Mela at ArAIEval Shared Task: Propagandistic Techniques Detection in Arabic with a Multilingual Approach

Md Abdur Razzaq Riyadh
Department of Artificial Intelligence
University of Malta
md.riyadh.23@um.edu.mt

Sara Nabhani
Department of Artificial Intelligence
University of Malta
sara.nabhani.23@um.edu.mt

Abstract

This paper presents our system submitted for Task 1 of the ArAIEval Shared Task on Unimodal (Text) Propagandistic Technique Detection in Arabic. Task 1 involves identifying all employed propaganda techniques in a given text from a set of possible techniques or detecting that no propaganda technique is present. Additionally, the task requires identifying the specific spans of text where these techniques occur. We explored the capabilities of a multilingual BERT model for this task, focusing on the effectiveness of using outputs from different hidden layers within the model. By fine-tuning the multilingual BERT, we aimed to improve the model's ability to recognize and locate various propaganda techniques. Our experiments showed that leveraging the hidden layers of the BERT model enhanced detection performance. Our system achieved competitive results, ranking second in the shared task, demonstrating that multilingual BERT models, combined with outputs from hidden layers, can effectively detect and identify spans of propaganda techniques in Arabic text.

1 Introduction

Propaganda is a powerful tool often used to influence public opinion and manipulate perceptions by spreading biased or misleading information. It employs a range of persuasive techniques to shape attitudes and beliefs, typically to benefit a particular agenda or cause. In today's digital age, the dissemination of propaganda has become more prevalent, particularly on social media and other online platforms. Detecting and analyzing propaganda in text is essential for mitigating its impact and promoting informed decision-making among the public.

This paper presents our approach to the Unimodal (Text) Propagandistic Technique Detection Task at ArAIEval 2024 (Hasanain et al., 2024b). Our approach leverages a multilingual BERT model (Devlin et al., 2018), known for capturing rich contextual information across multiple languages. The motivation for exploring mBERT was to examine the capabilities of a multilingual model for this task on Arabic. Inspired by the findings in (Liu et al., 2019) we experimented with both the multilingual capabilities of the model and the use of outputs from different hidden layers to explore their effectiveness for identifying propaganda techniques in Arabic text. To assess mBERT's strengths and weaknesses, we compared it to AraBERT. While our experiments revealed that AraBERT achieved superior overall performance compared to mBERT, interestingly, mBERT's earlier hidden layers exhibited significantly better performance than its later layers. Note that AraBERT was trained after the evaluation cycle, thus, only mBERT was submitted for the leaderboard.

2 Related Work

Propaganda detection has seen significant advancements through various shared tasks and studies focusing on fine-tuning transformer models and exploring multilingual approaches. The WANLP 2022 shared task (Alam et al., 2022), which centered on detecting propaganda techniques in Arabic tweets, had top-performing submissions predominantly using fine-tuned transformer models like AraBERT and MARBERT (Abdul-Mageed et al., 2021). The NGU_CNLP (Hussein et al., 2022) team achieved first place with an ensemble of AraBERT models combined with data augmentation techniques. Meanwhile, the IITD (Mittal and Nakov, 2022) team employed the multilingual XLM-R (Conneau et al., 2019) model but did not surpass the performance of the specialized monolingual models.

Similarly, in the ArAIEval 2023 shared task (Hasanain et al., 2023b), most submissions focused on fine-tuning advanced transformer models,

Proceedings of The Second Arabic Natural Language Processing Conference, pages 478–482

with common approaches including AraBERT and MARBERT. Many teams, such as HTE (Hadjer and Bouklouha, 2023), utilized multitask learning. Preprocessing techniques and data augmentation were also widely used to improve model performance. Some teams employed multilingual models in their systems. For example, the Legend team (Ojo et al., 2023) implemented XLM-RoBERTa, addressing class imbalance through weighted learning and dynamic learning rate adjustment, and the ReDASPersuasion (Qachfar and Verma, 2023) team combined a multilingual transformer model with a feature engineering module to extract language-agnostic features for persuasion detection. Despite these efforts, the performance of multilingual models like XLM-RoBERTa was generally moderate compared to top-performing systems that used monolingual models specifically fine-tuned for Arabic. These consistent results across both tasks highlight the effectiveness of monolingual models for detecting propaganda techniques in Arabic text, while multilingual models show promise but require further refinement to match their performance.

