Can GPT-4 Identify Propaganda?

Annotation and Detection of Propaganda Spans in News Articles

Maram Hasanain, Fatema Ahmad, Firoj Alam

{mhasanain, fakter, fialam}@hbku.edu.qa

QCRI
معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
جامعة حمد بن خليفة
HAMAD BIN KHALIFA UNIVERSITY

Introduction

- Propaganda techniques can influence readers opinions and actions.
- Need to design systems to detect them and the associated text spans.
- Research on Arabic content is relatively sparse and the datasets are limited in size.

وألقت القيادة الإيرانية باللوم على الغرب في هذه الاحتجاجات، وقالت إن الاضطرابات نتيجة ""مؤامرة"" تورطت فيها الولايات المتحدة وإسرائيل و""خونة إيرانيون في الخارج""

Overlap

Translation: The Iranian leadership blamed the West for these protests, and stated that the disturbances were the result of a "conspiracy" involving the United States, Israel, and "Iranian traitors abroad."

Overlap

Techniques: Smears, Loaded language, Name Calling

Contributions

- Release the largest dataset to date, ***ArPro***, for fine-grained propaganda detection
- Detailed insights on data collection and annotation, and comprehensive dataset statistics
- Investigate and compare the performance of GPT-4 for detecting and labeling spans with propagandistic techniques

ArPro VS. Existing Datasets

Reference	Lang	Content	# Items	# T
(Barrón-Cedeno et al., 2019)	En	News article	51,000	2
(Da San Martino et al., 2019)	En	News article	451	18
(Dimitrov et al., 2021b)	En	Memes	950	22
(Vijayaraghavan and Vosoughi, 2022)	En	Tweets	1,000	19
(Piskorski et al., 2023b)	En, Fr, de, It, Pl, Ru, Es, El, Ka	News article	2,049	23
(Alam et al., 2022b)	Ar		930	19
ArAIEval-23 (Hasanain et al., 2023a)	Ar	Paragraphs, Tweets	3,189	23
Ours	Ar	Paragraphs	8,000	23

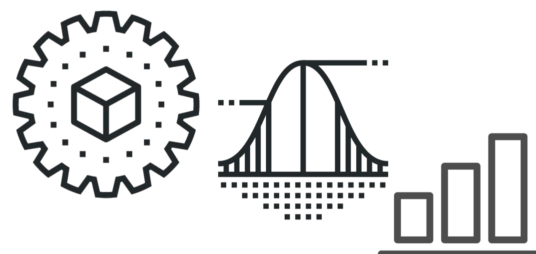
Constructing ArPro

Step 1

Acquire raw data

In-house dataset of over 600K news articles from ~400 Arabic news domains

Step 2



Prepare & sample

Parse articles, split into paragraphs, clean, remove duplicates and sampling

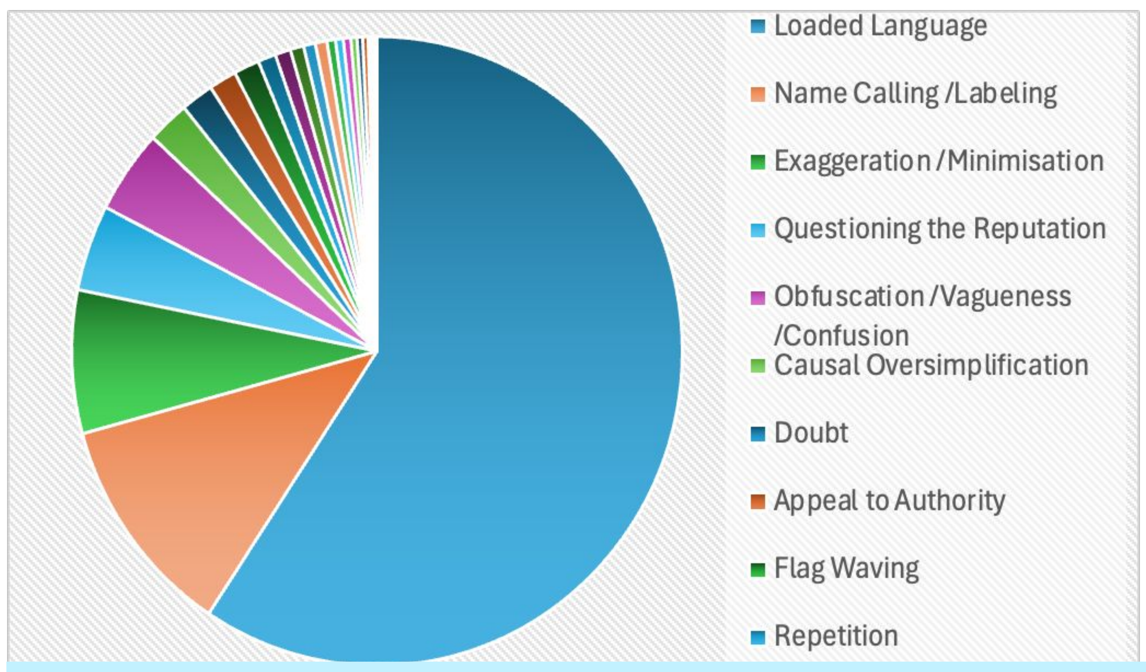
Step 3

Manual annotation

Adopt an existing two-tier taxonomy of six main categories, grouping 23 persuasion techniques.

ArPro Dataset Statistics

Content	Stat
# news articles	2,810
# paragraphs	8,000
# sentences	10,331
# words	277,952
avg sent. length	26.90
avg par. length	34.74
% Propagandistic paragraphs	63%



Distribution of all techniques with top ten shown on the right.

