PropXplain: Can LLMs Enable Explainable Propaganda Detection?

Maram Hasanain[1†], **Md Arid Hasan**[1*], **Mohamed Bayan Kmainasi**[1*], **Elisa Sartori**[2], **Ali Ezzat Shahroor**[1*], **Giovanni Da San Martino**[2], **Firoj Alam**[1†]
[1]Qatar Computing Research Institute, Qatar, [2]University of Padova, Italy
fialam@hbku.edu.qa, giovanni.dasanmartino@unipd.it

Abstract

There has been significant research on propagandistic content detection across different modalities and languages. However, most studies have primarily focused on detection, with little attention given to explanations justifying the predicted label. This is largely due to the lack of resources that provide explanations alongside annotated labels. To address this issue, we propose a multilingual (i.e., Arabic and English) explanation-enhanced dataset, the *first* of its kind. Additionally, we introduce an explanation-enhanced LLM for both label detection and rationale-based explanation generation. Our findings indicate that the model performs comparably while also generating explanations. We will make the dataset and experimental resources publicly available for the research community.[1]

1 Introduction

The proliferation of propagandistic content in online and social media poses a significant challenge to information credibility, shaping public opinion through manipulative rhetorical strategies (Da San Martino et al., 2019). Automatic propaganda detection has been an active area of research, with studies focusing on textual (Barrón-Cedeno et al., 2019), multimodal (Dimitrov et al., 2021a), and multilingual approaches (Piskorski et al., 2023b; Zhang and Zhang, 2022). However, most existing systems lack rational explanations that could improve media literacy and calibrate trust in predictions.

Yu et al. (2021) developed interpretable models for propaganda detection in news articles, combining qualitative features with pre-trained language models to enhance transparency. More recently, Zavolokina et al. (2024) conducted a user study in which GPT-4 was used for propaganda detection and explanation generation. They demonstrate that explanations foster critical thinking and highlight their importance. However, the current literature has paid little to no attention to developing datasets that include explanations alongside annotated propaganda labels. To address this gap, we propose a large multilingual (i.e., Arabic and English) explanation-enhanced dataset for propaganda detection. We build upon existing datasets, including ArPro (Hasanain et al., 2024a) and the SemEval-2023 English dataset (Piskorski et al., 2023a), enhancing them with explanations. Given the complexity of manually generating explanations and the higher reliability reported for GPT-4-based explanation generation (Wang et al., 2023), we opted to use a stronger LLM for explanation generation and manually checked for quality assurance. Figure 1 shows an example of a news sentence, its explanation, and human evaluation process. The developed dataset can be used to train specialized LLMs for propaganda detection and to provide explanations for their predictions. To this end, our contributions to this study are as follows:

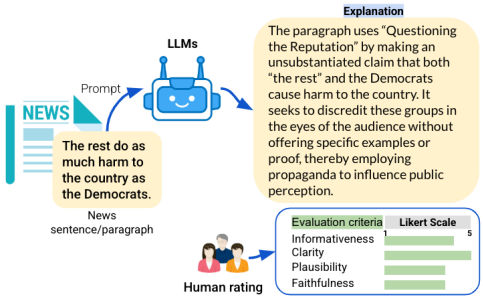

Figure 1: Example of a news sentence and its explanation and quality assessment process.

- We introduce an explanation-enhanced dataset for propaganda detection, consisting of approximately 21k and 6k news paragraphs and tweets for Arabic and English, respectively.
- To ensure the quality of the LLM-generated explanations, we manually evaluate explana-

[*] The contribution was made while the author was a contributor at the Qatar Computing Research Institute.
[†] Corresponding authors.
[1] https://github.com/firojalam/PropXplain

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 23855–23863
November 4-9, 2025

Split	# Articles	#items	Avg (W)	Avg Exp. (W)	% Prop.
			Arabic		
Train	8,103	18,453	32.4	48.1	63.8%
Dev	822	1,318	32.6	47.9	64.4%
Test	835	1,326	35.1	48.7	61.3%
Total	**8,913***	**21,097**	**32.6**	48.1	**63.7%**
			English		
Train	250	4,472	24.0	61.2	26.9%
Dev	204	621	23.9	61.6	27.9%
Test	225	922	23.7	61.2	27.9%
Total	**250***	**6,015**	**24.0**	**61.2**	**27.2%**

Table 1: Distribution of Arabic and English datasets. Exp.: explanation. Data items: annotated data elements including paragraphs and tweets. * Total unique articles. Prop.: Propagandistic. W: # Words

tions of the test set for each language.

