MIPD: Exploring Manipulation and Intention In a Novel Corpus of Polish Disinformation

Arkadiusz Modzelewski[1,2*], **Giovanni Da San Martino**[2], **Pavel Savov**[1], **Magdalena Anna Wilczyńska**[3†], **Adam Wierzbicki**[1]

[1]Polish-Japanese Academy of Information Technology, Poland
[2]University of Padova, Italy
[3]National Research Institute (NASK), Poland

Abstract

This study presents a novel corpus of 15,356 Polish web articles, including articles identified as containing disinformation. Our dataset enables a multifaceted understanding of disinformation. We present a distinctive multilayered methodology for annotating disinformation in texts. What sets our corpus apart is its focus on uncovering hidden intent and manipulation in disinformative content. A team of experts annotated each article with multiple labels indicating both disinformation creators' intents and the manipulation techniques employed. Additionally, we set new baselines for binary disinformation detection and two multiclass multilabel classification tasks: manipulation techniques and intention types classification.

1 Introduction

Mitigating the spread of disinformation on the web has become an important social challenge. Numerous significant events, including the COVID-19 pandemic and the Russo-Ukrainian conflict, highlight disinformation's negative impact on individuals and society (Springer and Özdemir, 2022; Dov Bachmann et al., 2023).

The high-level group of experts (HLEG) set up by the European Commission defines disinformation as *"false, inaccurate, or misleading information designed, presented, and promoted to intentionally cause public harm or for profit"* (de Cock Buning, 2018). There are two significant aspects in this definition: intention types ("the why") and misleading manipulations ("the how"). However, to our knowledge, no study in the literature examines intention types and manipulation in disinformation together, possibly due to a lack of quality annotated data. Therefore, we share with the research community the **M**anipulation and **I**ntention in **P**olish corpus of **D**isinformative web articles: the **MIPD** dataset. The MIPD dataset sheds light on the authors' intention and manipulation techniques in disinformation. Our high-quality open corpus, annotated by five professional fact-checkers and debunkers, will provide a multifaceted understanding of disinformation. Initially, we focus on Polish, the 5th most spoken language in the European Union (Ginsburgh et al., 2017). We chose this language because it is the largest of the V4 countries (Slovak Republic, Czech Republic, Poland, and Hungary), which have been particularly vulnerable to disinformation in recent years due to the Russo-Ukrainian conflict (Kuczyńska-Zonik, 2020; Bryjka, 2022).

Here are the main contributions of this work:

- We introduce the largest dataset[1] of online articles in Polish, annotated with intents, manipulation techniques and whether they are disinformative. To the best of our knowledge, this is the first corpus of its kind.
- We formulate two multiclass, multilabel tasks: a novel task for classifying intention types and a task for classifying manipulation techniques in Polish disinformative online articles. Additionally, we formulate a binary classification task for disinformation detection.
- We present experimental results using our dataset for three problems: disinformation detection, manipulation techniques, and intention types classification. We publish our models on *HuggingFace*[1] for full reproducibility.

2 Annotation Methodology and Guidelines

In order to ensure high-quality annotations, our annotation guidelines and methodology were created

* Corresponding author. Email: contact@amodzelewski.com, arkadiusz.modzelewski@pja.edu.pl † Work done as a researcher at Polish-Japanese Academy of Information Technology.

[1] Our MIPD dataset, along with links to fine-tuned models, the software used for our experiments, and the annotation guidelines and methodology used for dataset creation, is available at https://github.com/ArkadiusDS/MIPD.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 19769–19785
November 12-16, 2024

by fact-checking and debunking experts. We employed five Polish native-speaker experts with at least three years of fact-checking and debunking experience (on a one-year competitive salary). All debunking experts working on the project were previously employed in debunking organizations with the accreditation of the International Fact-Checking Network[2]. The same experts used the methodology to annotate the articles in MIPD. The methodology described here is also an educational tool for students who wish to learn how to detect disinformation.

The methodology is divided into five main steps:

1. Determining the article's thematic category.
2. Evaluating the credibility of the article's source and author (if known).
3. Determining the article's main class: *credible*, *disinformation*, *misinformation* or *hard-to-say*.
4. For disinformation, evaluating of manipulation.
5. For disinformation, evaluating of intention types and narratives.

Experts could return to previous steps and typically re-evaluate the main class after a detailed investigation of an article suspected to contain disinformation.

2.1 Annotation Scheme

In this section, we present a summary of the five different aspects considered in our annotation methodology.

2.1.1 Thematic Category

Given a web article, we start with an initial content analysis and determine the topic. Categorizing web articles into thematic domains enables future research on distinct features and patterns within different disinformation topics. Our assessment allows us to classify the articles into one of 10 detailed thematic categories. We base our taxonomy of thematic categories on a prior analysis of the work of fact-checking and debunking organizations, such as Snopes[3], "Counteracting Disinformation" Foundation[4], Demagog[5] Association, and Debunk EU[6]. We consider the following categories (if created, acronym in parentheses): *COVID-19 (COVID)*, *Migrations (MIG)*, *LGBT+*, *Climate Crisis (CLIM)*, *5G*, *War in Ukraine (WUKR)*, *Pseudomedicine (PSMED)*, *Women's rights (WOMR)*, *Paranormal Activities (PA)*, *News or Other (NEWS)*. The topics in our dataset significantly overlap with the most significant disinformation topics published in the recent EU DisinfoLab report (Sessa, 2023).

2.1.2 Evaluation of Source Credibility

For each article, the experts evaluate the credibility of the article's source (publishing portal or organization) and author (if known). Source and author credibility did not determine the overall evaluation of the article, but the experts maintain a list of sources with their credibility evaluation. The experts used this list to search for the next articles for evaluation. Sources were evaluated in three classes: *reliable*, *unreliable* or *mixed*. Articles from unreliable sources could be evaluated as credible, while articles from other sources could be evaluated as disinformation.

2.1.3 Main Credibility Evaluation

Given a web article, annotators identify from its content whether it contains *disinformation*, *misinformation*, or *credible information*. Annotators could also use a fourth category - *hard-to-say*.

In our annotation methodology, we adopt a disinformation definition provided by the European Commission's HLEG group (see Section 1). Disinformation is intentionally misleading or false. Unlike disinformation, misinformation is *misleading information shared by people who do not recognize it as such* (de Cock Buning, 2018).

We exclude articles with *misinformation* and *hard-to-say* labels from the primary published dataset. In this study, we wanted to focus on a binary classification: disinformation versus credible articles.

2.1.4 Manipulation Technique

Debunking experts identify the usage of manipulation techniques in disinformative articles. The annotation of manipulation techniques is a multiclass multilabel problem. The following presents our taxonomy and short descriptions of manipulation techniques adopted in our annotation methodology (acronym in parentheses).

Cherry Picking (CHP) Presenting information utilizing only data that supports a given hypothesis or argument, while ignoring the broader context (Morse, 2010).

Quote Mining (QM) Using a short fragment of someone's longer speech in a way that significantly

[2] The International Fact-Checking Network gives accreditation to debunking organizations that sign its code of principles. See https://www.poynter.org/ifcn/ [3] Snopes [4] Counteracting Disinformation [5] Demagog [6] Debunk EU

distorts its original tone (McGlone, 2005).

Anecdote (AN). The use of evidence in the form of personal experience or an isolated case, possibly rumor or hearsay, most often to discredit statistics (Cubitt, 2013).

Whataboutism (WH). Responding to a substantive argument not by addressing the heart of the matter, but by raising a new point that is unrelated to the topic at hand. (Little and Rogers, 2017).

Strawman (ST). It involves distorting someone else's argument in a way that makes it easier to refute it. It is often done by attributing a stance to opponents, who do not share it. (Talisse and Aikin, 2006).

Leading Questions (LQ). Flooding the recipient with a series of consecutive suggestive questions or putting them together leads the recipient to a predetermined thesis (Loftus, 1975).

Appeal to Emotion (AE). The use of words and phrases that are to arouse in the recipient extreme emotion and attitude to the presented matter (Brinton, 1988).

False Cause (FC). The individual employing this technique assumes a cause-and-effect relationship solely based on the observed correlation (Castagnoli, 2016).

