MemeMind at ArAIEval Shared Task: Generative Augmentation and Feature Fusion for Multimodal Propaganda Detection in Arabic Memes through Advanced Language and Vision Models

Uzair Shah[1], Md Rafiul Biswas[1], Marco Agus[1], Mowafa Househ[1], Wajdi Zaghouani[2]
[1] Hamad Bin Khalifa University, Qatar Foundation, Doha, Qatar,
[2] Northwestern University in Qatar, Education City, Doha, Qatar
Correspondence: wajdi.zaghouani@northwestern.edu

Abstract

Detecting propaganda in multimodal content, such as memes, is crucial for combating disinformation on social media. This paper presents a novel approach for the ArAIEval 2024 shared Task 2 on Multimodal Propagandistic Memes Classification, involving text, image, and multimodal classification of Arabic memes. For text classification (Task 2A), we fine-tune state-of-the-art Arabic language models and use ChatGPT4-generated synthetic text for data augmentation. For image classification (Task 2B), we fine-tune ResNet18, EfficientFormerV2, and ConvNeXt-tiny architectures with DALL-E-2-generated synthetic images. For multimodal classification (Task 2C), we combine ConvNeXt-tiny and BERT architectures in a fusion layer to enhance binary classification. Our results show significant performance improvements with data augmentation for text and image classification models and with the fusion layer for multimodal classification. We highlight challenges and opportunities for future research in multimodal propaganda detection in Arabic content, emphasizing the need for robust and adaptable models to combat disinformation.

1 Introduction

The rapid proliferation of social media has led to an unprecedented spread of information, including propagandistic content designed to manipulate public opinion (Chernyavskiy et al., 2024). Propaganda employs various techniques to influence the emotions, attitudes, and actions of its target audience (Jowett and O'donnell, 2018). The rise of multimodal content, such as memes, has added a new dimension to the challenge of detecting propaganda online (Manzoor et al., 2023; Dimitrov et al., 2021; Wilson et al., 2023). Memes, which combine text and images, have become a popular medium for spreading propaganda on social media platforms (Al-Saqqa and Awajan, 2021).

Detecting propaganda in memes is a complex task that requires understanding both textual and visual content. Previous work on multimodal propaganda detection has primarily focused on English-language content (Dimitrov et al., 2021; Kiela et al., 2020). However, research on detecting propaganda in Arabic memes is limited despite the increasing prevalence of such content on Arabic-language social media (Alhindi et al., 2019).

In this paper, we present a novel framework to tackle the ArAIEval 2024 shared task, focusing specifically on Task 2: Multimodal Propagandistic Memes Classification. For Task 2A, we fine-tune various pre-trained Arabic language models, including AraBERTv1, AraBERTv2, MARBERT, CAMeLBERT, AraGPT-2, qarib-bert, electra-gpt, and Arabic-BERT (Antoun et al., 2020; Abdul-Mageed et al., 2021; abd; Inoue et al., 2021; Alkhamissi et al., 2022). We augment these models with a classification layer and employ class weighting to handle the imbalanced dataset. We further enhance the models by fine-tuning on a combination of original and synthetic text data generated using ChatGPT4.

For Task 2B, we fine-tuned pre-trained ResNet18 (He et al., 2015), ConvNeXt-tiny (Liu et al., 2022), and EfficientNetV2 (Li et al., 2023) models for meme image classification. These models were augmented with additional synthetic images generated by DALL-E-2 to improve binary classification performance. For Task 2C, we combined the AraBERT model trained in Task 2A with the ConvNeXt-tiny model trained in Task 2B to jointly extract features from both text and images. These features were concatenated and fed into a fusion layer, followed by a binary classification layer.

The main contributions of our work are the following:

- We develop a novel holistic framework for detecting propaganda in Arabic memes, leverag-

Proceedings of The Second Arabic Natural Language Processing Conference, pages 467–472

ing and customizing with a fusion layer state-of-the-art pre-trained language models and multimodal architectures.

- We propose a data augmentation scheme for alleviating class imbalance by using synthetic propaganda text generated by ChatGPT4 for improving the performance of our text classification models.

