Raphael at ArAIEval Shared Task: Understanding Persuasive Language and Tone, an LLM Approach

Utsav Shukla, Manan Vyas, Shailendra Tiwari
Thapar Institute of Engineering and Technology
{ushukla_be17, mvyas_bemba, shailendra}@thapar.edu

Abstract

The widespread dissemination of propaganda and disinformation on both social media and mainstream media platforms has become an urgent concern, attracting the interest of various stakeholders such as government bodies and social media companies. The challenge intensifies when dealing with understudied languages like Arabic. In this paper, we outline our approach for detecting persuasion techniques in Arabic tweets and news article paragraphs. We submitted our system to ArAIEval 2023 Shared Task 1, covering both subtasks. Our main contributions include utilizing GPT-3 to discern tone and potential persuasion techniques in text, exploring various base language models, and employing a multi-task learning approach for the specified subtasks.

1 Introduction

In today's world, an average person encounters a plethora of information via social media platforms and online news resources. While this accessibility to up-to-date news and opinions is convenient and keeps individuals informed, it also raises concerns about disinformation, propaganda, hate speech, and political bias. Over the past decade, various efforts have been made to detect disinformation and fake news through the analysis of textual content in articles (Wang, 2017), social media content (Lu and Li, 2020), website metrics (Panayotov et al., 2022), and images (Zlatkova et al., 2019).

While significant progress has been made in the detection of disinformation, bias, and hate speech, the automated detection of propaganda and persuasion is a relatively newer domain. Propaganda and persuasion serve as instruments to influence public opinion or evoke emotional responses and has proved to be very harmful for society in last decade. Although there is substantial work in propaganda detection in English (Da San Martino et al., 2021), Arabic remain understudied for this problem. Collecting datasets and training models for Arabic becomes more challenging because of its numerous dialects spoken all around the world.

The ArAIEval 2023 Task 1 (Hasanain et al., 2023) is a collaborative effort that aims to address this gap by detecting persuasion techniques in Arabic. Task 1 is divided into two subtasks: Subtask 1A is a binary classification task that detects the presence of a persuasion technique in a given Arabic text, while Subtask 1B is a multi-label classification problem that identifies which of the 24 possible persuasion techniques are present. In our experiments, we designed and implemented multiple systems for both subtasks. We found that leveraging outputs from large language models (LLMs) for supervised training led to significant performance gains. Our systems ranked 4th and 3rd in Subtasks 1A and 1B respectively on development sets and 5th and 6th in Subtasks 1A and 1B, respectively on test sets. In the following sections, we discuss related work and elaborate on our experiments and submissions. We have made our code, prompts, and GPT-3.5 outputs publicly available for future reproducibility [1].

2 Related Work

Disinformation and fake news detection is a vibrant area of research within the NLP community, with methodologies ranging from text-based approaches to multimodal analyses that incorporate images and graphs. Propaganda detection has also garnered attention; (Da San Martino et al., 2020) introduced a shared task that identifies both the span and type of propaganda technique present. This was extended by (Dimitrov et al., 2021), who added a sub-task focused on recognizing persuasion techniques in memes, thus incorporating image modality.

To the best of our knowledge, apart from

[1] github.com/us241098/araieval_submission

Proceedings of the The First Arabic Natural Language Processing Conference (ArabicNLP 2023), pages 589–593

Label	Training	Development
true	1918	202
false	519	517

Table 1: Label distribution for subtask 1A

Label	Training	Development
Loaded Language	1574	176
Name Calling Labeling	692	77
No Technique	509	57
Questioning the Reputation	383	43
Exaggeration Minimisation	292	33
Obfuscation Vagueness Confusion	240	28
Doubt	143	16
Causal Oversimplification	128	15
Appeal to Fear Prejudice	108	12
Slogans	70	8
Flag Waving	63	7
Appeal to Hypocrisy	56	7
Appeal to Authority	48	5
Appeal to Values	37	4
Consequential Oversimplification	33	3
False Dilemma No Choice	32	3
Conversation Killer	28	3
Repetition	25	3
Guilt by Association	13	1
Appeal to Time	10	2
Whataboutism	9	1
Red Herring	8	1
Straw Man	6	1
Appeal to Popularity	2	1

Table 2: Label distribution for subtask 1B

ArAIEval 2023 (Hasanain et al., 2023), (Alam et al., 2022) is the only other work that specifically focuses on the detection of propaganda/persuasion techniques in the Arabic language.