Exaggeration (EG). The author overstates a phenomenon, making it appear larger, better, or worse, or oversimplifies a phenomenon making it seem less significant or smaller than it truly is (Kreider, 2022; Da San Martino et al., 2019).

Reference Error (RE). In this technique, the author refers to fake experts, propaganda statements made by politicians, anonymous entries published on social media, or false quotes from famous people to authenticate the presented thesis (Diethelm and McKee, 2009). It may present false choices and false analogies.

Misleading Clickbait (MC). A technique involves giving a title to the text that misrepresents or contradicts the content discussed within the article. Title created with a purpose to attract attention (Chen et al., 2015).

Manipulation and persuasion techniques have a lot in common. Detection of the latter has already been examined in previous studies, such as in work done by Da San Martino et al. (2019). Manipulation can be seen as distinct from persuasion in that it is concerned not with changing individuals' beliefs but with inducing them into choices that the manipulator desires (Paine, 1989). Therefore, we can assume that a manipulation technique is always used with malicious intent, which is also explored in our methodology and the MIPD dataset. On the other hand, persuasion techniques can be used without malicious intent (for example, persuading individuals to stop smoking or make other better health choices).

Our list of manipulation techniques includes techniques not considered in previous studies, e.g., in Da San Martino et al. (2019), such as *Cherry-Picking* and *Quote Mining*.

2.1.5 Intention Type

Debunking experts explore the intention types and narratives of creators of disinformative articles. Classifying the creator's intention in disinformative articles allows us to understand their characteristics and detect patterns in disinformation content. In our methodology, each intention corresponds to several narratives. An intention is a generalization of a narrative that we can define as a repeating pattern found in several disinformative articles (Sessa, 2023). Intention encapsulates the broader goal of the author, which guides specific narratives used to achieve that goal.

Figure 1 provides a breakdown of our taxonomy and brief explanations of the intention types. The annotation of intention is a multiclass multilabel problem.

2.2 Annotation Process

We can break down our annotation process into four stages:

1. **Methodology creation stage** - professional annotators and researchers collaborated to develop a methodology for annotating articles.
2. **Initial annotation stage** - the most experienced specialist and the leader of the annotators' group trained other less experienced participants. In this initial step, the annotators tested the methodologies on a small sample of articles.
3. **Article annotation stage** - each text was annotated independently by at least two annotators. We included articles in our dataset if the annotations from experts were the same. If not, the article passed to the fourth annotation stage.
4. **Annotation consensus** - if the evaluations of at least two experts did not match, the annotators met and discussed their evaluations, seeking consensus. The lack of consensus resulted in adding a *hard-to-say* label to an article. Article with *hard-to-say* label was excluded from our dataset. The discussion and development of

Statistic	PA	CLIM	COVID	5G	LGBT+	MIG	NEWS	PSMED	WUKR	WOMR	All
AVG_w	724	736	804	756	633	716	662	978	782	708	767
AVG_{ch}	5,062	5,280	5,764	5,471	4,552	5,091	4,672	7,085	5,517	5,046	5,485
$\#DOC$	1,046	1,011	6,049	1,048	1,036	1,030	1,033	1,013	1,026	1,064	15,356

Table 1: Data statistics per thematic category: average article length in number of words (AVG_w), average article length in number of characters (AVG_{ch}), number of articles (#DOC). Acronyms in columns provide information about topic (see subsection 2.1.1)

consensus have always occurred during face-to-face meetings.

Additionally, our annotators divided the labeling phases into subject areas. They labeled articles topic by topic. Each time, they underwent additional training provided by the most experienced person in a specific thematic area before the annotation process. The training ensured in-depth understanding and accurate identification of disinformation.

2.3 Impartiality and Bias Prevention

To avoid bias in the dataset, our methodology requires each article to be annotated independently by two experts. Due to the complexity and time cost of the evaluation (the evaluation of a disinformative article took 30 minutes on average), we could only assure two evaluations per article. Instead, in case of disparity in an article's annotations, the two evaluating experts attempted to reach a consensus. We removed all articles that did not reach consensus from our dataset. All article annotations in the dataset are the result of a consensus between two expert annotators. During the consensus building, annotators discussed their interpretation of the methodology. Therefore, double verification helped to avoid biases and human errors while also serving as a standardization of the methodology's application.

3 MIPD Corpus

MIPD is a novel dataset that includes 15,356 web articles in Polish. In addition to the article content, the data we publish contains four annotations: *(i)* whether an article is disinformative or credible; *(ii)* the intention types; *(iii)* the manipulation techniques used in the article; *(iv)* the thematic category of the article. Additionally, we publish the sources from which we derived our articles.

3.1 Data Sources

We selected articles from more than 400 sources, each being freely available and not requiring any subscription. Our articles partially come from general and common news sources operating within the public sector, i.e., official sources managed by the government and its institutions. We also incorporate articles from alternative and independent opinion-oriented media, websites, and blogs sharing scientific insights. Additionally, we collected articles from websites containing conspiracy theories and Russian propaganda. The list of sources is not exhaustive. We aimed to collect the least biased dataset possible. Therefore, we focused on including the broadest spectrum of views and beliefs. We publish our dataset and the sources from which we obtained the articles.

3.2 Data Quality

We evaluate inter-rater reliability using a consensus measure. Consensus estimates of inter-rater reliability assume that annotators can agree on their evaluations. It is most suited for nominal evaluations where different scale levels represent qualitatively different ideas (Stemler, 2019). In our case, the main annotation class includes categories: *credible*, *disinformation*, *misinformation*, and *hard-to-say*. The difference between credible information and disinformation is complex to describe. This complexity is evident in the subsequent steps of the methodology, which aim to illustrate different aspects of disinformation. Similarly, evaluating manipulation techniques and intention types requires using qualitatively different concepts for each rating level.

In total, 15,510 articles in our dataset had two independent annotations. After the double independent annotations, experts attempted to establish consensus. Our experts did not reach a consensus in 49 cases (we removed these 49 articles from the dataset). The percentage of articles that reached consensus is 99.69%. However, during the consensus-building process, annotators could agree that the annotation should be *hard-to-say*. Experts placed 105 articles in that category. We have removed these articles from the dataset, considering them as articles that did not reach a con-

Negating Scientific Facts [NSF]: Authors deny established scientific facts, such as challenging the existence and severity of COVID-19, promoting alternative treatments, questioning the safety of 5G, and denying the reality of climate change and human impact on the environment. The objective is to create skepticism and erode public trust in scientific consensus.

Undermining the Credibility of Public Institutions [UCPI]: Authors try to erode trust in public institutions by engaging in activities, i.e., discrediting pandemic control measures, reproaching human rights violations, negating defense capabilities, and undermining strategies addressing migration and climate crises. These actions weaken the trust and confidence in the reliability and authority of government bodies and public organizations.

Challenging an International Organization [CIO]: Involves a deliberate effort to erode confidence in international organizations like the EU, WHO, UN, and NATO by disseminating content that blames them for regional conflicts, accuses them of aggression against specific countries, undermines defense capabilities, and discredits international climate agreements.

Promoting Social Stereotypes/Antagonisms [PSSA]: Authors promote social stereotypes and antagonisms through tactics such as enhancing homophobia, transphobia, xenophobia (linked to economic, security, and health aspects), religious conflicts, and antisemitism.

Weakening International Alliances [WIA]: Authors disseminate false or misleading information to undermine the strength and unity of partnerships between countries. The goal is to create doubt among allied nations, undermining the trust and cooperation necessary for their mutual security and strategic interests.

Changing Electoral Beliefs [CEB]: Authors influence public opinion, especially during elections. Authors with this intention capitalize on exploiting public sentiments surrounding sensitive issues such as LGBT rights and migrations to sway voters, polarize opinions, and potentially impact political decisions during elections.

Undermining International Position of a Country [UIPC]: Authors spread claims aimed at deteriorating a nation's global standing by accusing it of meddling in the political processes of other countries. Authors may erode trust and confidence in the state's governance and humanitarian standards. It seeks to damage the state's reputation on the international stage through unfounded allegations.

Causing Panic [CP]: Authors spread false information to incite fear and unrest among the public. This strategy exploits readers' emotions to destabilize societal trust and order.

