Article

Limitations of Large Language Models in Propaganda Detection Task

Joanna Szwoch 1,*, Mateusz Staszkow 2, Rafal Rzepka 3 and Kenji Araki 3

Abstract: Propaganda in the digital era is often associated with online news. In this study, we focused on the use of large language models and their detection of propaganda techniques in the electronic press to investigate whether it is a noteworthy replacement for human annotators. We prepared prompts for generative pre-trained transformer models to find spans in news articles where propaganda techniques appear and name them. Our study was divided into three experiments on different datasets—two based on an annotated SemEval2020 Task 11 corpora and one on an unannotated subset of the Polish Online News Corpus, which we claim to be an even bigger challenge as an example of an under-resourced language. Reproduction of the results of the first experiment resulted in a higher recall of 64.53% than the original run, and the highest precision of 81.82% was achieved for gpt-4-1106-preview CoT. None of our attempts outperformed the baseline F1 score. One of the attempts with gpt-4-0125-preview on original SemEval2020 Task 11 achieved an almost 20% F1 score, but it was below the baseline, which oscillated around 50%. Part of our work that was dedicated to Polish articles showed that gpt-4-0125-preview had a 74% accuracy in the binary detection of propaganda techniques and 69% in propaganda technique classification. The results for SemEval2020 show that the outputs of generative models tend to be unpredictable and are hardly reproducible for propaganda detection. For the time being, these are unreliable methods for this task, but we believe they can help to generate more training data.

Keywords: propaganda detection; media bias; online news analysis; propaganda in online news; propaganda techniques

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 [1], spreads misinformation [2] and disinformation [3], or conveys propagandist overtones [4]. 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, we focused on Polish online news, as we 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 [5]. 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 [6]. In this study, we tackled another difficult problem of propaganda detection. The scarce research on it 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, we would like to propose a method that utilizes LLMs for this task and investigate whether it is a noteworthy replacement of human annotators.

1.1. Motivation

Research on propaganda in online news, especially in under-represented languages, is crucial due to the significant influence of online platforms on shaping public opinion and attitudes [7]. 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 [8]. 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 [9,10]. 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 [11]. 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 [12]. 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 [8]. 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, we think that research on propaganda in online news in under-represented 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.

1.2. Objective

In our work, we 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 de-politicizing news content, which would focus on summarizing the facts only, leaving behind the unnecessary content full of emotional load, opinions and subjective views. Another helpful tool could simultaneously consist of views of different political leanings, allowing the reader to compare the different stances.

The main goals of our 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 [5,13] subset (not annotated) to find possible examples and locations of propaganda techniques in the text.

In conclusion, the aim of our work was to check whether LLMs are capable of the automatic detection and annotation of propaganda techniques in online news. We 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.

1.3. Contributions

Below, we outline the key contributions of our work:

- We 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 we received significantly lower results.
- We 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. We 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.
- We 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 us 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.
- We discovered that GPT models can generate the output in the convenient form of a Python code, which enables faster processing and further analyses.

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

1.4. Paper Structure

This paper is structured as follows. In Section 2, we present existing works that focus on propaganda, including research dealing with propaganda in online news with reference to Poland. Section 3 describes the datasets we used in our experiments. In Section 4, we explain the methods used for our experiments. In Section 5, we show the results of the conducted experiments and Section 6 focuses on the error analysis of the obtained results. Section 7 is a discussion part, in which we consider the limitations of our work. Finally, Section 8 summarizes our work, which is followed up with the drawn conclusions and mentions future works that we plan to undertake.

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. We also mention related topics—media bias and fact checking. We 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.

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” [14]. 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 [15]. 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 [16] 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 [17]. 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 [18]. 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 [19]. 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. For the time being, it is impossible, as they use biased data.

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 [20–22]. One of the connoted works performed automatic political fact checking (true or false) by analyzing linguistic characteristics of news excerpts with distant supervision [23]. 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 [24]. 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 [25]. 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 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 [26].