3 Data

Our submitted system relies solely on the dataset provided by the organizers, without any additional data or augmentations. Subtask 1A is a binary classification task featuring two labels: 'true' and 'false,' which signify the presence or absence of persuasion techniques in the text. Subtask 1B, a multi-label classification problem, involves 24 labels representing various potential persuasion techniques. The data for these tasks come from two sources: tweets and paragraphs from news articles. During data preprocessing, we removed emojis and the text string "LINK" from all entries. Table 1 and Table 2 describe the label distribution of both subtasks.

4 System

For both subtasks, we conducted multiple experiments that included using various base models, employing large language models (LLMs) for reasoning, and adjusting both the architecture and loss functions. We discuss these major components and their applications in the subsequent sub-sections.

4.1 MARBERT

We leverage MARBERT, a state-of-the-art BERT-based model specifically pretrained on a large corpus of Arabic text, encompassing both Modern Standard Arabic and various dialects (Abdul-Mageed et al., 2021). The utilization of MARBERT allows us to capture intricate language features that are particularly pertinent to Arabic text.

Upon passing an arabic input text through the MARBERT encoder, the resulting contextual embeddings are generated. We specifically extract the embedding corresponding to the [CLS] token. This [CLS] token's embedding is then forwarded to binary classification head and multi label classification heads for subtask 1A and 1B respectively.

4.2 GPT 3.5

We utilize the Generative Pre-trained Transformer 3.5 (Brown et al., 2020) (GPT-3.5) for the task of generating description of Arabic texts in English and conducting tone and emotional analysis. The resultant English text and tone descriptions are subsequently encoded using either BERT (Devlin et al.,

Methodology	Micro F1	Macro F1
MARBERT	0.8145	0.7192
MARBERT+GPT 3.5(BERT)	0.8412	0.7490
MARBERT+GPT 3.5(RoBERTa)	0.8427	0.7571
MARBERT+GPT 3.5(RoBERTa)+Source as Feature Gate	0.8610	0.7922
MultiTask	0.8509	0.7698

Table 3: Results of our different systems on subtask 1A dev set

Methodology	Micro F1	Macro F1
MARBERT	0.6088	0.1996
MARBERT+GPT 3.5 (BERT)	0.6227	0.2056
MARBERT+GPT 3.5 (RoBERTa)	0.6399	0.2365
MARBERT+GPT 3.5 (RoBERTa)+Source as Feature Gate	0.6304	0.2287
MultiTask	0.5694	0.1602

Table 4: Results of our different systems on subtask 1B dev set

Task	Micro F1	Macro F1
Subtask 1A (Ours)	0.7475	0.7221
Subtask 1A (Best)	0.7634	0.7321
Subtask 1B (Ours)	0.5347	0.1772
Subtask 1B (Best)	0.5666	0.2156

Table 5: Our Submission to Subtask 1A and Subtask 1B compared to best performing systems

2019) or RoBERTa (Liu et al., 2019). When these encodings are concatenated with MARBERT encodings of original arabic texts before feeding them to respective classification heads, we observe a significant improvement over our MARBERT baseline performance.

4.3 MultiTask Training

Subtask 1A and 1B being on the same features allow us to formulate a multi task learning objective. During the forward pass [CLS] token encodings are passed through two separate fully-connected layers to produce logits for binary and multi-label classification. During backpropagation, the loss is calculated for both sub-tasks and weighted according to a learned parameter. The gradient of this total loss is then computed with respect to the model parameters. This dual-task learning enables the model to simultaneously optimize for binary and multi-label classification.

4.4 Source as Feature Gate

We use the Source provided in the datasets (Tweet or Paragraph) as feature gate for our concatenated encodings (MARBERT+BERT/RoBERTa). We have found the using the source as feature gate performs better in comparison to just concatenating the source vector to the embeddings.

5 Experiment Setting

All our experiments are done on single 12 GB GPU and our models take 5-15 minutes to be trained. We use "bert-base-cased", "roberta-base" and "MARBERTv2" variants from HuggingFace (Wolf et al., 2020) as our base models. We train our models upto 7 epochs and use AdamW optimizer (Andrew and Gao, 2007) with learning rate being set to 2e-5, and epsilon set to 1e-8.