Recent work on propaganda detection also includes studies on the capabilities of large language models. One such study (Hasanain et al., 2023a) investigated GPT-4's (OpenAI, 2024) ability to annotate propaganda spans in text, demonstrating its potential to act as both a general and expert annotator, with improved performance when given more information. Building on this, a more recent study (Hasanain et al., 2024a) introduced the ArPro dataset, the largest dataset for fine-grained propaganda detection in Arabic, annotated for 23 techniques.

Additionally, the ArMeme dataset (Alam et al., 2024) was introduced to explore propagandistic content in Arabic memes. This dataset focuses on visual and textual propaganda techniques within memes, providing a unique resource for understanding how propaganda operates in multimodal formats. This work adds to the growing body of research emphasizing the need for diverse datasets and advanced methodologies to tackle propaganda across different media and languages.

3 Data

For this task we have used the data provided for the task (Hasanain et al., 2024b), collected from tweets and news articles. The organizers provided the data in *jsonlines* format and already split into train, test and validation sets. There are almost 7000 samples in the training set, with each sample containing the text and the character level span of different propaganda techniques used in the text. It is a multi-label, multi-class dataset. Table 1 presents the number of document, tokens, the average number of tokens per document and the number of unique tokens in a corpus for training, validation and testing sets. There are 23 propaganda techniques found in the dataset with uneven distribution for each corpus. However, the training and validation distribution is similar for each label.

4 Task Overview

ArAIEval shared Task 1 (Hasanain et al., 2024b) involves identifying various propaganda techniques in Arabic text and pinpointing the exact character level spans where these techniques occur. Arabic, a low-resource language for this task, presents unique challenges for natural language processing (NLP) due to its rich morphology and syntactic complexity. Traditional NLP techniques often struggle with the diverse expressions and structures found in Arabic text.

The formal definition for the task is as follows: given an input sequence X of length N, denoted as $X = x_1, x_2, \ldots, x_N$, generate an output sequence Y of length 23, with one element for each class. Each element in the output sequence, y_c, is an N-dimensional vector where each element is either 1 or 0. Consecutive 1's in y_c define the span for that class.

5 Our Methodology

We approached the sequence tagging task as a multi-label, multi-class classification problem. The classes correspond to the original 23 classes from the dataset.

For tokenization, we used the pre-trained BERT subword tokenizer from Huggingface.[1] Since the subword tokenizer produces more tokens than words in the document, we trained and inferred on the first 256 tokens. We truncated or padded the input sequence to be 256 tokens, ensuring a uniform input length for the model.

The formal definition for the task is as follows: Given an input sequence X of length N, denoted as $X = x_1, x_2, \ldots, x_N$, generate an output sequence Y of length 23, with one element for each class.

[1] https://huggingface.co/docs/transformers/en/model_doc/bert#transformers.BertTokenizer

Data	Train	Validation	Test
Number of documents	6997	921	1046
Number of tokens	228373	27867	35204
Average tokens per document	32.63	30.25	33.65
Unique tokens	59193	13443	16108

Table 1: Corpus statistics

Each element in the output sequence, y_c, is an N-dimensional vector where each element is either 1 or 0. Consecutive 1's in y_c define the span for that class.

We focused on fine-tuning pre-trained BERT models (Devlin et al., 2018). Specifically, we ran our experiments on the base version of Multilingual BERT Cased with 110M parameters from HuggingFace.[2] The training loss function was a modified binary cross-entropy that ignores pad tokens.

We conducted our experiments using PyTorch on a single RTXA6000 GPU.[3] The experiments aimed to find the best hyperparameters and the optimal hidden representation from the BERT output for training the classifier layers, inspired by Liu et al. (2019).

We first performed a hyperparameter search using Optuna for 5 epochs on the dev set and then re-trained the system for additional epochs using the optimal hyperparameters found.[4] Optuna was configured to use NSGA-II algorithm (Deb et al., 2002). Our best-performing system on the leaderboard was trained on a single GPU for 400 epochs with a batch size of 32, a learning rate of 6.91×10^{-05}, a weight decay of 0.00358, and 226 warmup steps. We used feature representation from the 10th hidden layer of the BERT model.