As we can see, none of them dealt with propaganda as their core activity, but they all revolved 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 [27]. 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 [28]. 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 [29]. Authors also created QProp, which is a propaganda corpus prepared with distant supervi-
sion. 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.

Another work proposed a fine-grained analysis of information on the detection of propaganda techniques and spans in which they were used [30,31]. Based on previous works [32–40], 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 [41]. 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 [42,43]. 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 [44]. Propaganda tech-

### Table 1. Main organizations dealing with media bias and fact checking.

<table>
<thead>
<tr>
<th>Organization</th>
<th>Description</th>
<th>Website</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ad Fontes Media</td>
<td>Creator of Interactive Media Bias Chart®</td>
<td><a href="http://www.adfontesmedia.com">www.adfontesmedia.com</a> (accessed on 20 March 2024)</td>
</tr>
<tr>
<td>All Sides</td>
<td>Provides balanced news, media bias ratings and diverse perspectives—top news stories are presented from the left, center and right of the political spectrum.</td>
<td><a href="https://www.allsides.com/unbiased-balanced-news">https://www.allsides.com/unbiased-balanced-news</a> (accessed on 20 March 2024)</td>
</tr>
<tr>
<td>FactCheck.org</td>
<td>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.</td>
<td><a href="http://www.factcheck.org">www.factcheck.org</a> (accessed on 20 March 2024)</td>
</tr>
<tr>
<td>Media Bias/Fact Check</td>
<td>Promotes awareness of media bias and misinformation by rating the bias, factual accuracy and credibility of media sources.</td>
<td><a href="http://www.mediabiasfactcheck.com">www.mediabiasfactcheck.com</a> (accessed on 20 March 2024)</td>
</tr>
<tr>
<td>PolitiFact</td>
<td>Receiver of a Pulitzer award.</td>
<td><a href="http://www.politifact.com">www.politifact.com</a> (accessed on 20 March 2024)</td>
</tr>
<tr>
<td>Snopes</td>
<td>One of the first fact-checking services; initially investigated urban legends.</td>
<td><a href="http://www.snopes.com">www.snopes.com</a> (accessed on 20 March 2024)</td>
</tr>
</tbody>
</table>
nique classification was found to be the most successful when applying a RoBERTa-based model using a semi-supervised learning technique of self-training [45]. Later experiments with a fine-tuned RoBERTa model outperformed scores for the classification task [46].

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 [47–49]. 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 [50–55]. XGBoost and other classic methods were implemented only by two teams and performed poorly [56,57].

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, which were fine-tuned with the davinci model [58]. 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 [59] 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 language—Arabic [60]. 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.

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 [61] or the threat posed by Russian disinformation and propaganda in Poland [62]. 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 [63]. 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 [64].
When it comes to studies on propaganda in digital media, one study focused on computational propaganda in Poland [65]. 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 [66]. It proposed a mixed method of propaganda model theory [18] and ways of using propaganda techniques [42]. 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 [67]. 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 [47]. 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 [26]. However, just as in other countries, we 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 2 presents news outlets that were described in the MBFC website [68–72].

According to Similarweb, neither TVP Info nor TVN24 was in the top five most popular Polish online news websites [73], but they were one of the most important TV news providers in Poland [74].

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 3 [26,75].

