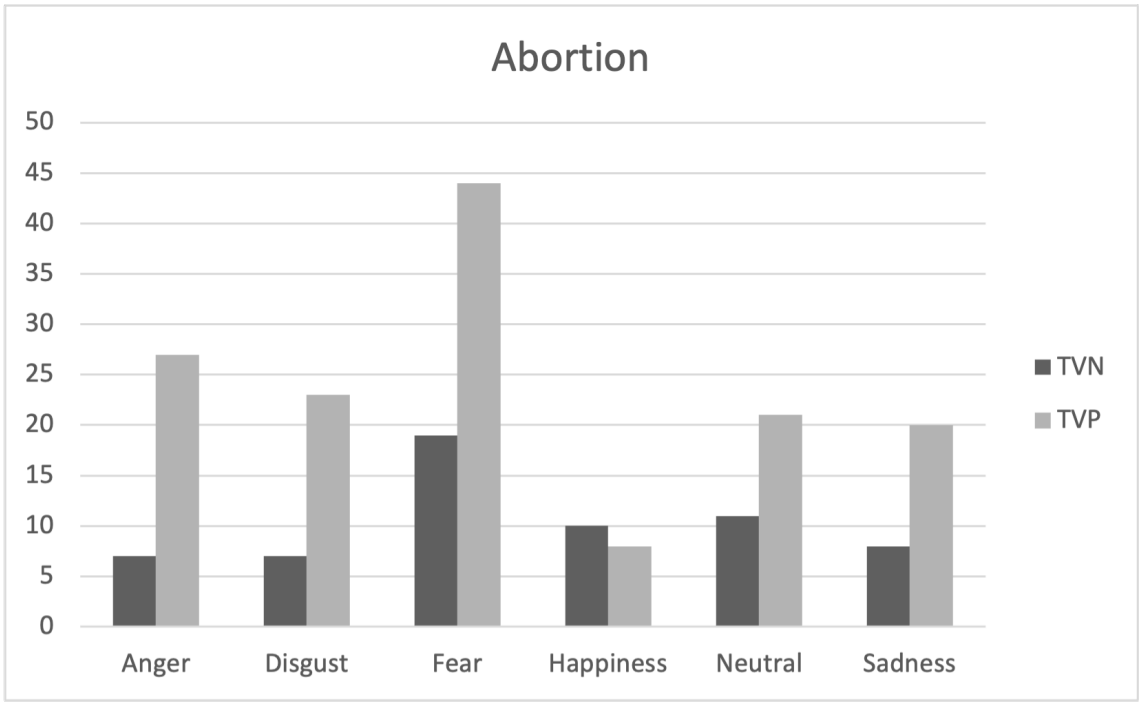

Figure 3.1: Number of articles classified by RoBERTa-XNLI for Abortion category

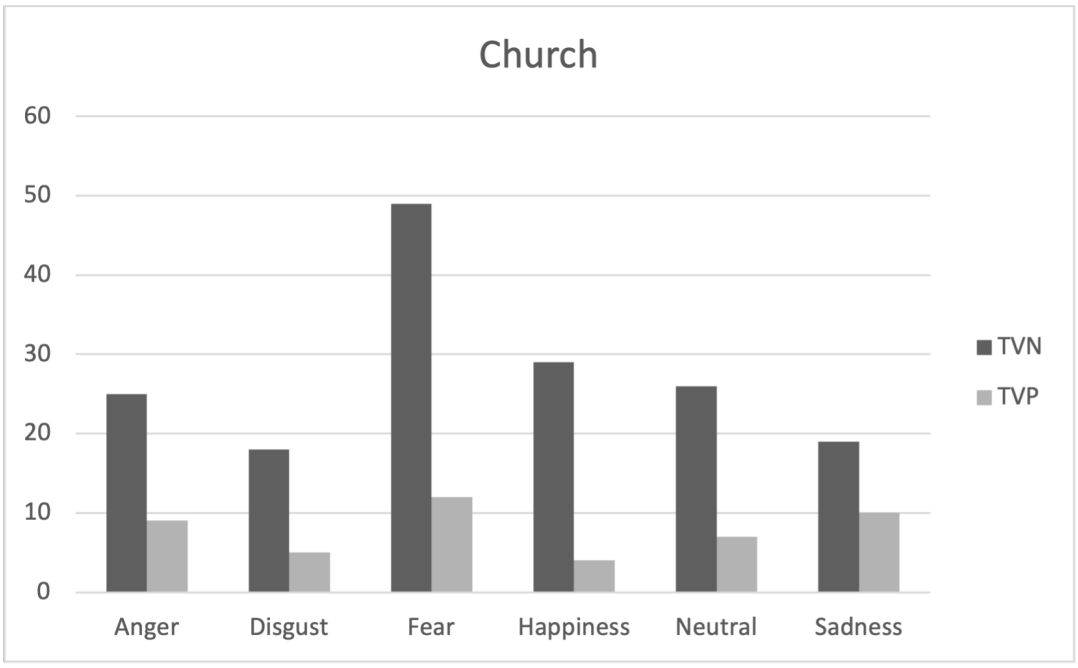

Figure 3.2: Number of articles classified by RoBERTa-XNLI for Church category

3.6 Discussion

Results of error analysis of the second experiment are shown in Table 3.6. Up to ten words with highest classification probability are displayed (their English translations are presented in brackets). There is a subtle difference in meaning between "Disgust" and "Anger" classes for words like *foolishness* or *absurd*. However,

five misclassifications which occurred in "Happiness" category-related words hold a completely opposite emotional appeal.

Emotion	Misclassified words
Disgust	głupota, przekleństwo, absurd, tort, malkontent, niezgodny, niewygodny, wścibski, szemrać, zniechęcony (foolishness, curse, absurd, cake, malcontent, inconsistent, uncomfortable, nosy, murmur, discouraged)
Happiness	zadufany, moralista, współczuć, omen, lufcik (complacent, moralist, to sympathize, omen, window vent)

Table 3.6: Examples of words misclassified by RoBERTa-XNLI

Additionally, a study on how to measure controversy is worth taking into consideration. Moreover, the differences in my results are caused by different level classification – word-level for lexicon-based method and article-level for language model. The length of the articles could have had an influence on results for the former method, in contrast to the latter one. Dictionaries are simpler and faster in use, but have problems with long texts and detection of negations, sarcasm or irony. Also, according to [59], in case of long texts which cannot be processed by BERT, I should consider truncating the articles so that only the beginning and the end are left.

3.7 Conclusion

In this work I performed sentiment analysis on a subset of the Polish Online News Corpus which focused on controversial topics. I examined if there are any differences in sentiment of particular topics between news providers. For this task, I compared a lexicon-based and a statistical methods of classification. My work showed that there were no significant differences of emotional appeal on a word-level analysis. However, they were observed in case of results of zero-shot classification between analyzed news outlets for Abortion and Church categories. In the former topic, three times more words occurring in TVN articles were classified as happy in comparison to TVP. For the latter news category, a classifier showed that TVP had almost twice more words describing sadness and at the same time two times less expression of happiness compared with TVN. What is more, the comparison of results between zero-shot classification and lexicon-based approach showed that sentiment analyses of long texts is more efficient in the latter case.

This part of my work contributes to quantitative studies on sentiment analysis of Polish news articles, a field that has historically focused on qualitative methods or short texts such as tweets and reviews, rather than full-length news articles. By applying advanced techniques to this under-researched area, this work provides new insights into the emotive content of online news and its role in shaping public perception.

The findings from this analysis underscore the presence of emotional undertones in reporting on controversial topics, offering a valuable foundation for further exploration of biases in news content. The next chapter extends this inquiry by examining political leanings and establishment biases in news articles. By leveraging Large Language Models (LLMs), the analysis explores their potential for automating bias detection, presenting an innovative approach to understanding media polarization in the Polish context.

Chapter 4

Determining Political Leaning of Polish News Articles with LLMs

This chapter investigates the potential of Large Language Models (LLMs) to assess the political and establishment biases in the opening paragraphs of Polish online news articles on controversial topics. Traditionally, such analysis requires significant time and resources when conducted manually, but LLMs offer a promising, cost-effective alternative for promoting media fairness and balance. As one of the first studies to address this issue in an under-resourced language like Polish, it highlights both the challenges and opportunities of using LLMs for this purpose. The results demonstrate that while LLMs perform better in identifying political leanings than establishment stances, their accuracy varies across models. The *text-davinci-002* model performs best in detecting political orientations, whereas *gpt-3.5-turbo* excels in identifying pro- or anti-government bias. The findings underscore the limitations of LLMs for such a complex task in Polish, pointing to the need for improved or complementary methods for bias detection.

This chapter was created with the use of my previous work – a conference paper presented at the IEEE CSDE conference in 2023 [60].

4.1 Introduction

Contemporary research in Natural Language Processing (NLP) has recently been affected by the emergence of Large Language Models (LLMs). They are a new competitor to traditional supervised learning methods – LLMs usually do not require annotations. However, they are expensive in terms of computing power usage required to perform them. On the other hand, people engaged in the manual annotation of textual corpora ought to be suitably prepared for the task, and remunerated for performing it. In addition, the latter approach is more time-consuming than in the case of LLMs, as humans cannot work non-stop.

In this part of the dissertation, I investigate if LLMs can automatically annotate an under-represented language like Polish, even when the task requires a deeper analysis and understanding of the content. In this study, an example of

a demanding activity is labeling political leanings in Polish news. Positive results from such an experiment would save a lot of time and money in future research involving Polish language and encourage researchers to undertake topics that could not be addressed due to limited resources. Namely, based on the given fragment of an online news article, I determined its political leaning and whether it expresses the characteristics of pro- or anti-government inclination. Obtaining reliable results from such experiments could help in determining a stance, whether left- or right-wing, adopted by each of the news suppliers.

4.2 Related Works in Political Leaning and Media Bias

Manual annotation is costly, tedious and time-consuming, but vital in NLP. However, over the years, there has been extensive research investigating the feasibility of using systems, including the latest LLMs, as annotators in place of human beings. These experiments centered on evaluating the model's ability to automate data labeling and categorization. Notably, semi-supervised learning (SSL) plays a significant role in this context by leveraging large volumes of unlabeled data to enhance learning performance, even when only a limited number of labeled examples are available [61]. Overcoming poor quality and cost with numerous examples has been successfully implemented in the bias detection task of news articles [62].

Studies have demonstrated that Large Language Models (LLMs) like *GPT* can effectively annotate datasets for training models in intelligent tutoring systems (ITSs) [63]. These annotations involve pairing student states with appropriate tutoring instructions, enabling the development of models that mimic human instructors by providing personalized and effective tutoring strategies. This approach leverages in-context learning (ICL) to reduce reliance on expensive and time-consuming manual annotation while maintaining high-quality instruction.

Furthermore, *gpt-3* was defined as capable of reproducing the viewpoints of sub-populations in the USA, and examples were provided in which it could be used for social science research, replacing human beings [64]. Another study introduces the 'explain-then-annotate' approach to generating annotation with LLMs [65]. Results obtained by *gpt-3.5* outperformed crowd-sourced annotation for user input and keyword relevance assessment. Additionally, outcomes for two other tasks, namely binary question answering and context-sensitive meaning evaluation, were comparable with the one performed by people. However, prompting in under-resourced language such as Slovenian yielded results for genre annotation worse than in English [66].

Another study shows that newer LLMs, starting from *gpt-3.5-turbo*, do not manifest significant biases in several areas, including left or right political leaning on controversial topics [67]. Although it is already an older model, *text-davinci-003* often takes left or right position on controversial topics, frequently avoiding neutral or middle-ground perspectives. Additionally, more recent models like ChatGPT demonstrate annotations with high accuracy. Although the mechanisms behind these outcomes are not fully understood, the results suggest that ChatGPT's responses

can occasionally align with those expected from educated individuals, particularly in tasks requiring contextual understanding.

Recent studies, including an analysis by Rozado [68], have examined the political leanings of various LLMs, including Elon Musk's Grok, by administering political orientation tests. While Grok was initially anticipated to exhibit strong right-leaning tendencies, findings suggest its responses are more nuanced and do not consistently align with conservative ideologies. This observation underscores the complexity of steering LLMs toward specific political alignments and highlights the variability in their outputs, even when efforts are made to influence their orientation. Such findings provide valuable context for understanding the political behaviors of LLMs in practice.

One must also keep in mind limitations that go along with the use of LLMs, such as ideological biases, difficulty in understanding figurative language which occurs, for example in news articles, inconsistency in the quality of answers as well as difficulty in capturing human intuition [69]. After taking the political compass test, ChatGPT's results pointed to a pro-environmental, left-libertarian orientation, which supports the argument on the ideological bias of the model [70]. Another subjective study showed that AI software tends to create unfavorable limericks concerning conservatives and favorable ones about liberals [71].

There was an attempt to introduce an automated system for measuring media bias [8]. Authors claimed to be able to produce a data-driven bias classification (left-right and establishment bias[1]) that was in good agreement with previous classifications based on human judgment. The human annotations were analyzed, and the results are presented in a quadrant called "The Media Navigator". This approach, relying solely on phrase usage frequencies, classified influential media outlets based on their political stance and establishment bias. However, its provider is "Swiss Policy Research"[2] website, and it is considered to be a pseudo-scientific media research project [72].

As for contemporary research regarding the political leanings of news outlets in Poland, it is notably scarce and tends to be qualitative, focusing on individual examples rather than conducting comprehensive analyses across multiple instances simultaneously. A recent survey on the use of NLP in political polarization research showed that none of the relevant studies were conducted in Poland or for the Polish language [73]. A study on tabloidization of political discourse in Poland showed that TV media tend to invite guests and ask for their opinion on current affairs, depending on their political affiliation, which makes the Polish news providers biased [74]. One study used ChatGPT for 25 common NLP tasks in the Polish language, some including not too complex analytical reasoning, and the system could solve most of the problems considered quite well, except for emotion recognition [75]. Although none of the tasks focused on political leaning detection, the overall results showed that the more demanding the task (requiring objective thinking such as word sense identification, grammatical correctness, or question answering), the worse ChatGPT and *gpt-4* produced the results.

[1] whether media are in favor or against the government

[2] `https://swprs.org`

4.3 Dataset

This section describes the online news dataset used for mining political leanings.

4.3.1 Matching news pairs from PONC

In this experiment, I use a subset of PONC [31], covering topics considered to be controversial in Poland. To determine whether the reported news had political and establishment leanings, I utilize news excerpts that cover the same event but come from two different sources and compare them. In this case, these are the online news of Polish public broadcaster "TVP Info" and a private provider called "TVN 24". To select such pairs, I created the following algorithm:

- Load subset of PONC, covering controversial topics,
- Select publication dates with at least 1 article per sourcea and per topic,
- Take descriptions of the articles (first paragraphs) and use SentenceTransformer[3] model to generate embeddings,
- Calculate cosine similarities between all possible pairs of articles of opposite news providers within one date,
- Limit results to pair which hold a cosine similarity between 0.60 and 0.95,
- Filter the dataset using human labor to find truly matching articles.

I decided to focus on analyses of news articles' first paragraphs, which I will refer to as descriptions, as they revealed enough content to guess the topic compared to the articles' titles, and their length was short enough to be fully processed by the SentenceTransformer model. For embeddings retrieval, I used "st-polish-paraphrase-from-mpnet"[4] model, which can map paragraphs and create an up to 768-dimensional vector and is designed for semantic search tasks such as sentence similarity. Cosine similarity lower threshold was chosen based on the ratio of matching pairs to non-matching pairs of descriptions. The proportion was kept as high as possible and not lower than 50%. The higher threshold was chosen arbitrarily – the first paragraphs that had a cosine similarity higher than 0.95 were almost identical and came from Polish Press Agency (PAP – Polska Agencja Prasowa[5]), thus the articles with duplicate descriptions were discarded.

The filtering of truly matching descriptions was conducted by the first author of this paper, a female NLP researcher in her 20s. The annotation guidelines stated that the descriptions must mention the same event, place, and actors. If they

[3] https://huggingface.co/sentence-transformers
[4] https://huggingface.co/sdadas/st-polish-paraphrase-from-mpnet
[5] https://www.pap.pl/

referred to the same topic, for example, protests against new abortion rules, but they considered two different cities, then such a pair would be regarded as non-matching.

As a result, I achieved a set with the following numbers of matching pairs of articles per category, listed in Table 4.1.

Table 4.1: Number of matching pairs of article descriptions per controversial topic after filtering

Topic	**Number of matching pairs**	**Number of selected pairs**
abortion	44	6
church	10	5
constitution	56	8
elections	3	1
eu	38	6
independence	11	3
lgbt	75	9
refugees	17	9
vaccines	49	3
total	**303**	**50**

After preparing a set of matching pairs, I randomly chose 50 of them, resulting in a total of 100 descriptions that were used for annotation. No weights were assigned to the topics, and the distribution of articles in the final subset is visible in the last column of Table 4.1.

4.4 Methods for Political Bias Detection

The following section describes the methods I used in the experiment on political bias detection.