We evaluated our system using a modified F1 measure provided by the shared task organisers. This measure accounts for the partial matching of spans between the gold standard and the hypothesis, ensuring a more accurate evaluation of the system's performance.

6 Results

We evaluated the performance of our system using different hidden layers from mBERT on both the development and test sets, and included results from AraBERTv2 (Antoun et al.) for comparison. The results, detailed in Table 3, show that the 8th layer of mBERT consistently outperformed the higher layers (10th and 12th) in terms of Micro F1, Macro F1, and Precision. This suggests that lower layers may provide better representations for propaganda detection.

Interestingly, the performance of the mBERT layers, especially the 8th layer, is not far behind that of AraBERTv2. This indicates that multilingual models like mBERT have significant potential, particularly when fine-tuning involves selecting the most effective hidden layers. AraBERTv2, while consistently performing better, highlights the benefits of a model specifically fine-tuned for Arabic. While the AraBERT and mBERT's 8th layer achieved better performances, we only submitted the result from *10th layer (mBERT)* to meet the submission deadline.

Both models demonstrated higher recall compared to precision, indicating a tendency to identify many relevant spans but also produce numerous false positives. The performance trends were consistent across both development and test sets, indicating robustness in the observed patterns.

Overall, these results align with the findings of the study by (Liu et al., 2019) on the effectiveness of different hidden layers. Their study also found that exploring hidden layers can significantly enhance model performance. These results suggest that multilingual models like mBERT have substantial potential, and that careful selection and experimentation with hidden layers can yield significant performance improvements.

7 Discussion

We evaluated the model's performance on the test set by categorizing spans into short (maximum 3 words) and long (more than 3 words) segments. The model performed better on short spans, with a precision, recall, and F1-score of 0.10. In contrast, for long spans, the precision, recall, and F1-score were 0.02. This indicates the model is more effective at identifying shorter spans, likely due to their

[2] https://huggingface.co/docs/transformers/en/model_doc/bert

[3] https://github.com/riyadhrazzaq/araieval24

[4] https://optuna.org

Reference	
AR	تتطلع تونس أن تكون افريقيا ثاني شريك تجاري، عبر إيجاد الحلول لكل الاشكاليات والمعوقات.
EN	Tunisia aspires for Africa to become its second-largest trading partner, by finding solutions to all the issues and obstacles.
Prediction	
AR	تتطلع تونس أن تكون افريقيا ثاني شريك تجاري، عبر إيجاد الحلول لكل الاشكاليات والمعوقات.
EN	Tunisia aspires for Africa to become its second-largest trading partner, by finding solutions to all the issues and obstacles.

(a)

Reference	
AR	قرر قاضي التحقيق في قضايا الإرهاب، تأييد قرار الإبقاء على ٣٩ شخصا
EN	The investigating judge in terrorism cases decided to uphold the decision to keep 39 people in custody.
Prediction	
AR	قرر قاضي التحقيق في قضايا الإرهاب ، تأييد قرار الإبقاء على ٣٩ شخصا
EN	The investigating judge in terrorism cases decided to uphold the decision to keep 39 people in custody.

(b)

Table 2: Refernce samples compared to the system prediction. Loaded Language Exaggeration Name Calling

	Test				Dev			
Model	μF1	mF1	Pre	Rec	μF1	mF1	Pre	Rec
12th layer (mBERT)	0.24	0.06	0.15	**0.62**	0.26	0.13	0.15	**0.59**
10th layer (mBERT)	0.28	0.09	0.19	0.53	0.29	0.14	0.20	0.50
8th layer (mBERT)	**0.33**	**0.10**	**0.24**	0.52	**0.32**	**0.15**	**0.24**	0.49
12th layer (AraBERTv2)	0.39	0.17	0.30	0.58	0.40	0.23	0.31	0.55
Shared Task Baseline	0.01	0.01	0.01	0.04	0.03	0.02	0.02	0.04

Table 3: Model performances on Dev and Test set. μF1 is Micro F1 and mF1 is Macro F1 Score

simpler structure and fewer contextual dependencies.