- Our comparative experiments show that the proposed LLM matches transformer-based models in performance while additionally providing explanations for its predictions.

2 Dataset

We investigate LLMs' ability for explainable propaganda detection in both a high-resource language (English) and a lower-resource language (Arabic). In this work, we extend existing datasets with natural language annotation explanations generated by OpenAI o1, and evaluated by humans.

2.1 Arabic Propaganda Dataset

Building upon the ArPro Arabic dataset (Hasanain et al., 2024a), we follow the same annotation approach to build a larger dataset by collecting and annotating $7K$ paragraphs. Furthermore, this extension includes collecting and annotating tweets, to examine propaganda use in social media. Eventually, our Arabic dataset comprises two types of annotated documents: tweets and news paragraphs. The news paragraphs are extracted from articles published by 300 distinct news agencies, capturing a broad spectrum of Arabic news sources. It covers a diverse range of writing styles and topics including 14 different topics such as news, politics, human rights, and science and technology. As for the tweets subset, we start from a manually constructed set of 14 keywords and phrases, covering the topic of Israeli-Palestinian war, targeting subtopics popular during October and early November 2023. We use Twitter's search API to search for tweets posted during the second week of November 2023 and matching the collected phrases, resulting in $5.7K$ tweets to annotate.

Data was annotated following a two-phase approach (Hasanain et al., 2024a). In the first phase, 3 annotators independently examine each data item (paragraph or tweet) and label it with propagandistic techniques. In the second phase, 2 expert annotators examine annotations from the first phase and resolve any conflicts. Finally, the dataset set was split into training, development, and testing subsets following a stratified sampling approach.

2.2 English Propaganda Dataset

The English dataset is composed of 250 articles, collected from 42 unique news sources, coming from all political positions. The articles are manually cleaned of any artifacts that are incorrectly included during collection, such as links. The articles include topics that trended in the late 2023 and early 2024, with discussions of politics and the Israeli-Palestinian war covering 60% of the articles. Each article is annotated by at least 2 annotators and reviewed by 1 curator, whose task is to resolve inconsistencies between annotations. During the whole process, random checks of the annotations are carried out to verify the quality and give feedback on inaccuracies. To create the dataset, the articles are divided into sentences and split into three subsets: training, development and testing.

Note that these datasets are annotated for fine-grained propaganda detection; however, for this study, we perform classification and explanation generation in a binary setup.

2.3 Explanation Generation

We use OpenAI o1 to generate natural language explanations for gold propaganda annotations. This LLM is designed to have superior reasoning capabilities[2] which we believe are required for the task at hand. During pilot studies, we experimented with another highly-effective LLM, GPT-4o and a variety of prompts. Our manual evaluation of different samples in English and Arabic revealed that explanations generated by OpenAI o1 are better on average (following the quality assessment described in the next section). Eventually, the following prompt is used for explanation generation: *"Generate one complete explanation shorter than 100 words on why the paragraph as a whole is [gold label (propagandistic/not propagandistic)]. Be very specific in this full explanation to the paragraph at hand. Your explanation must be fully in*

[2] https://openai.com/index/introducing-openai-o1-preview/

[language]." Note that we used the original coarse-grained technique labels when generating gold explanations. However, for the current study, we mapped these labels to binary categories to evaluate the impact of explanations on classification and interpretability. For propagandistic texts, we generated explanations for both the individual spans with technique labels and the entire text, incorporating the relevant techniques. For non-propagandistic texts, we provided explanations justifying the label (not-propagandistic).

Quality of Generated Explanations We verify the quality of the generated explanations by human evaluation. We used a 5-point Likert scale for various evaluation metrics selected from relevant studies on natural language explanation evaluation (Huang et al., 2024, 2023; Zavolokina et al., 2024), including *informativeness, clarity, plausibility, and faithfulness*. Evaluation was carried out for Arabic and English datasets on the full test set. We provided detailed annotation instructions guidelines (see in Appendix A) for the human evaluators and each explanation assessed by three evaluators (see in Appendix C).

