Large Language Models for Propaganda Span Annotation

Maram Hasanain, Fatema Ahmad, Firoj Alam

EMNLP 2024
12-16 November, 2024

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 associated text spans.

- Few datasets exist in any language, usually limited in size.

- Training specialized models requires large-scale annotated datasets, *can LLMs help develop such datasets?*

اضاف: "وبالتوازي مع الأجواء التفاؤلية التي تبثّها مصادرهم يتولون رمي الإشاعات بأن عملية التأليف انتهت، وباتت مسألة ساعات، وإيهام الآخرين بأنهم يريدون تشكيل الحكومة وغيرهم يريد التعطيل، تماماً كما فعلوا في الحكومات السابقة، وبالأخص في وزارة الطاقة التي استمرت لسنوات بعهدتهم وتكفلوا بتأمين الكهرباء 24 ساعة على 24، فأين الكهرباء؟"

Translation: He added: “In parallel with the optimistic atmosphere that their sources are spreading, they are spreading rumors that the formation process has ended, and that it's only a matter of hours. They deceive others into believing that they want to form the government while others want to obstruct it, just as they did with previous governments, particularly in the Ministry of Energy, which they held for years and promised to provide electricity 24/7. So where is the electricity?”

Techniques: Obfuscation, Intentional Vagueness, Confusion Smears

A news paragraph annotated by propaganda techniques at the text span level

Contributions

1. First attempt to explore GPT-4 as an annotator for propagandistic text spans detection and labelling.

2. Serving as a consolidator, GPT-4 provides labels that can be effectively used to train a specialized model for the task.

3. All scripts and annotations from human annotators and GPT-4 are released to the community.

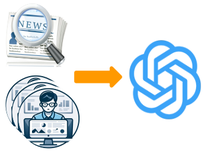
Approach

Acquire manually annotated data

Annotate news paragraphs in two stages:
(i) 3 annotators and
(ii) 2 expert annotators consolidate annotations.

Annotate with GPT-4

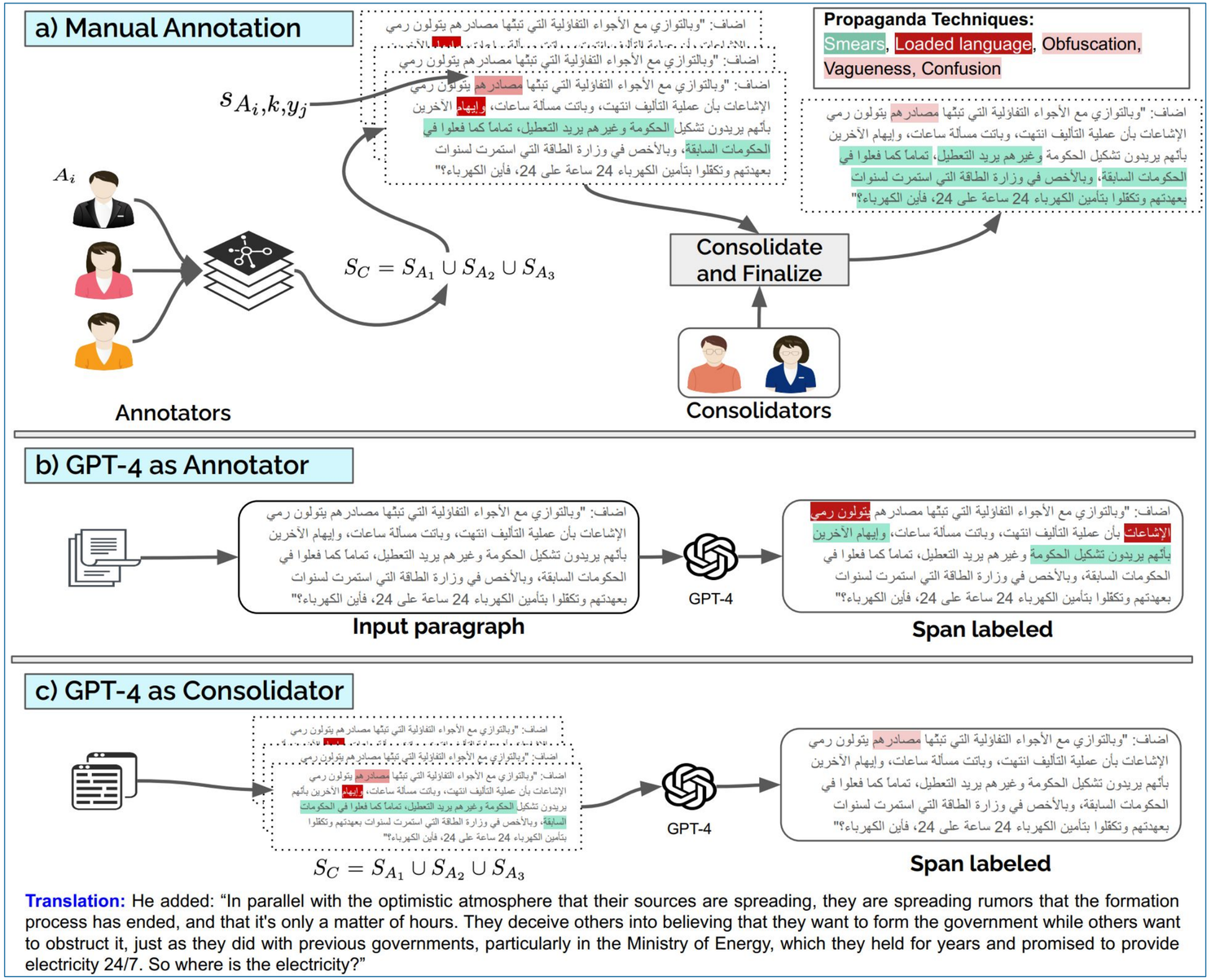
- ***Annotator***: Instruct GPT-4 to annotate paragraphs.
- ***Selector***: Select from all techniques given by annotators and extract matching text spans.
- ***Consolidator***: Consolidate all labels and spans from stage (i) annotators.



