THE WHY AND HOW OF DISINFORMATION: DATASETS, METHODS AND LANGUAGE MODELS EVALUATION

Author: Arkadiusz Modzelewski

NASK

Agenda

1. **Introduction** & Why Disinformation Cannot Be Ignored?
2. **Datasets**: Building the Foundation:
    a. MIPD - Polish Dataset
    b. MALINT - English Dataset
    c. Parliamentary Debates - Slavic Languages
3. **Persuasion and Intent**-Augmented Reasoning
    a. Psychological Motivation
    b. PCoT & ICoT for Disinformation Detection
4. **Disinformation Narratives**: Evaluating and Mining with LLMs
5. Conclusions and Further Reading

Why Disinformation Cannot Be Ignored

- Clear societal direct and indirect impact:
  - Changes real-world behaviour.
  - Shapes public opinion and trust in institutions.
  - Fuels violence and unrest.
- Dis/misinformation is a meta-risk

Misinformation is a threat to society – let's not pretend otherwise

Misinformation is a threat to society – let's not pretend otherwise

Sander van der Linden, Ullrich Ecker, & Stephan Lewandowsky argue misinformation has clear, measurable effects and that downplaying them risks public dialogue.

Zielony Ład zabroni uprawy własnych warzyw i owoców? Fałsz

Fałszywe informacje na temat tego, że Unia Europejska w ramach Zielonego Ładu zabroni uprawy własnych warzyw i owoców.

Demagog

Will the Green Deal ban growing your own vegetables and fruit?

Prezydent Nawrocki całuje swoją doradczynię? To przeróbka stworzona przez AI

W sieci pojawiły się rzekome zdjęcia, na których widać, jak prezydent Nawrocki całuje swoją doradczynię. Sprawdzamy, czy są prawdziwe.

Demagog

Is President Nawrocki kissing his advisor?

The Why and How of Disinformation

Disinformation

Disinformation:
defined as false, inaccurate, or **misleading** information designed, presented, and promoted to **intentionally** cause public harm or for profit. (de Cock Buning, 2018).

Disinformation vs. Misinformation

- Disinformation is defined as false, inaccurate, or misleading information designed, presented, and promoted to intentionally cause public harm or for profit (de Cock Buning, 2018).
- Misinformation is defined as unintentionally false, inaccurate, or misleading information.

The Why and How of Disinformation

Research Objective

Main Objective:
Improve the detection of disinformation, with a focus on identifying **manipulation** and malicious **intent** in a monolingual and multilingual textual data.

- Disinformation detection is a complex task that, according to most definitions, involve intent to harm and faulty or manipulative arguments.

- We propose to break the problem down and isolate two components:
  - persuasion techniques (faulty reasoning) - The How of Disinformation
  - and malicious intent - The Why of Disinformation

Manipulation Techniques

Manipulation and persuasion have many techniques in common, but manipulation is created with malicious intent (Paine et al. 1989). Persuasion also may affect the credibility of information.

Persuasion Techniques

The detection of persuasion overlaps to a large extent with work on the detection of propaganda. Persuasion techniques and propaganda techniques are often used interchangeably (Piskorski et al., 2023).

Propaganda Techniques

Disinformation Narratives

Disinformation narrative we define as a repeating pattern found in several disinformative articles.

Malicious Intent

Corresponds to several disinformation narratives. Intention encapsulates the broader goal of the author, which guides specific narratives used to achieve that goal.

Datasets: Building the Foundation

MIPD: Manipulation and Intention in Polish Disinformation

High quality, novel Polish dataset focused on understanding disinformation.

MultiDis: Multitopic English Disinformation Dataset

Comprises nearly 2,000 English articles on European and global disinformation.

Detecting the Genre, the Framing, and the Persuasion Techniques in Online News in a Multilingual Setup

News articles with persuasion techniques at paragraph level.

EUDisinfo: Pro-Kremlin disinformation

Collected with usage of the EUvsDisinfo database.

Automatic Detection and Characterization of Propaganda Techniques and Narratives from Diplomats of Major Powers

Tweets from diplomats with propaganda techniques.

MALINT: MALicious INTention Dataset

Comprises about 1600 English disinformation articles with malicious intent labels.

MIPD

Manipulation and Intention in Polish Disinformation

Our MIPD dataset is **the first high-quality corpus** specifically designed to analyze **disinformation** within the context of **malicious intent** and **manipulation** techniques.

Three-stage annotations:

- Disinformation: Identifying content as **disinformative** / **credible** / **hard-to-say**.
- Manipulation Techniques: 11 distinct techniques, e.g., Exaggeration, Whataboutism
- Malicious Intention: 9 types, including Undermining Institutions, Promoting Stereotypes

- Thematic Categories: Topics like COVID-19, Climate Crisis, Migration.
- We created annotation guidelines enabling extension of our dataset in other languages.

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

MIPD

Example Results and Article

Manipulation Techniques

Model	Manipulation Technique				Weighted F_1 Score
	Whataboutism	Appeal to Emotion	Exaggeration	Strawman	
HerBERT-B	0.19	0.40	0.64	0.27	0.42
HerBERT-L	**0.30**	0.44	0.66	**0.30**	**0.47**
PL-RoBERTa-B	0.20	0.38	0.64	0.25	0.41
PL-RoBERTa-L	0.26	**0.45**	**0.67**	0.28	**0.47**

Disinformation

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

- Paper presented at **EMNLP** Main Conference, Miami 2024

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

MALINT

MALicious INTention Dataset

Our MALINT dataset is the first English high-quality corpus designed to analyze **disinformation** within the context of **malicious intent.**

Statistic	Value
Total No. of Articles	1,600
Avg. Article Length (words)	963
Avg. Article Length (characters)	6,045

Malicious intent categories:

- Undermining the Credibility of Public Institutions [UCPI]
- Changing Political Views [CPV]
- Undermining International Organizations and Alliances [UIOA]
- Promoting Social Stereotypes/Antagonisms [PSSA]
- Promoting Anti-scientific Views [PASV]

Statistic	UCPI	UIOA	PASV	PSSA	CPV
Count	321	234	154	222	197
%	20.06	14.63	9.63	13.88	12.31

What is novel about this dataset?