In Table 2, we report the average scores for all evaluation metrics. We first compute the average across annotators for each explanation and then across all explanations. In addition, we also computed the annotation agreement on ordinal scales by adopting the agreement index $r^*_{wg(j)}$ (James et al., 1984), which compares the observed variance in ratings to the maximum possible variance under complete disagreement (see further details in Appendix C.2). As presented in Table 3, the values above 0.89 for Arabic and 0.94 for English suggest a strong agreement (O'Neill, 2017). The results also suggest that OpenAI o1 generally generates explanations that are of high quality, considering the metrics at hand (e.g., clarity).

Data	Faithfulness	Clarity	Plausibility	Informative
Arabic	4.35	4.49	4.42	4.26
English	4.72	4.76	4.71	4.71

Table 2: Average Likert scale value for each human evaluation metric across different sets of explanations.

3 LLM for Detection and Explanation

Model. For developing an explanation-enhanced LLM, we adapted Llama 3.1 8B Instruct, a robust open-source model with strong multilingual capabilities (Dubey et al., 2024). We selected the 8B variant over larger versions (70B, 45B) due to the high computational cost of fine-tuning and inference. Llama-3.1 8B has also shown strong performance in relevant multilingual tasks (Pavlyshenko, 2023; Kmainasi et al., 2024).

Dataset	Faithfulness	Clarity	Plausibility	Informative
Arabic	0.90	0.92	0.89	0.89
English	0.94	0.95	0.94	0.95

Table 3: Annotation agreement for each human evaluation (annotation) metric across datasets, computed using $r^*_{wg(j)}$ index.

Instruction-following dataset. We constructed the instruction-following datasets with the aim of enhancing the model's generalizability and to guide the LLM to follow user instructions, which is a standard approach to fine-tune an LLM (Zhang et al., 2023). To create versatile instructions, we prompt state-of-the-art LLMs including GPT-4o, and Claude-3.5-sonnet to generate instructions (See Appendix D). Using each LLM, we created ten diverse English instructions per language. Each instruction is uniformly distributed across dataset samples. Each sample is structured with system, user and assistant prompts, where user prompt is defined as *Instruction + {input_text}*, and assistant prompt is defined as *Label: {class_label} Explanation: {explanation}*.

Training. Due to limited computational resources, we adopted Low-rank Adaptation (LoRA) (Hu et al., 2022) for training as a parameter-efficient fine-tuning technique. LoRA captures task-specific updates through low-rank matrices that approximate full weight updates.

Parameters Setup. We fine-tune the `Meta-Llama-3.1-8B-Instruct` model for 24 epochs using mixed-precision training with bfloat16 (bf16). LoRA is applied with a rank of 8, scaling factor (α) of 32, and a dropout rate of 0.005. We adopt a learning rate of 1×10^{-4}, following a linear warmup with a ratio of 0.05. Optimization is carried out with AdamW (Loshchilov and Hutter, 2017), while checkpointing and evaluation are performed every 10 steps. Training employs a per-device batch size of 4, combined with gradient accumulation over 16 steps, yielding an effective batch size of 512. The maximum input sequence length is capped at 1024 tokens. All training runs are conducted on 8 NVIDIA H100 GPUs under Distributed Data Parallel (DDP) (Bai, 2022), and

Paragraph	Gold	Llama-Base	Llama-FT
Every single person when I came to them saying, 'Can I get a half billion dollar bond?'	The paragraph is *not propagandistic* because it merely conveys an individual's experience without attempting to manipulate the audience's perception or promote a specific agenda.	The text appears to be *propagandistic* because it presents a personal anecdote that is likely intended to persuade or influence the audience, rather than providing a neutral statement.	The paragraph is *not propagandistic* because it presents a straightforward statement without employing manipulative language or techniques.

Table 4: Generated explanations by different models.

Model	$F1_{Micro}$	$F1_{Macro}$	$F1_{BERT}$
Arabic			
AraBERT	0.762	0.749	–
GPT-4o	0.575	0.567	–
Llama 3.1 8B (label only)	0.794	0.780	
Llama 3.1 8B (Base)	0.588	0.588	0.507
Llama 3.1 8B (FT)	0.775	0.760	0.706
English			
BERT-base	0.772	0.691	–
GPT-4o	0.649	0.630	–
Llama 3.1 8B (label only)	0.766	0.686	
Llama 3.1 8B (Base)	0.572	0.562	0.596
Llama 3.1 8B (FT)	0.781	0.675	0.751

Table 5: Performance of the proposed model and baselines. $F1_{BERT}$ is the F1 score computed using BERTScore for the explanation.

model selection is based on the checkpoint with the lowest validation loss.