Raising Morale of a Conflict's Side [RMCS]: Authors intend to boost the spirit and confidence of a particular group involved in a conflict. It aims to positively influence supporter perceptions and commitment towards their side's objectives and actions.

Figure 1: Malicious intention types with a brief description. We give an acronym for each intention in brackets.

sensus. Therefore, the final percentage of articles that reached consensus as credible, disinformation, or misinformation is 99%.

3.3 Statistics on the Corpus

In Table 1, we present some statistics per different thematic categories, such as the number of articles, average number of words, and average number of characters in the article. Figure 2 shows the percentages of articles with credible information and disinformation per thematic category. We publish 10,359 credible articles and 4,997 articles with disinformation (details about the number of articles with specific intention type and manipulation in Appendix B).

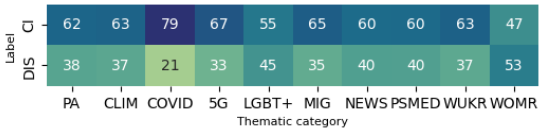

Figure 2: Percentage of disinformative (DIS) and credible (CI) articles per thematic category.

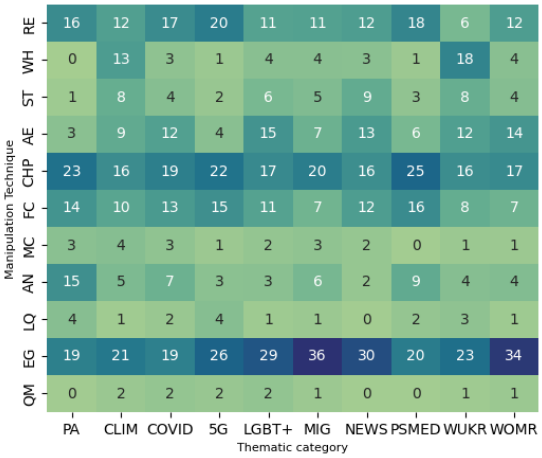

Figure 3: Percentage of different manipulation techniques per thematic category among articles with manipulation.

We present in Figure 3 and Figure 4 the percentages of articles with specific manipulation techniques and intention types per thematic category. These Figures show that neither manipulations nor intents are specific to the topic of the articles. One may consider the RMCS (*Raising Morale of a Conflict's Side*) intention type specific to the WUKR (*War in Ukraine*) topic. Nevertheless, the RMCS could likely be applicable when analyzing articles about other conflicts. In addition, we observe that a single manipulation technique can be used in articles with different intention types. Furthermore, an article designed with a particular intention may contain various manipulation techniques.

4 Experiments

Our experiments aim to test the data quality further and provide a baseline upon which future works can build.

4.1 Models and GPU

We fine-tuned two pre-trained Polish BERT-based Language Models: HerBERT (Mroczkowski et al., 2021), and Polish RoBERTa (Dadas et al., 2020). We chose these two models because they are the

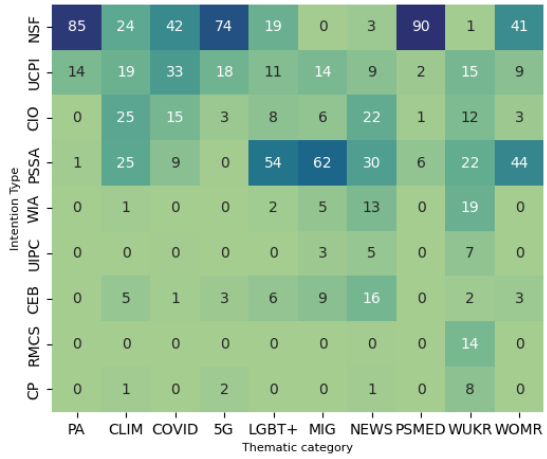

Figure 4: Percentage of different intention types per thematic categories among articles with malicious intention

ones that perform best on the KLEJ[7] Benchmark. KLEJ Benchmark is a comprehensive benchmark for Polish Language Understanding (Rybak et al., 2020). We used the HerBERT-base (HerBERT-B) and Polish-RoBERTa-v2-base (PL-RoBERTa-B) versions as well as the larger models: HerBERT-large (HerBERT-L) and Polish-RoBERTa-v2-large (PL-RoBERTa-L) versions. Appendix C presents more details about fine-tuned models.

For our computations to find the optimal hyperparameters and final fine-tuning of the models, we used the NVIDIA L40 GPU.

4.2 Experimental Setup

We began our experiments by dividing the data into train and validation in the proportions of 70%/30%. From the validation set, we randomly selected about 30% of the data as a test dataset. At the end of data preparation, we got datasets segmented into train/validation/test sets comprising 10,749 articles for training, 3,086 for validation, and an additional 1,521 for testing purposes. Next, we utilized our data and models to identify the optimal hyperparameters for training the model for disinformation binary classification and two multilabel multiclass tasks: manipulation and intention type classification. We accomplished this by performing a hyperparameter search for various learning rate values (ranging from 1e-6 to 1e-4) and weight decay (ranging from 0.005 to 0.2). Additionally, we implemented a linear warmup for the first 6% of the training steps. Appendix D shows the optimal hyperparameters we identified. Finally, we used trained models with optimal hyperparameters to predict the classes of the provided test dataset.

[7] KLEJ Benchmark leaderboard accessed on 15th April 2024 https://klejbenchmark.com/leaderboard/

Model	$Acc.$	F_{w}	F_1
HerBERT-B	0.94 ±0.004	0.94 ±0.004	0.91 ±0.007
HerBERT-L	0.95 ±0.003	0.95 ±0.003	0.93 ±0.004
PL-RoBERTa-B	0.94 ±0.005	0.94 ±0.005	0.91 ±0.008
PL-RoBERTa-L	**0.96** ±0.001	**0.96** ±0.001	**0.93** ±0.002

Table 2: Results for disinformation detection task. Table shows accuracy ($Acc.$), weighted F_1 score (F_{w}), and F_1 score on test data for pre-trained Polish BERT-based models. The results show the average metrics and their standard deviations, calculated from five different seeds.

4.3 Results

We computed results on test data that was unavailable during the fine-tuning process. The final result is an average of metric scores produced by models trained with five seeds. Tables 2, 3, and 4 show the final results with their corresponding standard deviations. For each classification task, we conducted paired t-tests to assess the statistical significance of the differences in weighted F_1 scores (F_{w}) across the various models. Detailed results from our statistical analysis are available in the Appendix E.

Disinformation Table 2 presents the results of four fine-tuned Polish BERT-based models on a disinformation detection task. This task was a binary classification to distinguish between disinformative and credible articles. Since the dataset is imbalanced, we adopted a weighted F_1 score as the primary evaluation metric. Notably, all models demonstrate high effectiveness. Evaluation metrics indicate minor variations across models. However, our analysis shows that the weighted F_1 scores are statistically different for all comparisons, except between PL-RoBERTa-B and HerBERT-B (see Appendix E). As for other evaluation metrics, the PL-RoBERTa-L model stands out with the highest weighted F_1 score.

Manipulation Technique Table 3 provides the performance of fine-tuned selected models, detailing results across individual manipulation techniques. Moreover, we show the models' overall effectiveness in the task using a weighted F_1 score. The HerBERT-L model and PL-RoBERTa-L performed best in this multilabel multiclass task.

Model	F_1 score for each manipulation technique class											F_w score
	CHP	QM	AN	WH	ST	LQ	AE	FC	EG	RE	MC	
HerBERT-B	0.45	0.00	0.14	0.19	0.27	**0.02**	0.40	0.31	0.64	0.43	0.00	0.42
	±0.01	±0.00	±0.05	±0.03	±0.03	±0.05	±0.02	±0.01	±0.01	±0.01	±0.00	±0.006
HerBERT-L	**0.48**	0.00	0.36	**0.30**	**0.30**	0.00	0.44	0.37	0.66	**0.50**	0.08	**0.47**
	±0.01	±0.00	±0.04	±0.03	±0.02	±0.00	±0.01	±0.03	±0.01	±0.00	±0.01	±0.008
PL-RoBERTa-B	0.44	0.00	0.08	0.20	0.25	0.00	0.38	0.33	0.64	0.38	0.00	0.41
	±0.02	±0.00	±0.08	±0.03	±0.03	±0.00	±0.01	±0.02	±0.01	±0.01	±0.00	±0.011
PL-RoBERTa-L	0.46	0.00	**0.39**	0.26	0.28	0.00	**0.45**	**0.38**	**0.67**	0.48	**0.15**	**0.47**
	±0.01	±0.00	±0.02	±0.05	±0.02	±0.00	±0.00	±0.02	±0.01	±0.02	±0.06	± 0.003

Table 3: Results for manipulation techniques classification. Table shows F_1 scores for pre-trained Polish BERT-based models in each manipulation type. Moreover, we present a weighted F_1 score (F_w) for the overall task. The results show the average metrics and their standard deviations, calculated from five different seeds. All evaluation metrics were computed for test data.