- First English human-annotated dataset with disinformation and malicious intent
- We present all annotation including intermediate annotations

MALINT

Example Results and Article

Malicious intent categories:

- Undermining the Credibility of Public Institutions [UCPI]
- Changing Political Views [CPV]
- Undermining International Organizations and Alliances [UIOA]
- Promoting Social Stereotypes/Antagonisms [PSSA]
- Promoting Anti-scientific Views [PASV]

Paper under review

Model	UCPI	UIOA	PASV	PSSA	CPV
Small Language Models					
BERT Base	0.562	0.484	0.500	**0.614**	0.293
BERT Large	0.528	0.437	0.543	0.529	0.306
DeBERTa V3 Base	0.675	0.505	0.580	0.523	0.400
DeBERTa V3 Large	**0.696**	**0.649**	**0.683**	0.547	0.460
RoBERTa Base	0.693	0.547	0.674	0.515	**0.486**
RoBERTa Large	0.682	0.630	0.680	0.505	0.444
DistilBERT Base	0.599	0.547	0.564	0.450	0.400
Large Language Models					
GPT 4o Mini	0.543	0.547	0.632	0.458	0.324
GPT 4.1 Mini	**0.702**	0.469	0.717	**0.479**	0.371
Gemini 2.0 Flash	0.639	**0.604**	0.722	0.452	0.444
Llama 3.3 70B	0.569	0.427	**0.738**	0.415	**0.496**
Gemma 3 27B it	0.682	0.395	0.667	0.424	0.407
Baselines					
Random	0.279	0.205	0.122	0.179	0.162
LR with BoW	0.581	0.477	0.595	0.424	0.376

Model	Micro F_1	Weighted F_1
Small Language Models		
BERT Base	0.421	0.414
BERT Large	0.578	0.521
DeBERTa V3 Base	0.812	0.804
DeBERTa V3 Large	**0.817**	0.815
RoBERTa Base	0.813	**0.821**
RoBERTa Large	0.775	0.808
DistilBERT Base	0.759	0.769
Large Language Models		
GPT 4o Mini	0.446	0.457
GPT 4.1 Mini	0.489	0.498
Gemini 2.0 Flash	0.410	0.404
Llama 3.3 70B	**0.542**	**0.570**
Gemma 3 27B it	0.440	0.485
Baselines		
Random	0.192	0.201
LR with BoW (OvR)	0.503	0.491

MALicious INTent Dataset and Inoculating LLMs for Enhanced Disinformation Detection

SlavicNLP 2025

Detection and Classification of Persuasion Techniques in Parliamentary Debates and Social Media

Dataset with Parliamentary Debates and Social Media for ACL co-located SlavicNLP Workshop

The task was structured into two subtasks:

- Detection, to determine whether a given text fragment contains persuasion techniques, and
- Classification, to determine for a given text fragment which persuasion techniques are present therein using a taxonomy of 25 persuasion technique taxonomy

SlavicNLP 2025 Shared Task: Detection and Classification of Persuasion Techniques in Parliamentary Debates and Social Media

Jakub Piskorski,[1] **Dimitar Dimitrov,**[2] **Filip Dobranić,**[3] **Marina Ernst,**[4] **Jacek Haneczok,**[5] **Ivan Koychev,**[2] **Nikola Ljubešić,**[6,3] **Michał Marcińczuk,**[7] **Arkadiusz Modzelewski,**[8,9] **Ivo Moravski,**[3] **Roman Yangarber**[10]

[1]Institute of Computer Science, Polish Academy of Science, Poland jpiskorski@gmail.com
[2] Sofia University "St. Kliment Ohridski", Bulgaria {ilijanovd, koychev, moravski}@fmi.uni-sofia.bg
[3] Institute for Contemporary History, Ljubljana, Slovenia filip.dobranic@inz.si
[4] University of Koblenz, Germany marinaernst@uni-koblenz.de
[5] Visa Technology Europe jacek.haneczok@gmail.com
[6] Jožef Stefan Institute, Ljubljana, Slovenia nikola.ljubesic@ijs.si
[7] CodeNLP, Poland marcinczuk@gmail.com
[8] University of Padua, Italy arkadiusz.modzelewski@unipd.it
[9] Polish-Japanese Academy of Information Technology, Poland arkadiusz.modzelewski@pja.edu.pl
[10] University of Helsinki first.last@helsinki.fi

Persuasion Detection

BG		HR		PL		RU		SI	
Team	F_1	Team	F_1	Team	F_1	Team	F_1	Team	F_1
FactUE	0.88	FactUE	0.96	oplot	0.90	INSAntive	0.87	UFAL4DEM	0.86
baseline	0.88	baseline	0.94	syntax_squad	0.90	Gradient-Flush	0.86	FactUE	0.85
oplot	0.87	UFAL4DEM	0.94	FactUE	0.90	UFAL4DEM	0.86	baseline	0.85
syntax_squad	0.87	oplot	0.92	baseline	0.90	FactUE	0.84	oplot	0.85
UFAL4DEM	0.86	INSAntive	0.89	UFAL4DEM	0.89	baseline	0.83	syntax_squad	0.82
Gradient-Flush	0.84	Gradient-Flush	0.85	Gradient-Flush	0.88	oplot	0.83	Gradient-Flush	0.81
PSAL_NLP	0.82	PSAL_NLP	0.83	INSAntive	0.88	syntax_squad	0.80	INSAntive	0.65
INSAntive	0.81			PSAL_NLP	0.83	PSAL_NLP	0.73	PSAL_NLP	0.62

Persuasion Classification

BG		HR		PL		RU		SI	
Team	F_1^m	Team	F_1^m	Team	F_1^m	Team	F_1^m	Team	F_1^m
PSAL_NLP	0.41	Gradient-Flush	0.49	PSAL_NLP	0.42	INSAntive	0.30	Gradient-Flush	0.32
INSAntive	0.34	PSAL_NLP	0.44	Gradient-Flush	0.41	PSAL_NLP	0.29	PSAL_NLP	0.30
Gradient-Flush	0.34	baseline	0.44	INSAntive	0.41	oplot	0.21	baseline	0.27
dutir	0.28	UFAL4DEM	0.36	FactUE	0.39	Gradient-Flush	0.19	INSAntive	0.20
FactUE	0.23	dutir	0.30	dutir	0.36	UFAL4DEM	0.13	dutir	0.19
UFAL4DEM	0.21	INSAntive	0.30	UFAL4DEM	0.25	FactUE	0.11	oplot	0.18
oplot	0.20	oplot	0.27	baseline	0.24	baseline	0.02	UFAL4DEM	0.17
baseline	0.16	FactUE	0.17	oplot	0.20			FactUE	0.08