Evaluation. For the evaluation, we used a zero-shot approach and selected a random instruction from our instruction sample as a prompt, which is a common approach reported in a prior study (Kmainasi et al., 2024). The temperature parameter was set to zero to ensure result reproducibility. Additionally, we implemented post-processing function to extract the labels and corresponding explanations.

Evaluation Metrics. To assess classification performance, we used macro and micro F_1 scores. For evaluating explanations, we used BERTScore (Zhang et al., 2020), which leverages contextual embeddings. Specifically, we computed the F_1 score using AraBERT (v2) (Antoun et al., 2020) for Arabic and BERT-base-uncased (Devlin et al., 2019) for English.[3]

4 Results and Discussion

We compare our proposed fine-tuned Llama 3.1 8B Instruct model to baseline models: fine-tuned transformer models using AraBERT (as reported in Hasanain et al. (2024a)) and BERT-base for Arabic and English, respectively. These models are commonly-used for the task (Hasanain et al., 2023). Note that BERT based models are used for label prediction only. We have also fine-tuned LLaMA without explanations to ensure a fair comparison. For Arabic, the macro-F1 score achieved is 0.780, representing a 3.0% improvement over the fine-tuned AraBERT. For English, the results differ slightly: fine-tuned LLaMA attains a macro-F1 of 0.686, which is 0.5% lower than that of AraBERT. Additionally, we compare the model's performance to two LLMs: GPT-4o and un-finetuned Llama 3.1 8B Instruct. As Table 5 shows, the performance of our fine-tuned Llama model achieves a micro F1 score that is on par or better than other models. Specifically, the model significantly outperforms the other LLMs tested.

As for its performance in explanation, in reference to the gold explanations, we observe a 25% and 40% improvements over the base model for English and Arabic, respectively. The fine-tuned model shows better alignment with gold explanations as demonstrated by the example in Table 5.

5 Related Work

Automatic detection of misinformation and propagandistic content has gained significant attention over the past years. Research has explored various problems, including cross-lingual propaganda analysis (Barrón-Cedeno et al., 2019), news article propaganda detection (Da San Martino et al., 2019), and misinformation and propaganda related to politics and war. Building on the seminal work of Da San Martino et al. (2019), resources have been developed for multilingual (Piskorski et al., 2023a; Hasanain et al., 2023) and multimodal setups (Dimitrov et al., 2021b; Hasanain et al., 2024b). Reasoning-based explanations in NLP have advanced fact-checking (Russo et al., 2023), hate speech detection (Huang et al., 2024), and propaganda detection (Zavolokina et al., 2024). While binary classifiers effectively identify propaganda, they often lack transparency, making interpretation

[3]BERTScore was chosen over BLEU and ROUGE as it captures semantics, better reflecting explanation quality.

difficult (Atanasova, 2024). Yu et al. (2021) showed that qualitative reasoning aids deception detection, while Atanasova (2024) emphasized explanation generation for better interpretability. Yet, explicit prediction reasoning for propaganda detection remains under-explored, particularly in multilingual settings. Our work addresses the gap by developing a multilingual explanation-enhanced dataset and proposing a specialized LLM.

6 Conclusions and Future Work

In this study, we introduce a multilingual dataset for propaganda detection and explanation, which is the *first* large dataset accompanied by explanations for the task. For Arabic, we have created a new propaganda-labeled dataset of size $13K$ samples, consisting of tweets and news paragraphs. Using OpenAI o1, we generated explanations for this dataset, as well as for ArPro (consisting of $8K$ instances), and for English starting from the SemEval-2023 dataset. To ensure quality, we manually evaluated the explanations and found they can serve as gold-standard references. We propose an explanation-enhanced LLM based on Llama-3.1 (8B) that matches strong baselines in performance while providing high-quality explanations. For future work, we plan to extend it to multilabel classification and span-level propaganda detection.