Statistical tests showed no statistically significant difference between these two larger models. We got similar results comparing smaller versions of these models (see Appendix E for details). PL-RoBERTa-L achieved the highest F_1 score for five manipulation techniques. Importantly, *Quote Mining*, *Leading Questions*, and *Misleading Clickbait* were particularly challenging. Specifically, none of the models could detect the *Quote Mining* technique. The decreased performance observed in classifying these three techniques is likely due to their relatively rare occurrence in our dataset.

Intention Type In the task of intention classification, PL-RoBERTa-L exhibits the best results, reaching a weighted F_1 score of 0.71. In the multilabel multiclass intention classification task, similar to the disinformation binary classification, statistical tests show a significant difference in all comparisons except for one between the smaller model versions (see Appendix E). A closer examination of the performance across distinct intention categories presented in Table 4 reveals that PL-RoBERTa-L outperforms other models in 8 out of 9 categories of intention types.

5 GPT and Disinformation Classification

In addition, we decided to explore the efficacy of generative models in disinformation classification. Specifically, we experimented with two OpenAI generative models that are accessible via their APIs: GPT-3.5 and GPT-4[8]. Our objective was to assess the ability of these models to classify articles as containing disinformation using a zero-shot classification approach. We employed two zero-shot strategies for each model: one without defining disinformation and the other including the definition. The definition we utilized was proposed by the High-Level Expert Group (HLEG) established by the European Commission (see Section 1). Although our findings are preliminary and warrant further in-depth analysis, we present them to demonstrate the potential of generative models in classifying disinformation.

First, we randomly drew a sample of 10% of the articles from our entire dataset. Then, we used a prompt to classify articles with generative models (our prompts are available in Appendix F). We repeated these steps for two approaches: (i) zero-shot classification with a disinformation definition included in the prompt; (ii) and zero-shot classification without a definition. Finally, we calculated various evaluation metrics. Table 5 presents the result of these calculations.

Our investigation reveals that Polish BERT-based models fine-tuned on the MIPD dataset significantly outperform chosen generative models: GPT-4 and GPT-3.5. The GPT models, when applied in a zero-shot approach without a definition of disinformation, achieved weighted F_1 scores of 0.84 for GPT-4 and 0.61 for GPT-3.5, respectively. In the zero-shot approach with the given definition of disinformation, both the GPT-4 and GPT-3.5 models improved their results. Nevertheless, these results are inferior to any Polish BERT-based models. Our findings highlight the effectiveness of HerBERT and Polish RoBERTa in handling Polish disinformation, likely due to their specialized training and fine-tuning utilizing Polish datasets. In contrast, the results of GPT models suggest that generative models may require domain-specific fine-tuning to reach the performance of language-specific BERT variants

[8] Details on the models used: We utilized a snapshot of GPT-4 from June 13th, 2023, named `gpt-4-0613`, and `gpt-3.5-turbo-instruct`, which has capabilities similar to GPT-3 era models. The last access to these models was on 28th May 2024.

Model	F_1 score for each intention class									F_w score
	UCPI	CEB	UIPC	CIO	WIA	PSSA	NSF	CP	RMCS	
HerBERT-B	0.56	0.19	0.31	0.52	0.46	0.69	0.81	0.22	0.42	0.65
	±0.01	±0.03	±0.07	±0.03	±0.03	±0.01	±0.01	±0.12	±0.04	±0.006
HerBERT-L	0.62	0.27	**0.38**	0.60	0.46	0.71	0.84	0.24	0.51	0.69
	±0.01	±0.05	±0.04	±0.01	±0.03	±0.01	±0.01	±0.08	±0.05	±0.006
PL-RoBERTa-B	0.56	0.17	0.38	0.55	0.48	0.67	0.81	0.25	0.46	0.65
	±0.02	±0.01	±0.04	±0.02	±0.02	±0.00	±0.01	±0.06	±0.04	±0.009
PL-RoBERTa-L	**0.62**	**0.30**	0.37	**0.63**	**0.49**	**0.74**	**0.86**	**0.27**	**0.56**	**0.71**
	±0.01	±0.02	±0.05	±0.01	±0.01	±0.01	±0.00	±0.07	±0.04	±0.005

Table 4: Results for malicious intention type classification. Table shows F_1 scores for pre-trained Polish BERT-based models in each intention type. Moreover, we present a weighted F_1 score (F_w) for the overall task. The results show the average metrics and their standard deviations, calculated from five different seeds. All evaluation metrics were computed for test data.

Model	Prompt Type	*Acc.*	F_w	F_1
GPT-4	Without Definition	0.85	0.84	0.73
	With Definition	0.86	0.86	0.77
GPT-3.5	Without Definition	0.60	0.61	0.51
	With Definition	0.70	0.70	0.56

Table 5: Results of the disinformation detection task for GPT-4 and GPT-3.5, showing accuracy (*Acc.*), F_1 score, and weighted F_1 score (F_w). The results present Zero-Shot Classification with and without a definition of disinformation.

in the disinformation classification task.

6 Related Work

Below we discuss previous work related to each of the three classification tasks we present.

6.1 Disinformation Detection

In recent years, we have observed a rapid increase in research aiming to understand and detect fake news. However, it must be stressed that fake news and disinformation are not equivalent and should not be used as synonyms. Fake news can be defined as *"fabricated information that mimics news media content in form but not in organizational process or intent"* (Lazer et al., 2018). The term *fake news* is insufficient to capture the complex problem of disinformation. Nonetheless, fake news is a type of disinformation, leading to overlap in research between the two areas (Broda and Strömbäck, 2024).

Several recent surveys (Capuano et al., 2023; Rastogi and Bansal, 2023; Kondamudi et al., 2023; Aïmeur et al., 2023) summarize a significant body of research on disinformation and fake news detection methods, utilizing various models and feature sets, adopting different definitions and focusing on diverse aspects of fake news. However, most of the described methods focus only on binary classification (*real* vs. *fake*, *disinformation* vs. *reliable information*, etc.). Some notable works leveraging BERT-based models for such classification are demonstrated in studies conducted by Kaliyar et al. (2021), Heidari et al. (2021), and (Kula et al., 2021a,b).

According to Capuano et al. (2023), most human-labeled datasets contain political news in English, e.g. political data provided by Wang (2017). Due to the global impact of the COVID-19 pandemic and the accompanying spread of disinformation, several datasets have been published that focus specifically on COVID-19 and vaccines (Patwa et al., 2021). Other examples of medical information datasets labeled by experts include Nabożny et al. (2023).

Our work is different from the existing ones in various aspects: (i) we propose disinformation detection in Polish, which is a less explored language; (ii) we create a novel and largest Polish disinformation corpus of 15,356 web articles; (iii) our dataset includes articles of 10 different thematic categories; (iv) and our multifaceted annotation scheme enable enhanced understanding of disinformation by uncovering used manipulation and authors' intention types.

6.2 Manipulation Techniques Detection

Although manipulation and persuasion have distinct characteristics (see Section 2.1.4), they also share significant commonalities. In addition, there is an overlap between persuasion and propaganda (Piskorski et al., 2023; Da San Martino et al., 2020b). Jin et al. (2022) introduced LOGIC, a dataset consisting of 13 classes of fallacious arguments of general domain drawn from several online educational sources, designed to teach and assess students' knowledge about logical fallacies. LOGIC does not have examples without fallacies.