Persuasion and Intent-Augmented Reasoning for Enhanced Disinformation Detection

Persuasion and Intent-Augmented Reasoning

Motivation

- Psychological studies show that teaching individuals to recognize persuasive fallacies improves their ability to distinguish between real and fake news (Hruschka et al. 2023, PLoS One)

- Inoculation theory suggests that, just as people can be protected against viruses through vaccines, they can also be “vaccinated” to resist persuasive messages (McGuire et al., 1964).
- An inoculation message has two parts:
  - a threat - alerts individuals that a persuasive attack is coming (Lewandowsky et al., 2017)
  - refutational preemption (prebunking) - involves providing people with arguments or tools to resist persuasive attacks, helping them better recognize and respond to such attempts (Pfau et al., 2005)

PCoT

Persuasion-Augmented Chain of Thought for Detecting Fake News and Social Media Disinformation

Main Objective:
Does incorporating knowledge of persuasion strategies into LLMs enhance their disinformation detection performance?

- A critical aspect of disinformation is its coexistence with manipulation and persuasion to mislead audiences.

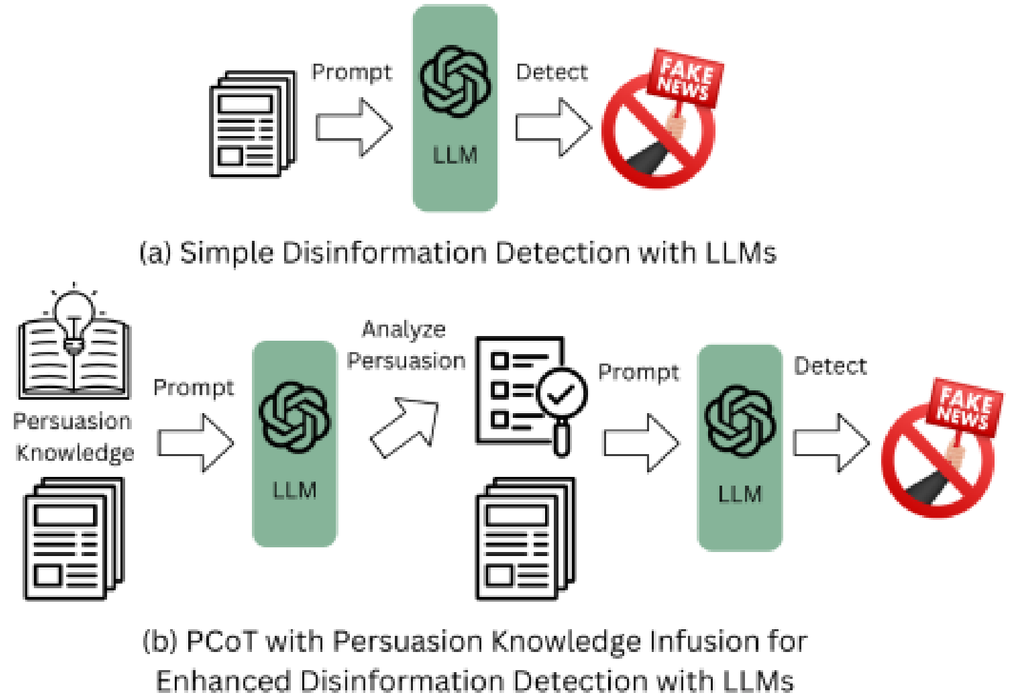
"PCoT: Persuasion-Augmented Chain of Thought for Detecting Fake News and Social Media Disinformation",
Modzelewski et al., ACL 2025, Vienna, Austria

Persuasion-Augmented Chain of Thought

A Two-Stage Method

I Stage:
An LLM is prompted to perform multi-faceted reasoning by analyzing persuasion strategies within the text.

II Stage:
The second stage performs the disinformation detection task, enriched by the previously generated analysis of persuasion strategies.



PCoT: A Two-Stage Method

I Stage:

For the first stage, we developed prompts that incorporate knowledge of persuasion strategies. Persuasion strategies used in our study presented by Piskorski et al. at ACL 2023:

- Attack on reputation
- Justification
- Simplification
- Distraction
- Call
- Manipulative wording

Method	F_1 Micro
DMT	↑9% 0.722 ±0.035
DTAT	↑4% 0.689 ±0.042
Base MT	0.664±0.030

Objective: Multi-label zero-shot detection of persuasion strategies with contextual explanations per each persuasion strategy.

Finding: Using a single prompt to identify all persuasion strategies was more effective than separate prompts for each strategy's binary classification.

PCoT: A Two-Stage Method

II Stage:
The previously generated analysis of persuasion strategies, which includes the determination of whether a persuasion strategy is present in the text and a contextual explanation for its presence, is passed to the second stage, along with a prompt for the task of detecting disinformation.

We have chosen three competitive methods from Lucas et al. (2023) that served as baselines, allowing us to evaluate the effectiveness of PCoT: VaN, Z-CoT, DeF-SpeC

Objective: Zero-shot binary classification of disinformation.

Datasets Used in our Study

High-quality datasets used in the study:

- **MultiDis** - English dataset that includes multiple thematic categories.
- **EUDisinfo** is a newly constructed dataset containing recent and up-to-date examples.
- **CoAID** - fake news and social media disinformation.
- **ECTF** - social media disinformation, posts from X platform (former Twitter).
- **ISOT Fake News** - A dataset of 44k+ fake and reliable articles from reputable and unreliable sources, identified via Politifact.