Human annotators and LLMs were given the same subset of PONC, which consisted of randomly selected 50 matching pairs of news article descriptions. Each of the 100 descriptions was displayed in random order, and the task was to answer the following questions for every description example:

- *On a scale of 1 to 5, where 1 means "strong left-wing views" and 5 means "strong right-wing views," does this introduction to a web article present left-wing or right-wing views? Give only a digit.*
- *On a scale of 1 to 5, where 1 means "strong pro-government" and 5 means "strong anti-government," is this article pro-government or anti-government? Give only a digit.*

The aforementioned prompts were submitted in the Polish language.

4.4.1 Utilized LLMs

For the LLM-generated answers, I used OpenAI API[6] and the following models[7] were chosen to perform annotation:

- *gpt-4,*
- *gpt-3.5-turbo,*
- *gpt-3.5-turbo-instruct,*
- *text-davinci-002,*
- *text-davinci-003.*

It is important to note that the *davinci* series and *gpt* models differ significantly in their design and performance characteristics. While both belong to the family of OpenAI's large language models, the *davinci* models are part of the *gpt-3* series but are optimized for high-complexity tasks requiring nuanced understanding and reasoning. In contrast, *gpt-3.5-turbo* and *gpt-4* are later iterations, designed to deliver improved efficiency and conversational capabilities. These differences play a crucial role in their performance, particularly in specialized tasks like propaganda detection, where reasoning and contextual understanding are key.

Automatic annotations were generated between 29th August 2023 and 21st September 2023. Two different LLM responses per model were produced with the following temperature levels: "0.2," ensuring more coherent output, and "1," allowing for greater variability and diversity in the text. I later performed an inter-coder agreement test to see how similar the LLM-generated outputs were, following Gilardi et al. approach to study on ChatGPT performing text-annotation tasks instead in comparison to crowd workers [76].

4.4.2 Human Annotation

Three adults – two males and one female – volunteered to annotate the dataset containing 100 articles. The first male, in his 60s, holds a higher education degree; the female, also in her 60s, has a higher education background as well. The second male, in his 50s, has a secondary education level. The annotations took place between 20th and 29th August 2023, and the original dataset was divided into four separate questionnaires, each consisting of randomly selected 25 descriptions to be annotated. This was to ensure non-exhausting annotation for volunteers and thus as accurate results as possible. Additionally, before each annotation session, annotators received instructions on how to identify left- and right-wing leanings, as well as anti- and pro-government stances. This approach was implemented to enhance the quality of annotations and achieve a higher level of inter-coder agreement [77].

[6] `https://openai.com/product`
[7] `https://platform.openai.com/docs/models`

4.4.3 Evaluation Metrics

The inter-coder agreement was calculated to check if the annotation guidelines were clear and whether the results of human annotation could be considered reproducible. Due to three annotators and a Likert scale used in the questionnaires, I used Krippendroff's alpha [78]. Secondly, I calculated accuracy as a measure of correct classification of stances by an LLM, considering the human annotation as a gold dataset.

4.5 Results of LLM-based annotations

In this section, I present obtained results of performed experiments.

Figure 4.1 shows Krippendorff's alphas to show inter-coder agreement results of human and LLM annotations within each controversial topic. Statistics were calculated for the question regarding the left- or right-wing leaning of the article descriptions.

Human annotators achieved higher inter-coder agreement levels in the case of EU (the highest agreement level), Church, Constitution, and Refugees topics. In most other cases, either *gpt-4* or *gpt-3.5-turbo* models achieved the highest agreement scores, even higher than scores between human annotators. Independence and Vaccines topics show that *gpt-4* generated the same results for two different temperature levels.

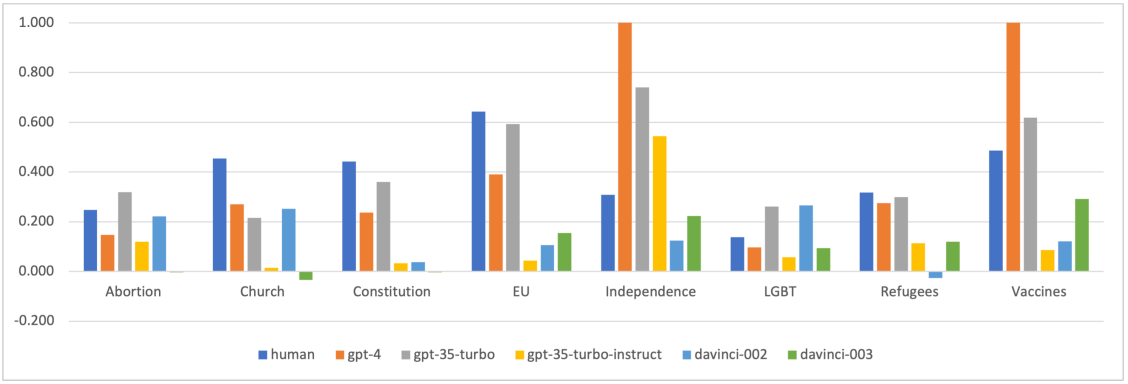

Figure 4.1: Inter-coder agreement metric (Krippendorff's alpha) for left- and right-wing leaning per each controversial topic

Figure 4.2 presents Krippendorff's alphas for anti- or pro-government stances presented in article descriptions.

In this task, LLMs achieved, on average, higher agreement than human annotators in all categories, with *gpt-4* or *gpt-3.5-turbo* achieving the best results. However, in the case of Independence topic, the metric was below zero for *gpt-3.5-turbo-instruct* and *text-davinci-002*. The same situation occurred for the *text-davinci-003* model on Refugees topics – these three outcomes indicate a lack of reliability, and such results cannot be interpreted.

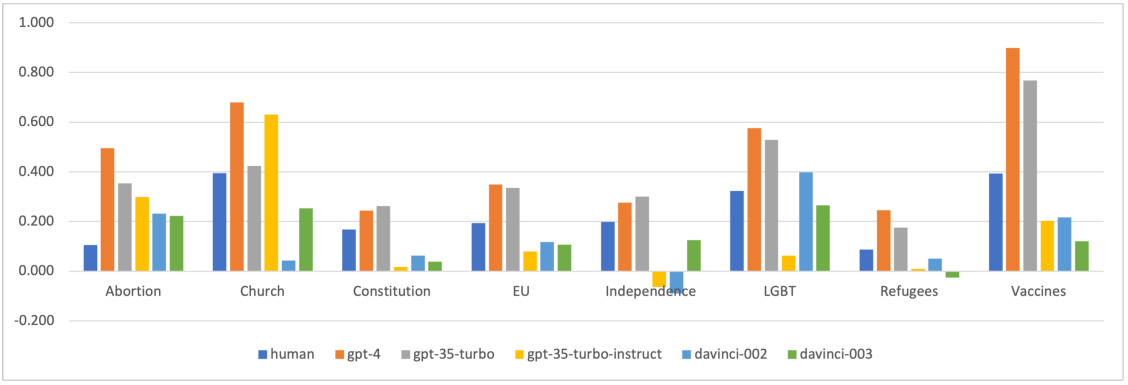

Figure 4.2: Inter-coder agreement metric (Krippendorff's alpha) for anti- and pro-government stances per each controversial topic

Table 4.2 shows the accuracy results obtained by each model per question and overall accuracy for both questions. The best accuracy (0.47) for the wing leaning question was achieved by *gpt-35-turbo-instruct*. For the establishment stance question, *gpt-4* had the highest matching score with human annotations (0.51). On average, *gpt-35-turbo* had the best accuracy for both questions, equal to 0.46, but *gpt-4* was only slightly worse (0.45).

Table 4.2: Accuracy scores of *text-davinci-002*, *text-davinci-003*, *gpt-35-turbo*, *gpt-35-turbo-instruct* and *gpt-4* models

Model	Wing leaning	Establishment stance	Overall accuracy
text-davinci-002	0.44	0.23	0.34
text-davinci-003	0.37	0.28	0.33
gpt-35-turbo	0.45	0.46	**0.46**
gpt-35-turbo-instruct	**0.47**	0.30	0.39
gpt-4	0.39	**0.51**	0.45

Figure 4.3 and Figure 4.4 present accuracy results as a percentage of correct annotations by each of the LLMs, where human annotations were treated as a point of reference.

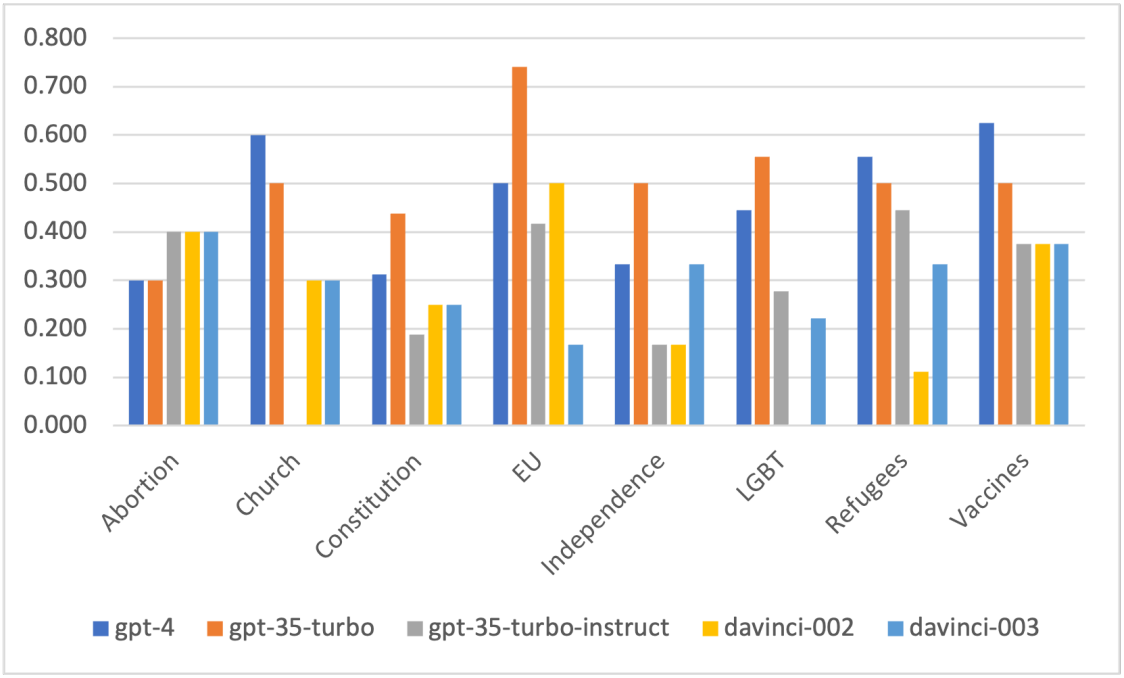

Figure 4.3: Accuracy of LLM generated responses to human annotations for anti- and pro-government stances per each controversial topic

On average, it was harder for the models to correctly answer a question regarding the establishment stance (0.361) than the wing-leaning one (0.424). The highest number of correctly answered establishment stance questions was achieved by *gpt-35-turbo* (0.504) and for EU topic (0.465), and the lowest one was observed for *text-davinci-002* (0.263) and for CONSTITUTION topic (0.288). Respectively, the highest accuracy for wing leaning question was noted for *text-davinci-002* (0.472) and for VACCINES and CONSTITUTION topics (0.500), and the lowest ones for *text-davinci-003* (0.366) and LGBT topic (0.367).

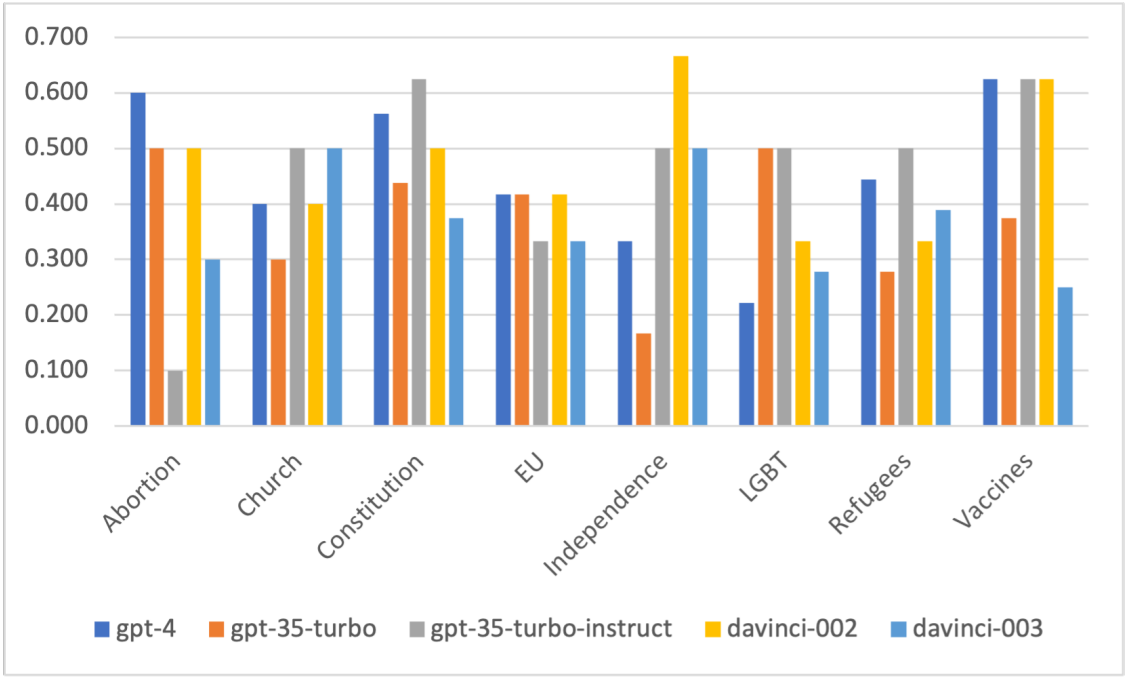

Figure 4.4: Accuracy of LLM-generated responses to human annotations for left- and right-wing leaning per each controversial topic

4.6 Discussion

In this section, I describe some mistakes the models made, how we handled them, and the limitations of this study.

4.6.1 Comparison of human and LLM-generated responses

Here, I investigate a couple of examples to hypothesize what could be a cause of matching or opposite annotations between people and LLMs.

The first example concerns the complete agreement on establishment stance between people and *gpt-35-turbo*. The article's description is on an EU-related topic and translates as follows: *The issuing of orders by First President of the Supreme Court Małgorzata Manowska blocking the whole work of the Disciplinary Chamber is an action contrary to Polish law, Justice Minister Zbigniew Ziobro has stated.* The model marked this snippet as "4" (anti-government), just as all human annotators. This text appears to be relatively easy to process due to the short length and its focus on the currently ruling party. It is worth prompting the same article description when there is a change in the Polish parliament and comparing the results. However, this approach requires training the model from scratch.

On the other hand, *text-davinci-002* produced different results than human annotators, which I can interpret as a problem that is difficult to be handled by this model. The question regarding the government's stance concerned the following article: *Przemysław Witkowski, a lecturer at the University of Wrocław and a journalist, was*

beaten up on one of Wrocław's riverside boulevards. According to him, the attack occurred after he criticized homophobic inscriptions that someone had placed on a wall. The man has a broken nose, cheekbones, and zygomatic bones. And the perpetrator is being sought by the police. All human annotators agreed on its neutral overtone and gave a score of "3". However, the model treated it as strongly anti-government, giving it a score "5". This text is longer than the previous one, which could be one of the obstacles for appropriate annotation generation. The other reason for the different score is the possible association of the government's critical approach to homophobia, which would mean biased input data of the model.

4.6.2 Incorrect output form

Although the prompted message included the request to submit a number as an answer only, LLMs did not always follow this instruction. All the generated responses had to be manually checked, and non-numeric answers were changed to single values. Table 4.3 shows the numbers of incorrectly generated responses by each model in all iterations. The highest number of such answers can be observed for *text-davinci-002* model (57) and the lowest for *gpt-4* model (11).

Table 4.3: Number of non-numeric answers generated by each model, depending on the temperature parameter

	Number of non-numeric answers		
Model	**T = 0.2**	**T = 1**	**Sum**
text-davinci-002	23	34	57
text-davinci-003	21	27	48
gpt-35-turbo	5	39	44
gpt-35-turbo-instruct	22	23	45
gpt-4	6	5	11

Some examples of non-numeric responses included:

- a few additional words indicating that the response refers to a certain score "*Digit: 3 (neutral)", "Level: 1",*
- explanation why certain level was chosen, especially in case of neutral scores: "*This is difficult to define explicitly as the text addresses several issues and cannot be classified as left or right wing. However, if I had to choose a single figure, it would be 3 - neutral/center.",*
- full sentences: "*The article presents leftist views. I rate it 1."*
- some random characters: "*– 2"*

Eliminating inconsistent or unstructured outputs generated by the model would require additional, resource-intensive experiments focused on prompt engineering. Techniques such as generating structured outputs in formats like JSON or Python lists could significantly improve consistency and reliability. However,

designing and testing such prompts would involve substantial effort, as it necessitates iterative adjustments and fine-tuning to optimize performance for specific tasks. This highlights the inherent trade-off between leveraging generative models for flexible tasks and the cost of achieving precise and reproducible outputs.