<table>
<thead>
<tr>
<th>Media Outlet</th>
<th>Bias Rating</th>
<th>Factual Reporting</th>
<th>Media Type</th>
<th>Traffic/Popularity</th>
<th>MBFC Credibility Rating</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>TVP Info</td>
<td>Right</td>
<td>Mixed</td>
<td>TV station</td>
<td>High traffic</td>
<td>Medium</td>
<td>[<a href="http://www.tvp.info">www.tvp.info</a>](accessed on 20 March 2024)</td>
</tr>
<tr>
<td>TVN24</td>
<td>Left-center</td>
<td>High</td>
<td>TV station</td>
<td>High traffic</td>
<td>High</td>
<td>[<a href="https://tvn24.pl/">https://tvn24.pl/</a>](accessed on 20 March 2024)</td>
</tr>
<tr>
<td>Gazeta Wyborcza</td>
<td>Left-center</td>
<td>High</td>
<td>Newspaper</td>
<td>High traffic</td>
<td>High</td>
<td>[<a href="http://www.wyborcza.pl">www.wyborcza.pl</a>](accessed on 20 March 2024)</td>
</tr>
<tr>
<td>FL24.net</td>
<td>Far-right</td>
<td>Mixed</td>
<td>Website</td>
<td>Minimal traffic</td>
<td>Low; questionable reasoning: propaganda, poor sourcing, lack of transparency</td>
<td>No longer exists; operated in France, owned by Polish media outlet</td>
</tr>
<tr>
<td>Political Critique</td>
<td>Left-center</td>
<td>High</td>
<td>Website</td>
<td>No information</td>
<td>No information</td>
<td>No longer exists; Polish version: [<a href="https://krytykapolityczna.pl/">https://krytykapolityczna.pl/</a>](accessed on 20 March 2024)</td>
</tr>
</tbody>
</table>
Table 3. List of Polish organizations that investigate media bias and perform fact checking of news in Polish.

<table>
<thead>
<tr>
<th>Organization</th>
<th>Year of Creation</th>
<th>Website</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fundacja Reporterów</td>
<td>2010</td>
<td><a href="http://www.fundacjareporterow.org">www.fundacjareporterow.org</a> (accessed on 20 March 2024)</td>
<td>indEX and Ukraine Monitor projects observe influence of disinformation on extremist groups</td>
</tr>
<tr>
<td>Konkretn.pl</td>
<td>2018</td>
<td><a href="http://www.konkretn.pl">www.konkretn.pl</a></td>
<td>Created by the TVN group</td>
</tr>
<tr>
<td>Sprawdźam AFP</td>
<td>2019</td>
<td><a href="http://www.sprawdzam.afp.com/list">www.sprawdzam.afp.com/list</a></td>
<td>Part of Agence France-Presse (AFP; a multilingual, multicultural news agency)</td>
</tr>
<tr>
<td>AntyFAKE</td>
<td>2019</td>
<td><a href="http://www.artyfake.pl">www.artyfake.pl</a></td>
<td>No longer updated</td>
</tr>
<tr>
<td>Odejuk.pl</td>
<td>2020</td>
<td><a href="http://www.odejuk.pl">www.odejuk.pl</a></td>
<td>No longer updated</td>
</tr>
<tr>
<td>Pravda</td>
<td>2020</td>
<td><a href="http://www.pravda.org.pl">www.pravda.org.pl</a></td>
<td>Fact checking of information, statements and digital content</td>
</tr>
<tr>
<td>#FakeHunter</td>
<td>2020</td>
<td><a href="http://www.fakenews.pl">www.fakenews.pl</a></td>
<td>Created to check news regarding SARS-CoV-2</td>
</tr>
</tbody>
</table>

List may not be exhaustive.

3. Datasets

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

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 [30,42], which consisted of 536 news articles in the final version. Table 4 shows the exact distribution of articles.

Table 4. PTC-SemEval2020 corpus distribution.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>News Article Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training set</td>
<td>371</td>
</tr>
<tr>
<td>Development set</td>
<td>75</td>
</tr>
<tr>
<td>Test set</td>
<td>90</td>
</tr>
</tbody>
</table>

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:

- Appeal to authority;
- Appeal to fear/prejudice;
Table 5 shows columns in training dataset.