- We offer insights into the challenges and opportunities for future research on multimodal propaganda detection in Arabic content.

2 Related work

The ArAIEval shared task, organized by Hasanain et al. (2024b), focuses on detecting propagandistic techniques in unimodal and multimodal Arabic content. This shared task builds upon the previous edition, ArAIEval 2023 (Hasanain et al., 2023b), which targeted persuasion techniques and disinformation detection in Arabic text. The organizers have also explored the use of large language models for propaganda span annotation (Hasanain et al., 2023a) and investigated the ability of GPT-4 to identify propaganda spans in news articles (Hasanain et al., 2024a). The ArAIEval shared task is a continuation of the efforts made in the WANLP 2022 shared task on propaganda detection in Arabic (Alam et al., 2022). The spread of propaganda on social media platforms is not limited to English-language content. Arabic-language social media has also seen a rise in the dissemination of propagandistic content, including memes (Kougia et al., 2023). Alhindi et al. (2019) presented a corpus of Arabic propaganda posts on social media, highlighting the need for research on detecting propaganda in Arabic content. Rosso et al. (2018) conducted a survey on author profiling, deception, and irony detection for the Arabic language, discussing various aspects of Arabic natural language processing that could be relevant to the task of identifying propaganda in Arabic memes.

Rangel et al. (2019) focused on detecting deceptive tweets in Arabic for cyber-security purposes, demonstrating the importance of identifying manipulative content on Arabic social media platforms. Zaghouani and Charfi (2018) presented a large multi-dialect Twitter corpus for Arabic, which could be useful for understanding the linguistic characteristics of Arabic social media content and developing models for detecting propaganda in memes. While not directly related to propaganda detection, other studies have demonstrated the application of text analysis techniques to Arabic content in various social contexts. For example, Zaghouani (2018) presented a large-scale Arabic social media corpus for detecting youth depression, highlighting the importance of analyzing Arabic social media content for mental health monitoring purposes. In this work, we alleviated the class imbalance through the automatic generation of synthetic texts by chat-GPT, and we improved multimodal classification by using a fusion layer concatenating image and text features to feed a dense neural layer for classification.

3 Data

The ArAIEval 2024 shared task consists of two main tasks focusing on the detection of propagandistic content in Arabic media. Task 1, Unimodal (Text) Propagandistic Technique Detection, is a sequence tagging task that involves identifying propaganda techniques and their corresponding spans within a given text snippet. The input can be either a news paragraph or a tweet. The goal is to detect the specific propaganda techniques used in the text and the exact spans where each technique appears. Task 2, Multimodal Propagandistic Memes Classification, focuses on the classification of propagandistic content in multimodal memes. This task is further divided into three Tasks. Task 2A, Text-based Propaganda Classification, aims to classify whether the text extracted from a meme is propagandistic or not. Task 2B, Image-based Propaganda Classification, involves determining whether the content of a meme (an image with overlaid text) is propagandistic. Task 2C, Multimodal Propaganda Classification, requires detecting whether the content of a meme is propagandistic based on both the extracted text and the meme image itself. Our focus is solely on Task 2, specifically the Multimodal Propagandistic Memes Classification task. The dataset (Alam et al., 2024) consists of 2143, 312, and 607 pairs of text and images for the training, validation, and test splits, respectively. The training set is imbalanced, with 603 pairs in the propaganda class and 1540 pairs in the non-propaganda class.

4 Methodology

In this section, we present the various techniques used for developing the final model for the task. The objective is to classify memes as either propagandistic or non-propagandistic. The task involves multimodal classification of propagandistic memes, subdivided into three distinct Tasks: **Task 2A**, **Task 2B**, and **Task 2C** (description in Sec 3). Each of these Tasks entails a binary classification objective.