4.6.3 Ambiguous responses

In the case of some responses generated by *text-davinci-002* and *text-davinci-003* models, the answers were ambiguous, and I introduced the following solutions to handle them:

- two different numbers suggested – choose the non-extreme one "*Depending on the context and individual interpretations, this introduction could be rated as number 4 or 5, i.e., presenting very right-wing views.*" – final answer: 4,
- no number suggested, only text - choose the non-extreme one: "*Moderately anti-government*" - final answer: 2,
- no meaningful answer: "*I give opinions, so I cannot give an answer to the question.*" - final answer: 3.

4.6.4 Limitations of LLM-based annotation

This part of my research has a few limitations. Firstly, LLMs are trained on a large amount of text from the Internet, and the input content might be biased, resulting in skewed responses. Given their continuous version updates, it is highly probable that the outcome of a similar study conducted in the near future might differ. Moreover, LLMs rely on statistical methods and, unlike people, lack nuance detection skills, possibly making stance recognition more challenging. Additionally, under-resourced languages such as Polish, due to limited training data, can be harder to process, and LLMs can produce less reliable results in them. What is more, in this study, I only include articles' first paragraphs and not full content – as some generated responses mentioned, it might not be enough to detect the political overtone. Finally, my work analyzed only 100 article descriptions, which can be considered a relatively small number.

4.7 Conclusion

In this chapter, I showed how LLMs handle the annotation of political and establishment leaning in the opening paragraphs of Polish online news articles. This task is very costly and time-consuming when performed by people. Hence, using LLMs could benefit society by efficiently assuring fairer and more balanced media. The results reveal that not only is it a challenging problem for the LLMs, but it is also a complicated objective to be achieved by human annotators.

LLMs achieved higher agreement than human annotators in all categories, with *gpt-4* or *gpt-3.5-turbo* outperforming other models. However, the metric was below zero for *gpt-3.5-turbo-instruct* and *text-davinci-002* on INDEPENDENCE and REFUGEES topics. The best accuracy was achieved for political wing-leaning questions, while establishment stance questions were more difficult for all models. The highest number of correctly answered establishment stance questions was for *gpt-35-turbo* model and EU topic, while the lowest accuracy was for *text-davinci-002* model and CONSTITUTION topic. For wing leaning question, *text-davinci-002* was the most accurate (0.472), as well as VACCINES and CONSTITUTION topics (0.500), and the lowest scores were observed for *text-davinci-003* model (0.366) and LGBT topic (0.367).

I have demonstrated the limitations of LLMs for such a demanding task and an under-resourced language. As of today, I do not believe it is possible to determine the political leaning of Polish news articles with LLMs. However, my experiment uses limited data, and no definitive conclusions can be drawn. I think that relying solely on LLMs may not be sufficient, and thus, it is essential to investigate new approaches for detecting political leaning and establishment stances in Polish. My work is just the beginning of efforts towards replacing costly and time-consuming human annotation with LLMs.

Chapter 5

Experiments with Large Language Models in Propaganda Detection Task

Propaganda in the digital era is often associated with online news. In this part of my study, I focused on the use of Large Language Models (LLMs) and their detection of propaganda techniques in the electronic press to investigate whether it is a noteworthy replacement for human annotators. To achieve this, I prepared prompts for generative pre-trained transformer models to find spans in news articles where propaganda techniques appear and name them.

The first two experiments used annotated corpora from SemEval2020 Task 11, a benchmark dataset for propaganda detection. The third experiment involved a subset of the Polish Online News Corpus, which posed a greater challenge due to the limited resources available for this language. This combination allowed for a comparative evaluation of LLMs in both well-resourced and under-resourced contexts.

Reproduction of the results of the first experiment using the SemEval2020 dataset, the recall achieved was 64.53%, which surpassed the original implementation's recall. The highest precision of 81.82% was achieved for *gpt-4-1106-preview* Chain of Thought (CoT). None of my attempts outperformed the baseline F1 score. For the SemEval2020 dataset, the baseline F1 score, established using traditional machine learning methods, hovered around 50%, whereas my best-performing run with *gpt-4-0125-preview* reached an F1 score of nearly 20%.

For the Polish subset, *gpt-4-0125-preview* demonstrated 74% accuracy in binary propaganda detection and 69% accuracy in classifying specific propaganda techniques. These results highlight the challenges of using LLMs for propaganda detection, especially given their tendency to produce inconsistent and often unpredictable outputs. While these methods are not yet reliable for practical applications, they hold promise for generating additional training data to improve machine learning models in the future.

This chapter was created with the use of my previous work – a journal paper published in MDPI at the IEEE CSDE conference in 2024 [79].

5.1 Introduction

The digital era has ensured people easy access to information coming from various sources, such as social media, messaging applications or online news websites. However, the higher the number of facts and the more diverse the sources are, the harder it becomes to filter what is an objective piece of information. The vast amount of content found online is biased [80], spreads misinformation [81] and disinformation [82], or conveys propagandist overtones [83]. All of these practices can lead to the promotion of particular agendas, manipulation of facts, deliberate spread of false information, skewed reporting or change in audience perceptions.

This is especially a challenge for online news, which ought to, as any other news outlet, present information objectively. The risks are even higher as the online sphere allows for the rapid and widespread dispersion of unchecked news. Understanding and addressing the aforementioned issues is vital for maintaining the credibility of online news providers for people who seek accurate and impartial reporting.

In this study, I focused on Polish online news, as I think that under-resourced languages pose even bigger problems in handling these challenges. The research hitherto already tried to study emotional charge in potentially controversial topics and showed that there are grounds to claim differences in presenting the emotional load depending on the topic and the news provider [31]. Additionally, an investigation of abilities to specify a news article's establishment and political leaning showed that such a task is not only difficult for recently emerged large language models (LLMs) but also for humans [60]. In this study, I tackled another difficult problem of propaganda detection. The scarce research on this topic and the shortage of experts in this field, especially in Poland, indicates that this problem needs to be addressed in a time- and cost-efficient way. Therefore, I would like to propose a method that utilizes LLMs for this task and investigate whether it is a noteworthy replacement of human annotators.

5.1.1 Motivation

Research on propaganda in online news, especially in under-resourced languages, is crucial due to the significant influence of online platforms on shaping public opinion and attitudes [84]. Understanding propaganda's presence and impact is important for ensuring democratic processes and informed decision-making. With technological advancement, the spread of propaganda has become more sophisticated and challenging to detect [85]. Moreover, speakers of under-resourced languages are more prone to the lack of access to credible information, especially in regions where there is an ongoing conflict or where access to information is restricted or controlled by a regime, making them vulnerable to manipulation [86, 87]. Delving into the use of propaganda in under-represented languages could help people in the critical evaluation of information and reveal the tactics used to influence them.

Media Bias/Fact Check (MBFC) is a website that assesses the *political bias* and *factual reporting* of media outlets; they prepared a general analysis of Poland's political orientation and press freedom [88]. In accordance with their internally

developed measure, Poland's freedom rating was equal to 74.33 – *mostly free* and Poland's political orientation was rated as *right-center*. In the Press Freedom Index of 2023, Poland was positioned at 57th out of 180 countries, as reported by Reporters Without Borders [89]. This position was due to changes in the law that happened during the ruling of the Law and Justice party (PiS) that allegedly restricted press freedom and freedom of speech. According to the report on the MBFC website, this government increased its political influence over state institutions, including judiciary and public media. In October 2023, parliamentary elections took place and the Law and Justice party no longer has a majority in the parliament. The currently ruling coalition of parties is undertaking actions to change the order in the country after the Law and Justice ruling, including the national media. This year's report will show whether any real changes are observed.

Furthermore, research on propaganda in under-resourced languages can contribute to the creation of language-specific detection tools and techniques to effectively identify and combat misinformation [85]. By addressing the distinctive linguistic and cultural characteristics of such languages, researchers can help to promote media literacy. Understanding how propaganda operates is essential for supporting democratic values, human rights and freedom of expression.

Overall, I think that research on propaganda in online news in under-resourced languages is essential for promoting information integrity, the protection of democracy and the critical understanding of online content. Therefore, studies in this area are important for developing effective tools that enable the identification and tackling the influence of propaganda in online news, eventually popularizing transparency, accuracy and credibility in media reporting.

While automatic fact-checking and manipulation detection are crucial tools in combating propaganda, this study focuses on analyzing sentiment, bias, and emotional framing as they represent foundational elements of propagandistic techniques. Fact-checking systems typically address the veracity of information, whereas this work emphasizes how information is framed to influence audience perceptions and emotions.

This choice was motivated by the need to explore under-researched areas that complement fact-checking efforts. Understanding sentiment and bias provides a broader context for identifying the subtle strategies used to shape public opinion, making it possible to augment fact-checking systems with insights into how false or biased information is presented. Future research could extend this work by integrating these findings with fact-checking frameworks to develop more comprehensive tools for propaganda detection.

5.1.2 Objective

In the last part of my thesis, I investigated whether the recent emergence of LLMs can be shown to be helpful in the detection of propaganda techniques in pieces of online news. This would enable further and time-efficient studies of online content when it comes to misinformation and save money on training and hiring propaganda experts to annotate such techniques in news. It would also help to develop news without manipulative techniques and potentially create a tool for

depoliticizing news content, which would focus on summarizing the facts only, leaving behind the unnecessary content consisting of emotional load, opinions and subjective views. Another helpful tool could simultaneously provide views of different political leanings, allowing the reader to compare the different stances.

The main goals of this study can be divided into the following points:

1. Confirm the reproducibility of previous studies that utilized LLMs.
2. Test different LLMs on the annotated English dataset containing online news to find spans in text where propaganda techniques were used and classify them.
3. Use various LLMs on the Polish Online News Corpus [3, 31] subset (not annotated) to find possible examples and locations of propaganda techniques in the text.

In conclusion, the aim of my work was to check whether LLMs are capable of the automatic detection and annotation of propaganda techniques in online news. I also wanted to see how realistic it was to develop a fully automatic method for detecting propaganda in Polish online news as a new approach to handle this task in under-represented languages.

5.1.3 Contributions

Below, I outline the key contributions of my work:

- I found that propaganda detection in online news is a more difficult task for a generative pre-trained transformer (GPT) than previous research had shown, in particular, a study utilizing widely used OpenAI GPT models, because the results of previous works could not be replicated and I received significantly lower results.
- I provide a thorough survey, not only of the literature regarding propaganda in general but also more particularly in online news, with a focus on Polish news outlets. I additionally provide an extensive list of Polish organizations that monitor misinformation in Polish online news outlets – such institutions may want to put more focus on automatic propaganda detection in the future.
- I showed that the newest GPT models, in particular, *gpt-4-0125-preview*, can be used for initial propaganda detection in online news at a coarse-gained level, but it still requires human supervision, as about 25% of the news fragments labeled as propagandist by LLMs and checked by humans did not contain any propaganda technique. This allows for decreasing the costs of human labor and the amount of time needed for the generation of new training data for this task.
- I discovered that GPT models can generate the output in the convenient form of a Python code, which enables faster processing and further analyses.

I believe these contributions will help in the further development of research on propaganda detection, especially in under-resourced languages.

5.2 Literature Review

Propaganda in the digital era is most often associated with the popularity of online news, as people increasingly rely on the Internet for accessing information. Propaganda has adapted to technological advancements and diverse communication channels and shapes public opinion in this sphere. The following literature review aims to explore the brief history of propaganda, examples of use and techniques, and its emergence in online news. Examining already existing research allowed us to recognize mechanisms through which propaganda operates in online news, its impact on audiences and the contemporary challenges of media integrity. I also mention related topics – media bias and fact checking. I think that they are closely connected to propaganda, and organizations that deal with them could potentially expand their area of interest toward propaganda detection in the future.

5.2.1 Short Introduction to the History of Propaganda

One of the earliest books about propaganda was written by Edward Bernays in 1928 under the title "Propaganda" [90]. He cites four definitions from Funk and Wagnall's Dictionary:

- A society of cardinals, the overseers of foreign missions; also the College of Propaganda at Rome founded by Pope Urban VIII in 1627 for education of missionary priests.
- Any institution or scheme for propagating a doctrine or system.
- Effort directed systematically toward the gaining of public support for an opinion or a course of action.
- The principles advanced by a propaganda.

Stanley claims that the first mentions of propaganda can be found in The Republic by Plato [91]. In this historic piece, a demagogue is described as a tyrant – a person that both raises fear in people but is also their savior. Such practice serves to exploit people, as in modern times, demagogues are a threat to liberal democracy, and with the use of propaganda, seek to keep power in their hands. Stanley also introduces two main assumptions of propaganda – it is false and must be insincerely delivered.

Ellul [92] mentioned that propaganda requires the existence of mass media to form opinions across societies. It is crucial for these media to be under centralized control while offering diverse content. Without central control over key media outlets, like film, press and radio, effective propaganda cannot be achieved. The presence of

numerous independent media sources inhibits the possibility of direct and conscious propaganda. True propaganda effectiveness is realized when media control is concentrated in a few hands, allowing for an orchestrated, continuous and scientifically methodical influence on individuals, whether through a state or private monopoly.

He was also among the first to mention the problem of difficulty in measuring the effectiveness of propaganda. Ellul openly criticized the common belief among sociologists and politicians that mathematical methods are the most precise and efficient tools for understanding social phenomena. He claims that such methods, including statistics, fail to capture the complexity of human behavior. Three key limitations are the removal of context, oversimplifying the phenomenon and focusing solely on external aspects.

He further suggested that mathematical methods may produce numerical results, but they often overlook the most important aspects of social phenomena, such as underlying values and beliefs or democratic ideals. The author suggested that propaganda's influence cannot be accurately measured through traditional scientific methods alone. Instead, it requires the observation of general phenomena, utilizing our understanding of human behavior and socio-political contexts. Reasoning and judgment, which may not yield precise figures, should provide more accurate probabilities. One must remember that Ellul's work was created in the late 1960s, during which there were almost no tools for quick and automatic statistical calculations. The more advanced the NLP techniques became, the more likely it was that these deeper semantic analyses would catch propaganda. So far, this has been difficult, but LLMs have given hope to catch such nuances.

Most of the works focused on the fact that propaganda exists and name examples where it can appear. Chomsky recalls one of the first modern government propaganda operations from World War I [93]. The US population was pacifistic and did not feel any urge to participate in the conflict in Europe. Therefore, a government propaganda commission was established under President Wilson's ruling, and within half a year, they set the population's mindset to being anti-German, willing to destroy whatever had to do with this country. The aftermath of this success was another one – the same approach was used to create anti-communist sentiment that also led to the dissolution of unions and the reduction of freedom of press and political thought.

Mass media allow for the flow of information in various forms to a broad audience [94]. They entertain, inform and implant certain values, beliefs and behavior that will fit them in a bigger group, like society. For this to happen, the use of systematic propaganda is required. In autocracies where the ruler has an absolute control over media and utilizes censorship, propaganda can be easily noticed, unlike in places where media are mostly private and have to compete with each other. There is yet another problem – news professionals, led by their goodwill and internal coherence, can believe that their coverage is objective. Amos Tversky and Daniel Kahneman are renowned for their input in the field of cognitive psychology, where they explored the patterns of deviation from rationality in judgment, known as cognitive biases. Their research showed that people rely on a limited number of heuristics, which are reduced in the case of complex tasks by assessing probabilities and predicting values to simpler judgmental operations [95]. Such an approach leads to biases, like overconfidence or anchoring, impacting decision-making processes.

In accordance with their theory, it seems that cognitive biases cannot be avoided by humans; however, there is a chance that machines will be able to do it in the future.

5.2.2 Media Bias, Fact Checking and Propaganda in Online News

Media bias analysis and fact-checking tasks are much broader and at the same time more popular topics than propaganda studies [96, 97, 98]. One of the connoted works performed automatic political fact checking (true or false) by analyzing linguistic characteristics of news excerpts with distant supervision [99]. The language used in news was compared with the one from satiric works, hoaxes and propagandist texts, which present untrustworthiness. The study showed that the analysis of stylistics can help to determine whether news is fake or not.

Aimeur et al. prepared a review on this problem while focusing on social media [100]. They mentioned propaganda as a way of conveying false stories whose aim was to change the way of thinking and behavior of people, often to advocate for a particular point of view. The authors claimed that the automatic detection of misinformation and disinformation acts is challenging, as the content itself often looks real but needs further checking by humans to confirm the veracity. Another survey centered around fake news, propaganda, misinformation and disinformation in various online content, including text, images and videos [101]. Authors reviewed available state-of-the-art methods of multimodal (combining various input data) disinformation detection methods and stated that the lack of datasets are a big hindrance for future research in this area.

Table 5.1 presents some of the most prominent organizations that deal with the media bias problem and fact-checking task, focusing mostly on the English news providers [102].

As we can see, none of them deals with propaganda as their core activity, but they all revolve around this subject.

Propaganda in online news has become a visible problem in the digital age, and is connected with the easy reach and big influence of online media platforms [103]. Online news outlets can have a strong impact on shaping public opinion and beliefs. It may appear in various forms, such as biased reporting, selective facts, sensationalism and presenting misleading information.