Habernal et al. (2017) created *Argotario*, a game for learning argumentation fallacies. Furthermore, they collected a dataset of five fallacies (plus non-fallacious sentences) and reported classification experiments based on neural networks and feature-based models (Habernal et al., 2018).

Alhindi et al. (2022) proposed the CLIMATE dataset, which contains 679 text segments derived from 92 articles about climate change. Fallacious arguments were identified by domain experts. It comprises 10 categories, including a *No Fallacy* class (Musi et al., 2022).

Finally, Da San Martino et al. (2019, 2020a) introduced a more fine-grained analysis by creating a corpus of news articles annotated with 18 persuasion techniques. Piskorski et al. (2023) further extended work done by Da San Martino et al. (2019) and presented 23 persuasion techniques in a multilingual dataset and categorized them into 6 coarse-grained types of techniques.

In our work, we propose a different annotation scheme including techniques not considered in previous studies such us *Cherry-Picking*. Additionally, we focus on manipulation techniques used in Polish texts. Finally, we are the first to explore manipulation and intents and their usage in disinformation together.

6.3 Intention Types Detection

Web articles containing disinformation can take the form of both news and non-news content, with the malicious intent to mislead the public (Zhou and Zafarani, 2020). The intention behind creating news has been recognized as a crucial aspect of news understanding (Sharma et al., 2019; Rashkin et al., 2017). As a result, Wang et al. (2023) proposed a formal definition and a systematic analytical framework of news creation intent.

The first assessment of intentional vs. unintentional dissemination of fake news presents work done by Zhou et al. (2022). The author proposed an *influence graph*, which was utilized to assess the intent of fake news spreaders. Guo et al. (2023) further evaluates a spreading intent type of fake news by categorizing intent into five classes. However, the mentioned research focuses on intent in news and fake news articles.

Despite the increasing amount of research dedicated to disinformation (Zannettou et al., 2019; Nakov and Da San Martino, 2021; Xu et al., 2021; Yuan et al., 2023; Lucas et al., 2023), there remains a lack of focus on understanding the intent behind the creation of disinformation. To the best of our knowledge, our work is the first to examine disinformation in terms of the hidden intentions of its creators. We are the first to introduce a multilabel multiclass classification of intention types in disinformation proposed by debunking and fact-checking experts.

7 Conclusion and Future Work

Our study introduces a novel multifaceted Polish MIPD dataset, enabling a comprehensive approach to understand disinformation. The MIPD dataset includes a rich set of annotations, such as article topics and disinformation vs. credible information annotation. Beyond the binary annotation of disinformation, we explore manipulation techniques and malicious intention types utilized by creators of disinformation. In order to ensure high-quality annotations, our annotation guidelines and methodology were created by fact-checking and debunking experts. Moreover, our dataset was annotated by experts with at least three years of experience in debunking organizations with the accreditation of the International Fact-Checking Network. We publish our annotation guidelines and methodology, which are not language-specific and may be used in any other language.

Utilizing our dataset, we established new baselines for disinformation classification and two multilabel multiclass tasks: manipulation techniques and intention types classification. In our experiments, we used four different Polish BERT-based models. Additionally, for disinformation classification, we also present a preliminary analysis utilizing two LLMs, namely GPT-4 and GPT-3.5. Our analysis reveals that Polish BERT-based models fine-tuned on the MIPD dataset significantly outperform chosen generative models.

Annotating articles by experts is costly and time-consuming. Employing a comprehensive annotation methodology makes this process even more expensive. Accordingly, we plan to focus our future efforts on developing cost-effective methods using a semi-weakly supervised approach to detect misleading content. These methods could be invaluable for professional fact-checkers and debunkers and reduce manual efforts. Moreover, we plan to collect and annotate a dataset containing more languages and cultural contexts.

8 Limitations

Dataset Our annotation methodology assumed that the articles could fall into one from ten different subject categories presented in Section 2. When an article did not belong to any of the categories it was possible to select 11th option: *"Not related to the topic"* (articles with this category were excluded from further exploration and our dataset). Nevertheless, despite covering many topics, we cannot conclude that the data is representative and covers all possible thematic categories. Moreover, it is important to highlight that we created dataset only for Polish. We are actively working on resolving this issue by annotating English articles. In the future, we will publish full dataset with all articles in a more comprehensive publication.

Models We conducted our experiments using small transformer encoder architectures. Additionally, we examined how larger generative models perform in the context of disinformation classification. However, we have not conducted experiments on how LLMs handle intentions and manipulative techniques. Furthermore, our results on detecting disinformation using LLMs are based solely on a zero-shot approach. We leave these experiments and the exploration of alternative architectures for future studies.

Biases Human articles annotation can be prone to subjectivity. To remedy this, our articles were annotated by experienced debunkers and fact-checkers specializing in various fields. In addition, we created a comprehensive annotation methodology document that was at the same time flexible and could change to accommodate emerging issues (e.g. the presence of a new topic of articles extensively available online). Each article was annotated by two experts who needed to reach consensus.

9 Ethics

Dataset All the data gathered by our experts is from the public domain and is not copyrighted. Our data, does not contain information that could uniquely identify individuals. We utilized this data for research purpose only and it will be available on the CC BY-NC-ND 4.0 licence. Moreover, our data collection protocol was approved by an ethics review board.

Models We are researching disinformation detection, manipulation techniques, and intentions classification solely for the benefit of society. Our models and data can be helpful for fact-checkers and debunkers. Despite our good intentions, unfortunately, our models could also be used by disinformation creators and other malicious actors. However, since the models that we fine-tuned and publish are not generative, they can only be used to test disinformation authored by someone else.

Annotation We did not use crowdsourcing at any stage of data collection and annotation. Our experts involved in data annotation were employed by university and paid a fair salary as part of their professional duties. Expert's annotations were not influenced by any political or business decisions. Moreover, they worked in a self-governed team.

Computational resources Employing extensive language models frequently demands significant computational resources. This could have an impact on climate changes (Strubell et al., 2019). However, our models required little computing power, because we did not train the model from scratch, but performed fine-tuning. Moreover, the computer equipment used for this research was purchased by the university for research and educational purposes solely.

Acknowledgements

This work has been supported from the grant of the Polish National Center for Research and Development titled: "Development and verification of original methods of vertical artificial intelligence for automatic and precise detection of disinformation"(project number: INFOSTRATEG-I/0010/2021-00).

Giovanni Da San Martino would like to thank the Qatar National Research Fund, part of Qatar Research Development and Innovation Council (QRDI), for funding this work by grant NPRP14C0916-210015. He also would like to thank the European Union under the National Recovery and Resilience Plan (NRRP), Mission 4 Component 2 Investment 1.3 - Call for tender No. 341 of March 15, 2022 of Italian Ministry of University and Research – NextGenerationEU; Code PE00000014, Concession Decree No. 1556 of October 11, 2022 CUP D43C22003050001, Progetto "SEcurity and RIghts in the CyberSpace (SERICS) - Spoke 2 Misinformation and Fakes - DEcision supporT systEm foR cybeR intelligENCE (Deterrence) for also funding this work.

References

Esma Aïmeur, Sabrine Amri, and Gilles Brassard. 2023. Fake news, disinformation and misinformation in social media: a review. *Social Network Analysis and Mining*, 13(1):30.

Tariq Alhindi, Tuhin Chakrabarty, Elena Musi, and Smaranda Muresan. 2022. Multitask instruction-based prompting for fallacy recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8172–8187, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Alan Brinton. 1988. Pathos and the" appeal to emotion": An aristotelian analysis. *History of Philosophy Quarterly*, 5(3):207–219.

Elena Broda and Jesper Strömbäck. 2024. Misinformation, disinformation, and fake news: lessons from an interdisciplinary, systematic literature review. *Annals of the International Communication Association*, 48(2):139–166.

Filip Bryjka. 2022. Russian disinformation regarding the attack on ukraine. PISM Polski Instytut Spraw Międzynarodowych.

Nicola Capuano, Giuseppe Fenza, Vincenzo Loia, and Francesco David Nota. 2023. Content based fake news detection with machine and deep learning: a systematic review. *Neurocomputing*.

Luca Castagnoli. 2016. Aristotle on the non-cause fallacy. *History and Philosophy of Logic*, 37(1):9–32.