PCoT was evaluated on 4 groups of texts:

- Fake News
- Social Media Disinformation
- Prior Cutoff Data
- Post Cutoff Data

Persuasion-Augmented Chain of Thought

Results

Competitive methods presented by Lucas et al. at EMNLP 2023 and taken as baselines:

- VaN
- Z-CoT
- DeF-SpeC

Models used for evaluation of PCoT:

- GPT 4o Mini,
- Llama 3.1 8B,
- Claude 3 Haiku,
- Llama 3.3 70B,
- Gemini 1.5 Flash

Evaluation Metric for all experiments:
F1 score

	Overall		Articles		Posts		Prior Cutoff		Post Cutoff	
	Base	PCoT	Base	PCoT	Base	PCoT	Base	PCoT	Base	PCoT
GPT 4o Mini										
VaN	0.759	0.845 ↑11%	0.788	0.885 ↑12%	0.700	0.762 ↑9%	0.742	0.830 ↑12%	0.790	0.874 ↑11%
Z-CoT	0.765	0.846 ↑11%	0.801	0.884 ↑10%	0.696	0.767 ↑10%	0.747	0.835 ↑12%	0.801	0.869 ↑8%
DeF-SpeC	0.772	0.834 ↑8%	0.816	0.867 ↑6%	0.690	0.766 ↑11%	0.742	0.813 ↑10%	0.832	0.875 ↑5%
Gemini 1.5 Flash										
VaN	0.681	0.810 ↑19%	0.673	0.843 ↑25%	0.695	0.748 ↑8%	0.683	0.778 ↑14%	0.679	0.875 ↑29%
Z-CoT	0.689	0.808 ↑17%	0.681	0.838 ↑23%	0.703	0.752 ↑7%	0.670	0.777 ↑16%	0.687	0.872 ↑27%
DeF-SpeC	0.744	0.834 ↑12%	0.764	0.876 ↑15%	0.708	0.754 ↑6%	0.721	0.810 ↑12%	0.790	0.884 ↑12%
Claude 3 Haiku										
VaN	0.710	0.797 ↑12%	0.714	0.820 ↑15%	0.702	0.747 ↑6%	0.728	0.797 ↑9%	0.677	0.796 ↑18%
Z-CoT	0.588	0.774 ↑32%	0.601	0.800 ↑33%	0.550	0.716 ↑30%	0.565	0.767 ↑36%	0.626	0.786 ↑26%
DeF-SpeC	0.780	0.795 ↑2%	0.806	0.810 ↑0%	0.727	0.763 ↑5%	0.809	0.812 ↑0%	0.727	0.766 ↑5%
Llama 3.3 70B										
VaN	0.740	0.845 ↑14%	0.747	0.881 ↑18%	0.727	0.768 ↑6%	0.733	0.839 ↑14%	0.752	0.856 ↑14%
Z-CoT	0.722	0.843 ↑17%	0.725	0.878 ↑21%	0.718	0.770 ↑7%	0.707	0.837 ↑18%	0.750	0.855 ↑14%
DeF-SpeC	0.732	0.832 ↑14%	0.740	0.863 ↑17%	0.717	0.768 ↑7%	0.719	0.806 ↑12%	0.755	0.880 ↑17%
Llama 3.1 8B										
VaN	0.627	0.792 ↑26%	0.565	0.802 ↑42%	0.736	0.773 ↑5%	0.649	0.788 ↑21%	0.585	0.801 ↑37%
Z-CoT	0.660	0.791 ↑20%	0.623	0.804 ↑29%	0.725	0.764 ↑5%	0.670	0.789 ↑18%	0.638	0.795 ↑25%
DeF-SpeC	0.697	0.773 ↑11%	0.688	0.784 ↑14%	0.712	0.752 ↑6%	0.683	0.767 ↑12%	0.724	0.785 ↑8%
Average	0.711	0.815 ↑15%	0.715	0.842 ↑18%	0.700	0.758 ↑8%	0.705	0.803 ↑14%	0.721	0.838 ↑16%

Persuasion-Augmented Chain of Thought

Further Evaluation

Model	Z-CoT	RaR	CoVe	PCoT
GPT 4o Mini	0.765	0.698	0.790	**0.846**
Gemini 1.5 Flash	0.689	0.573	0.736	**0.808**
Claude 3 Haiku	0.588	0.768	0.441	**0.774**
Llama 3.3 70B	0.722	0.657	0.835	**0.843**
Llama 3.1 8B	0.660	0.566	0.764	**0.791**

Overall F1 scores of different prompting methods on five datasets.

Model	PCoT BV	Base
GPT 4o Mini	0.814 ↑ ± 0,007	0.765 ± 0,007
Gemini 1.5 Flash	0.790 ↑ ± 0,014	0.705 ± 0,034
Claude 3 Haiku	0.736 ↑ ± 0,013	0.693 ± 0,097
Llama 3.3 70B	0.831 ↑ ± 0,007	0.731 ± 0,009
Llama 3.1 8B	0.785 ↑ ± 0,011	0.661 ± 0,035

Comparison of average F1 scores and standard deviations between Base prompts and PCoT without persuasion strategy augmentation.

Model	Overall
GPT 4o Mini + PCoT	**0.846**
Llama 3.1 8B + PCoT	0.791
o3-mini	0.770
o1-mini	0.634

Overall F1 scores for PCoT-enhanced models vs. OpenAI reasoning models on five datasets.

Persuasion-Augmented Chain of Thought

Findings

- Four specific persuasion strategies are particularly characteristic for disinformation: **Attack on reputation, Simplification, Distraction, and Manipulative wording**.
- **Persuasion is more commonly used in disinformation** than in credible information, though a significant proportion of credible content also contains persuasion.

- Infusing persuasion knowledge **improves** generative LLMs' **disinformation detection**, **especially for long texts and data not seen** during training.
- Detecting disinformation is particularly challenging in texts where no persuasion strategy has been predicted.

Model	Persuasion		No Persuasion	
	PCoT	Base	PCoT	Base
GPT 4o Mini	0.872 ↑ ± 0,006	0.824 ± 0,008	0.342 ↑ ± 0,025	0.305 ± 0,009
Gemini 1.5 Flash	0.844 ↑ ± 0,014	0.738 ± 0,036	0.444 ↑ ± 0,013	0.430 ± 0,007
Claude 3 Haiku	0.831 ↑ ± 0,014	0.756 ± 0,101	0.177 ↓ ± 0,043	0.295 ± 0,084
Llama 3.3 70B	0.871 ↑ ± 0,007	0.781 ± 0,007	0.409 ↑ ± 0,010	0.343 ± 0,006
Llama 3.1 8B	0.812 ↑ ± 0,008	0.679 ± 0,050	0.536 ↑ ± 0,014	0.494 ± 0,059
Average	0.847 ↑	0.753	0.392 ↑	0.368

MALINT

MALicious INTent Dataset and Inoculating LLMs for Enhanced Disinformation Detection

One of the main objectives:
Does incorporating knowledge of intent into LLMs enhance their disinformation detection performance?