7 Limitations

Generating manual explanations is inherently complex. However, providing a rationale alongside the predicted label enhances trust and reliability in automated systems. Given the challenges of manual explanation creation, we relied on OpenAI's o1 – the most capable model at the time of writing – for generating explanations in this study. To ensure the reliability of these explanations, we conducted a manual evaluation based on four criteria: informativeness, clarity, plausibility, and faithfulness. The preliminary evaluation scores suggest that we can use them as gold explanation. For both label prediction and explanation generation, we focused on a binary classification task. However, future work should extend this to multiclass and multilabel settings. Additionally, for fine-tuning, we explored a multilingual model (Llama 3.1 8B), leaving room for further investigations into other models, including language-centric models.

Ethics and Broader Impact

We enhanced existing datasets by incorporating explanations. To the best of our knowledge, the dataset does not include any personally identifiable information, eliminating privacy concerns. For the explanations, we provided detailed annotation guidelines. It is important to acknowledge that annotations are inherently subjective, which may introduce biases into the evaluation process. We encourage researchers and users of this dataset to critically assess these factors when developing models or conducting further studies.

References

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based model for Arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools*, OSACT '20, pages 9–15, Marseille, France.

Pepa Atanasova. 2024. Generating fact checking explanations. In *Accountable and Explainable Methods for Complex Reasoning over Text*, pages 83–103. Springer.

Hao Bai. 2022. Modern distributed data-parallel large-scale pre-training strategies for nlp models. In *Proceedings of the 6th International Conference on High Performance Compilation, Computing and Communications*, pages 44–53.

Alberto Barrón-Cedeno, Giovanni Da San Martino, Israa Jaradat, and Preslav Nakov. 2019. Proppy: A system to unmask propaganda in online news. In *Proceedings of the AAAI Conference on Artificial Intelligence*, AAAI '19, pages 9847–9848.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. 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, Hong Kong, China. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT '19, Minneapolis, Minnesota, USA.

Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav

Nakov, and Giovanni Da San Martino. 2021a. Detecting propaganda techniques in memes. In *ACL-IJCNLP*.

Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021b. Task 6 at SemEval-2021: Detection of persuasion techniques in texts and images. In *Proceedings of the 15th International Workshop on Semantic Evaluation*, SemEval '21, Bangkok, Thailand.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Maram Hasanain, Fatema Ahmad, and Firoj Alam. 2024a. Can GPT-4 Identify Propaganda? Annotation and Detection of Propaganda Spans in News Articles. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2724–2744.

Maram Hasanain, Firoj Alam, Hamdy Mubarak, Samir Abdaljalil, Wajdi Zaghouani, Preslav Nakov, Giovanni Da San Martino, and Abed Freihat. 2023. ArAIEval shared task: Persuasion techniques and disinformation detection in Arabic text. In *Proceedings of ArabicNLP 2023*, pages 483–493, Singapore (Hybrid). Association for Computational Linguistics.

Maram Hasanain, Md. Arid Hasan, Fatema Ahmed, Reem Suwaileh, Md. Rafiul Biswas, Wajdi Zaghouani, and Firoj Alam. 2024b. ArAIEval shared task: Propagandistic techniques detection in unimodal and multimodal arabic content. In *Proceedings of the Second Arabic Natural Language Processing Conference (ArabicNLP 2024)*. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Fan Huang, Haewoon Kwak, and Jisun An. 2023. Is chatgpt better than human annotators? potential and limitations of chatgpt in explaining implicit hate speech. In *Companion Proceedings of the ACM Web Conference 2023*, WWW '23 Companion, page 294–297, New York, NY, USA. Association for Computing Machinery.

Fan Huang, Haewoon Kwak, Kunwoo Park, and Jisun An. 2024. ChatGPT rates natural language explanation quality like humans: But on which scales? In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3111–3132, Torino, Italia. ELRA and ICCL.

Lawrence R James, Robert G Demaree, and Gerrit Wolf. 1984. Estimating within-group interrater reliability with and without response bias. *Journal of applied psychology*, 69(1):85.