Task 2A: We fine-tuned various pre-trained models trained on Arabic content, including AraBERTv1, AraBERTv2, MARBERT, CAMeL-BERT, AraGPT-2, qarib-bert, electra-gpt, and Arabic-BERT (Antoun et al., 2020; Abdul-Mageed et al., 2021; abd; Inoue et al., 2021; Alkhamissi et al., 2022). To accomplish this, we augmented these models by attaching a classification layer with random weights. Given the imbalanced nature of the dataset, we employed a class weighting scheme during the fine-tuning process. Initially, we fine-tuned the pre-trained model using the original dataset. Subsequently, we optimized the model by adjusting the dropout rate. Further enhancement was achieved by fine-tuning the last layer of the model using both synthetic and original text, resulting in improved performance over the original dataset (see Table 1). To further augment the dataset, we leveraged ChatGPT4 to analyze themes and generated synthetic text for both classes. This augmentation effectively doubled the size of the dataset. Following this, we fine-tuned the last layer of the model once again and increased the dropout rate (refer to Section 4). The incorporation of additional synthetic text into the original dataset led to an observable improvement in model performance (see Table 1).

Task 2B: We fine-tuned ResNet18 (He et al., 2015), ConvNeXt-tiny (Liu et al., 2022), and EfficientNetV2 (Li et al., 2023), augmenting them by replacing the classification layer with a binary classification layer. Given the dataset's imbalance, we employed a class weighting scheme during training. Initially, we froze the feature extraction layers and trained only the classification layer to stabilize its weights. Additionally, we optimized the model by adjusting the dropout rate (for ResNet18) and stochastic drop path rate (for ConvNeXt-tiny). After these adjustments, we trained the model for a few epochs before unfreezing it for further training. To balance the dataset, we leveraged DALL-E-2 to generate synthetic images. Through its image-to-image pipeline, we created synthetic propaganda images by providing propaganda meme images and associated text. These augmentations significantly improved model performance (see Table 1). With these enhancements, we fully trained the model to accurately categorize memes. Incorporating additional synthetic images into the original dataset notably boosted performance (see Table 1).

Task 2C: Initially, we utilized pretrained AraBERT and ConvNeXt-tiny models to extract text and visual features from meme text and images. Subsequently, various machine learning algorithms were employed for classification. Additionally, we incorporated the trained text model from Task 2A (see subsection 4) and the visual model from Task 2B (see subsection 4). Features were extracted from these models and concatenated, leading to improved performance compared to the baseline. Later, we combined the text and visual models in a deep learning setup to extract feature embeddings. A fusion layer was attached to combine these features, followed by a binary linear classification layer. We optimized the dropout rate for the text model and the stochastic drop path rate for the visual model. Both the text and visual models were frozen, and only the fusion and classification layers were trained. The incorporation of text and visual models together in a deep learning setup, along with the use of a fusion layer, resulted in a noticeable improvement in model performance (see Table 1).

Experimental Setup. We conducted all experiments using PyTorch 2.1.2 on an Ubuntu 22.04.4 operating system, training the models on an Nvidia-RTX 2080 GPU. The AdamW optimizer was utilized, with an initial learning rate of $1e^{-5}$ for text models (Task 2A) and the multimodal network of Task 2C, and $1e^{-4}$ for visual models (Task 2B). A batch size of 32 was maintained across different tasks, while for GPT text models, the batch size was set to 1. The sentence length for tokenization was fixed at 128, and images were resized to 224x224 across all tasks. In Task 2A, we initially set the dropout rate to 0.25 for the original dataset and trained for 50 epochs. Subsequently, we increased the dropout rate to 0.5 to balance the dataset through synthetic data and further trained for 30 epochs. In a later stage, when the dataset size doubled due to synthetic data augmentation, the dropout rate was set to 0.75, and

training was extended to 100 epochs. For Task 2B, the stochastic drop path rate was set to 0.1 for training the last layer and trained for 10 epochs. During full model training, the rate was increased to 0.15, and training was extended to 100 epochs. Additionally, the dropout rate was set to 0.5 for text models, while the stochastic drop path rate was set to 0.2 for Task 2C. Model training for Task 2C was conducted for 100 epochs. The model achieving the highest F1-micro score is saved.