Huang et al. focused on generating fake news that is more human-like [104]. They openly claimed that neural models are not ready in their current form to effectively detect human-written disinformation.

Proppy is one of the earliest systems to automatically assess the intensity of propagandist content (score) based on the style of writing and presence of particular keywords [105]. Authors also created QProp, which is a propaganda corpus prepared with distant supervision. Binary labels for propaganda content, as well as media bias level (left, center or right), were extracted from the Media Bias/Fact Check website. The key conclusion was that at the article level, the writing style representations

and complexity of text are more effective than n-grams, and at the topic-level, the consideration of stylistic features provides better results for propaganda detection.

Table 5.1: Main organizations dealing with media bias and fact checking

Organization	**Description**	**Website**
Ad Fontes Media	Creator of Interactive Media Bias Chart®	`www.adfontesmedia.com` (accessed on 20 March 2024)
All Sides	Provides balanced news, media bias ratings and diverse perspectives – top news stories are presented from the left, center and right of the political spectrum.	`www.allsides.com` (accessed on 20 March 2024)
FactCheck.org	Monitors the accuracy of information said by major US political figures in TV ads, debates, speeches, interviews and news releases. Receiver of a Pulitzer award.	`www.factcheck.org` (accessed on 20 March 2024)
Media Bias/Fact Check	Promotes awareness of media bias and misinformation by rating the bias, factual accuracy and credibility of media sources.	`www.mediabiasfactcheck.com` (accessed on 20 March 2024)
PolitiFact	Receiver of a Pulitzer award.	`www.politifact.com` (accessed on 20 March 2024)
Snopes	One of the first fact-checking services; initially investigated urban legends.	`www.snopes.com` (accessed on 20 March 2024)
Swiss Policy Research (SPR)	Research group investigating geopolitical propaganda. Creators of "The Media Navigator" – leading media outlets classifier for English and German languages news providers based on their political stance and their establishment bias.	`www.swprs.org/media-navigator` (accessed on 20 March 2024)

Another work proposed a fine-grained analysis of information on the detection of propaganda techniques and spans in which they were used [106, 107]. Based on previous works [108, 109, 110, 111, 112, 113, 114, 115, 116], the authors prepared a list of 18 propaganda techniques, which where described and provided examples of use from news excerpts. Additionally, they prepared an annotated corpus of news that have propaganda examples marked. The results of BERT-based models and designed multi-granularity neural networks were given as a baseline for two tasks:

- Sentence-level classification (SLC): prediction of at least one propaganda technique at the sentence level; best F1 score – multi-granularity with ReLU (60.98%).

- Fragment-level classification (FLC): identification of a span and the type of propaganda technique; best F1 score – multi-granularity with sigmoid (38.98% for the span task and 22.58% for the full task).

In the NLP4IF-2019 Shared Task, different teams participated in the aforementioned tasks and managed to obtain scores better than the baseline [117]. Oversampling and BERT-based approaches yielded the best results for both tasks. In general, the fragment detection and technique-naming tasks were more difficult than the binary classification of propagandist content.

The continuation and expansion of the previous problem was conducted during SemEval2020 Shared Task 4 on the detection of propaganda techniques in news articles [118, 119]. Due to limited examples of certain propaganda techniques, after deleting and merging some of them, the final number was limited to 14. The best results for the span identification task were obtained by employing a heterogeneous pre-trained model [120]. Propaganda technique classification was found to be the most successful when applying a RoBERTa-based model using a semi-supervised learning technique of self-training [121]. Later experiments with a fine-tuned RoBERTa model outperformed scores for the classification task [122].

Further development of a propaganda corpus and automatic propaganda detection field was part of SemEval2023 Task 3, which focused on category (opinion, reporting, or satire), framing (14 generic frames, including economic, morality and political) and persuasion (propaganda) technique detection in online news in different languages [123, 124, 125]. The list of techniques was enlarged to 23, forming six coarse-grained categories. Articles in six European languages, including Polish, were collected and annotated. The best yielding results for the third subtask on persuasion techniques classification were obtained with fine-tuned transformer models, such as XLNet, RoBERTa, XLM-RoBERTa-large or MarianMT [126, 127, 128, 129, 130, 131]. XGBoost and other classic methods were implemented only by two teams and performed poorly [132, 133].

Recently, in connection with the emergence and growing popularity of LLMs, there were several attempts to use commercial models for such a complex task, like propaganda detection. Sprenkamp et al. tested two OpenAI models, namely, *gpt-4* and *gpt-3-davinci*, having the latter one fine-tuned with the input dataset [1]. They reformulated the original task and used the annotated development set as a test set for easier evaluation of the results. The authors claimed to achieve results similar to the state-of-the-art (SOTA) RoBERTa results with *gpt-4*.

Jones [134] delved into the topic of prompt engineering and used a *gpt-3.5-turbo* model to find propaganda techniques in the SemEval2020 dataset, as well as in a new, unannotated set of articles from the Russia Today online news website. His approach focused on a multiclass binary technique detection with an accompanying explanation of LLM and binary propaganda detection based on the appearance of techniques and percentage rating provided by gpt-3.5-turbo. One of the problems is that LLMs were found not to be able to detect the same techniques as humans, but the study suggested they can be used for the initial recognition of possible propaganda.

Lastly, Hasanain et al. prepared a new large annotated dataset for propaganda

detection in an under-resourced langauge – Arabic [135]. They used fine-tuned versions of AraBERT and XLM-RoBERTa, as well as GPT-4, for the detection of 23 propaganda techniques. The fine-tuned models achieved better results than GPT-4 in a zero-shot setting. Additionally, the LLM did not handle the propaganda span identification well.

5.2.3 Propaganda, Media Bias and Fact Checking in Poland

Little do we know about propaganda in Polish media, especially online ones, as it is an unpopular topic, rarely covered by scientific works. There are some examples of qualitative works that studied attitudes of Polish Internet users toward Islamic refugees [136] or the threat posed by Russian disinformation and propaganda in Poland [137]. Closely related, but not in the online sphere, the study of mediatization of politics analyzes popular weekly magazines in Poland and their influence on political processes [138]. The author claimed that in the analyzed period of time, Newsweek magazine could be characterized as a balanced, non-radical and objective medium that avoided propaganda bias, including ideological ones. On the other hand, such magazines like Polityka, which tried to avoid political and propaganda bias, failed to do it at the ideological level. It openly promoted values like a common Europe, equal rights for minorities, supported weaker and underprivileged groups, and was against xenophobia, as well as conservative values. Another example was Wprost, which showed ideological, propaganda and political bias – it criticized all political actors, but varied in intensity. Considering all this, the author still thought that the weekly opinion magazine market in Poland was a good example of external media pluralism due to the representation of various political preferences, ideologies, norms and values from the left to the right. However no such similar work was done on online news outlets. Additionally, propagandist narrative could also be found in the daily newspaper Gazeta Wyborcza, which clearly stood in opposition to the previously ruling Law and Justice party and the President of Poland [139].

When it comes to studies on propaganda in digital media, one study focused on computational propaganda in Poland [140]. This phenomenon can be described as a collection of social media platforms, autonomous agents and big data, whose task is to manipulate public opinion. The quantitative part of this study focused on the analysis of Polish Twitter data and reports that a very small amount of accounts were accountable for a vast spread of fake news. What is more, there were two times more right-wing bots than left-wing ones. One thesis analyzed propaganda in the online news regarding a controversial media law from 2021 called *Lex TVN* [141]. The study proposed a mixed method of propaganda model theory [94] and ways of using propaganda techniques [118]. Content indeed consisted of propaganda across different online news platforms. Due to the limited number of articles checked, as the methods were not automatic, the author suggested further investigation of TVP Info articles, as no propagandist examples were found. Another study focused on fake news appearing regarding COVID-19 in both online news outlets and traditional media in Poland [142]. A rising amount of fake news on the Internet was observed and one of the conclusions included the high need for professional fact checkers in the professional media.

One of the subtasks during SemEval-2023 Task 3 was the detection of persuasion techniques in Online News in different languages, including Polish [123]. It was an extension of the SemEval-2020 Task 11 scope, which was expanded to 23 fine-grained propaganda techniques, which could be grouped into six coarse classes.

The number of works in Polish regarding propaganda, especially in news and online news, is low [102]. However, just as in other countries, one can observe that there is a growth in interest in media bias and fact checking of online news. Media Bias/Fact Check website (MBFC), although being based in America and focusing primarily on their local media, also provides reports of political bias and factual reporting of foreign media outlets, including Polish ones. Table 5.2 presents news outlets that were described in the MBFC website [143, 144, 145, 146, 147].

Table 5.2: Comparison of news outlets in Poland based on Media Bias/Fact Check (MBFC) data

Media Outlet	**Bias Rating**	**Factual Reporting**	**Media Type**	**Traffic / Popularity**	**MBFC Credibility Rating**	**Notes**
TVP Info	Right	Mixed	TV station	High	Medium	`www.tvp.info` (accessed on 20 March 2024)
TVN24	Left-center	High	TV station	High	High	`www.tvn24.pl` (accessed on 20 March 2024)
Gazeta Wyborcza	Left-center	High	Newspaper	High	High	`www.wyborcza.pl` (accessed on 20 March 2024)
FL24.net	Far-right	Mixed	Website	Minimal	Low; questionable reasoning: propaganda, poor sourcing, lack of transparency	No longer exists; operated in France, owned by Polish media outlet
Political Critique	Left-center	High	Website	No information	No information	No longer exists; Polish version: `www.krytykapolityczna.pl` (accessed on 20 March 2024)

According to Similarweb, neither TVP Info nor TVN24 was in the top five most popular Polish online news websites [148], but they were one of the most important TV news providers in Poland [149].

Polish organizations that dealt or are dealing with media bias and fact checking, which could be possibly interested with propaganda detection in the future, are presented in Table 5.3 [102, 150].

Table 5.3: List of Polish organizations that investigate media bias and perform fact checking of news in Polish

Organization	**Year of Creation**	**Website**	**Notes**
Fundacja Reporterów	2010	`www.fundacjareporterow.org` (accessed on 20 March 2024)	indEX and Ukraine Monitor projects observe influence of disinformation on extremist groups
Demagog	2014	`www.demagog.org.pl` (accessed on 20 March 2024)	First Polish fact-checking organization
OKO.press	2016	`www.oko.press` (accessed on 20 March 2024)	Does not allow readers to report suspicious content
Demaskator24	2018	`www.web.archive.org/web/20190615000000*/Demaskator24.pl` (accessed on 20 March 2024)	Only Facebook account exists, no longer updated
Konkret24.pl	2018	`www.konkret24.tvn24.pl` (accessed on 20 March 2024)	Created by the TVN group
Sprawdzam AFP	2019	`www.sprawdzam.afp.com/list` (accessed on 20 March 2024)	Part of Agence France-Presse (AFP; a multilingual, multicultural news agency)
AntyFAKE	2019	`www.antyfake.pl` (accessed on 20 March 2024)	No longer updated
Odfejkuj.info	2020	`www.odfejkuj.info` (accessed on 20 March 2024)	No longer updated
Pravda	2020	`www.pravda.org.pl` (accessed on 20 March 2024)	Fact checking of information, statements and digital content
FakeNews.pl	2020	`www.fakenews.pl` (accessed on 20 March 2024)	International Fact-Checking Network member
#FakeHunter	2020	`www.fake-hunter.pap.pl` (accessed on 20 March 2024)	Created to check news regarding SARS-CoV-2
Zgłoś Trolla	2022	`www.web.archive.org/web/20190615000000*/zglostrolla.pl` (accessed on 20 March 2024)	Created to check news regarding war in Ukraine

List may not be exhaustive.

5.3 Datasets

This subsection describes in detail the datasets that were used in the conducted experiments, namely, the Propaganda Techniques Corpus and a subset of the Polish Online News Corpus.

5.3.1 Propaganda Techniques Corpus

The International Workshop on Semantic Evaluation in 2020 (SemEval 2020) consisted of twelve tasks. As part of the "Societal Applications of NLP" section, Task 11 concerned the "Detection of Propaganda Techniques in News Articles". The main goal of this workshop was to develop the automatic tools to detect the aforementioned techniques. The organizers created the PTC-SemEval20 corpus [118, 106], which consisted of 536 news articles in the final version. Table 5.4 shows the exact distribution of articles.

Table 5.4: PTC-SemEval2020 corpus distribution

Dataset	**News Article Count**
Training set	371
Development set	75
Test set	90

All of articles were annotated by experts. The annotation included a span in which a propaganda technique was used – span identification (SI), which is a binary sequence tagging task, as well as the technique's name – technique classification (TC), which is a multiclass classification problem. Initially, there were 18 propaganda techniques, but due to the scarce appearance of certain categories, similar underrepresented ones were joined together or removed, leaving 14 categories as a result[1]:

- Appeal to authority;
- Appeal to fear/prejudice;
- Bandwagon, reductio ad hitlerum;
- Black-and-white fallacy;
- Causal oversimplification;
- Doubt;
- Exaggeration, minimization;
- Flag-waving;

[1]Descriptions of each category is presented in Appendix A.1.2, in one of the used prompts.

- Loaded language;
- Name calling, labeling;
- Repetition;
- Slogans;
- Thought-terminating cliches;
- Whataboutism, straw men, red herring.

Table 5.5 shows columns in training dataset.

Table 5.5: Training dataset columns

Column Name	**Column Description**
id	Article identification number
technique	Propaganda technique
begin_offset	Beginning of the span (inclusive)
end_offset	End of the span (exclusive)

5.3.2 Polish Online News Corpus (PONC)

For the final experiment presented in this paper, I used a subset of PONC that covered contemporary controversial topics in Poland [3, 31]. The PONC is a collection of online news articles from two leading Polish TV news sources: TVN24 and TVP Info. Controversial topics tend to have more intense emotional charge [151, 152] and I decided to focus on them to look for news articles with examples of propaganda techniques. I prepared another subset of the PONC, i.e., a high-emotional-charge subset, in which I selected the highest value for each of the five emotions, i.e., anger, disgust, fear, happiness and sadness, and I selected the top 9 articles per emotion and per news provider. In total, I select 90 articles, with 45 from TVP Info and 45 from TVN24.

The decision to select the top nine articles per emotion from each news provider was motivated by the need to ensure a balanced and focused analysis of emotionally charged content. By evenly distributing the articles between TVP Info and TVN24, this approach aimed to capture a representative sample of contrasting perspectives, given the ideological differences between the two outlets. Limiting the selection to the most emotionally salient articles allowed for a detailed examination of how each provider frames emotionally charged topics, ensuring that the analysis remains both manageable and relevant to the study's objectives.

5.4 Methods

In this section, I describe the experiments I conducted on the datasets described in Section 5.3. The chosen methods utilized different approaches to SI and

TC tasks.

5.4.1 LLM on SemEval2020 Task 3 – English Data, CoT approach for TC

First, I attempted to reproduce the results obtained by Sprenkamp et al. [1]. I followed the guidelines and ran the code from the authors' Github to confirm the claimed output of *gpt-4* using the *chain of thought* (CoT) method for TC on the specially prepared variant of the SemEval2020 Task 3 dataset [153]. I used prompts provided by the authors – *base*, which only asked for an answer, and *chain of thought*, which required the model to show the reasoning behind the given answer. Both prompts included examples for each of the propaganda techniques, applying the *few-shot* approach. Then, I conducted new experiments with the following models:

- *gpt-4-0125-preview* – as of 26.03.2024, same as *gpt-4-turbo-preview*.
- *gpt-4-1106-preview* – according to OpenAI, should ensure reproducible outputs [154].
- *gpt-3.5-turbo-0125* – as of 26.03.2024, same as *gpt-3.5-turbo*.
- *gpt-3.5-turbo-1106* – according to OpenAI, should ensure reproducible outputs [154].

All the models were run five times – once using the basic prompt type and once including the chain of thought instruction [155]. I compared the three metrics proposed by the authors, namely, F1 score, precision and recall. I also present the F1 scores obtained for all propaganda techniques per model in Section 5.5.2.

5.4.2 LLM on SemEval2020 – English Data

In the second experiment, I used several newer LLMs, namely, *gpt-3.5-turbo-0125* and *gpt-4-0125-preview*, for the SemEval2020 Task 11 subtasks – SI and TC. During the shared task in 2020, only *gpt-2* was used [119].

The instructions of the tasks were as follows [118]:

- Subtask 1 (SI) – given an article, identify specific fragments that contain at least one propaganda technique.
- Subtask 2 (TC) – given a text fragment identified as propaganda and its document context, identify the applied propaganda technique [117].

Having tried the instructions from above, the responses generated were not satisfying since they often did not contain all the requested information and the format required further transformation. Therefore, I prepared my own prompts, which I found to be the most effective in obtaining the desired results. The prompt for TC

can be found in Appendices A.1.1 and A.1.2. I kept the names of the techniques in a format that was required by the organizers of the shared task. Having run the models, I evaluated their performance on the test set and I calculated the F1 score, precision and recall for the SI task, as well as the F1 scores for all propaganda techniques for the TC task.