Mohamed Bayan Kmainasi, Ali Ezzat Shahroor, Maram Hasanain, Sahinur Rahman Laskar, Naeemul Hassan, and Firoj Alam. 2024. LlamaLens: Specialized multilingual llm for analyzing news and social media content. *Preprint*, arXiv:2410.15308.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Thomas A O'Neill. 2017. An overview of interrater agreement on likert scales for researchers and practitioners. *Frontiers in psychology*, 8:777.

Bohdan M. Pavlyshenko. 2023. Analysis of disinformation and fake news detection using fine-tuned large language model. *Preprint*, arXiv:2309.04704.

Jakub Piskorski, Nicolas Stefanovitch, Giovanni Da San Martino, and Preslav Nakov. 2023a. SemEval-2023 task 3: Detecting the category, the framing, and the persuasion techniques in online news in a multi-lingual setup. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2343–2361, Toronto, Canada. Association for Computational Linguistics.

Jakub Piskorski, Nicolas Stefanovitch, Nikolaos Nikolaidis, Giovanni Da San Martino, and Preslav Nakov. 2023b. 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.

Daniel Russo, Serra Sinem Tekiroğlu, and Marco Guerini. 2023. Benchmarking the generation of fact checking explanations. *Transactions of the Association for Computational Linguistics*, 11:1250–1264.

Han Wang, Ming Shan Hee, Md Rabiul Awal, Kenny Tsu Wei Choo, and Roy Ka-Wei Lee. 2023. Evaluating GPT-3 generated explanations for hateful content moderation. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 6255–6263.

Seunghak Yu, Giovanni Da San Martino, Mitra Mohtarami, James Glass, and Preslav Nakov. 2021. Interpretable propaganda detection in news articles. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, RANLP '21.

Liudmila Zavolokina, Kilian Sprenkamp, Zoya Katashinskaya, Daniel Gordon Jones, and Gerhard Schwabe. 2024. Think fast, think slow, think critical: Designing an automated propaganda detection tool. In *Proceedings of the 2024 CHI Conference*

on Human Factors in Computing Systems, CHI '24, New York, NY, USA. Association for Computing Machinery.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, et al. 2023. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Wenshan Zhang and Xi Zhang. 2022. Cross-lingual propaganda detection. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 4330–4336. IEEE.

A Annotation Guideline

You will be shown a news paragraph, a label assigned to it, and an explanation for the assigned label. As an annotator, your task is to carefully examine each news paragraph, label, and explanation. Then assess the quality of the explanation provided for the assigned label. Follow the steps below to ensure a thorough evaluation:

Analyze the News Paragraph

- Read the news paragraph, sentence and/or social media post.
- Understand the overall message and potential implications.

Check the Assigned Label

- Check the given label. The label is the result of annotation done by multiple human annotators.

Evaluate the Explanation

- Read the explanation provided for why the news paragraph has been assigned its label.
- Assess the explanation based on the metrics below. Each metric is scored on a Likert scale from 1-5.

Metrics

Informativeness Measures the extent to which the explanation provides relevant and meaningful information for understanding the reasoning behind the label. A highly informative explanation offers detailed insights that directly contribute to the justification, while a low-informative explanation may be vague, incomplete, or lacking key details.

As an annotator, you are judging if the explanation is providing enough information to explain the label assigned.

- 1 = Not informative: The explanation lacks relevant details and does not help understand why the news paragraph is labeled as such.
- 2 = Slightly informative: The explanation provides minimal information, but key details are missing or unclear.
- 3 = Moderately informative: The explanation contains some useful details but lacks depth or supporting reasoning.
- 4 = Informative: The explanation is well-detailed, providing a clear and meaningful justification for the label.
- 5 = Very informative: The explanation is thorough, insightful, and fully justifies the label with strong supporting details.

Clarity Assesses how clearly the explanation conveys its meaning. A clear explanation is well-structured, concise, and easy to understand without requiring additional effort. It should be free from ambiguity, overly complex language, or poor phrasing that might hinder comprehension.

As an annotator, you are judging the language and the structure of the explanation. Spelling mistakes, awkward use of language, and wrong translation will affect this metric negatively.

- 1 = Very unclear: The explanation is confusing, vague, or difficult to understand.
- 2 = Somewhat unclear: The explanation has some clarity but includes ambiguous or poorly structured statements.
- 3 = Neutral: The explanation is somewhat clear but may require effort to fully grasp.
- 4 = Clear: The explanation is well-structured and easy to understand with minimal ambiguity.
- 5 = Very clear: The explanation is highly readable, precise, and effortlessly understandable.