5 Results and Discussion

The official metric used for each task is the Macro F1-score. Table 1 illustrates the tangible impact of data augmentation and classification methods on model performance (Macro F1) during the development phase of Task 2A, Task 2B, and Task 2C of ArAIEval. Notably, the augmentation techniques have led to significant improvements across different tasks. In Task 2A, augmenting both propaganda and non-propaganda classes (denoted as "Both, 2x") doubled the dataset size and resulted in a notable increase in Macro F1-score, with AraBERT achieving an impressive F1-score of 0.80. Task 2B showcased a similar trend, where augmenting only the propaganda class enhanced model performance. ConvNeXt-tiny, for instance, achieved an F1-score of 0.736 with augmentation, compared to 0.729 without augmentation. Task 2C demonstrated the effectiveness of utilizing a fusion approach (Shah et al., 2024), combining AraBERT and ConvNeXt-tiny models. This fusion approach yielded the highest Macro F1-score among all tasks, with the ensemble model achieving an F1-score of 0.837. These results underscore the importance of data augmentation strategies and innovative classification methods in enhancing model performance in multimodal classification tasks.

Table 2 presents the leaderboard standings for Tasks 2A, 2B, and 2C of the ArAIEval competition hosted at ArabicNLP24. Notably, our team, Meme_mind, secured commendable positions across all tasks, showcasing competitive performance in the multimodal classification challenge. In Task 2A, our team achieved a Macro F1-score of 0.7464, securing a notable position among other participants. Additionally, in Task 2B, our model exhibited robust performance with a Macro F1-score of 0.6634, further underlining our team's proficiency in meme image classification.

Task	Model	Augmentation	F1
2A	AraBERT	No	0.763
	AraBERT	Yes (prop.)	0.789
	AraBERT	Yes (both, 2x)	0.80
2B	ConvNeXt	No	0.729
	ConvNeXt	Yes (prop.)	0.736
2C	AraBERT + ConvNeXt	Random Forest	0.819
	AraBERT + ConvNeXt	Fusion	0.837

Table 1: Impact of Data Augmentation and Classification Methods on Model Performance (Macro F1) during the Development Phase of Task 2A, Task 2B, and Task 2C of ArALEval. "Prop." denotes augmentation for only the propaganda class, while "Both, 2x" indicates augmentation for both classes, doubling the dataset size.

Furthermore, in Task 2C, our team's ensemble approach garnered significant recognition, achieving a Macro F1-score of 0.7972 and securing a prominent position on the leaderboard. These results underscore the effectiveness of our approach in addressing the challenges posed by propagandistic content detection across diverse modalities.

Task	Team	Macro F1
Task 2A	MZ	0.7869
	CLTL	0.7796
	Meme_mind (Ours)	0.7464
	DLRG	0.7394
	One_by_zero	0.6742
	Z-Index	0.6330
	Baseline (Random)	0.4529
Task 2B	CLTL	0.7104
	Meme_mind (Ours)	0.6634
	MZ	0.6592
	Baseline (Random)	0.4755
Task 2C	MZ	0.8051
	ASOS	0.7987
	CLTL	0.7980
	Meme_mind (Ours)	0.7972
	Team Engima	0.7526
	MODOS	0.7290
	Z-Index	0.7120
	Baseline (Random)	0.4923

Table 2: Leaderboard for Tasks 2A, 2B, and 2C of the ArAIEval competition at ArabicNLP24.

6 Conclusion

We presented our contribution to the ArAIEval Shared Task, utilizing advanced language and vision models. Our approach includes augmentation through generated texts and layers to fuse multimodal features. This initiative and our proposed solutions highlight the increasing significance of developing effective methods for analyzing Arabic social media content, especially in detecting propaganda and manipulative content. By leveraging state-of-the-art language models and multimodal

architectures, our goal is to advance propaganda detection in Arabic memes and the broader field of Arabic natural language processing.

7 Acknowlegdement

The research reported in this paper is funded by grant NPRP14C0916-210015 from the Qatar National Research Fund (QNRF) part of Qatar Research Development and Innovation Council (QRDI).

References

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021. ArBERT: Bert for arabic language understanding. *arXiv preprint arXiv:2110.05609*.

Samer Al-Saqqa and Arafat Awajan. 2021. Memeganda: Identifying propaganda memes using encoding and decoding deep features. *IEEE Access*, 9:153670–153682.