<table>
<thead>
<tr>
<th>Column Name</th>
<th>Column Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>id</td>
<td>Article identification number</td>
</tr>
<tr>
<td>technique</td>
<td>Propaganda technique</td>
</tr>
<tr>
<td>begin_offset</td>
<td>Beginning of the span (inclusive)</td>
</tr>
<tr>
<td>end_offset</td>
<td>End of the span (exclusive)</td>
</tr>
</tbody>
</table>

### 3.2. Polish Online News Corpus (PONC)

For the final experiment presented in this paper, we used a subset of the Polish Online News Corpus (PONC) that covered contemporary controversial topics in Poland [5,13]. 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 [76,77] and we decided to focus on them to look for news articles with examples of propaganda techniques. We prepared another subset of the PONC, i.e., a high-emotional-charge subset, in which we selected the highest value for each of the five emotions, i.e., anger, disgust, fear, happiness and sadness, and we selected the top 9 articles per emotion and per news provider. In total, we select 90 articles, with 45 from TVP Info and 45 from TVN24.

### 4. Methods

In this section, we describe the experiments we conducted on the datasets described in Section 3. Our methods utilized different approaches to SI and TC tasks.

#### 4.1. LLM on SemEval2020—English Data, Sprenkamp et al.’s Approach

First, we attempted to reproduce the results obtained by Sprenkamp et al. [58]. We 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 [78]. We 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, we 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—should ensure reproducible outputs [79].
- gpt-3.5-turbo-0125—as of 26.03.2024, same as gpt-3.5-turbo.
- gpt-3.5-turbo-1106—should ensure reproducible outputs [79].

All the models were run five times—once using the basic prompt type and once including the chain of thought instruction [80]. We compared the three metrics proposed by
the authors, namely, F1 score, precision and recall. We also present the F1 scores obtained for all propaganda techniques per model.

4.2. LLM on SemEval2020—English Data

In the second experiment, we 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 [43].

The instructions of the tasks were as follows [42]:

• 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 [41].

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, we prepared our own prompts, which we found to be the most effective in obtaining the desired results. The prompt for TC can be found in Appendices A.1 and A.2. We kept the names of the techniques in a format that was required by the organizers of the shared task. Having run the models, we evaluated their performance on the test set and we 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.

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

Our third experiment was based on our original data, which was a subset of the PONC that covered controversial topics. We believe that contentious issues tend to have more debatable descriptions in news, and therefore, we decided to look for examples of propaganda techniques in them. We 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. We 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, A.3 and A.4. From the 90 articles with high emotional charge, we randomly took 100 examples of detected propaganda techniques and manually checked to see whether they were correct. We performed the following:

• Binary classification task—whether there was propaganda in the chosen news excerpt; if there was no propaganda technique being used, we marked it as “no propaganda”.
• Propaganda technique classification to check whether the correct technique was chosen; if not, we added our comment of the suggested technique.

The annotation was performed by the first author of this article based on her best knowledge on the topic.

5. Results

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

5.1. LLM on SemEval2020—English Data, Sprenkamp et al. Approach

Table 6 presents the results of all the model calculations for the first experiment. As for the first run, we 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 our attempt to reproduce the original results, the difference in performance was between 7 and up to 10 percentage points, and we managed to outperform the baseline run with the best recall (64.53%).

Next, we 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,
we did not notice any stability of the results and we 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, we noticed that for \textit{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>
<thead>
<tr>
<th>Model</th>
<th>Precision</th>
<th>Recall</th>
<th>F1 Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline (gpt-4 CoT [58])</td>
<td>0.56868</td>
<td>0.57821</td>
<td>0.57340</td>
</tr>
<tr>
<td>Baseline—reproduce attempt</td>
<td>0.46479</td>
<td>0.64525</td>
<td>0.54055</td>
</tr>
<tr>
<td>gpt-3.5-turbo-0125 base</td>
<td>0.58923</td>
<td>0.48883</td>
<td>0.53435</td>
</tr>
<tr>
<td>gpt-3.5-turbo-0125 CoT</td>
<td>0.60700</td>
<td>0.43575</td>
<td>0.50732</td>
</tr>
<tr>
<td>gpt-3.5-turbo-1106 base</td>
<td>0.65104</td>
<td>0.34916</td>
<td>0.45455</td>
</tr>
<tr>
<td>gpt-3.5-turbo-1106 CoT</td>
<td>0.63934</td>
<td>0.32682</td>
<td>0.43253</td>
</tr>
<tr>
<td>gpt-4-0125-preview base</td>
<td>0.53292</td>
<td>0.47486</td>
<td>0.50222</td>
</tr>
<tr>
<td>gpt-4-0125-preview CoT</td>
<td>0.58419</td>
<td>0.47486</td>
<td>0.52388</td>
</tr>
<tr>
<td>gpt-4-1106-preview base</td>
<td>0.70833</td>
<td>0.04749</td>
<td>0.08901</td>
</tr>
<tr>
<td>gpt-4-1106-preview CoT</td>
<td>0.81818</td>
<td>0.05028</td>
<td>0.09474</td>
</tr>
</tbody>
</table>