- The second critical aspect of disinformation is its maliciously intentional promotion

Intent-Augmented Chain of Thought

Results

Competitive methods presented by Lucas et al. at EMNLP 2023 and taken as baselines:

- VaN
- Z-CoT
- DeF-SpeC

Models used for evaluation of PCoT:

- GPT 4o Mini,
- GPT 4.1 Mini
- Llama 3.3 70B,
- Gemini 2.0 Flash
- Gemma 3 27B it

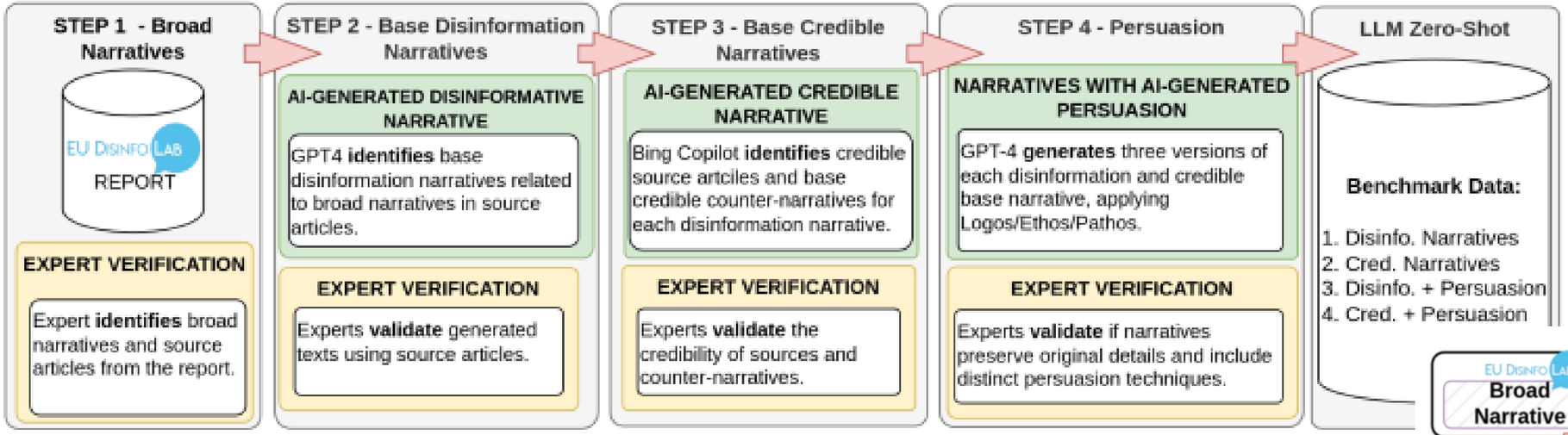
Evaluation Metric for all experiments:
F1 score

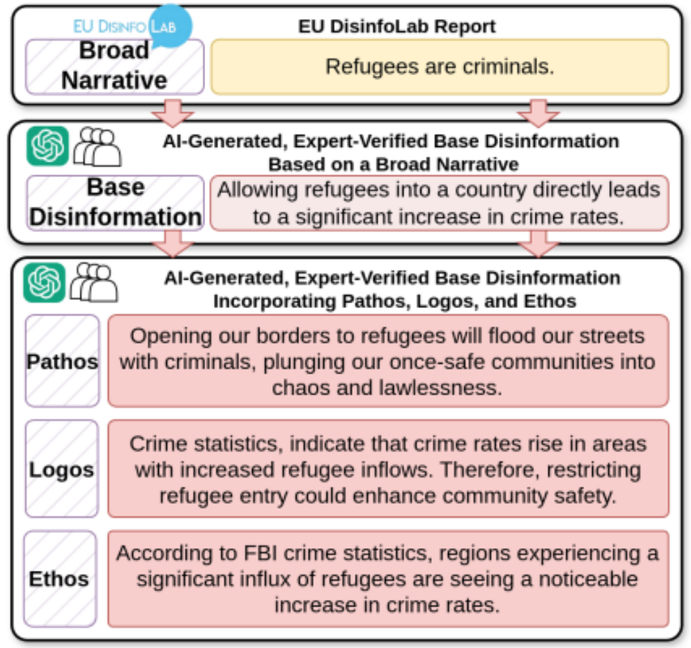
	Overall		Articles		Posts		Prior Cutoff		Post Cutoff	
	Base	IBI	Base	IBI	Base	IBI	Base	IBI	Base	IBI
GPT 4o Mini										
VaN	0.736	0.828 ↑13%	0.754	0.862 ↑14%	0.703	0.755 ↑7%	0.727	0.821 ↑13%	0.762	0.846 ↑11%
Z-CoT	0.740	0.826 ↑12%	0.764	0.854 ↑12%	0.692	0.766 ↑11%	0.724	0.823 ↑14%	0.786	0.833 ↑6%
DeF-SpeC	0.746	0.792 ↑6%	0.782	0.817 ↑4%	0.682	0.742 ↑9%	0.712	0.771 ↑8%	0.843	0.850 ↑1%
GPT 4.1 Mini										
VaN	0.698	0.751 ↑8%	0.718	0.772 ↑8%	0.659	0.705 ↑7%	0.672	0.709 ↑6%	0.767	0.862 ↑12%
Z-CoT	0.673	0.748 ↑11%	0.685	0.765 ↑12%	0.649	0.712 ↑10%	0.640	0.710 ↑11%	0.757	0.849 ↑12%
DeF-SpeC	0.748	0.780 ↑4%	0.780	0.803 ↑3%	0.686	0.732 ↑7%	0.720	0.752 ↑4%	0.828	0.856 ↑3%
Gemini 2.0 Flash										
VaN	0.701	0.762 ↑9%	0.703	0.803 ↑14%	0.699	0.677 ↓3%	0.682	0.731 ↑7%	0.754	0.851 ↑13%
Z-CoT	0.670	0.733 ↑9%	0.667	0.763 ↑14%	0.675	0.670 ↓1%	0.646	0.694 ↑7%	0.736	0.838 ↑14%
DeF-SpeC	0.767	0.803 ↑5%	0.795	0.835 ↑5%	0.710	0.738 ↑4%	0.749	0.787 ↑5%	0.814	0.847 ↑4%
Gemma 3 27b it										
VaN	0.694	0.773 ↑11%	0.684	0.801 ↑17%	0.711	0.710 ↓0%	0.662	0.750 ↑13%	0.782	0.830 ↑6%
Z-CoT	0.622	0.767 ↑23%	0.671	0.793 ↑18%	0.516	0.711 ↑38%	0.561	0.746 ↑33%	0.775	0.822 ↑6%
DeF-SpeC	0.739	0.791 ↑7%	0.742	0.825 ↑11%	0.734	0.720 ↓2%	0.712	0.769 ↑8%	0.815	0.851 ↑4%
Llama 3.3 70B										
VaN	0.756	0.770 ↑2%	0.762	0.796 ↑4%	0.744	0.717 ↓4%	0.730	0.738 ↑1%	0.824	0.856 ↑4%
Z-CoT	0.736	0.781 ↑6%	0.739	0.804 ↑9%	0.730	0.733 ↑0%	0.714	0.748 ↑5%	0.793	0.867 ↑9%
DeF-SpeC	0.716	0.762 ↑6%	0.723	0.788 ↑9%	0.702	0.707 ↑1%	0.684	0.720 ↑5%	0.798	0.872 ↑9%
Average	0.716	0.778 ↑9%	0.731	0.805 ↑10%	0.686	0.720 ↑5%	0.689	0.751 ↑9%	0.789	0.849 ↑8%