Yimin Chen, Niall J Conroy, and Victoria L Rubin. 2015. Misleading online content: recognizing clickbait as" false news". In *Proceedings of the 2015 ACM on workshop on multimodal deception detection*, pages 15–19.

Sean Cubitt. 2013. Anecdotal evidence. *NECSUS. European Journal of Media Studies*, 2(1):5–18.

Giovanni Da San Martino, Alberto Barrón-Cedeño, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. 2020a. SemEval-2020 task 11: Detection of propaganda techniques in news articles. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1377–1414, Barcelona (online). International Committee for Computational Linguistics.

Giovanni Da San Martino, Stefano Cresci, Alberto Barrón-Cedeño, Seunghak Yu, Roberto Di Pietro, and Preslav Nakov. 2020b. A survey on computational propaganda detection. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 4826–4832. International Joint Conferences on Artificial Intelligence Organization. Survey track.

Giovanni Da San Martino, Yu Seunghak, Alberto Barrón-Cedeno, Rostislav Petrov, Preslav Nakov, et al. 2019. Fine-grained analysis of propaganda in news article. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 5636–5646. Association for Computational Linguistics.

Sławomir Dadas, Michał Perełkiewicz, and Rafał Poświata. 2020. Pre-training polish transformer-based language models at scale. In *Artificial Intelligence and Soft Computing: 19th International Conference, ICAISC 2020, Zakopane, Poland, October 12-14, 2020, Proceedings, Part II 19*, pages 301–314. Springer.

Madeleine de Cock Buning. 2018. *A multi-dimensional approach to disinformation: Report of the independent high level group on fake news and online disinformation*. Publications Office of the European Union.

Pascal Diethelm and Martin McKee. 2009. Denialism: what is it and how should scientists respond? *The European Journal of Public Health*, 19(1):2–4.

Sascha-Dominik Dov Bachmann, Dries Putter, and Guy Duczynski. 2023. Hybrid warfare and disinformation: A ukraine war perspective. *Global Policy*, 14(5):858–869.

Victor Ginsburgh, Juan D Moreno-Ternero, and Shlomo Weber. 2017. Ranking languages in the european union: Before and after brexit. *European Economic Review*, 93:139–151.

Zhen Guo, Qi Zhang, Xinwei An, Qisheng Zhang, Audun Josang, Lance M Kaplan, Feng Chen, Dong H Jeong, and Jin-Hee Cho. 2023. Uncertainty-aware reward-based deep reinforcement learning for intent analysis of social media information. In *1st AAAI Workshop on Uncertainty Reasoning and Quantification in Decision Making (UDM-AAAI'23)*.

Ivan Habernal, Raffael Hannemann, Christian Pollak, Christopher Klamm, Patrick Pauli, and Iryna Gurevych. 2017. Argotario: Computational argumentation meets serious games. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 7–12, Copenhagen, Denmark. Association for Computational Linguistics.

Ivan Habernal, Patrick Pauli, and Iryna Gurevych. 2018. Adapting serious game for fallacious argumentation to german: Pitfalls, insights, and best practices. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Maryam Heidari, Samira Zad, Parisa Hajibabaee, Masoud Malekzadeh, SeyyedPooya HekmatiAthar, Ozlem Uzuner, and James H Jones. 2021. Bert model for fake news detection based on social bot activities in the covid-19 pandemic. In *2021 IEEE 12th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, pages 0103–0109. IEEE.

Zhijing Jin, Abhinav Lalwani, Tejas Vaidhya, Xiaoyu Shen, Yiwen Ding, Zhiheng Lyu, Mrinmaya Sachan, Rada Mihalcea, and Bernhard Schoelkopf. 2022. Logical fallacy detection. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 7180–7198, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Rohit Kumar Kaliyar, Anurag Goswami, and Pratik Narang. 2021. Fakebert: Fake news detection in social media with a bert-based deep learning approach. *Multimedia tools and applications*, 80(8):11765–11788.

Medeswara Rao Kondamudi, Somya Ranjan Sahoo, Lokesh Chouhan, and Nandakishor Yadav. 2023. A comprehensive survey of fake news in social networks: Attributes, features, and detection approaches. *Journal of King Saud University-Computer and Information Sciences*, 35(6):101571.

AJ Kreider. 2022. Argumentative hyperbole as fallacy. *Informal Logic*, 42(2):417–437.

Aleksandra Kuczyńska-Zonik. 2020. Propaganda, disinformation, strategic communication–how to improve cooperation in cee region? *Bulletin of Lviv Polytechnic National University*, 4:160–164.

Sebastian Kula, Michał Choraś, and Rafał Kozik. 2021a. Application of the bert-based architecture in fake news detection. In *13th International Conference on Computational Intelligence in Security for Information Systems (CISIS 2020) 12*, pages 239–249. Springer.

Sebastian Kula, Rafał Kozik, and Michał Choraś. 2021b. Implementation of the bert-derived architectures to tackle disinformation challenges. *Neural Computing and Applications*, pages 1–13.

David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, et al. 2018. The science of fake news. *Science*, 359(6380):1094–1096.

Adrian Little and Juliet Brough Rogers. 2017. The politics of 'whataboutery': The problem of trauma trumping the political in conflictual societies. *The British Journal of Politics and International Relations*, 19(1):172–187.

Elizabeth F Loftus. 1975. Leading questions and the eyewitness report. *Cognitive psychology*, 7(4):560–572.

Jason Lucas, Adaku Uchendu, Michiharu Yamashita, Jooyoung Lee, Shaurya Rohatgi, and Dongwon Lee. 2023. Fighting fire with fire: The dual role of llms in crafting and detecting elusive disinformation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14279–14305.

Matthew S McGlone. 2005. Quoted out of context: Contextomy and its consequences. *Journal of Communication*, 55(2):330–346.

Janice M Morse. 2010. "cherry picking": Writing from thin data.

Robert Mroczkowski, Piotr Rybak, Alina Wróblewska, and Ireneusz Gawlik. 2021. HerBERT: Efficiently pretrained transformer-based language model for Polish. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, pages 1–10, Kiyv, Ukraine. Association for Computational Linguistics.

E. Musi, M. Aloumpi, E. Carmi, S. Yates, and K. O'Halloran. 2022. Developing fake news immunity: fallacies as misinformation triggers during the pandemic. *Online Journal of Communication and Media Technologies*, 12(3). Copyright © 2022 by authors; licensee OJCMT by Bastas, CY. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/).

Aleksandra Nabożny, Bartłomiej Balcerzak, Mikołaj Morzy, Adam Wierzbicki, Pavel Savov, and Kamil Warpechowski. 2023. Improving medical experts' efficiency of misinformation detection: an exploratory study. *World Wide Web*, 26(2):773–798.

Preslav Nakov and Giovanni Da San Martino. 2021. Fake news, disinformation, propaganda, media bias, and flattening the curve of the covid-19 infodemic. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 4054–4055.

Scott C Paine. 1989. Persuasion, manipulation, and dimension. *The Journal of Politics*, 51(1):36–49.

Parth Patwa, Shivam Sharma, Srinivas Pykl, Vineeth Guptha, Gitanjali Kumari, Md Shad Akhtar, Asif Ekbal, Amitava Das, and Tanmoy Chakraborty. 2021. Fighting an infodemic: Covid-19 fake news dataset. In *Combating Online Hostile Posts in Regional Languages during Emergency Situation: First International Workshop, CONSTRAINT 2021, Collocated with AAAI 2021, Virtual Event, February 8, 2021, Revised Selected Papers 1*, pages 21–29. Springer.

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

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 2931–2937.

Shubhangi Rastogi and Divya Bansal. 2023. A review on fake news detection 3t's: Typology, time of detection, taxonomies. *International Journal of Information Security*, 22(1):177–212.

Piotr Rybak, Robert Mroczkowski, Janusz Tracz, and Ireneusz Gawlik. 2020. Klej: Comprehensive benchmark for polish language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1191–1201.

M.G Sessa. 2023. Connecting the disinformation dots: insights, lessons, and guidance from 20 eu member states. https://www.disinfo.eu/publications/connecting-the-disinformation-dots/.

Karishma Sharma, Feng Qian, He Jiang, Natali Ruchansky, Ming Zhang, and Yan Liu. 2019. Combating fake news: A survey on identification and mitigation techniques. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(3):1–42.