Table 6. Results from Sprenkamp et al. experiment [58].

Table 7 shows the F1 score for all the propaganda techniques per model.

<table>
<thead>
<tr>
<th>Technique\Model</th>
<th>Baseline—gpt-4 CoT</th>
<th>Baseline—gpt-4 CoT, Our Attempt</th>
<th>gpt-3.5-turbo-0125 Base</th>
<th>gpt-3.5-turbo-0125 CoT</th>
<th>gpt-3.5-turbo-1106 Base</th>
<th>gpt-3.5-turbo-1106 CoT</th>
<th>gpt-4-0125-preview base</th>
<th>gpt-4-0125-preview CoT</th>
<th>gpt-4-1106-preview CoT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Appeal to authority</td>
<td>0.19048</td>
<td>0.24000</td>
<td>0.32432</td>
<td>0.22857</td>
<td>0.11765</td>
<td>0.31579</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.40000</td>
</tr>
<tr>
<td>Appeal to fear/prejudice</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.48276</td>
<td>0.00000</td>
<td>0.54545</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.53333</td>
</tr>
<tr>
<td>Bandwagon, reductio ad hitlerum</td>
<td>0.00000</td>
<td>0.12500</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
</tr>
<tr>
<td>Black-and-white fallacy</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.18182</td>
<td>0.00000</td>
<td>0.50000</td>
<td>0.35088</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
</tr>
<tr>
<td>Causal oversimplification</td>
<td>0.50000</td>
<td>0.37500</td>
<td>0.29630</td>
<td>0.31579</td>
<td>0.11111</td>
<td>0.12500</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.41509</td>
</tr>
<tr>
<td>Doubt</td>
<td>0.54545</td>
<td>0.53465</td>
<td>0.24390</td>
<td>0.56757</td>
<td>0.00000</td>
<td>0.41509</td>
<td>0.06667</td>
<td>0.06897</td>
<td>0.55172</td>
</tr>
<tr>
<td>Exaggeration, minimization</td>
<td>0.64000</td>
<td>0.64706</td>
<td>0.35088</td>
<td>0.057142</td>
<td>0.05128</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.56250</td>
</tr>
<tr>
<td>Flag-waving</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.15385</td>
<td>0.00000</td>
<td>0.10811</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
</tr>
<tr>
<td>Loaded language</td>
<td>0.93617</td>
<td>0.93617</td>
<td>0.90625</td>
<td>0.89660</td>
<td>0.77876</td>
<td>0.71698</td>
<td>0.30380</td>
<td>0.34568</td>
<td>0.84034</td>
</tr>
<tr>
<td>Name calling, labeling</td>
<td>0.74286</td>
<td>0.72165</td>
<td>0.79630</td>
<td>0.73684</td>
<td>0.76636</td>
<td>0.76364</td>
<td>0.07407</td>
<td>0.00000</td>
<td>0.57895</td>
</tr>
<tr>
<td>Repetition</td>
<td>0.65789</td>
<td>0.67327</td>
<td>0.60274</td>
<td>0.68235</td>
<td>0.59459</td>
<td>0.53846</td>
<td>0.10000</td>
<td>0.15000</td>
<td>0.37500</td>
</tr>
<tr>
<td>Slogans</td>
<td>0.10000</td>
<td>0.23529</td>
<td>0.00000</td>
<td>0.09523</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
</tr>
<tr>
<td>Thought-terminating cliches</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
</tr>
<tr>
<td>Whataboutism, straw men, red herring</td>
<td>0.33333</td>
<td>0.45283</td>
<td>0.43243</td>
<td>0.22723</td>
<td>0.08333</td>
<td>0.09091</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.16667</td>
</tr>
</tbody>
</table>