Disinformation Narratives:
Evaluating and Mining with LLMs

EU DisinfoTest

Benchmark for Evaluating LLMs' Ability to Detect Disinformation Narratives



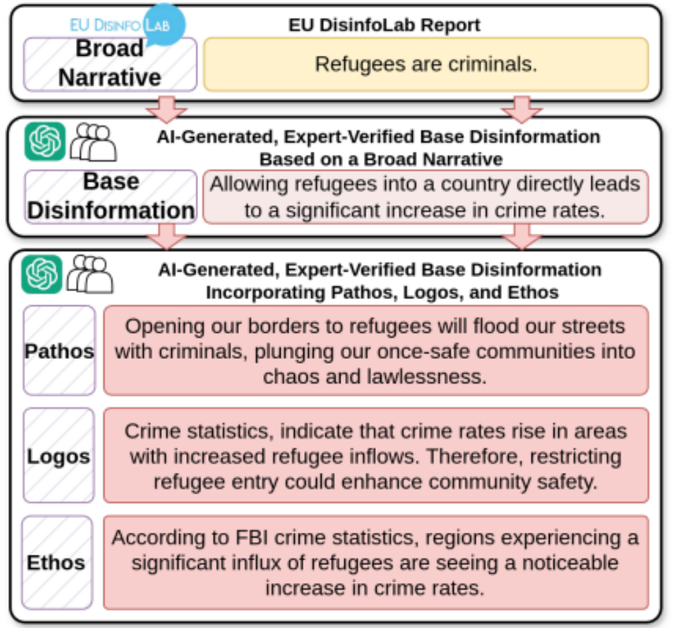


Reminder: We define disinformation narrative as a repeating pattern found in several disinformative articles.

EU DisinfoTest Benchmark informed by EU DisinfoLab research and created through a Human-in-the-Loop approach includes over 1,300 narratives that incorporate three modes of persuasion

EU DisinfoTest

Impact of Modes of Persuasion on Disinformation Narrative Detection

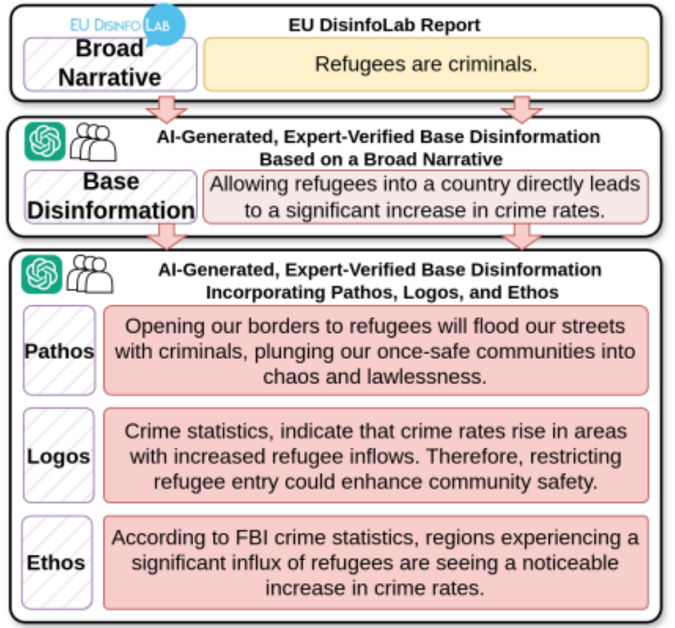


- In the **Base** scenario, most models reliably identify disinformation, as indicated by an average TNR of 0.95
- The ability to detect disinformation narratives under **Pathos** slightly improves, with the TNR at 0.97
- Under influence of the **Logos** the ability to detect disinformation shows significant drop, with the TNR dropping by an average of 15 p.p. from the Base.
- The introduction of **Ethos** also affect the ability to detect disinformation, with a substantial 28 p.p. decrease in the TNR from the Base.

	TNR			
Model	Base	Pathos	Logos	Ethos
L3-70b	0.99	0.99 ↑0%	0.86 ↓12%	0.70 ↓29%
Opus	0.98	0.99 ↑0%	0.88 ↓9%	0.86 ↓11%
L3-8b	0.98	0.98 ↓0%	0.82 ↓16%	0.69 ↓29%
Mixtral	0.98	0.91 ↓6%	0.59 ↓39%	0.27 ↓72%
GPT4o	0.96	0.99 ↑3%	0.86 ↓10%	0.83 ↓13%
GPT3.5	0.92	0.99 ↑6%	0.73 ↓21%	0.66 ↓28%
Sonnet	0.91	0.96 ↑5%	0.89 ↓2%	0.86 ↓6%
Haiku	0.90	0.96 ↑6%	0.80 ↓11%	0.57 ↓36%
Average	0.95	0.97 ↑2%	0.80 ↓15%	0.68 ↓28%

EU DisinfoTest

Impact of Modes of Persuasion on Disinformation Narrative Detection



- In the **Base** scenario, an average TPR is equal to 0.91.
- In the **Pathos** scenario, TPR notably decreases to an average of 0.69 for all models, a 23 p.p. reduction from the Base scenario.
- Under the influence of **Logos** persuasion, there is a moderate impact on the ability of models to detect credible statements, with the TPR decreasing by an average of 6 p.p. compared to the Base scenario.
- The introduction of **Ethos** improves the models' ability to identify credible statements, with the TPR rising by an average of 1 p.p. compared to the Base