Plausibility Refers to the extent to which an explanation logically supports the assigned label and appears reasonable given the news paragraph's content. A plausible explanation should be coherent, factually consistent, and align with the expected reasoning behind the label. While it does not require absolute correctness, it should not contain obvious contradictions or illogical claims.

As an annotator, you are judging if the explanation actually supports the label assigned to it. For example, if a text is labeled as "Not Propaganda," the explanation given should be for that label.

- 1 = Not plausible at all: The explanation does

not align with the label and seems completely incorrect.
- 2 = Weakly plausible: The explanation has some relevance but lacks strong justification or contains logical inconsistencies.
- 3 = Moderately plausible: The explanation somewhat supports the label but may be incomplete or partially flawed.
- 4 = Plausible: The explanation logically supports the label and is mostly reasonable.
- 5 = Highly plausible: The explanation is fully aligned with the label and presents a strong, logical justification.

Faithfulness Measures how accurately an explanation reflects the reasoning behind the assigned label. A faithful explanation correctly represents the key factors and logical steps that justify the label, without adding misleading or unrelated details. High faithfulness means the explanation stays true to the actual reasoning used for classification, ensuring reliability and consistency.

As an annotator, you are judging how well the explanation reflects the logic behind the label. For example, if the explanation claims an implication of the text, it should also present the logical reasoning behind it.

- 1 = Not faithful at all: The explanation is completely unrelated to the given label and does not reflect a valid reasoning process.
- 2 = Weakly faithful: Some elements of the explanation are relevant, but much of it is misleading, inconsistent, or lacks proper justification.
- 3 = Moderately faithful: The explanation captures parts of the reasoning but includes unrelated, unclear, or unnecessary justifications.
- 4 = Faithful: The explanation aligns well with the reasoning behind the label and includes relevant, logical details.
- 5 = Highly faithful: The explanation fully and accurately reflects the correct reasoning, without any misleading or irrelevant information.

B Annotation Platform

We present the screenshot of the interface designed for the evaluation of LLM generated explanation, which consisted of a paragraph, label, and explanation for the label, annotation guidelines, and four different evaluation metrics including informativeness, clarity, plausibility, and faithfulness. 5-point Likert scale is used for each evaluation metric and the annotator is asked to follow the annotation guideline to select an appropriate Likert scale value for each metric.

C Annotation Details

C.1 Annotation Setup

We recruited annotators who are native Arabic speakers and fluent in English, all holding at least a bachelor's degree. Since they were proficient in English, they also worked on English news paragraphs. We provided annotation guidelines and necessary consultation. All annotators had prior experience with similar tasks. A total of six annotators participated in the evaluation task. In accordance with institutional requirements, each signed a Non-Disclosure Agreement (NDA). For their compensation, we hired a third-party company to manage payments at standard hourly rates based on location.

C.2 Annotation Agreement

To assess the consistency of human ratings, we also computed inter-annotator agreement for each evaluation metric – informativeness, clarity, plausibility, and faithfulness – based on 5-point Likert scale annotations. We adopted the $r^*_{wg(j)}$ index (James et al., 1984), a widely used measure for inter-annotator agreement on ordinal scales, which compares observed variance in ratings to the maximum possible variance under complete disagreement. For each item, the agreement score is computed as:

$$r^*_{wg(j)} = 1 - \frac{S^2_X}{\sigma^2_{\text{mv}}},$$

where S^2_X is the observed variance across annotators and σ^2_{mv} is the maximum variance possible given the scale (computed as $\sigma^2_{\text{mv}} = 0.5(X^2_U + X^2_L) - [0.5(X_U + X_L)]^2$, with $X_U = 5$ and $X_L = 1$ for a 5-point scale). This method allows us to capture the degree of consensus among annotators while accounting for the bounded nature of Likert ratings. We report the average $r^*_{wg(j)}$ per metric. In Figure 3, we report the agreement scores for both datasets. The average agreement scores for Arabic and English are above 0.89 and 0.94, respectively, for all metrics. These values indicate a strong agreement (O'Neill, 2017).