This was more proof that the GPT-generated results were irreproducible. First of all, our 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, we do not think that GPT models are a sufficient tool for processing propaganda techniques detection because the results we obtained seemed to be random and no reliable conclusions could be drawn from them.

5.2. LLM on SemEval2020—English Data

First, we ran gpt-3.5-turbo-0125 and gpt-4-0125-preview three times on the SemEval2020 Task 11 test set. We present the results of the first subtask (SI) in Table 8.

None of our 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, we checked our annotation for the second subtask (TC) with the golden set and show the results in Table 9.

Table 8. Task span identification (SI)—results.

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

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

Table 9. Task technique classification (TC)—F1 scores.

<table>
<thead>
<tr>
<th>Technique—Model</th>
<th>Baseline</th>
<th>gpt-4-0125-Preview</th>
<th>gpt-3.5-turbo-0125</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1 score</td>
<td>0.25196</td>
<td>0.23352</td>
<td>0.30000</td>
</tr>
<tr>
<td>Appeal to authority</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
</tr>
<tr>
<td>Appeal to fear/prejudice</td>
<td>0.03681</td>
<td>0.05128</td>
<td>0.01449</td>
</tr>
<tr>
<td>Bandwagon, reductio ad hitlerum</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
</tr>
<tr>
<td>Black-and-white fallacy</td>
<td>0.00000</td>
<td>0.04211</td>
<td>0.03125</td>
</tr>
<tr>
<td>Causal oversimplification</td>
<td>0.11561</td>
<td>0.03922</td>
<td>0.00000</td>
</tr>
<tr>
<td>Doubt</td>
<td>0.29143</td>
<td>0.03265</td>
<td>0.00995</td>
</tr>
<tr>
<td>Exaggeration, minimization</td>
<td>0.14420</td>
<td>0.05882</td>
<td>0.06329</td>
</tr>
<tr>
<td>Flag-waving</td>
<td>0.06195</td>
<td>0.04878</td>
<td>0.01869</td>
</tr>
<tr>
<td>Loaded language</td>
<td>0.46477</td>
<td>0.43468</td>
<td>0.47970</td>
</tr>
<tr>
<td>Name calling, labeling</td>
<td>0.00000</td>
<td>0.08458</td>
<td>0.02778</td>
</tr>
<tr>
<td>Repetition</td>
<td>0.19262</td>
<td>0.02643</td>
<td>0.03550</td>
</tr>
<tr>
<td>Slogans</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
</tr>
<tr>
<td>Thought-terminating cliches</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
</tr>
<tr>
<td>Whataboutism, straw men, red herring</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
</tr>
</tbody>
</table>
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 to detect, and the best result was obtained by gpt-3.5-turbo-0125. We 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, we 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 [42].

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, we 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.

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, we 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. We 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, we 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 we think it should be noted as name calling, labeling (original: być może bez „Bołków”, 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.
7. Discussion

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

- As a general observation, we 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. [58], it was not mentioned whether the models were run several times, but we 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 golden labels were not released to the public.

The second experiment was also limited in one aspect—golden 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, we 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 our experimental procedures [59]. Due to the complicated nature of the task of propaganda detection, we can expect that the results of his study are also not reproducible. What is more, the author used the annotated data that was possibly used for pre-training GPTs by OpenAI; therefore, we 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.

8. Conclusions and Future Work

Our 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. We 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 [81].

One of the biggest obstacles for current propaganda detection studies is the lack of datasets that have full open access and are reliably annotated. We 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, we 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, we 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 our findings, we 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. We additionally discovered that GPT models can generate concise outputs in Python code that is easy to process for analyses.