	TPR			
Model	Base	Pathos	Logos	Ethos
Haiku	0.97	0.80 ↓17%	0.89 ↓8%	0.96 ↓0%
GPT4o	0.96	0.86 ↓10%	0.90 ↓6%	0.91 ↓4%
Mixtral	0.95	0.89 ↓6%	0.96 ↑1%	0.98 ↑2%
L3-70b	0.94	0.65 ↓30%	0.86 ↓8%	0.95 ↑0%
Opus	0.93	0.69 ↓25%	0.82 ↓11%	0.98 ↑5%
Sonnet	0.93	0.76 ↓18%	0.88 ↓5%	0.96 ↑4%
GPT3.5	0.81	0.44 ↓45%	0.76 ↓6%	0.86 ↑5%
L3-8b	0.73	0.45 ↓38%	0.67 ↓8%	0.70 ↓4%
Average	0.91	0.69 ↓23%	0.86 ↓6%	0.92 ↑1%

EU DisinfoTest

Impact of Modes of Persuasion on Disinformation Narrative Detection

- **Ethos** enhances the perceived credibility of the analyzed narratives - Notable decrease in the models' ability to detect basic disinformation by an average of 28%, alongside a slight improvement in detecting credible narratives.
- **Pathos** shows an opposite influence, reducing the overall perceived credibility of the narratives. This suggests that infusing texts with emotional content makes them more susceptible to being misidentified as disinformation.
- **Logos** primarily reduces the TNR, but it can also lower the TPR. It may introduce the complexities that confuse the model in both scenarios.

	TPR			
Model	Base	Pathos	Logos	Ethos
Haiku	0.97	0.80 ↓17%	0.89 ↓8%	0.96 ↓0%
GPT4o	0.96	0.86 ↓10%	0.90 ↓6%	0.91 ↓4%
Mixtral	0.95	0.89 ↓6%	0.96 ↑1%	0.98 ↑2%
L3-70b	0.94	0.65 ↓30%	0.86 ↓8%	0.95 ↑0%
Opus	0.93	0.69 ↓25%	0.82 ↓11%	0.98 ↑5%
Sonnet	0.93	0.76 ↓18%	0.88 ↓5%	0.96 ↑4%
GPT3.5	0.81	0.44 ↓45%	0.76 ↓6%	0.86 ↑5%
L3-8b	0.73	0.45 ↓38%	0.67 ↓8%	0.70 ↓4%
Average	0.91	0.69 ↓23%	0.86 ↓6%	0.92 ↑1%

	TNR			
Model	Base	Pathos	Logos	Ethos
L3-70b	0.99	0.99 ↑0%	0.86 ↓12%	0.70 ↓29%
Opus	0.98	0.99 ↑0%	0.88 ↓9%	0.86 ↓11%
L3-8b	0.98	0.98 ↓0%	0.82 ↓16%	0.69 ↓29%
Mixtral	0.98	0.91 ↓6%	0.59 ↓39%	0.27 ↓72%
GPT4o	0.96	0.99 ↑3%	0.86 ↓10%	0.83 ↓13%
GPT3.5	0.92	0.99 ↑6%	0.73 ↓21%	0.66 ↓28%
Sonnet	0.91	0.96 ↑5%	0.89 ↓2%	0.86 ↓6%
Haiku	0.90	0.96 ↑6%	0.80 ↓11%	0.57 ↓36%
Average	0.95	0.97 ↑2%	0.80 ↓15%	0.68 ↓28%

Paper presented at **EMNLP 2024**
Finidngs Conference in Miami

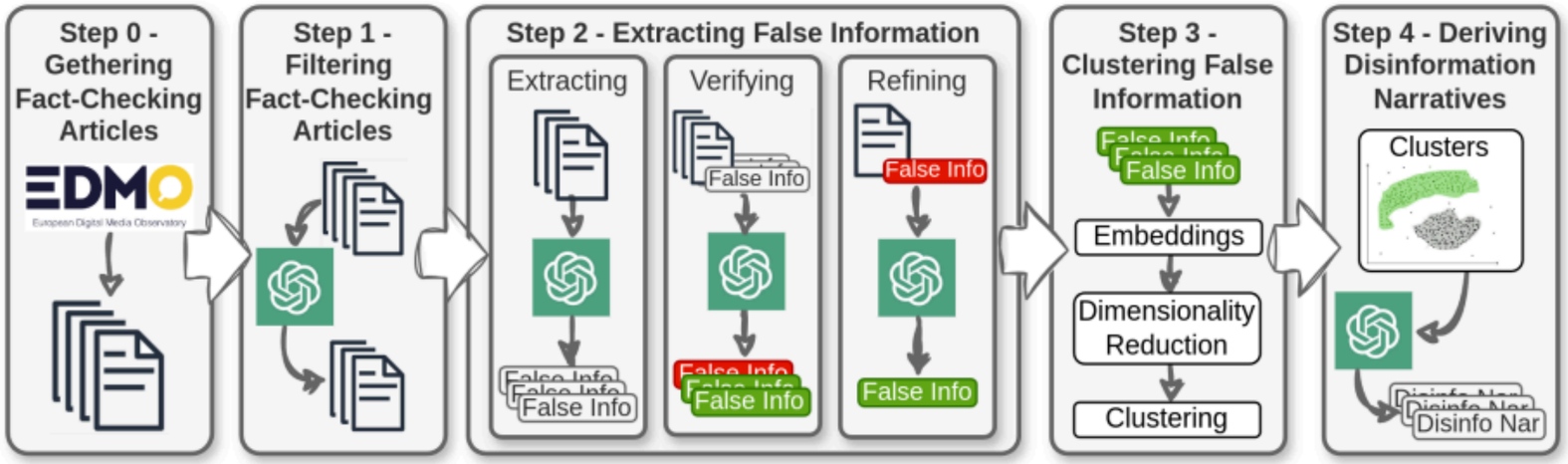
EU DisinfoTest: a Benchmark for Evaluating Language Models' Ability to Detect Disinformation Narratives

Witold Sosnowski[1], Arkadiusz Modzelewski[1], Kinga Skorupska[1], Jahna Otterbacher[2], Adam Wierzbicki[1]