5.4.3 Propaganda Technique Detection on the PONC subset with the Use of an LLM

My third experiment was based on my original data, which was a subset of the PONC that covered controversial topics. I believe that contentious issues tend to have more debatable descriptions in news, and therefore, I decided to look for examples of propaganda techniques in them. I prompted *gpt-4-0125-preview* to provide us with spans of the text and the full text and the name of propaganda technique included in the given part of the news. I chose *gpt-4-0125-preview* and prompted in Polish (few-shot approach, propaganda technique names are in English, but the propaganda technique examples are in Polish) to provide the most accurate answer. The prompts are listed in Appendices A.1.1, A.1.3 and A.1.4. From the 90 articles with high emotional charge, I randomly took 100 examples of detected propaganda techniques and manually checked to see whether they were correct. I performed the following:

- Binary classification task – whether there was propaganda in the chosen news excerpt; if there was no propaganda technique being used, I marked it as "no propaganda".
- Propaganda technique classification to check whether the correct technique was chosen; if not, I added my comment of the suggested technique.

The annotation was performed by the first author of this article based on her best knowledge on the topic. The guidelines for annotation were taken from *News Categorization, Framing and Persuasion Techniques: Annotation Guidelines* Technical Report [125].

5.5 Results of Propaganda Detection Experiments

This section describes the results of each of the conducted experiments.

5.5.1 LLM on SemEval2020 Task 3 – English Data, experiments with Sprenkamp et al. approach

Table 5.6 presents the results of all the model calculations for the first experiment.

Table 5.6: Results from Sprenkamp et al. experiment [1]

Model	Precision	Recall	F1 Score
Baseline (*gpt-4* CoT [1])	0.56868	0.57821	**0.57340**
Baseline – reproduce attempt	0.46479	**0.64525**	0.54035
gpt-3.5-turbo-0125 base	0.58923	0.48883	0.53435
gpt-3.5-turbo-0125 CoT	0.60700	0.43575	0.50732
gpt-3.5-turbo-1106 base	0.65104	0.34916	0.45455
gpt-3.5-turbo-1106 CoT	0.63934	0.32682	0.43253
gpt-4-0125-preview base	0.53292	0.47486	0.50222
gpt-4-0125-preview CoT	0.58419	0.47486	0.52388
gpt-4-1106-preview base	0.70833	0.04749	0.08901
gpt-4-1106-preview CoT	**0.81818**	0.05028	0.09474

As for the first run, I tried to recreate the result obtained by Sprenkamp et al. to support their stance that the results generated by GPTs are reproducible. As we can see, after my attempt to reproduce the original results, the difference in performance was between 7 and up to 10 percentage points, and I managed to outperform the baseline run with the best recall (64.53%).

Next, I performed the same experiment but with the newer versions of OpenAI models that were not used by the author to see whether the results could be improved and reproducible, as the provider states. After running all the models twice for both prompts, I did not notice any stability of the results and I did not obtain repeated metrics. Moreover, none of the newer models outperformed the score obtained by the authors of this approach in the cases of F1 score and recall. However, I noticed that for *gpt-4-1106-preview chain of thought*, the precision was equal to 81.82%, which was much higher than for the other models, but at the same time, the recall and F1 score were below 10%.

Table 5.7 shows the F1 score for all the propaganda techniques per model.

This was more proof that the GPT-generated results were irreproducible. First of all, the F1 scores did not match the results that were obtained by Sprenkamp et al. The easiest one to detect seemed to be *loaded language*, as the scores were high for every model. The same seemed to be true for *name calling, labeling* and *repetition.* None of the models could correctly predict any instance of the *thought-terminating cliches* category. In conclusion, I do not think that GPT models are a sufficient tool for processing propaganda techniques detection because the results I obtained seemed to be random and no reliable conclusions could be drawn from them.

5.5.2 LLM on SemEval2020 – English Data, results of gpt-3.5-turbo-0125 and gpt-4-0125-preview for SI and TC

First, I ran *gpt-3.5-turbo-0125* and *gpt-4-0125-preview* three times on the SemEval2020 Task 11 test set. I present the results of the first subtask (SI) in Table 5.8.

Technique\Model	Baseline—gpt-4 CoT	Baseline—gpt-4 CoT, Our Attempt	gpt-3.5-turbo-0125 Base	gpt-3.5-turbo-0125 CoT	gpt-3.5-turbo-1106 Base	gpt-3.5-turbo-1106 CoT	gpt-4-1106-Preview Base	gpt-4-1106-Preview CoT	gpt-4-0125-Preview base	gpt-4-0125-Preview CoT
Appeal to authority	0.19048	0.24000	0.32432	0.22857	0.11765	**0.31579**	0.00000	0.00000	0.40000	0.24000
Appeal to fear/prejudice	0.00000	0.00000	0.48276	0.00000	**0.54545**	0.00000	0.00000	0.00000	0.53333	0.00000
Bandwagon, reductio ad hitlerum	0.00000	**0.12500**	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
Black-and-white fallacy	0.00000	0.00000	0.18182	0.00000	0.00000	0.00000	0.00000	0.00000	**0.28571**	0.00000
Causal oversimplification	**0.50000**	0.37500	0.29630	0.31579	0.11111	0.12500	0.00000	0.00000	0.41509	0.37500
Doubt	0.54545	0.53465	0.24390	0.56757	0.00000	0.41509	0.06667	0.06897	0.55172	**0.57447**
Exaggeration, minimization	0.64000	**0.64706**	0.35088	0.057142	0.05128	0.00000	0.00000	0.00000	0.56250	0.60274
Flag-waving	0.00000	0.00000	0.15385	0.00000	0.10811	0.00000	0.00000	0.00000	**0.23256**	0.00000
Loaded language	**0.93617**	0.93617	0.90625	0.89600	0.77876	0.71698	0.30380	0.34568	0.84034	0.93233
Name calling, labeling	0.74286	0.72165	**0.79630**	0.73684	0.76636	0.76364	0.07407	0.00000	0.57895	0.66667
Repetition	0.65789	0.67327	0.60274	**0.68235**	0.59459	0.53846	0.10000	0.15000	0.37500	0.47273
Slogans	0.10000	**0.23529**	0.10000	0.09523	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
Thought-terminating cliches	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
Whataboutism, straw men, red herring	0.33333	**0.45283**	0.43243	0.27273	0.08333	0.09091	0.00000	0.00000	0.16667	0.33333

Table 5.7: F1 scores for each of the propaganda techniques

Table 5.8: Task span identification (SI) – results

Model	**Precision**	**Recall**	**F1**
Hitachi	0.56544	0.47368	0.51551
ApplicaAI	0.59954	0.41650	0.49153
gpt-4-0125-preview, try 1	0.16059	0.20109	0.17857
gpt-4-0125-preview, try 2	0.16659	0.23174	0.19384
gpt-4-0125-preview, try 3	0.16595	0.23174	0.19340
gpt-3.5-turbo-0125, try 1	0.15132	0.06707	0.09294
gpt-3.5-turbo-0125, try 2	0.17226	0.06583	0.09525
gpt-3.5-turbo-0125, try 3	0.16172	0.07142	0.09908
baseline	0.00320	0.13045	0.00162

Due to limited space, only the top 2 results from the SemEval2020 Task 11 shared task are presented.

None of my attempts came close to the best results from the shared task. *gpt-4-0125-preview* performed better than *gpt-3.5-turbo-0125*, and both models were better than the baseline.

Next, I checked the annotation for the second subtask (TC) with the golden set and show the results in Table 5.9.

The overall F1 score for *gpt-3.5-turbo-0125* turned out to be the best and outperformed the baseline's value, but all the values oscillated between 20% and 30%. Again, as in the first experiment, *loaded language* proved to be the easiest

Table 5.9: Task technique classification (TC) – F1 scores

Technique – Model	**Baseline**	**gpt-4-0125-Preview**	**gpt-3.5-turbo-0125**
F1 score	0.25196	0.23352	**0.30000**
Appeal to authority	0.00000	0.00000	0.00000
Appeal to fear/prejudice	0.03681	**0.05128**	0.01449
Bandwagon, reductio ad hitlerum	0.00000	0.00000	0.00000
Black-and-white fallacy	0.00000	**0.04211**	0.03125
Causal oversimplification	0.11561	0.03922	0.00000
Doubt	0.29143	0.03265	0.00995
Exaggeration, minimization	0.14420	0.05882	0.06329
Flag-waving	0.06195	0.04878	0.01869
Loaded language	0.46477	0.43468	**0.47970**
Name calling, labeling	0.00000	0.08458	0.02778
Repetition	0.19262	0.02643	0.03550
Slogans	0.00000	0.00000	0.00000
Thought-terminating cliches	0.00000	0.00000	0.00000
Whataboutism, straw men, red herring	0.00000	0.00000	0.00000

to detect, and the best result was obtained by *gpt-3.5-turbo-0125*. I could also observe a slight improvement in the detection of *appeal to fear/prejudice* and *causal oversimplification* for *gpt-4-0125-preview*, and for both LLMs for the *black-and-white fallacy* and *name calling, labeling* categories. None of the approaches could handle *appeal to authority*; *bandwagon, reductio ad hitlerum*; *slogans*; *thought-terminating cliches*; nor *whataboutism, straw men, red herring* techniques. Such F1 scores were the result of unbalanced data, in which some techniques had a larger number of occurrences. In other words, I did not see much improvement in comparison with the baseline results, and they were notably worse than the ones obtained by the shared task participants [118].

5.5.3 Propaganda Technique Detection in PONC Subset with the Use of LLM

Having evaluated the news fragments that were chosen by *gpt-4-0125-preview* as being propaganda techniques examples, I concluded the following based on the sample of 100 randomly selected excerpts:

- A total of 26 out of 100 fragments were marked by the annotator as not propaganda (accuracy = 74%).
- A total of 23 out of 74 examples of propaganda were marked as the *wrong propaganda technique classification* (accuracy = 69%).
- The most popular techniques were *appeal to fear/prejudice* (22) and *loaded language* (21).
- There were no examples of *repetition* nor *whataboutism, straw men, red herring.*

5.6 Error Analysis

For the second experiment, it is worth noting that both the *gpt-3.5-turbo-0125* and *gpt-4-0125-preview* models required error handling due to undesired outputs. No matter how precise the instructions in the prompt were, at times, the generated text did not match the appropriate label format. For the span identification (SI) and technique classification (TC) tasks, I encountered the following errors:

- The generated technique name was not included in the provided list.
- The generated technique name was more granulated, e.g., *whataboutism* instead of *whataboutism, straw men, red herring.*
- The output was a description of the used technique instead of the label.

Additionally, for the TC task, the format of the output required by the submission website with gold labels was strict and needed to have all the technique

names filled, even if the model did not find any. I replaced the incorrect techniques and missing values with *loaded language*, as this was the technique with the highest frequency of occurrence in the training set.

Below, I present a couple of examples of mistakes made by *gpt-4-0125-preview*:

- "I appeal to the government, to the Prime Minister, to all those who make decisions." (original: *Apeluję do rządu, do premiera, do wszystkich, którzy podejmują decyzje.*) – mistakenly marked as *appeal to authority* due to the use of the word *appeal* and mentioning examples of authorities, such as prime minister or the government.
- Fragment "perhaps without the 'Boleks', i.e., without the opposition activists who secretly collaborated with the political police, many revolutions would have had a bloodier course" was marked as *bandwagon*, but I think it should be noted as *name calling, labeling* (original: *być może bez 'Bolków', czyli bez działaczy opozycyjnych, którzy tajnie współpracowali z policją polityczną, wiele rewolucji miałoby bardziej krwawy przebieg*).
- One fragment concerned the Polish and Belarusian border crisis and included emojis of flags of both countries. It was mistakenly marked as *flag-waving*.

5.7 Discussion

Having obtained the results from the first experiment, I can raise the following open points for further discussion:

- As a general observation, I can say that *gpt-4-0125-preview* was often unable to output an accurate span for a propaganda technique – some selected fragments were too long and the additional text did not include any valuable context to better understand the detected propaganda technique.
- Although the temperature was set to "0", which should provide more deterministic results, for the given prompt and task, various GPT models were unable to generate the same results; therefore, the blue method should be considered not reproducible.
- In the original paper by Sprenkamp et al. [1], it was not mentioned whether the models were run several times, but I can assume it was done only once, and thus, the results are not trustworthy.
- In the same paper, there was no mention about error analysis nor any specific mistakes that the models made when predicting the propaganda techniques.
- Reformulation of the original SemEval2020 Task 3 and the use of the annotated development set as a test set was an example of data contamination – there could be a high risk that the models were trained on this data and it was the reason for significantly better results. The experiment should be conducted on the original test set for which the gold labels were not released to the public.

The second experiment was also limited in one aspect – gold datasets were not publicly available and the submission system had a strict input data format requirement; post-processing was required in cases where the models did not detect any propaganda technique. In such instances, I decided to replace the missing values with the most numerous technique from the training dataset, that is, *loaded language*. This also raised a problem of a scarce number of annotated datasets in this field.

The methodology proposed by Jones was not implemented within the scope of the experimental procedures [134]. Due to the complicated nature of the task of propaganda detection, I can expect that the results of this study are not reproducible. What is more, the author used the annotated data that was possibly used for pre-training GPTs by OpenAI; therefore, I could have another example of data contamination.

The third experiment gave some hope for LLMs, such as *gpt-4-0125-preview*, to be a possible propaganda technique detection tool, but the results are limited on some levels. First of all, a larger number of examples than 100 should be manually checked by human annotators (preferably experts) to verify the credibility. Additionally, more annotators would allow for comparing the results and reducing the bias, for example, by calculating the inter-coder agreement. Second, fine-tuning the models with examples of Polish propaganda in news could enhance the results of detection and classification.

5.8 Conclusion

My work shows the results of various experiments with the use of LLMs in propaganda detection tasks. The results show that the outputs of generative models were unpredictable, even if the parameters were set in such a way that should ensure reproducibility. I believe that at this stage, it is too soon to confidently use LLMs for such complex tasks, and other methods should be used for problems that require deeper reasoning. At the same time, further enhancements of LLMs can bring new capabilities, as just scaling such models have shown visible improvement on various NLP benchmarks [156].

One of the biggest obstacles for current propaganda detection studies is the lack of datasets that have full open access and are reliably annotated. I see potential in the further annotation of the PONC subset that could be beneficial for future studies. In order to reduce the cost and workload of such a task, I think that it is possible to use LLMs, such as *gpt-4-0125-preview*, as a method for selecting more examples for the training and testing of a propaganda detection task. Finally, I believe that this approach at a fine-grained level (detecting the exact span with the propaganda technique) might still be difficult for LLMs, but the coarse-grained approach of the binary detection of propagandist news could already be implemented in organizations that fight against misinformation and disinformation. As part of my findings, I also provide an extensive list of organizations that deal with misinformation in online news in Poland that could potentially be interested in automatic propaganda detection. I additionally discovered that GPT models can generate concise outputs in Python code that is easy to process for analyses.

Chapter 6

Conclusion of the Thesis

In this chapter, I will begin by providing a brief summary of the thesis, including its findings and contributions. Following that, I will discuss potential for expanding the work presented in the thesis, leading to its conclusion.

6.1 Summary of Contributions

This dissertation makes several significant contributions to the study of political bias, sentiment analysis, and propaganda detection in Polish online news, particularly in the context of under-resourced languages:

1. Creation of the Polish Online News Corpus (PONC)

A comprehensive dataset of online news articles from two major Polish news outlets, TVP Info and TVN24, was developed. The dataset includes articles spanning controversial topics, enabling detailed analysis of sentiment, bias, and propaganda. This corpus represents a foundational resource for future research in Polish computational linguistics.

2. Sentiment Analysis of Controversial Topics

A comparative analysis was conducted using lexicon-based and statistical approaches to examine emotive content in articles on controversial topics. This work provides insights into the differences in sentiment between news providers and highlights the applicability of these methods to longer texts in Polish.

3. Bias Detection Using Large Language Models (LLMs)

Experiments with LLMs, such as *GPT-3.5-turbo* and *GPT-4*, demonstrated their potential for detecting political leanings and establishment biases in online news. The findings shed light on the challenges of applying LLMs to under-resourced languages and their effectiveness compared to human annotation.

4. Propaganda Detection

The study explored LLMs' capability to identify propaganda techniques

in news articles. A novel methodology for dataset annotation and analysis was introduced, providing a foundation for automating propaganda detection tasks in Polish.

5. Methodological Advancements for Under-Resourced Languages

The dissertation highlights the limitations and potential of LLMs for tasks in under-resourced languages like Polish, offering methodological insights for adapting state-of-the-art NLP techniques to such contexts.

These contributions advance the understanding of media bias and polarization in Polish news and establish a robust framework for future research in computational analysis of online news.

6.2 Future Work

Several potential directions for extending and deepening the research presented in this thesis have been identified. One possibility is the expansion of the corpus by incorporating articles from other online news outlets from the same period, allowing for a more comprehensive analysis of the political spectrum within Polish media. This broader dataset would provide a fuller understanding of the political gradation present in the media landscape.