D Prompts

To generate instructions for the instruction-following dataset, we prompt the LLMs using the

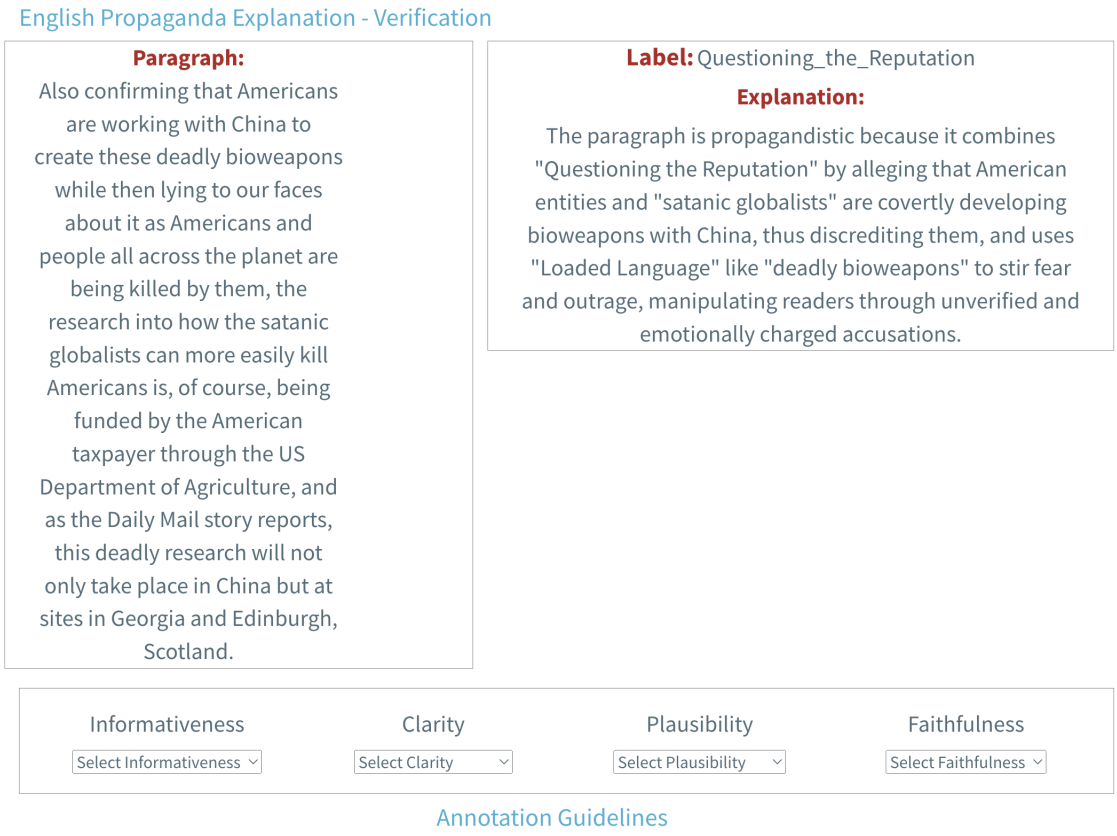

Figure 2: A screenshot of the annotation platform for the explanation evaluation of English propaganda.

following prompt: *We are creating an English instruction-following dataset for an [language] dataset covering the task of propaganda detection with explanation. The user defined the task as follows: Detecting propaganda in a piece of text and explaining why this piece of text is propagandistic. Propaganda can be defined as a form of communication aimed at influencing people's opinions or actions toward a specific goal, using well-defined rhetorical and psychological techniques. For that task, the labels include: ['non-propagandistic', 'propagandistic']. Write 10 very diverse and concise English instructions making sure the labels provided above are part of the instruction. Only return the instructions without additional text.*

E Data Release

Our proposed dataset PropXplain will be released under the CC BY-NC-SA 4.0 – Creative Commons Attribution 4.0 International License: https://creativecommons.org/licenses/by-nc-sa/4.0/.

F Potential Applications

LLMs capable of detecting propaganda with explanations have several real-world applications. They can enhance social media moderation by identifying manipulative content, support fact-checkers with transparent justifications, and serve as educational tools for improving media literacy. Additionally, such models can aid NGOs and government agencies in monitoring disinformation campaigns, while offering tools to understand bias in online content. By providing interpretable outputs, these systems foster trust, accountability, and informed decision-making in digital environments.