Firoj Alam, Abul Hasnat, Fatema Ahmed, Md Arid Hasan, and Maram Hasanain. 2024. ArMeme: Propagandistic content in arabic memes. *arXiv: 2406.03916*.

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.

Tariq Alhindi, Savvas Petridis, and Smaranda Muresan. 2019. Arpr: Arabic propaganda detection using transfer learning. In *Proceedings of the 2nd Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda (NLP4IF)*, pages 1–6.

Belal Alkhamissi, Mohamed Gabr, Marwan Khaled, and Rania Essam. 2022. Adapting multilingual neural language models for arabic dialects. *arXiv preprint arXiv:2202.06741*.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based model for arabic language understanding. *arXiv preprint arXiv:2003.00104*.

Alexander Chernyavskiy, Dmitry Ilvovsky, and Preslav Nakov. 2024. Unleashing the power of discourse-enhanced transformers for propaganda detection. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1452–1462.

Dimitar Dimitrov, Bishr Baran, Pavlos Fafalios, Ran Yu, Xiaojun Zhu, Matthäus Zloch, and Stefan Steinmetz. 2021. Detecting propaganda techniques in memes. *arXiv preprint arXiv:2109.08013*.

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.

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.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*.

Naoya Inoue, Pontus Stenetorp, and Kentaro Inui. 2021. Interplay between preferences and machine learning in language model fine-tuning. *arXiv preprint arXiv:2110.08413*.

Garth S Jowett and Victoria O'donnell. 2018. *Propaganda & persuasion*. Sage Publications.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. Hateful memes challenge: Detecting hate speech in multimodal memes. In *Advances in Neural Information Processing Systems*, volume 33, pages 2611–2624.

Vasiliki Kougia, Simon Fetzel, Thomas Kirchmair, Erion Çano, Sina Moayed Baharlou, Sahand Sharifzadeh, and Benjamin Roth. 2023. Memegraphs: Linking memes to knowledge graphs. In *International Conference on Document Analysis and Recognition*, pages 534–551. Springer.

Yanyu Li, Ju Hu, Yang Wen, Georgios Evangelidis, Kamyar Salahi, Yanzhi Wang, Sergey Tulyakov, and Jian Ren. 2023. Rethinking vision transformers for mobilenet size and speed. In *Proceedings of the IEEE international conference on computer vision*.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. 2022. A convnet for the 2020s. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Muhammad Arslan Manzoor, Sarah Albarri, Ziting Xian, Zaiqiao Meng, Preslav Nakov, and Shangsong Liang. 2023. Multimodality representation learning: A survey on evolution, pretraining and its applications. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(3):1–34.

Francisco Rangel, Paolo Rosso, Anis Charfi, and Wajdi Zaghouani. 2019. Detecting deceptive tweets in arabic for cyber-security. In *2019 IEEE International Conference on Intelligence and Security Informatics (ISI)*, pages 86–91. IEEE.

Paolo Rosso, Francisco Rangel, Iraisy Hernández Far'ıas, Leticia Cagnina, Wajdi Zaghouani, and Anis Charfi. 2018. A survey on author profiling, deception, and irony detection for the arabic language. *Language and Linguistics Compass*, 12(4):e12275.

Uzair Shah, Muhammad Tukur, Mahmood Alzubaidi, Giovanni Pintore, Enrico Gobbetti, Mowafa Househ, Jens Schneider, and Marco Agus. 2024. Multi-panowise: Holistic deep architecture for multi-task dense prediction from a single panoramic image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1311–1321.

Anna Wilson, Seb Wilkes, Yayoi Teramoto, and Scott Hale. 2023. Multimodal analysis of disinformation and misinformation. *Royal Society Open Science*, 10(12):230964.

Wajdi Zaghouani. 2018. A large-scale social media corpus for the detection of youth depression (project note). *Procedia Computer Science*, 142:347–351.

Wajdi Zaghouani and Anis Charfi. 2018. Arap-tweet: A large multi-dialect twitter corpus for gender, age and language variety identification. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.