We picked a sample from the test set for a qualitative analysis to understand the model's performance in detail. We found several instances where the model identified propaganda techniques that were not present in the gold labels but seemed logical and potentially correct. In Sample 2a, the phrase "إيجاد الحلول لكل الاشكاليات والمعوقات" (finding solutions to all issues and obstacles) could be seen as exaggeration, implying a comprehensive and possibly unrealistic promise. Another example in Sample 2b, the word "تأييد" (uphold) within the context of terrorism cases may suggest a bias or specific stance, potentially serving a propagandistic purpose. The model identifies this as loaded language, highlighting a possible implicit bias.

The model also showed segmentation issues, identifying propaganda with larger spans than necessary. For example, in 2b, the reference label marked "الإرهاب" (terrorism) as Name Calling, while the model predicted "قضايا الإرهاب" (terrorism cases). Although the model's prediction is not necessarily incorrect, as it relates to the context, it doesn't match the reference label precisely.

8 Conclusion

In this study, we explored the use of a multilingual BERT model to detect propaganda techniques in Arabic text. Our approach involved fine-tuning the BERT model and utilizing representations from different hidden layers to improve detection accuracy. Despite the challenges posed by Arabic's complex morphology and syntax, our system performed well, securing the second position in the ArAIEval Shared Task. The results indicate that multilingual BERT models, when properly fine-tuned, can effectively identify and locate various propaganda techniques in Arabic texts. Future work could focus on further refining these models and addressing class imbalance to enhance detection performance.

Acknowledgments

We acknowledge the assistance of the LT-Bridge Project (GA 952194) and DFKI for the use of their Virtual Laboratory. Also, authors have been supported financially by the EMLCT [5] programme.

[5] https://mundus-web.coli.uni-saarland.de/

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, Abul Hasnat, Fatema Ahmed, Md Arid Hasan, and Maram Hasanain. 2024. Armeme: Propagandistic content in arabic memes.

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 The Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 108–118, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Wissam Antoun, Fady Baly, and Hazem Hajj. Arabert: Transformer-based model for arabic language understanding. In *LREC 2020 Workshop Language Resources and Evaluation Conference 11–16 May 2020*, page 9.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan. 2002. A fast and elitist multiobjective genetic algorithm: Nsga-ii. *IEEE Transactions on Evolutionary Computation*, 6(2):182–197.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Khaldi Hadjer and Taqiy Bouklouha. 2023. HTE at ArAIEval shared task: Integrating content type information in binary persuasive technique detection. In *Proceedings of ArabicNLP 2023*, pages 502–507, Singapore (Hybrid). Association for Computational Linguistics.

Maram Hasanain, Fatema Ahmed, and Firoj Alam. 2023a. Large language models for propaganda span annotation. *arXiv preprint arXiv:2311.09812*.

Maram Hasanain, Fatema Ahmed, 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, Torino, Italy.

Maram Hasanain, Firoj Alam, Hamdy Mubarak, Samir Abdaljalil, Wajdi Zaghouani, Preslav Nakov, Giovanni Da San Martino, and Abed Alhakim Freihat. 2023b. 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.

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)*, Bangkok. Association for Computational Linguistics.

Ahmed Samir Hussein, Abu Bakr Soliman Mohammad, Mohamed Ibrahim, Laila Hesham Afify, and Samhaa R. El-Beltagy. 2022. NGU CNLP atWANLP 2022 shared task: Propaganda detection in Arabic. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 545–550, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019. Linguistic knowledge and transferability of contextual representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1073–1094, Minneapolis, Minnesota. Association for Computational Linguistics.

Shubham Mittal and Preslav Nakov. 2022. IITD at WANLP 2022 shared task: Multilingual multigranularity network for propaganda detection. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 529–533, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Olumide Ojo, Olaronke Adebanji, Hiram Calvo, Damian Dieke, Olumuyiwa Ojo, Seye Akinsanya, Tolulope Abiola, and Anna Feldman. 2023. Legend at ArAIEval shared task: Persuasion technique detection using a language-agnostic text representation model. In *Proceedings of ArabicNLP 2023*, pages 594–599, Singapore (Hybrid). Association for Computational Linguistics.

OpenAI. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Fatima Zahra Qachfar and Rakesh Verma. 2023. ReDASPersuasion at ArAIEval shared task: Multilingual and monolingual models for Arabic persuasion detection. In *Proceedings of ArabicNLP 2023*, pages 549–557, Singapore (Hybrid). Association for Computational Linguistics.