Another direction involves analyzing the dynamics of emotional load within the articles, not just focusing on the dominant sentiment but also considering the overall emotional intensity. This could be achieved by exploring other sentiment dictionaries available in Polish, which would enhance the lexicon-based sentiment analysis used in this work.

In addition, extending the analysis of political and establishment bias to full-length online articles, rather than limiting the examination to the initial paragraphs, would offer deeper insights. To further improve the reliability and breadth of the analysis, increasing the number of annotators through crowd-sourcing could prove beneficial.

Moreover, usage of other large language models (LLMs) such as Gemini, Llama, Claude, Falcon, or QRA for the propaganda detection task could yield different perspectives and enhance the accuracy of the detection. Engaging an expert in Polish propaganda detection to assist with labeling the PONC articles would provide a valuable resource for future use, potentially transforming the dataset into a valuable training set for further studies.

Applying explainable artificial intelligence (XAI) techniques would allow for more detailed analysis of the model outputs. XAI could help researchers better understand and interpret the decisions made by the models, particularly in the context of detecting propaganda, thus enriching the overall contribution of the research.

Lastly, recent advancements in fine-tuning techniques [157], such as Direct Preference Optimization (DPO) [158], offer promising avenues for enhancing model

performance in tasks involving bias detection and sentiment analysis. These methods enable models to better align with specific objectives by incorporating human preference data, thereby improving contextual understanding and reducing inconsistencies in output. While this dissertation primarily leveraged zero-shot and in-context learning approaches, integrating fine-tuning techniques like DPO could significantly enhance model accuracy and robustness, particularly in under-resourced languages like Polish. Future work could explore these methodologies to refine the models' ability to identify subtle biases and complex propaganda techniques, further advancing the field.

Bibliography

[1] Kilian Sprenkamp, Daniel Gordon Jones, and Liudmila Zavolokina. Large Language Models for Propaganda Detection, 2023.

[2] Brussels European Parliament. Flash Eurobarometer 3153 (Media & News Survey 2023). GESIS, Cologne. ZA8777 Data file Version 1.0.0, https://doi.org/10.4232/1.14244, 2024.

[3] Joanna Szwoch, Mateusz Staszkow, Rafal Rzepka, and Kenji Araki. Creation of Polish Online News Corpus for Political Polarization Studies. In *Proceedings of the LREC 2022 workshop on Natural Language Processing for Political Sciences*, June 2022.

[4] Kaarle Nordenstreng, Svetlana Pasti, Tao Zhang, Savyasaachi Jain, Giuliano Bobba, Henry Wolgast, Aaron Hyzen, and Liziane Guazina. Coverage of the Russia-Ukraine War by Television News. *International Journal of Communication*, pages 6857–6873, 11 2023.

[5] Eli Pariser. *The filter bubble: What the Internet is hiding from you.* Penguin UK, 2011.

[6] Sabina Cisek and Monika Krakowska. The filter bubble: a perspective for information behaviour research, 10 2018.

[7] Kiran Garimella, Tim Smith, Rebecca Weiss, and Robert West. Political Polarization in Online News Consumption, 04 2021.

[8] Samantha D'Alonzo and Max Tegmark. Machine-Learning media bias. *CoRR*, abs/2109.00024, 2021.

[9] Howard D. Wactlar, Alexander Hauptmann, and M. Witbrock. InformediaTM: News-on-Demand experiments in speech recognition. 1998.

[10] Mari Ostendorf, Patti Price, and Stefanie Shattuck-Hufnagel. Boston University Radio speech Corpus LDC96S36. Web Download, 1996.

[11] Jayson Victoriano, Jaime Pulumbarit, and Luisito Lacatan. Data analysis of BulSU faculty research engagement based on Google Scholar data using web data scrapping technique. *International Journal of Computing Sciences Research*, 26:1–12, 01 2022.

[12] Ikechukwu E. Onyenwe, Ebele G. Onyedinma, Chidinma A. Nwafor, and Obinna Agbata. Developing Products Update-Alert System for e-Commerce Websites Users Using HTML Data and Web Scraping Technique. *CoRR*, abs/2109.00656, 2021.

[13] Daniel Hladek, Jan Stas, and Jozef Juhar. The Slovak Categorized News Corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1705–1708, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA).

[14] Wisam A. Qader, Musa M. Ameen, and Bilal I. Ahmed. An Overview of Bag of Words; Importance, Implementation, Applications, and Challenges. In *2019 International Engineering Conference (IEC)*, pages 200–204, 2019.

[15] Bhartendoo Vimal. Application of Logistic Regression in Natural Language Processing. *International Journal of Engineering Research and*, V9, 06 2020.

[16] Aviel Stein, Janith Weerasinghe, Spiros Mancoridis, and Rachel Greenstadt. News Article Text Classification and Summary for Authors and Topics. pages 1–12, 11 2020.

[17] Yohan Muliono and Fidelson Tanzil. A Comparison of Text Classification Methods k-NN, Naïve Bayes, and Support Vector Machine for News Classification. *Jurnal Informatika: Jurnal Pengembangan IT*, 3:157–160, 05 2018.

[18] Seppe Van Den Broucke and Bart Baesens. *Practical Web Scraping for Data Science: Best Practices and Examples with Python. Examples*, pages 197–298. Apress, Berkeley, CA, 2018.

[19] Leen Al Qadi, Hozayfa El Rifai, Safa Obaid, and Ashraf Elnagar. Arabic Text Classification of News Articles Using Classical Supervised Classifiers. pages 1–6, 10 2019.

[20] Peiyang Yu, Victor Cui, and Jiaxin Guan. Text Classification by using Natural Language Processing. *Journal of Physics: Conference Series*, 1802:042010, 03 2021.

[21] Dimitris Liparas, Yaakov HaCohen-Kerner, Anastasia Moumtzidou, Stefanos Vrochidis, and Ioannis Kompatsiaris. News Articles Classification Using Random Forests and Weighted Multimodal Features. volume 8849, 11 2014.

[22] Tej Bahadur Shahi and Ashok Kumar Pant. Nepali news classification using Naïve Bayes, Support Vector Machines and Neural Networks. In *2018 International Conference on Communication information and Computing Technology (ICCICT)*, pages 1–5, 2018.

[23] Aye Hninn Khine and Khin Thandar Nwet. Automatic Myanmar News Classification Using Naiive Bayes Classifier. 2016.

[24] Kathrin Blagec, Georg Dorffner, Milad Moradi, and Matthias Samwald. A critical analysis of metrics used for measuring progress in artificial intelligence. *CoRR*, abs/2008.02577, 2020.

[25] Rafał Klepka. *Ewolucja Wiadomości TVP1: od medialnej stronniczości do propagandy politycznej?*, pages 244–253. 03 2017.

[26] Agnieszka Węglińska, Łukasz Szurmiński, and Maria Wąsicka-Sroczyńska. Politicization as a Factor of Shaping News in the Public Service Media : A Case Study on Public Television in Poland. *Athenaeum Polskie Studia Politologiczne*, 72:29–51, 12 2021.

[27] Hanna Batorowska, Olga Wasiuta, and Rafał Klepka. *Media jako instrument wpływu informacyjnego i manipulacji społeczeństwem*. 02 2019.

[28] Mohammed Alenezi and Zainab Alqenaei. Machine Learning in Detecting COVID-19 Misinformation on Twitter. *Future Internet*, 13:244, 09 2021.

[29] David Bracewell, Jiajun Yan, Fuji Ren, and Shingo Kuroiwa. Category Classification and Topic Discovery of Japanese and English News Articles. *Electr. Notes Theor. Comput. Sci.*, 225:51–65, 01 2009.

[30] Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. Deep learning–based text classification: a comprehensive review. *ACM Computing Surveys (CSUR)*, 54(3):1–40, 2021.

[31] Joanna Szwoch, Mateusz Staszkow, Rafal Rzepka, and Kenji Araki. Sentiment Analysis of Polish Online News Covering Controversial Topics – Comparison Between Lexicon and Statistical Approaches. In *Proceedings of the 10th Language & Technology Conference (LTC 2023): Human Language Technologies as a Challenge for Comoputer Science and Linguistics - 2023*, pages 277–281, Poznań, Poland, April 2023. Wydawnictwo Naukowe UAM.

[32] Laura Burbach, Patrick Halbach, Martina Ziefle, and André Calero Valdez. Opinion Formation on the Internet: The Influence of Personality, Network Structure, and Content on Sharing Messages Online. *Frontiers in Artificial Intelligence*, 3, 2020.

[33] Erick Elejalde, Leo Ferres, and Eelco Herder. On the nature of real and perceived bias in the mainstream media. *PLoS ONE*, 13, 2018.

[34] Swati Aggarwal, Tushar Sinha, Yash Kukreti, and Siddarth Shikhar. Media bias detection and bias short term impact assessment. *Array*, 6:100025, 2020.

[35] Pansy Nandwani and Rupali Verma. A review on sentiment analysis and emotion detection from text. *Social Network Analysis and Mining*, 11, 12 2021.

[36] Paul Machete and Marita Turpin. The Use of Critical Thinking to Identify Fake News: A Systematic Literature Review. *Responsible Design, Implementation and Use of Information and Communication Technology*, 12067:235 – 246, 2020.

[37] Mika Mäntylä, Daniel Graziotin, and Miikka Kuutila. The Evolution of Sentiment Analysis - A Review of Research Topics, Venues, and Top Cited Papers. *Computer Science Review*, 27, 12 2016.

[38] Jingfeng Cui, Zhaoxia Wang, Seng Ho, and Erik Cambria. Survey on sentiment analysis: evolution of research methods and topics. *Artificial intelligence review*, pages 1–42, 01 2023.

[39] Vishal S. Shirsat, Rajkumar S. Jagdale, and S. N. Deshmukh. Document Level Sentiment Analysis from News Articles. In *2017 International Conference on Computing, Communication, Control and Automation (ICCUBEA)*, pages 1–4, 2017.

[40] Caio Mello, Gullal Cheema, and Gaurish Thakkar. Combining sentiment analysis classifiers to explore multilingual news articles covering London 2012 and Rio 2016 Olympics. *International Journal of Digital Humanities*, pages 1–27, 11 2022.

[41] Tim Loughran. When Is a Liability NOT a Liability? Textual Analysis, Dictionaries, and 10-Ks. *The Journal of Finance*, 66:35 – 65, 02 2011.

[42] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May 2010. European Language Resources Association (ELRA).

[43] Finn Årup Nielsen. A new ANEW: evaluation of a word list for sentiment analysis in microblogs. In Matthew Rowe, Milan Stankovic, Aba-Sah Dadzie, and Mariann Hardey, editors, *Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages*, volume 718 of *CEUR Workshop Proceedings*, pages 93–98, May 2011.

[44] Michelle Terblanche and Vukosi Marivate. Loughran McDonald-SA-2020 Sentiment Word List, April 2021.

[45] Lori Young and Stuart Soroka. Affective News: The Automated Coding of Sentiment in Political Texts. *Political Communication*, 29(2):205–231, 2012.

[46] Krzysztof Tomanek. *Sentiment analysis: history and development of the method within CAQDAS (in Polish: Analiza sentymentu: historia i rozwój metody w ramach CAQDAS.* 01 2014.

[47] Jan Kocoń, Maciej Piasecki, and Monika Zaśko-Zielińska. PlWordNet as a Basis for Large Emotive Lexicons of Polish. In *Human language technologies as a challenge for computer science and linguistics: 8th Language & Technology Conference*, pages 189–193, 11 2017.

[48] Lingyun Zhao, Lin Li, and Xinhao Zheng. A BERT based Sentiment Analysis and Key Entity Detection Approach for Online Financial Texts. *2021 IEEE 24th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, pages 1233–1238, 01 2020.

[49] Rosario Catelli, Serena Pelosi, and Massimo Esposito. Lexicon-Based vs. BERT-Based Sentiment Analysis: A Comparative Study in Italian. *Electronics*, 11:374, 01 2022.

[50] Eryka Probierz and Adam Galuszka. Emotion Detection Based on Sentiment Analysis: An Example of a Social Robots on Short and Long Texts Conversation. *European Research Studies Journal*, XXV:135–144, 05 2022.

[51] Saif Mohammad and Peter Turney. NRC emotion lexicon. *National Research Council of Canada*, pages 1–234, 11 2013.

[52] Andraž Pelicon, Marko Pranjić, Dragana Miljkovic, Blaž Škrlj, and Senja Pollak. Zero-Shot Learning for Cross-Lingual News Sentiment Classification. *Applied Sciences*, 10:5993, 08 2020.

[53] Monika Riegel, Małgorzata Wierzba, Marek Wypych, Łukasz Żurawski, Katarzyna Jednoróg, Anna Grabowska, and Artur Marchewka. Nencki Affective Word List (NAWL): the cultural adaptation of the Berlin Affective Word List–Reloaded (BAWL-R) for Polish. *Behavior Research Methods*, 01 2015.

[54] Sara Dellantonio, Remo Job, and Claudio Mulatti. Imageability: now you see it again (albeit in a different form). *Frontiers in Psychology*, 5:279, 04 2014.

[55] A. Paivio. *Mental Representations: A Dual Coding Approach*. Oxford Psychology Series. Oxford University Press, 1990.

[56] Melissa Võ, Markus Conrad, Lars Kuchinke, Karolina Urton, Markus Hofmann, and Arthur Jacobs. The Berlin affective word list reloaded (bawl-r). *Behavior research methods*, 41:534–8, 05 2009.

[57] Christy Horner, Dennis Galletta, Jennifer Crawford, and Abhijeet Shirsat. Emotions: The unexplored fuel of fake news on social media. *Journal of Management Information Systems*, 38:1039–1066, 10 2021.

[58] Joanna Baran, Michał Kajstura, Maciej Ziolkowski, and Krzysztof Rajda. Does Twitter know your political views? POLiTweets dataset and semi-automatic method for political leaning discovery. In *Proceedings of the LREC 2022 workshop on Natural Language Processing for Political Sciences*, June 2022.

[59] Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. How to Fine-Tune BERT for Text Classification? In Maosong Sun, Xuanjing Huang, Heng Ji, Zhiyuan Liu, and Yang Liu, editors, *Chinese Computational Linguistics*, 2019.

[60] Joanna Szwoch, Mateusz Staszkow, Rafal Rzepka, and Kenji Araki. Can LLMs Determine Political Leaning of Polish News Articles? In *In the proceedings of the 10th IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE'23)*, Yanuca Island, Fiji, 2024.

[61] Xiangli Yang, Zixing Song, Irwin King, and Zenglin Xu. A Survey on Deep Semi-Supervised Learning. *IEEE Transactions on Knowledge and Data Engineering*, 35(9):8934–8954, September 2023.

[62] Qin Ruan, Brian Mac Namee, and Ruihai Dong. Bias Bubbles: Using Semi-Supervised Learning to Measure How Many Biased News Articles Are Around Us. In *Irish Conference on Artificial Intelligence and Cognitive Science*, 2021.

[63] Aleksandar Vujinović, Nikola Luburić, Jelena Slivka, and Aleksandar Kovacevic. Using ChatGPT to Annotate a Dataset: A Case Study in Intelligent Tutoring Systems, 07 2023.

[64] Taylor Sorensen, Joshua Robinson, Christopher Rytting, Alexander Shaw, Kyle Rogers, Alexia Delorey, Mahmoud Khalil, Nancy Fulda, and David Wingate. An Information-theoretic Approach to Prompt Engineering Without Ground Truth Labels. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2022.

[65] Xingwei He, Zhenghao Lin, Yeyun Gong, A-Long Jin, Hang Zhang, Chen Lin, Jian Jiao, Siu Ming Yiu, Nan Duan, and Weizhu Chen. AnnoLLM: Making Large Language Models to Be Better Crowdsourced Annotators. In Yi Yang, Aida Davani, Avi Sil, and Anoop Kumar, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, pages 165–190, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[66] Taja Kuzman, Igor Mozetič, and Nikola Ljubešić. ChatGPT: Beginning of an End of Manual Linguistic Data Annotation? Use Case of Automatic Genre Identification, 2023.

[67] Vahid Ghafouri, Vibhor Agarwal, Yong Zhang, Nishanth Sastry, Jose Such, and Guillermo Suarez-Tangil. AI in the Gray: Exploring Moderation Policies in Dialogic Large Language Models vs. Human Answers in Controversial Topics. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, CIKM '23. ACM, October 2023.

[68] David Rozado. The Political Preferences of LLMs, 2024.

[69] Partha Pratim Ray. ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*, 3:121–154, 2023.

[70] Jochen Hartmann, Jasper Schwenzow, and Maximilian Witte. The political ideology of conversational AI: Converging evidence on ChatGPT's pro-environmental, left-libertarian orientation. *SSRN Electronic Journal*, 01 2023.

[71] Robert McGee. Is Chat Gpt Biased Against Conservatives? An Empirical Study. *SSRN Electronic Journal*, 01 2023.

[72] Daniel Vogler. Medienstrukturen. In Forschungszentrum Öffentlichkeit und Gesellschaft (fög), editor, *Qualität der Medien. Schweiz - Suisse - Svizzera. Jahrbuch 2017*, pages 23–39. Schwabe, Basel, October 2017. Herausgegeben vom Forschungszentrum Öffentlichkeit und Gesellschaft (fög) im Auftrag der Kurt Imhof Stiftung für Medienqualität, Zürich.