Simon Springer and Vural Özdemir. 2022. Disinformation as covid-19's twin pandemic: False equivalences, entrenched epistemologies, and causes-of-causes. *OMICS: A Journal of Integrative Biology*, 26(2):82–87.

Steven E Stemler. 2019. A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research, and Evaluation*, 9(1):4.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in nlp. *arXiv preprint arXiv:1906.02243*.

Robert Talisse and Scott F Aikin. 2006. Two forms of the straw man. *Argumentation*, 20:345–352.

William Yang Wang. 2017. " liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648*.

Zhengjia Wang, Danding Wang, Qiang Sheng, Juan Cao, Silong Su, Yifan Sun, Beizhe Hu, and Siyuan Ma. 2023. Understanding news creation intents: Frame, dataset, and method. *arXiv preprint arXiv:2312.16490*.

Fan Xu, Victor S Sheng, and Mingwen Wang. 2021. A unified perspective for disinformation detection and truth discovery in social sensing: a survey. *ACM Computing Surveys (CSUR)*, 55(1):1–33.

Xin Yuan, Jie Guo, Weidong Qiu, Zheng Huang, and Shujun Li. 2023. Support or refute: Analyzing the stance of evidence to detect out-of-context mis-and disinformation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4268–4280.

Savvas Zannettou, Tristan Caulfield, Emiliano De Cristofaro, Michael Sirivianos, Gianluca Stringhini, and Jeremy Blackburn. 2019. Disinformation warfare: Understanding state-sponsored trolls on twitter and their influence on the web. In *Companion proceedings of the 2019 world wide web conference*, pages 218–226.

Xinyi Zhou, Kai Shu, Vir V Phoha, Huan Liu, and Reza Zafarani. 2022. "this is fake! shared it by mistake": Assessing the intent of fake news spreaders. In *Proceedings of the ACM Web Conference 2022*, pages 3685–3694.

Xinyi Zhou and Reza Zafarani. 2020. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR)*, 53(5):1–40.

A Manipulation Techniques Examples

While our dataset is in Polish and the annotators are native Polish speakers, we have developed our annotation guidelines and methodology in English. This approach facilitates the application of our methodology to various languages. Consequently, our annotation guidelines feature explanations of manipulation techniques and examples in English. Below, we provide some of these examples, along with their explanations.

Cherry Picking Presenting information using only data that supports a given thesis while ignoring the broader context. It may include the slothful induction (rejecting inconvenient evidence that challenges our beliefs) or the Texas sharpshooter error (ignoring differences and emphasizing similarities, using from among an extensive dataset a small slice that supports our thesis).
Example: *Data show that last winter was the coldest in 10 years, indicating that the climate is not warming at all.*
Explanation: Using a single case of data taken out of context, ignoring the trend seen over a longer time frame. Even an exceptionally cold winter is not evidence of a change in trend.

Quote Mining Using a short excerpt from someone's longer speech/text in a way that significantly distorts its actual, original meaning.
Example: *It is as though fossils were just planted there, without any evolutionary history.*
Explanation: The quote comes from R. Dawkins' book, "Blind Watchmaker". It is used by proponents of creationism as if it were evidence to support their beliefs. In fact, as the full statement shows, this short quote completely distorts the meaning of Dawkins' statement, which criticizes creationists and disagrees with their hypothesis. The quote in full reads as follows: *It is as though*

fossils were planted there, without any evolutionary history. Needless to say this appearance of sudden planting has delighted creationists. Both schools of thought (Punctuationists and Gradualists) despise so-called scientific creationists equally, and both agree that the major gaps are real, that they are true imperfections in the fossil record. The only alternative explanation of the sudden appearance of so many complex animal types in the Cambrian era is divine creation and (we) both reject this alternative.

Anecdote The use of evidence in the form of personal experience or an isolated case, possibly rumor or hearsay, most often to discredit statistics.
Example: *In Bavaria, Germany, 25 cm of snow has just fallen. Global warming is the biggest lie in human history.*
Explanation: Authors use a singular example they experienced personally to discredit the statistic. Individual experiences cannot be translated into an entire population or multi-year studies.

Whataboutism Responding to a substantive argument not by addressing the heart of the matter but by raising a new point unrelated to the topic. Often referred to as dropping a false lead to divert attention from the topic.
Example: *You talk about LGBT+ people being persecuted. What about hungry children? No one thinks about them!*
Explanation: In this case, the author is trying to change the object of discussion by redirecting attention to something else. Often, as in this case, this is to hide the author's prejudice by expressing concern for another, usually worldview-neutral thing.

Strawman Misrepresenting someone's argument in a way that makes it easier to refute. It usually boils down to attributing to an opponent a position the opponent does not share.
Example: *Since you criticize Russia's actions, that means you are Russophobic.*
Explanation: The author of this statement assumes that criticism implies prejudice. It makes it easy to portray oneself or someone as a victim.

Leading Questions Flooding the target audience with consecutive questions or false arguments/studies that are suggestive. Guiding the recipient to a preconceived thesis. A statement consisting of a plethora of poorly related information, half-truths, and misinterpretations designed to overwhelm by their sheer volume.
Example: *I do not think we even know the actual mortality rate for monkeypox. Has a Westerner ever died from it? Could this possibly be the same money pox that occurs in Africa? If so, how did it suddenly appear in so many countries at once? (. . .) Perhaps it is here just to nudge us to get another shot?*
Explanation: The questions are intended to lead the recipient to a specific conclusion. The author wants to avoid accusations of spreading conspiracy theories, so instead, he will try to shape the viewer's opinion through suggestive questions. He can always defend himself by saying he is "just asking questions."

Appeal to Emotion The use of words and phrases arouses in the recipient a strong emotion and attitude toward the presented matter. The person using this technique tries to resonate with the recipient's prejudices (Appeal to Fear/Prejudice) or their values and traditions (Appeal to Values). They may also use short, vital phrases, including stereotyping or labeling (Slogans) and offensive and hateful language (Loaded Language). It can also use group affiliation (Flag Waving) or suggest a time for action (Appeal to Time) to mobilize the recipient to take specific actions.
Example: *They are MURDERING our children with vaccines. They are DEVILS who are just trying to harm the innocent! All doctors who administer vaccines will face cruel punishment.*
Explanation: The above statements are highly emotionally charged. Keywords are emphasized to evoke negative emotions in the recipient. A specific group of culprits to be held accountable is also indicated.

False Cause Assuming a cause-and-effect relationship solely based on the observed correlation. Among the manipulative statements used are those relating to time, such as those assuming that two events happening at the same time must be related (Cum Hoc Ergo Propter Hoc) or one following the other must be cause and effect (Post Hoc Ergo Propter Hoc).
Example: *Russia's military intervention in Ukraine is the culmination of U.S. and NATO aggression, dating back to their blitzkrieg against Serbia 23 years ago.*
Explanation: The above statement simplifies the whole issue, leaving out many events. Authors

choose a convenient cause for them and omit all other aspects.

Exaggeration The simplification and misrepresentation of a phenomenon or issue. For example, an author manipulating an audience may present a vision in which one decision can lead to unwanted negative consequences (Slippery Slope). Another way is to exaggerate minor or irrelevant aspects of an issue or the attitudes of individuals to denigrate an entire group or issue (Blowfish). One can also be used to manipulatively exaggerate the importance of a small group of people with different opinions than the rest of their community (Magnified Minority).
Example: *If we allow same-sex marriages, the next step will be to legalize pedophilia.*
Explanation: The author uses Slippery Slope. It is intended to falsely show the consequences of a decision or prove that one decision leads to another negative one. This technique is primarily intended to cause fear or reluctance in the recipient.

Reference Error It is a reference to unreliable sources or people. It can involve passing on knowledge from anonymous individuals, such as from social media, citing propaganda claims by politicians or media, primarily from authoritarian countries. It can also involve using untrue quotes circulating online to prove a point. This technique often cites fake experts or others to pretend to be a supposed authority (Appeal to Authority).
Example: *Vladmir Putin said Western sanctions are not working against Russia, and the economy is only growing. Once again, Putin proves how good a president he is, and the West has shown its weakness.*
Explanation: This is an example of invoking the propaganda claims of politicians and considering them a reliable source of information. However, it serves to manipulate public opinion.