6 Results

Table 3 and Table 4 shows our performance on subtask 1A and 1B development sets respectively. We observe that when GPT 3.5 outputs are used in training and inference we get significant gains over our MARBERT baseline in both sub-task 1A and 1B. For encoding the English outputs from GPT 3.5, RoBERTa is found to be better than BERT. We also observe that using source as feature gate give us gains in subtask 1A but not in subtask 1B. MARBERT+GPT 3.5(RoBERTa)+Source as Feature Gate is our submitted system for subtask 1A and MARBERT+GPT 3.5 (RoBERTa) is our sub-

mitted system for subtask 1B for both development and test sets. On development sets our systems ranked 4th and 3rd on subtask 1A and 1B respectively.

Table 5 shows our performance on test set. Here our submitted system in subtask 1A ranked 5th in terms of both macro and micro F1. While in subtask 1B we ranked 6th in terms of micro F1 and 4th in terms of macro F1.

7 Discussions and Limitations

Our experiments indicate that using prompts with large language models (LLMs) and leveraging their outputs as features in supervised training environments show promise, especially for understudied languages like Arabic. In the future, we plan to explore additional LLMs, with a preference for open-source options. Another exciting avenue we aim to investigate is fine-tuning these LLMs on Arabic-specific data to enhance performance. We also aim to Benchmark the only LLM performance without using their outputs for supervised models.

One limitation we've identified is the high computational/financial cost associated with closed LLM inference. However, this challenge may be mitigated as more open-source LLMs become available and as optimization techniques such as PEFT (Liu et al., 2022), QLoRA (Dettmers et al., 2023), and quantization continue to evolve.

8 Conclusion

The widespread dissemination of propaganda and misinformation through various media channels, including social media and mainstream outlets, has garnered considerable attention from key players like government agencies and social media companies. In this study, we outline our methodology for identifying persuasive tactics employed in Arabic-language tweets and text segments. For the 2023 ArAIEval shared task 1, we have used GPT-3.5 as the cornerstone of our system to analyze the tone and potential persuasion strategies in the text. We have also discussed the limitations of the system proposed and suggested to incorporate Open Source LLMs and multiple optimization techniques in our future work.

References

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021. ARBERT & MARBERT: Deep bidirectional transformers for Arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7088–7105, Online. Association for Computational Linguistics.

Firoj Alam, Hamdy Mubarak, Wajdi Zaghouani, Giovanni Da San Martino, and Preslav Nakov. 2022. Overview of the WANLP 2022 shared task on propaganda detection in Arabic. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 108–118, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Galen Andrew and Jianfeng Gao. 2007. Scalable training of L_1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Giovanni Da San Martino, Alberto Barrón-Cedeño, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. 2020. 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. 2021. A survey on computational propaganda detection. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, IJCAI'20.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021. SemEval-2021 task 6: Detection of persuasion techniques in texts and images. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 70–98, Online. Association for Computational Linguistics.

Maram Hasanain, Firoj Alam, Hamdy Mubarak, Samir Abdaljalil, Wajdi Zaghouani, Preslav Nakov, Giovanni Da San Martino, and Abed Alhakim Freihat. 2023. ArAIEval Shared Task: Persuasion Techniques and Disinformation Detection in Arabic Text. In *Proceedings of the First Arabic Natural Language Processing Conference (ArabicNLP 2023)*, Singapore. Association for Computational Linguistics.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin Raffel. 2022. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Yi-Ju Lu and Cheng-Te Li. 2020. Gcan: Graph-aware co-attention networks for explainable fake news detection on social media.

Panayot Panayotov, Utsav Shukla, Husrev Taha Sencar, Mohamed Nabeel, and Preslav Nakov. 2022. GREENER: Graph neural networks for news media profiling. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7470–7480, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

William Yang Wang. 2017. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426, Vancouver, Canada. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing.

Dimitrina Zlatkova, Preslav Nakov, and Ivan Koychev. 2019. Fact-checking meets fauxtography: Verifying claims about images. 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 2099–2108, Hong Kong, China. Association for Computational Linguistics.