[73] Renáta Németh. A scoping review on the use of natural language processing in research on political polarization: trends and research prospects. *Journal of Computational Social Science*, 6(1):289–313, April 2023.

[74] Dorota Piontek. The Tabloidization of Political Discourse. The Polish Case. *"Central European Journal of Communication"*, 4:275–292, 01 2011.

[75] Jan Kocoń, Igor Cichecki, Oliwier Kaszyca, Mateusz Kochanek, Dominika Szydło, Joanna Baran, Julita Bielaniewicz, Marcin Gruza, Arkadiusz Janz, Kamil Kanclerz, Anna Kocoń, Bartłomiej Koptyra, Wiktoria Mieleszczenko-Kowszewicz, Piotr Miłkowski, Marcin Oleksy, Maciej Piasecki, Łukasz Radliński, Konrad Wojtasik, Stanisław Woźniak, and Przemysław Kazienko. ChatGPT: Jack of all trades, master of none. *Information Fusion*, 99:101861, 2023.

[76] Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30), jul 2023.

[77] Timo Spinde, Manuel Plank, Jan-David Krieger, Terry Ruas, Bela Gipp, and Akiko Aizawa. Neural Media Bias Detection Using Distant Supervision With BABE - Bias Annotations By Experts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1166–1177, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[78] Klaus Krippendorff. Computing Krippendorff's Alpha-Reliability. 2011.

[79] Joanna Szwoch, Mateusz Staszkow, Rafal Rzepka, and Kenji Araki. Limitations of Large Language Models in Propaganda Detection Task. *Applied Sciences*, 14(10):1–22, 2024.

[80] Tim Groeling. Media Bias by the Numbers: Challenges and Opportunities in the Empirical Study of Partisan News. *Annual Review of Political Science*, 16(Volume 16, 2013):129–151, 2013.

[81] Zoë Adams, Magda Osman, Christos Bechlivanidis, and Björn Meder. (why) Is Misinformation a Problem? *Perspectives on Psychological Science*, 18(6):1436–1463, 2023. PMID: 36795592.

[82] Tomislav Levak. Disinformation in the New Media System – Characteristics, Forms, Reasons for its Dissemination and Potential Means of Tackling the Issue; Dezinformacije u novomedijskom sustavu – značajke, oblici, razlozi širenja i potencijalni načini njihova suzbijanja. *Medijska istraživanja*, 26:29–58, 01 2021.

[83] Elena Kotelenets and Viktor Barabash. Propaganda and Information Warfare in Contemporary World: Definition Problems, Instruments and Historical Context. In *Proceedings of the International Conference on Man-Power-Law-Governance: Interdisciplinary Approaches (MPLG-IA 2019)*, pages 374–377. Atlantis Press, 2019/12.

[84] Tamar Mitts, Gregoire Phillips, and Barbara F. Walter. Studying the Impact of ISIS Propaganda Campaigns. *The Journal of Politics*, 84(2):1220–1225, 2022.

[85] Miroslava Pavlíková, Barbora Šenkýřová, and Jakub Drmola. *Propaganda and Disinformation Go Online*, pages 43–74. Springer International Publishing, Cham, 2021.

[86] Irina Khaldarova and Mervi Pantti. Fake News. *Journalism Practice*, 10(7):891–901, 2016.

[87] Samuel Woolley. Digital Propaganda: The Power of Influencers. *Journal of Democracy*, 33:115–129, 07 2022.

[88] Media Bias/Fact Check - Poland Government and Media Profile, 2023. `https://mediabiasfactcheck.com/poland-media-profile/` Accessed on 20.03.2024.

[89] Reporters Without Borders - Poland, 2023. `https://rsf.org/en/country/poland` Accessed on 20.03.2024.

[90] E.L. Bernays. *Propaganda*. H. Liveright, 1928.

[91] J. Stanley. *How Propaganda Works*. Princeton University Press, 2015.

[92] J. Ellul. *Propaganda: The Formation of Men's Attitudes*. A Borzoi book. Knopf Doubleday Publishing Group, 1965.

[93] N. Chomsky. *Media Control: The Spectacular Achievements of Propaganda*. Open Media Series. Seven Stories Press, 2002.

[94] E.S. Herman and N. Chomsky. *Manufacturing Consent: The Political Economy of the Mass Media*. Pantheon Books, 2002.

[95] Amos Tversky and Daniel Kahneman. *Judgment under Uncertainty: Heuristics and Biases*, pages 141–162. Springer Netherlands, Dordrecht, 1975.

[96] Felix Hamborg, Karsten Donnay, and Bela Gipp. Automated identification of media bias in news articles: an interdisciplinary literature review. *International Journal on Digital Libraries*, pages 1–25, 2018.

[97] Preslav Nakov, Husrev Taha Sencar, Jisun An, and Haewoon Kwak. A Survey on Predicting the Factuality and the Bias of News Media, 2021.

[98] Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. A Survey on Automated Fact-Checking. *Transactions of the Association for Computational Linguistics*, 10:178–206, 2022.

[99] Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. Truth of Varying Shades: Analyzing Language in Fake News and Political Fact-Checking. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel, editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2931–2937, Copenhagen, Denmark, 09 2017. Association for Computational Linguistics.

[100] Esma Aimeur, Sabrine Amri, and Gilles Brassard. Fake news, disinformation and misinformation in social media: a review. *Social Network Analysis and Mining*, 13, 02 2023.

[101] Firoj Alam, Stefano Cresci, Tanmoy Chakraborty, Fabrizio Silvestri, Dimiter Dimitrov, Giovanni Da San Martino, Shaden Shaar, Hamed Firooz, and Preslav Nakov. A Survey on Multimodal Disinformation Detection, 2022.

[102] Czym Jest Fact-Checking? – Zarys Inicjatyw na Świecie i w Polsce (English: What is Fact-Checking? - Outline of Initiatives in the World and in Poland), 2019. `https://cyberpolicy.nask.pl/czym-jest-fact-checking-zarys-inicjatyw-na-swiecie-i-w-polsce/` Accessed on 20.03.2024.

[103] R. Paul, L. Elder, and Foundation for Critical Thinking. *The Thinker's Guide for Conscientious Citizens on how to Detect Media Bias & Propaganda in National and World News: Based on Critical Thinking Concepts & Tools.* Thinker's guide series. Rowman & Littlefield, 2004.

[104] Kung-Hsiang Huang, Kathleen McKeown, Preslav Nakov, Yejin Choi, and Heng Ji. Faking Fake News for Real Fake News Detection: Propaganda-loaded Training Data Generation, 2023.

[105] Alberto Barrón-Cedeño, Israa Jaradat, Giovanni Da San Martino, and Preslav Nakov. Proppy: Organizing the news based on their propagandistic content. *Information Processing & Management*, 56(5):1849–1864, 2019.

[106] Giovanni Da San Martino, Seunghak Yu, Alberto Barr'on-Cede no, Rostislav Petrov, and Preslav Nakov. Fine-Grained Analysis of Propaganda in News Articles. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, EMNLP-IJCNLP 2019, Hong Kong, China, November 2019.

[107] Seunghak Yu, Giovanni Da San Martino, and Preslav Nakov. Experiments in Detecting Persuasion Techniques in the News, 2019.

[108] Anthony Weston. *A Rulebook for Arguments.* Hackett, Indianapolis, 2009.

[109] Robyn Torok. Symbiotic radicalisation strategies: Propaganda tools and neuro linguistic programming. 2015.

[110] Gabriel Teninbaum. Reductio ad Hitlerum: Trumping the Judicial Nazi Card. *Michigan State Law Review*, page 541, 01 2009.

[111] G.S. Jowett and V. O'Donnell. *Propaganda and Persuasion.* Advances in Political Science. SAGE Publications, 1986.

[112] Renee Hobbs. Teaching about Propaganda: An Examination of the Historical Roots of Media Literacy. *Journal of Media Literacy Education*, 6:56–67, 11 2014.

[113] Jean Goodwin and Raymie McKerrow. Accounting for the Force of the Appeal to Authority. 2011.

[114] Monika L. Richter. The Kremlin's platform for 'useful idiots' in the West: An overview of RT's editorial strategy and evidence of impact. Technical report, Kremlin Watch, 2017.

[115] J. Hunter. Brainwashing in a Large Group Awareness Training?: The Classical Conditioning Hypothesis of Brainwashing, 2015.

[116] Lavinia Dan. Techniques for the Translation of Advertising Slogans. In *International Conference Literature, Discourse and Multicultural Dialogue, LDMD '15*, pages 12–23, 2015.

[117] Giovanni Da San Martino, Alberto Barrón-Cedeño, and Preslav Nakov. Findings of the NLP4IF-2019 Shared Task on Fine-Grained Propaganda Detection. In Anna Feldman, Giovanni Da San Martino, Alberto Barrón-Cedeño, Chris Brew, Chris Leberknight, and Preslav Nakov, editors, *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 162–170, Hong Kong, China, 11 2019. Association for Computational Linguistics.

[118] Giovanni Da San Martino, Alberto Barrón-Cedeño, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. SemEval-2020 Task 11: Detection of Propaganda Techniques in News Articles. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, 2020.

[119] Giovanni Da San Martino, Stefano Cresci, Alberto Barrón-Cedeño, Seunghak Yu, Roberto Di Pietro, and Preslav Nakov. A Survey on Computational Propaganda Detection. *CoRR*, abs/2007.08024, 2020.

[120] Gaku Morio, Terufumi Morishita, Hiroaki Ozaki, and Toshinori Miyoshi. Hitachi at SemEval-2020 Task 11: An Empirical Study of Pre-Trained Transformer Family for Propaganda Detection. In Aurelie Herbelot, Xiaodan Zhu, Alexis Palmer, Nathan Schneider, Jonathan May, and Ekaterina Shutova, editors, *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1739–1748, Barcelona (online), 12 2020. International Committee for Computational Linguistics.

[121] Dawid Jurkiewicz, Łukasz Borchmann, Izabela Kosmala, and Filip Graliński. ApplicaAI at SemEval-2020 Task 11: On RoBERTa-CRF, Span CLS and Whether Self-Training Helps Them. In Aurelie Herbelot, Xiaodan Zhu, Alexis Palmer, Nathan Schneider, Jonathan May, and Ekaterina Shutova, editors, *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1415–1424, Barcelona (online), 12 2020. International Committee for Computational Linguistics.

[122] Malak Abdullah, Ola Altiti, and Rasha Obiedat. Detecting Propaganda Techniques in English News Articles using Pre-trained Transformers. In *2022 13th International Conference on Information and Communication Systems (ICICS)*, pages 301–308, 2022.

[123] Jakub Piskorski, Nicolas Stefanovitch, Giovanni Da San Martino, and Preslav Nakov. SemEval-2023 Task 3: Detecting the Category, the Framing, and the Persuasion Techniques in Online News in a Multi-lingual Setup. In Atul Kr. Ojha, A. Seza Doğruöz, Giovanni Da San Martino, Harish Tayyar Madabushi, Ritesh Kumar, and Elisa Sartori, editors, *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2343–2361, Toronto, Canada, July 2023. Association for Computational Linguistics.

[124] Jakub Piskorski, Nicolas Stefanovitch, Nikolaos Nikolaidis, Giovanni Da San Martino, and Preslav Nakov. Multilingual multifaceted understanding of online news in terms of genre, framing, and persuasion techniques. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3001–3022, Toronto, Canada, 07 2023. Association for Computational Linguistics.

[125] Jakub Piskorski, Nicolas Stefanovitch, Valerie-Anne Bausier, Nicolo Faggiani, Jens Linge, Sopho Kharazi, Nikolaos Nikolaidis, Giulia Teodori, Bertrand De Longueville, Brian Doherty, Jason Gonin, Camelia Ignat, Bonka Kosteva, Eleonora Mantica, Lorena Marcaletti, Enrico Rossi, Alessio Spadaro, Marco Verile, Giovanni Da San Martino, Firoj Alam, and Preslav Nakov. News categorization, framing and persuasion techniques: Annotation guidelines. Technical report. Technical report, European Commission Joint Research Centre, 03 2023.

[126] Yuta Koreeda, Ken-ichi Yokote, Hiroaki Ozaki, Atsuki Yamaguchi, Masaya Tsunokake, and Yasuhiro Sogawa. Hitachi at SemEval-2023 task 3: Exploring cross-lingual multi-task strategies for genre and framing detection in online news. In Atul Kr. Ojha, A. Seza Doğruöz, Giovanni Da San Martino, Harish Tayyar Madabushi, Ritesh Kumar, and Elisa Sartori, editors, *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1702–1711, Toronto, Canada, 07 2023. Association for Computational Linguistics.

[127] Amalie Pauli, Rafael Sarabia, Leon Derczynski, and Ira Assent. TeamAmpa at SemEval-2023 task 3: Exploring multilabel and multilingual RoBERTa models for persuasion and framing detection. In Atul Kr. Ojha, A. Seza Doğruöz, Giovanni Da San Martino, Harish Tayyar Madabushi, Ritesh Kumar, and Elisa Sartori, editors, *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 847–855, Toronto, Canada, 07 2023. Association for Computational Linguistics.

[128] Ben Wu, Olesya Razuvayevskaya, Freddy Heppell, João A. Leite, Carolina Scarton, Kalina Bontcheva, and Xingyi Song. SheffieldVeraAI at SemEval-2023 task 3: Mono and multilingual approaches for news genre, topic and persuasion technique classification. In Atul Kr. Ojha, A. Seza Doğruöz, Giovanni Da San Martino, Harish Tayyar Madabushi, Ritesh Kumar, and Elisa Sartori, editors, *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1995–2008, Toronto, Canada, 07 2023. Association for Computational Linguistics.

[129] Neele Falk, Annerose Eichel, and Prisca Piccirilli. NAP at SemEval-2023 task 3: Is less really more? (back-)translation as data augmentation strategies for detecting persuasion techniques. In Atul Kr. Ojha, A. Seza Doğruöz, Giovanni Da San Martino, Harish Tayyar Madabushi, Ritesh Kumar, and Elisa Sartori, editors, *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1433–1446, Toronto, Canada, 07 2023. Association for Computational Linguistics.

[130] Timo Hromadka, Timotej Smolen, Tomas Remis, Branislav Pecher, and Ivan Srba. KInITVeraAI at SemEval-2023 task 3: Simple yet powerful multilingual fine-tuning for persuasion techniques detection. In Atul Kr. Ojha, A. Seza Doğruöz, Giovanni Da San Martino, Harish Tayyar Madabushi, Ritesh Kumar, and Elisa Sartori, editors, *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 629–637, Toronto, Canada, 07 2023. Association for Computational Linguistics.

[131] Francisco-Javier Rodrigo-Ginés, Laura Plaza, and Jorge Carrillo-de Albornoz. UnedMediaBiasTeam @ SemEval-2023 task 3: Can we detect persuasive techniques transferring knowledge from media bias detection? In Atul Kr. Ojha, A. Seza Doğruöz, Giovanni Da San Martino, Harish Tayyar Madabushi, Ritesh Kumar, and Elisa Sartori, editors, *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 787–793, Toronto, Canada, 07 2023. Association for Computational Linguistics.

[132] Arkadiusz Modzelewski, Witold Sosnowski, Magdalena Wilczynska, and Adam Wierzbicki. DSHacker at SemEval-2023 Task 3: Genres and Persuasion Techniques Detection with Multilingual Data Augmentation through Machine Translation and Text Generation. In Atul Kr. Ojha, A. Seza Doğruöz, Giovanni Da San Martino, Harish Tayyar Madabushi, Ritesh Kumar, and Elisa Sartori, editors, *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1582–1591, Toronto, Canada, July 2023. Association for Computational Linguistics.

[133] Jiaxu Dao, Jin Wang, and Xuejie Zhang. YNU-HPCC at SemEval-2020 Task 11: LSTM Network for Detection of Propaganda Techniques in News Articles. In Aurelie Herbelot, Xiaodan Zhu, Alexis Palmer, Nathan Schneider, Jonathan May, and Ekaterina Shutova, editors, *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1509–1515, Barcelona (online), December 2020. International Committee for Computational Linguistics.

[134] D. G. Jones. Detecting Propaganda in News Articles Using Large Language Models. *Engineering: Open Access*, pages 01–12, 2024.

[135] Maram Hasanain, Fatema Ahmed, and Firoj Alam. Can GPT-4 Identify Propaganda? Annotation and Detection of Propaganda Spans in News Articles, 2024.

[136] Marzena Kutt. Zderzenie cywilizacji — skuteczna propaganda czy strach? Stosunek polskich internautów do islamskich uchodźców (English: Clash of civilisations - effective propaganda or fear? Attitudes of Polish Internet users towards Islamic refugees). *Roczniki Kulturoznawcze*, 8:25–38, 01 2017.

[137] Piotr Pogorzelski. Zagrożenie Rosyjską Dezinformacją w Polsce i Formy Przeciwdziałania (English: The Threat of Russian Disinformation in Poland and the Forms of Counteraction). Technical report, Kolegium Europy Wschodniej im. Jana Nowaka-Jeziorańskiego we Wrocławiu, 2017.