Top Topics	#pars (%propagandistic)
News	2993 (73)
Politics	2330 (62)
Health	594 (47)
Social	473 (56)
Sports	403 (58)
Miscellaneous	286 (68)
Arts and Culture	215 (47)
Religion	210 (39)
Science and Technology	175 (40)

Experiments

Aims

- Strong baselines on our ArPro dataset.
- Evaluation of the most powerful closed LLM to-date, GPT-4

Classification Tasks

1. Binary propaganda detection (***Binary***)
2. Coarse-grained propaganda detection (***Multilabel, 6 labels***)
3. Propaganda techniques detection (***Multilabel, 23 labels***)
4. Propaganda text spans identification (***Multilabel + Multiclass + Sequence tagging***)

Data Splits

- 75% train, 8.5% dev, and 16.5% test.

Models

- Transformer/pre-trained language models (**PLMs**): AraBERT, XLM-RoBERTa
- **GPT-4**

Evaluation Measures

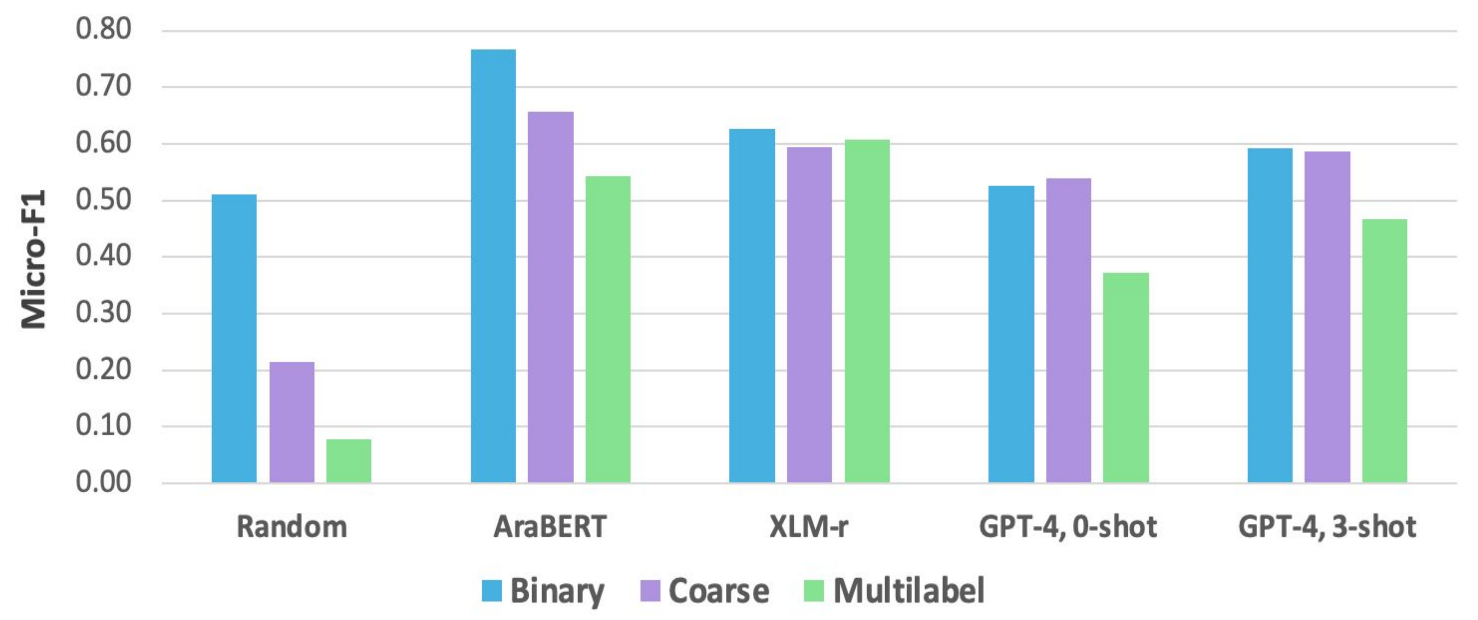
- **Tasks 1-3**: Micro-F1
- **Task 4**: modified F1 (considers partial matches)

Distribution:
Binary and coarse grained

Label	Train	Dev	Test
Binary			
Propagandistic	3,777	425	832
Non-Propagandistic	2,225	247	494
Total	**6,002**	**672**	**1,326**
Coarse-grained			
Call	176	21	40
Distraction	74	9	16
Justification	471	48	102
Manipulative_Wording	3,460	387	757
no_technique	2,225	247	494
Reputation	1,404	163	314
Simplification	384	42	82
Total	**8,194**	**917**	**1,805**

Results

How does fine-tuned PLMs perform in propaganda detection in different granularities (Tasks 1-3) compared to GPT-4?



- AraBERT, Arabic-specific PLM model, outperforms XLM-r
- GPT-4, 0-shot lags behind PLMs in all 3 tasks
- GPT-4, 3-shot closes the gap especially for the coarser classification granularities

How effective is GPT-4 for detecting and labeling propagandistic spans in text?

- Investigate GPT-4 0-shot performance over Arabic
- Compare to six other languages from a multilingual dataset (SemEval23 shared task 3)

Lang.	#Samples	Micro-F1 (Random)
Arabic	1,326	0.117 (0.010)
English	3,127	0.111 (0.008)
French	610	0.138 (0.017)
German	522	0.057 (0.012)
Italian	882	0.115 (0.015)
Polish	800	0.071 (0.011)
Russian	515	0.073 (0.011)

- GPT-4, significantly outperforms a random baseline, but still underperforming
- Results on Arabic are in-line with other languages

مجلس قطر للبحوث والتطوير والابتكار
QRDI COUNCIL

Acknowledgments: The contributions of this work were funded by the NPRP grant 14C-0916-210015, which is provided by the Qatar National Research Fund (a member of Qatar Foundation).