Although we decided to use OpenAI’s GPT-4, as it is a popular benchmark in recent studies, there are many other LLMs, both open source and paid, that were not tested in our research due to limited GPU resources, such as Gemini, Llama 2, Bloom, Claude, Falcon 180B, OPT-175B, XGen-7B, GPT-NeoX, GPT-J, Gemma, Mistral 7B, Zephyr-7B, Vicuna 13-B or Polish Llama version—QRA. It would be interesting to see the differences in the quality of results between these models in the future. We also plan to experiment with the BERT-based models mentioned as SOTA in previous works to see their performance on the Polish online news. Another idea is to use the updated SemEval2023 dataset from Task 3 on “Detecting the Category, the Framing, and the Persuasion Techniques in Online News in a Multi-lingual Setup” [47]. The third subtask involved detecting persuasion techniques in paragraphs of news articles. It used the findings from SemEval2020 Task 11 and proposed a different approach to the problem, namely, a multi-label task at the paragraph level. In other words, instead of finding very specific spans where the propaganda technique was used, it focused more on a general context and asked to find these techniques in whole paragraphs. Additionally, more than one technique could be found in the paragraph. This method seems to be less detailed, but still focuses on the problem and can possibly yield better results than the initial approach. Jones’ approach is also worth considering [59]—although the method is simplified, the latest LLMs could be used to do a binary check regarding whether a given article is propaganda or not, estimate a probability of this as a percentage and also name the techniques that are used. Furthermore, we could perform further analysis of labeling the PONC articles by an expert in the field of propaganda. GPT results could be used as a suggestion to be verified by a professional so that the task could be performed in a time-efficient manner. It would also be interesting to investigate the overlap of patterns, as well as the differences between the Polish and English languages. Finally, we believe that further analyses with the use of explainable artificial intelligence (XAI) could help to understand and interpret the generated outputs.

**Author Contributions:** Conceptualization, J.S. and R.R.; methodology, J.S. and R.R.; software, J.S. and M.S.; validation, J.S., R.R. and K.A.; formal analysis, J.S.; investigation, J.S.; resources, J.S.; data curation, J.S.; writing—original draft preparation, J.S.; writing—review and editing, J.S., M.S., R.R. and K.A.; visualization, J.S.; supervision, R.R. and K.A.; project administration, R.R. All authors read and agreed to the published version of this manuscript.

**Funding:** This work was partially supported by JSPS Kakenhi Grant Number 22K12160.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** This work utilized the SemEval2020 Task 3 datasets available at https://propaganda.math.unipd.it/ptc/index.html (accessed on 31 March 2024). Polish Online News Corpus (PONC) is available upon request.

**Conflicts of Interest:** The authors declare no conflicts of interest.

**Abbreviations**

The following abbreviations are used in this manuscript:

- **GPT** Generative pre-trained transformer
- **LLM** Large language model
- **MBFC** Media Bias/Fact Check
- **PiS** Prawo i Sprawiedliwość (Law and Justice)
- **PONC** Polish Online News Corpus
Appendix A. Prompts

Appendix A.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>”

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

You are a multi-label text classifier identifying 14 propaganda techniques within newspaper 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>

Appendix A.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

Appendix A.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 szczęścić. 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 zwierzęć".—Bez węganiżmu taka perspektywa nie będzie możliwa—oceniła."
6. **Doubt**—"Zadziwiające, że PiS nie potrafi sięgnąć po pieniądze z Funduszu Odbudowy, a mąmi nam oczy nierelernalnym odszkodowaniem od Berlina"
7. **Exaggeration,Minimisation**—"Aborcja to tylko zabieg medyczny"
8. **Flag-Waving**—"Już nigdy nie pozwolimy, by na polskiej ziemi stała 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**


* Translation of a prompt from Appendix A.2 into Polish. Examples of techniques are original for Polish language.

References


Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.