Misleading Clickbait Giving the text a title that does not reflect the information presented in the article, often even contradicting it.
Example: *Covidian Church. Once again, the "Carol visit" will not take place because of the "pandemic"*
Explanation: The title uses the phrase "Covidian Church" and puts the word "pandemic" in quotes, implying skepticism or mocking the situation, which suggests people are overreacting or they are using the pandemic as an excuse to cancel the carol visit. However, the article itself explains in the first sentence *Due to the coronavirus pandemic, there will be no "carol visit" in the traditional form in the Archdiocese of Poznan this year either.*, which shows that "Carol visit" will take place, but not in a traditional form. The title is misleading, creating a sense of controversy or conspiracy not reflected in the content, presenting a straightforward explanation. This technique is used to attract readers by distorting or exaggerating the facts.

B Additional Data Statistics

Article category	Number of Articles
Credible information	10359
Disinformation	4997

Table 6: Number of articles categorized by credibility.

Manipulation Technique	Number of Articles
Cherry Picking	1526
Quote Mining	125
Anecdote	442
Whataboutism	426
Strawman	434
Leading Questions	127
Appeal to Emotion	909
False Cause	915
Exaggeration	2153
Reference Error	1108
Misleading Clickbait	177

Table 7: Number of articles with specific manipulation technique.

Intention Type	Number of Articles
NSF	2879
UCPI	1522
CIO	915
PSSA	1887
WIA	296
CEB	294
UIPC	113
CP	96
RMCS	122

Table 8: Number of articles with specific intention type

C Fine-Tuned Models

For all tasks, we utilized pre-trained BERT-based models specifically created for the Polish language (Mroczkowski et al., 2021). As of 03.11.2024, these models are available on HuggingFace[9] under the following names:

[9] https://huggingface.co/models

- `sdadas/polish-roberta-large-v2`
- `sdadas/polish-roberta-base-v2`
- `allegro/herbert-base-cased`
- `allegro/herbert-large-cased`

After fine-tuning models for text classification, we named the resulting models **PolBERT** when the base model was HerBERT, and **PolBERTa** when the base model was Polish RoBERTa.

Our fine-tuned models are publicly available on *HuggingFace* under the MIPD collection. Additionally, direct links to each model can be found in the README of our GitHub repository: https://github.com/ArkadiusDS/MIPD.

D Optimal Hyperparameters

We performed hyperaprameters tuning for all versions of chosen models, namely *HerBERT-base-cased*, *HerBERT-large-cased*, *Polish-RoBERTa-base-v2* and *Polish-RoBERTa-large-v2*. Batch size was not tuned for optimal value. We assumed 16 for train and evaluation batch size. To check all optimal values for learning rate and weight decay see Table 9, Table 10 and Table 11.

Model	Hyperparameters	
	learning rate	weight decay
HerBERT-B	3e-5	0.1
PL-RoBERTa-B	2e-5	0.2
HerBERT-L	1e-5	0.03
PL-RoBERTa-L	1e-5	0.02

Table 9: Optimal hyperparameters for disinformation detection

Model	Hyperparameters	
	learning rate	weight decay
HerBERT-B	1e-5	0.03
PL-RoBERTa-B	1e-5	0.1
HerBERT-L	1e-5	0.02
PL-RoBERTa-L	2e-5	0.01

Table 10: Optimal hyperparameters for manipulation classification

Model	Hyperparameters	
	learning rate	weight decay
HerBERT-B	1e-5	0.2
PL-RoBERTa-B	3e-5	0.03
HerBERT-L	1e-5	0.03
PL-RoBERTa-L	2e-5	0.1

Table 11: Optimal hyperparameters for intention classification

E Statistical Tests

For each classification task, we performed paired t-tests to evaluate the statistical significance of the differences in weighted F_1 scores (F_w) among the various models. The rationale for using a paired t-test is that the results for each model within a task were calculated using the same 5 seeds, making it crucial to account for the paired nature of the data. Below, we present the null and alternative hypotheses that were tested:

Null Hypothesis (H_0): There is no difference in means between the two paired groups (i.e., the mean difference is zero).

Alternative Hypothesis (H_1): There is a significant difference in means between the two paired groups.

The paired t-test evaluates whether to reject the null hypothesis, indicating a statistically significant difference in means between the two models. In our analysis, we used the commonly accepted p-value threshold of 5% to determine statistical significance. The detailed results from our statistical analysis are presented in tables: 12, 13, 14. All p-values below 0.05 in presented tables indicate the rejection of null hypothesis. **Note:** Abbreviations used in tables are as follows: HB for HerBERT-B, HL for HerBERT-L, PB for PL-RoBERTa-B, and PL for PL-RoBERTa-L.

Comparison	Avg. Diff.	Std. Deviation	p-value
HB vs HL	0.0095	0.0036	0.0041
HB vs PB	0.0000	0.0047	0.9882
HL vs PB	0.0095	0.0045	0.0090
HL vs PL	0.0048	0.0036	0.0415
HB vs PL	0.0143	0.0043	0.0017
PB vs PL	0.0143	0.0056	0.0047

Table 12: The table presents the results of model comparisons for the disinformation binary detection task, showing the average difference (Avg. Diff.) across five different seeds, along with the corresponding standard deviation and p-value for paired t-test.

F Prompt for GPT-based Models

We utilized English prompts for OpenAI's generative models to ensure reproducibility and understanding across different languages. Below we show prompts used in our experiments.

1. Zero-shot classification with GPT-3.5 without definition of disinformation in prompt:

   > You are an assistant who detects disinformation. Answer the question

Comparison	Avg. Diff.	Std. Deviation	p-value
HB vs HL	0.0486	0.0114	0.0007
HB vs PB	0.0114	0.0106	0.0732
HL vs PB	0.0600	0.0188	0.0020
HL vs PL	0.0023	0.0066	0.5364
HB vs PL	0.0467	0.0071	0.0009
PB vs PL	0.0553	0.0066	0.0006

Table 13: The table presents the results of model comparisons for the manipulation multiclass multilabel classification task, showing the average difference (Avg. Diff.) across five different seeds, along with the corresponding standard deviation and p-value for paired t-test.

Comparison	Avg. Diff.	Std. Deviation	p-value
HB vs HL	0.0420	0.0076	0.0003
HB vs PB	0.0078	0.0085	0.1089
HL vs PB	0.0342	0.0106	0.0020
HL vs PL	0.0201	0.0068	0.0027
HB vs PL	0.0622	0.0066	0.0000
PB vs PL	0.0544	0.0116	0.0005

Table 14: The table presents the results of model comparisons for the manipulation malicious intention type multilabel classification task, showing the average difference (Avg. Diff.) across five different seeds, along with the corresponding standard deviation and p-value for paired t-test.

> of whether the text contains disinformation. Answer using only one word: Yes or No. If the text contains disinformation, answer Yes, and if the text does not contain disinformation, answer No. Text: "<*Here we passed article for classification*>"
> Answer:

2. Zero-shot classification with GPT-3.5 with definition of disinformation in prompt:

> You are an assistant who detects disinformation. Disinformation is defined as false, inaccurate, or misleading information designed, presented, and promoted to intentionally cause public harm or for profit. Answer the question of whether the text contains disinformation. Answer using only one word: Yes or No. If the text contains disinformation, answer Yes, and if the text does not contain disinformation, answer No. Text: "<*Here we passed article for classification*>"
> Answer:

3. Zero-shot classification with GPT-4 without definition of disinformation in prompt:

   - For system role:

     > You are an assistant who detects disinformation.

   - For user role:

     > If the text contains disinformation, answer Yes, and if the text does not contain disinformation, answer No. Text:<*Here we passed article for classification*>. Answer:"

4. Zero-shot classification with GPT-4 with definition of disinformation in prompt:

   - For system role:

     > You are an assistant who detects disinformation. Disinformation is defined as false, inaccurate, or misleading information designed, presented, and promoted to intentionally cause public harm or for profit.

   - For user role:

     > If the text contains disinformation, answer Yes, and if the text does not contain disinformation, answer No. Text:<*Here we passed article for classification*>. Answer:"