Train and annotate with SLMs

Train a span extraction and annotation SLM (AraBERT) using: manual annotations and each of the 3 sets of GPT-4 annotations ⇒ **4 SLMs**

Approach



Translation: He added: “In parallel with the optimistic atmosphere that their sources are spreading, they are spreading rumors that the formation process has ended, and that it's only a matter of hours. They deceive others into believing that they want to form the government while others want to obstruct it, just as they did with previous governments, particularly in the Ministry of Energy, which they held for years and promised to provide electricity 24/7. So where is the electricity?”

ArPro Dataset: Manual Annotation Stats

Dataset covers 23 propaganda techniques

Content	Stat
# news articles	2,810
# paragraphs	8,000
avg par. length	34.74
% Propagandistic paragraphs	63%
# Techniques	17,521

Distribution of manually annotated dataset (ArPro)

Technique	%
Loaded Language	59.3
Name Calling-Labeling	11.5
Exaggeration-Minimisation	7.4
Questioning the Reputation	4.4
Obfuscation-Vagueness-Confusion	4.3

Distribution of top 5 techniques in ArPro

Experiments

Aims

- Study GPT-4's performance in its different roles.
- Investigate quality of generated labels in training SLMs

Tasks

Propaganda text spans identification and labelling (***Multilabel + Multiclass + Sequence tagging***)

Datasets

- **ArPro:** 75% train, 8.5% dev, and 16.5% test.
- **ArAIEvalT1:** SOTA *test* subset

Experiments

Models

- Transformer models (**PLM**): AraBERT
- **GPT-4**

Evaluation Measures

- Modified F1 (considers partial matches)
- Inter-rater agreement (γ)

Results

How does GPT-4 perform in annotation?

Role	Micro-F1	Macro-F1	Span (γ)
Annotator	0.050	0.045	0.247
Selector	0.137	0.144	0.477
Consolidator	0.671	0.570	0.609

Wrong indices predicted!

Apply a correction heuristic!

Role	Micro-F1 (orig)	Micro-F1 (correct)
Annotator	0.050	0.117
Selector	0.137	0.297
Consolidator	0.671	0.670

- In cases where GPT-4 has to predict the text spans (annotator and selector), it predicts wrong text indices but correct text spans.
- A simple heuristic to find first appearance of text span in a paragraph significantly improves performance

Results

How useful are GPT-4 annotations for SLM training?

- Train a SLM using the 3 sets of GPT-4 annotations
- Compare to SLM trained on gold labels

Model	Train Set	Micro-F1
Random	-	0.010
GPT-4	-	0.117
AraBERT	$\text{GPT-4}_{\text{Annotator}}$	0.127
AraBERT	$\text{GPT-4}_{\text{Selector}}$	0.236
AraBERT	$\text{GPT-4}_{\text{Consolidator}}$	*0.335*
AraBERT	$\text{ArPro}_{\text{train}}$	**0.387**

Test on $\text{ArPro}_{\text{test}}$
which is the test split of of the same gold dataset

Model	Train Set	Micro-F1
CUET_sstm	-	0.300
AraBERT	$\text{GPT-4}_{\text{Annotator}}$	0.124
AraBERT	$\text{GPT-4}_{\text{Selector}}$	0.257
AraBERT	$\text{GPT-4}_{\text{Consolidator}}$	*0.334*
AraBERT	$\text{ArPro}_{\text{train}}$	**0.406**

Test on $\text{ArAIEval24T1}_{\text{test}}$
which is the Arabic SOTA for the task

- SLM shows significantly better performance when using GPT-4 as consolidator labels
- SLM with GPT-4 annotations outperforms SOTA on $\text{ArAIEval24T1}_{\text{test}}$

Results

Does GPT-4 annotate some techniques more effectively?

- Compute per-technique annotation agreement (γ) with gold labels

Technique	Annotator
Causal Oversimplification	0.889
Consequential Oversimplification	0.835
Doubt	0.815
Obfuscation/Vagueness/Confusion	0.791
	Selector
Doubt	0.802
Flag Waving	0.705
Appeal to Hypocrisy	0.66
Loaded Language	0.654
	Consolidator
False Dilemma /No Choice	0.872
Loaded Language	0.774
Straw Man	0.697
Doubt	0.695

Across different roles, GPT-4 shows High agreement with gold labels for some techniques.

Conclusions

- This is the first attempt at investigating GPT-4's performance as an annotator for propaganda span identification and labelling.
- We the model's performance when provided with sets of information of varied richness, which represents an increased cost and effort in hiring human annotators.
- We study the effectiveness of GPT-4's labels when used to train specialized models for the task.

- **Results:**
  - Providing more information significantly improves the model's annotation performance and agreement with human expert consolidators.
  - We can train effective models using labels provided by GPT-4 when acting as a consolidator.

- Future research will explore additional LLMs and learning setups (e.g., few shot learning).

Thank You

https://github.com/MaramHasanain/llm_prop_annot

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