[138] Boguslawa Dobek-Ostrowska. Mediatyzacja polityki w tygodnikach opinii w Polsce – między polityzacją a komercjalizacją (English: Mediatisation of

politics in weekly opinion magazines in Poland - between politicisation and commercialisation). *Zeszyty Prasoznawcze*, 61:224–246, 01 2018.

[139] Paulina Olechowska. Stopień stronniczości polskich dzienników ogólnoinformacyjnych (wybrane wyznaczniki) / degree of bias in Polish general-information newspapers (selected determinants). *Political Preferences*, 16:107–130, 01 2017.

[140] R Gorwa. Computational propaganda in Poland: false amplifiers and the digital public sphere. Technical report, Computational Propaganda Research Project, 2017.

[141] Patrycja Treichel. Master's Thesis: News Propaganda in Poland: Mixed Methods Analysis of the Online News Coverage About the Media Law Proposal Lex TVN, 2022.

[142] Monika Hapek & Marzena Barańska Malwina Popiołek. Infodemia – an Analysis of Fake News in Polish News Portals and Traditional Media During the Coronavirus Pandemic. *Communication & Society*, 34:81–89, 2021.

[143] Media Bias/Fact Check - TVP Info – Bias and Credibility, 2023. `https://mediabiasfactcheck.com/tvp-info-bias/` Accessed on 20.03.2024.

[144] Media Bias/Fact Check - TVN24 – Bias and Credibility, 2023. `https://mediabiasfactcheck.com/tvn24-bias/` Accessed on 20.03.2024.

[145] Gazeta Wyborcza – Bias and Credibility, 2023. `https://mediabiasfactcheck.com/gazeta-wyborcza-bias/` Accessed on 20.03.2024.

[146] Political Critique – Bias and Credibility, 2023. `https://mediabiasfactcheck.com/political-critique/` Accessed on 20.03.2024.

[147] FL24.net – Bias and Credibility, 2024. `https://mediabiasfactcheck.com/fl24-net/` Accessed on 20.03.2024.

[148] Top Websites Ranking - Most Visited News & Media Publishers Websites in Poland, 2024. `https://www.similarweb.com/top-websites/poland/news-and-media/` Accessed on 20.03.2024.

[149] „Wiadomości" liderem programów informacyjnych. Wszystkie dzienniki ze spadkiem oglądalności (English: „Wiadomości" the leader of news programmes. All daily tv news with falling viewing figures.), 2024. `https://www.wirtualnemedia.pl/artykul/propaganda-wiadomosci-prowadzacy-fakty-wydarzenia-wrzesien-2023-rok` Accessed on 20.03.2024.

[150] Małgorzata Kowalska-Chrzanowska and Przemysław Krysiński. Polskie projekty fact-checkingowe demaskujące fałszywe informacje na temat wojny w Ukrainie (English: Polish fact-checking projects exposing false information about the war in Ukraine). *Media i Społeczeństwo*, 17:51–71, 12 2022.

[151] Yelena Mejova, Amy X. Zhang, Nicholas Diakopoulos, and Carlos Castillo. Controversy and Sentiment in Online News. *CoRR*, abs/1409.8152, 2014.

[152] Ernest Jakaza and Marianna Visser. 'Subjectivity' in newspaper reports on 'controversial' and 'emotional' debates: An appraisal and controversy analysis. *Language Matters*, 47(1):3–21, 2016.

[153] Large language models for propaganda detection - github project, 2023. `https://github.com/sprenkamp/LLM_propaganda_detection` Accessed on 26.03.2024.

[154] The OpenAI API - Documentation - Models, 2024. `https://platform.openai.com/docs/models` Accessed on 26.03.2024.

[155] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc., 2022.

[156] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. A Survey of Large Language Models, 2023.

[157] Shantanu Kumar and Shruti Singh. Fine-Tuning Llama 3 for Sentiment Analysis: Leveraging AWS Cloud for Enhanced Performance. *SN Computer Science*, 5, 12 2024.

[158] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct Preference Optimization: Your Language Model is Secretly a Reward Model, 2024.

Appendix A

A.1 Prompts

A.1.1 Prompt for Task 2 (LLM on SemEval2020 – English Data) and Task 3 (Propaganda Technique Detection in the PONC Subset with the Use of LLM, First Try) – In English

"Having the following list of 14 propaganda techniques categories:

1. Appeal_to_Authority
2. Appeal_to_fear-prejudice
3. Bandwagon,Reductio_ad_hitlerum
4. Black-and-White_Fallacy
5. Causal_Oversimplification
6. Doubt
7. Exaggeration,Minimisation
8. Flag-Waving
9. Loaded_Language
10. Name_Calling,Labeling
11. Repetition
12. Slogans
13. Thought-terminating_Cliches
14. Whataboutism,Straw_Men,Red_Herring

find all examples of propaganda techniques used in the article below. Prepare a table with the result in Python two-dimensional array format (we found the array format to be the easiest one for post-processing and GPT models generated the most accurate results, with as little residue as possible). Return it as a variable "propaganda_techniques". Do not add any other comments. First column is "begin_offset" – list span's beginning character (included). Second column is "end_offset" – list span's ending character (excluded). Third column is "technique" – write the name of the utilized propaganda technique, using the categories listed above. Fourth column is "text" – write down the text from the found span. Article:*<inserted_article_text>* "

A.1.2 Prompt for Task 2 (LLM on SemEval2020 – English Data – Technique Classification)

You are a multi-label text classifier identifying 14 propaganda techniques within news paper articles. These are the 14 propaganda techniques you classify with definitions and examples:

- Loaded_Language – Uses specific phrases and words that carry strong emotional impact to affect the audience, e.g., a lone lawmaker's childish shouting.'
- Name_Calling,Labeling – Gives a label to the object of the propaganda campaign as either the audience hates or loves, e.g., 'Bush the Lesser.'
- Repetition – Repeats the message over and over in the article so that the audience will accept it, e.g., 'Our great leader is the epitome of wisdom. Their decisions are always wise and just'.
- Exaggeration,Minimisation – Either representing something in an excessive manner or making something seem less important than it actually is, e.g., 'I was not fighting with her; we were just playing'.
- Appeal_to_fear-prejudice – Builds support for an idea by instilling anxiety and/or panic in the audience towards an alternative, e.g., 'stop those refugees; they are terrorists.'
- Flag-Waving; Playing on strong national feeling (or with respect to a group, e.g., race, gender, political preference) to justify or promote an action or idea, e.g., 'entering this war will make us have a better future in our country'.
- Causal_Oversimplification – Assumes a single reason for an issue when there are multiple causes, e.g., 'If France had not declared war on Germany, World War II would have never happened'.
- Appeal_to_Authority – Supposes that a claim is true because a valid authority or expert on the issue supports it, 'The World Health Organisation stated, the new medicine is the most effective treatment for the disease.'
- Slogans – A brief and striking phrase that contains labeling and stereotyping, e.g., "Make America great again"!

- Thought-terminating_Cliches – Words or phrases that discourage critical thought and useful discussion about a given topic, e.g., "it is what it is"
- Whataboutism,Straw_Men,Red_Herring – Attempts to discredit an opponent's position by charging them with hypocrisy without directly disproving their argument, e.g., 'They want to preserve the FBI's reputation'.
- Black-and-White_Fallacy – Gives two alternative options as the only possibilities, when actually more options exist, e.g., 'You must be a Republican or Democrat'.
- Bandwagon,Reductio_ad_hitlerum – Justify actions or ideas because everyone else is doing it, or reject them because it's favored by groups despised by the target audience, e.g., "Would you vote for Clinton as president? 57% say yes."
- Doubt – Questioning the credibility of someone or something, e.g., 'Is he ready to be the Mayor'?

You will be given a list of starting (inclusive) and ending indexes (exclusive) of characters in the article, which represent a fragment of the article. Indicate which one of the propaganda techniques from the list above is present in the given fragments. Respond just with the propaganda technique name, start index number and end index number, separated by comma, in the sequence represented by indexes. Use only propaganda techniques from the list above. If no propaganda technique was identified return "no propaganda detected". Here is the list of indexes: *<inserted_indexes>* Here is the article: *<inserted_article_text>*

A.1.3 Prompt for Task 3, Second Prompt – In Polish, Techniques in English

"Poniższa lista zawiera 14 kategorii technik propagandowych:

1. Appeal_to_Authority
2. Appeal_to_fear-prejudice
3. Bandwagon,Reductio_ad_hitlerum
4. Black-and-White_Fallacy
5. Causal_Oversimplification
6. Doubt
7. Exaggeration,Minimisation
8. Flag-Waving
9. Loaded_Language
10. Name_Calling,Labeling

11. Repetition
12. Slogans
13. Thought-terminating_Cliches
14. Whataboutism,Straw_Men,Red_Herring

Znajdź wszystkie przykłady technik propagandowych użytych w poniższym artykule. Przygotuj dwuwymiarową tablicę (array) w języku Python. Zwróć ją jako zmienną "propaganda_techniques". Nie dodawaj żadnych innych komentarzy. Pierwsza kolumna to "begin_offset" – lista znaków początku zakresu (włącznie). Druga kolumna to "end_ offset" – końcowy znak zakresu listy (wyłącznie). Trzecia kolumna to "technique" – wpisz nazwę wykorzystanej techniki propagandowej, korzystając z kategorii wymienionych powyżej. Czwarta kolumna to "tekst" – wpisz tekst ze znalezionego zakresu. Artykuł: *<inserted_ article_ text>*"

* Translation of a prompt from Appendix A.1.1 into Polish.

A.1.4 Prompts for Task 3, Third Prompt – In Polish, Techniques In English, Few-Shot – Examples of Propaganda Techniques in Polish

"Poniższa lista zawiera 14 kategorii technik propagandowych i ich przykłady:

1. Appeal_to_Authority – "Nie 'zbawiajmy' świata kosztem Polski, pięknie pisał' Prymas Tysiąclecia"
2. Appeal_to_fear-prejudice – "Według najnowszych danych agencji badawczej Inquiry, aż 47 proc. respondentów w tej grupie deklaruje, że nie będzie się szczepić. Czy naprawdę w Polsce jesteśmy gotowi ryzykować życiem i zdrowiem naszych dzieci?"
3. Bandwagon, Reductio_ad_hitlerum – "Aż 65% badanych uważa, że niepełnoletność matki nie jest argumentem zezwalającym na aborcję"
4. Black-and-White_Fallacy – "Była zastępczyni rzecznika praw obywatelskich w rozmowie z Interią stwierdziła, że „potrzebna jest partia, która w sposób pryncypialny podejdzie do kwestii walki z katastrofą klimatyczną i bezkompromisowo do praw zwierząt". – Bez weganizmu taka perspektywa nie będzie możliwa – oceniła."
5. Causal_Oversimplification – "Dzis Wielki Dzień Pszczół. Ginie ich miliony przez zmiany klimatyczne. A jeśli nadal będziemy je zabijać, np. używając neonikotynoidów to wkrótce będziemy obchodzić Dzień Wspomnienia o Pszczołach."
6. Doubt – "Zadziwiające, że PiS nie potrafi sięgnęć po pieniądze z Funduszu Odbudowy, a mami nam oczy nierealnym odszkodowaniem od Berlina"

7. Exaggeration,Minimisation – "Aborcja to tylko zabieg medyczny"
8. Flag-Waving – "Już nigdy nie pozwolimy, by na polskiej ziemi stanęła noga rosyjskiego żołnierza"
9. Loaded_Language – "Oni się chcą tylko nachapać i nakraść."
10. Name_Calling, Labeling – "Ci zaś, którzy nie pamiętają PRL, mogą sobie skojarzyć styl telewizji Jacka Kurskiego z Chinami albo innymi krajami Wschodu."
11. Repetition
12. Slogans – "Stop Ukrainizacji Polski!."
13. Thought-terminating_Cliches – "Taka jest prawda i koniec."
14. Whataboutism, Straw_Men, Red_Herring

Znajdź wszystkie przykłady technik propagandowych użytych w poniższym artykule. Przygotuj dwuwymiarową tablicę (array) w języku Python. Zwróć ją jako zmienną "propaganda_techniques". Nie dodawaj żadnych innych komentarzy. Pierwsza kolumna to "begin_offset" – lista znaków początku zakresu (włącznie). Druga kolumna to "end_offset" – końcowy znak zakresu listy (wyłącznie). Trzecia kolumna to "technique" – wpisz nazwę wykorzystanej techniki propagandowej, korzystając z kategorii wymienionych powyżej. Czwarta kolumna to "tekst" – wpisz tekst ze znalezionego zakresu. Artykuł: "*<inserted_article_text>*"

* Translation of a prompt from Appendix A.1.2 into Polish. Examples of techniques are original for Polish language.

Acknowledgements

This dissertation would not have been possible without the support and encouragement I received from many sides. I am deeply grateful to all those who contributed to the completion of this work through their guidance, expertise, and unwavering belief in my abilities.

First of all, I would like to express my sincere appreciation to the Japanese Ministry of Education, Culture, Sports, Science and Technology (MEXT) for awarding me the scholarship, which provided me with the invaluable opportunity to further my education in Japan. I am equally grateful to Hokkaido University for allowing me to conduct my research at their institution.

I would like to express my deepest gratitude to my supervisor, Associate Professor Rafal Rzepka, for his invaluable guidance, support, and encouragement throughout my doctoral journey. Your insightful feedback, unwavering patience, and mentorship have shaped both this research and my academic growth. I am profoundly grateful for the opportunities and wisdom you have shared with me, which have been essential to the completion of this dissertation.

I also extend my thanks to Professor Kenji Araki for his time, expertise, and constructive criticism, which significantly contributed to refining the direction and quality of my work.

I would like to express my sincere gratitude to Professor Yuji Sakamoto, Professor Miki Haseyama, Professor Yoshinori Dobashi, Professor Takahiro Ogawa and Associate Professor Toshihiko Itoh for agreeing to review my dissertation.

My education would not have been possible without the unwavering support, both financial and emotional, from my beloved parents, Ewa and Piotr. Thank you for being there through all my ups and downs, for allowing me to explore the world, including my time studying in Japan, and for being my most loyal supporters. Your love and encouragement mean everything to me.

I would also like to express my gratitude to my brother Jakub, his wife Izabela, and their son Piotr. Your support throughout my studies in Japan meant the world to me. I could always count on you, and sharing my experiences with you made this journey all the more meaningful.

A heartfelt thank you to my grandparents, who cheered me on every step of the way throughout my schooling and studies. I wish they could be here now to receive my gratitude in person. You will always hold a special place in my heart.

And finally, to my dearest Mateusz, thank you from the bottom of my heart. Your love, help, patience, and belief in me were the foundation that made all of this possible. I could not have achieved any of this without You.

Publications

Journal Papers

I. Joanna Szwoch, Mateusz Staszkow, Rafal Rzepka, Kenji Araki. "Limitations of Large Language Models in Propaganda Detection Task." Applied Sciences, vol. 14, no. 10, 4330, 2024, pp. 1-22. [Online]. Available: https://doi.org/10.3390/app14104330. (IF=2.7, TC=2).

Cited by:

1. Ivo Verhoeven, Mishra Pushkar, and Ekaterina Shutova. "Yesterday's News: Benchmarking Multi-Dimensional Out-of-Distribution Generalisation of Misinformation Detection Models." arXiv preprint, vol. 2410.18122, October 2024. [Online]. Available: https://arxiv.org/abs/2410.18122.

2. Василий Юрьевич Осипов et al. "Интеллектуальное нейроуправление новостными потоками с непрерывным обучением." Информационно-управляющие системы, vol. 6, pp. 35–45, December 2024.

Conference Papers

I. Joanna Szwoch, Mateusz Staszkow, Rafal Rzepka, Kenji Araki. "Can LLMs Determine Political Leaning of Polish News Articles?" Proceedings of the 2023 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE), Nadi, Fiji, December 2023, pp. 1-6.

II. Joanna Szwoch, Mateusz Staszkow, Rafal Rzepka, Kenji Araki. "Sentiment Analysis of Polish Online News Covering Controversial Topics – Comparison Between Lexicon and Statistical Approaches." Proceedings of the Language Technology Conference (EDO 2023 Workshop), LTC 2023, Poznań, Poland, April 2023, pp. 277-281. [BEST PAPER AWARD]

III. Joanna Szwoch, Mateusz Staszkow, Rafal Rzepka, and Kenji Araki. "Creation of Polish Online News Corpus for Political Polarization Studies." Proceedings of the First Workshop on Natural Language Processing for Political Sciences (PoliticalNLP), Marseille, France, 24 June 2022, pp. 86-90. European Language Resources Association (ELRA).

Cited by:

1. Michele Joshua Maggini, Davide Bassi, and Pablo Gamallo Otero. "A systematic review of Hyperpartisan News Detection: A Comprehensive Framework for Definition, Detection, and Evaluation." Research Square Preprint, version 1, January 2024. [Online]. Available: https://doi.org/10.21203/rs.3.rs-3893574/v1.