[1]Polish-Japanese Academy of Information Technology
[2]Open University of Cyprus

Disinformation Narrative Mining with LLMs

DiNaM



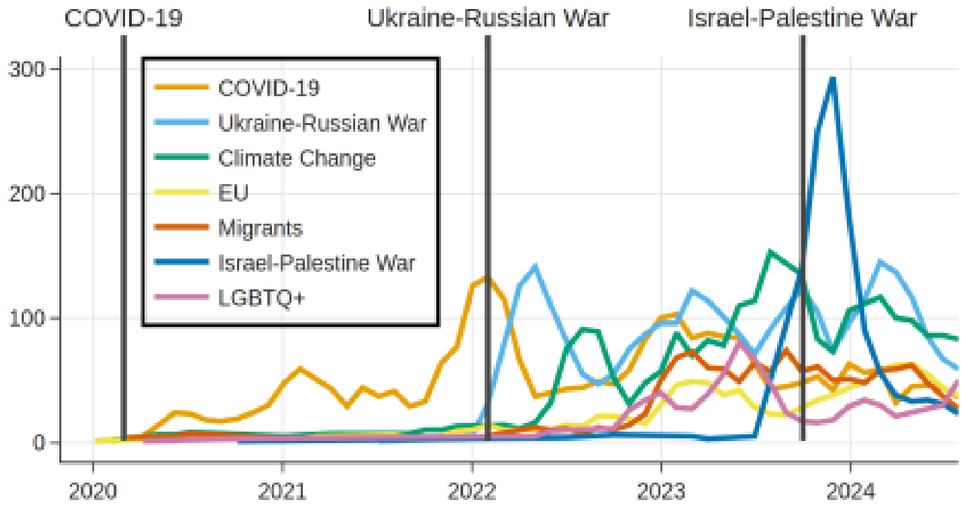
Reminder: We define disinformation narrative as a repeating pattern found in several disinformative articles.

- **Identify false information**: DiNaM identifies false information by first filtering fact-checking articles that conclude the claim being reviewed is false, and then extracting instances of the false information from these articles.
- **Cluster false information**: DiNaM groups identified false information into clusters based on semantic similarities.
- **Derive narratives**: For each cluster, DiNaM synthesizes patterns of information into disinformation narratives.

Disinformation Narrative Mining with LLMs

DiNaM

- We manually categorized the narratives discovered by DiNaM into **seven main topics.**
- This categorization enabled us to track the evolution of disinformation topics in response to real-world events.
- In 2020, as the COVID-19 pandemic unfolded, disinformation narratives related to the virus emerged.
- In 2022, as the war between Ukraine and Russia escalated, disinformation surrounding the conflict intensified.
- A significant spike in Israeli-Palestinian disinformation occurred in late 2023, coinciding with the onset of the Israel-Palestine war.



DiNaM: Disinformation Narrative Mining with Large Language Models

Witold Sosnowski[1], Arkadiusz Modzelewski[1,2,3], Kinga Skorupska[1], Adam Wierzbicki[1]

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

Paper soon to be presented at **EMNLP 2025**
Main Conference in Suzhou, China

Conclusions.. and it is soon the end, but

still quite a lot of things to do

- What language tools do disinformation agents employ?
- Can incorporating persuasion and intent knowledge enhance the ability of LLMs in disinformation detection?
- We have begun exploring a new direction in disinformation research, focusing on the study of disinformation narratives.

My final goals for PhD studies and dissertation:

- Unifying research and experiments on persuasion and intention-augmented reasoning.
- Publication of a corpus of approximately 7,000 texts from various sources annotated with disinformation, intentions, and manipulative techniques in English:
  - Dataset with improved methodology and guidelines
  - annotated by 4 different universities across Europe
- Parliamentary debates dataset and different ideas

Further Reading

Additional References

For those interested in **Disinformation and Disinformation Narratives** Detection:

- **MIPD**: Exploring manipulation and intention in a novel corpus of polish disinformation, Arkadiusz Modzelewski, Giovanni Da San Martino, Pavel Savov, Magdalena Wilczyńska, and Adam Wierzbicki, EMNLP Main 2024
- ***PolyNarrative***: *A Multilingual, Multilabel, Multi-domain Dataset for Narrative Extraction from News Articles, Nikolaos Nikolaidis, Nicolas Stefanovitch, Purificação Silvano, Dimitar Iliyanov Dimitrov, Roman Yangarber, Nuno Guimarães, Elisa Sartori, Ion Androutsopoulos, Preslav Nakov, Giovanni Da San Martino, Jakub Piskorski, ACL Main 2025*
- ***EU DisinfoTest***: *a Benchmark for Evaluating Language Models' Ability to Detect Disinformation Narratives*, Witold Sosnowski, Arkadiusz Modzelewski, Kinga Skorupska, Jahna Otterbacher, Adam Wierzbicki, EMNLP Findings 2024
- ***DiNaM***: *Disinformation Narrative Mining with Large Language Models*, Witold Sosnowski, Arkadiusz Modzelewski, Kinga Skorupska, Adam Wierzbicki, EMNLP Main 2025

For those interested in **Persuasion Techniques**:

- **PCoT**: Persuasion-Augmented Chain of Thought for Detecting Fake News and Social Media Disinformation, Arkadiusz, Modzelewski, Witold Sosnowski, Tiziano Labruna, Adam Wierzbicki, and Giovanni Da San Martino, ACL Main 2025
- *Fine-Grained* ***Analysis of Propaganda in News*** *Articles*, Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, Preslav Nakov, EMNLP 2019
- *SemEval-2020 Task 11: Detection of* ***Propaganda Techniques in News*** *Articles*, Giovanni Da San Martino, Alberto Barrón-Cedeño, Henning Wachsmuth, Rostislav Petrov, Preslav Nakov, SemEval 2020
- *SemEval-2023 Task 3: Detecting the Category, the Framing, and the* ***Persuasion Techniques*** *in Online News in a* ***Multi-lingual Setup***, Jakub Piskorski, Nicolas Stefanovitch, Giovanni Da San Martino, Preslav Nakov, SemEval 2023,

Any Questions?

ARKADIUSZ MODZELEWSKI

Website: amodzelewski.com

arkadiusz.modzelewski@pja.edu.pl
arkadiusz.modzelewski@unipd.it
arkadiusz.modzelewski@nask.pl