ASOS at ArAIEval Shared Task: Integrating Text and Image Embeddings for Multimodal Propaganda Detection in Arabic Memes

Yasser Alhabashi[1], Abdullah I. Alharbi[2], Samar Ahmed, Serry Sibaee[1], Omer Nacar[1], Lahouari Ghouti[1], Anis Koubaa[1]
[1]Robotics and Internet-of-Things Lab, Prince Sultan University, Riyadh 12435, Saudi Arabia
[2]Faculty of Computing and Information Technology in Rabigh
King Abdulaziz University, Jeddah, Saudi Arabia
{yalhabashi,ssibaee,onajar,lghouti, akoubaa}@psu.edu.sa,
Samar.sass6@gmail.com, aamalharbe@kau.edu.sa

Abstract

This paper describes our participation in the ArAIEval Shared Task 2024, focusing on Task 2C, which challenges participants to detect propagandistic elements in multimodal Arabic memes. The challenge involves analyzing both the textual and visual components of memes to identify underlying propagandistic messages. Our approach integrates the capabilities of MARBERT and ResNet50, top-performing pre-trained models for text and image processing, respectively. Our system architecture combines these models through a fusion layer that integrates and processes the extracted features, creating a comprehensive representation that is more effective in detecting nuanced propaganda. Our proposed system achieved significant success, placing second with an F1-score of 0.798.

1 Introduction

With the rise of social media platforms and the growing number of users worldwide, people have found a more accessible way to express themselves. They have turned to humour and wit, creating jokes and memes to succinctly convey thoughts, emotions, and behaviours.

Memes are compositions of images and text, merging them to collectively communicate ideas. They have become a prevalent form of communication on various social media platforms such as Twitter and Facebook, where content often combines text with images or videos. With the widespread use of memes on social media platforms across all age groups, it is important to categorize and understand the sentiment and purpose behind them. This involves determining whether memes convey positive or negative implications and whether they are used to spread hatred, ridicule, and misinformation (Dupuis and Williams, 2019). It's also crucial to identify whether they contain hateful or non-hateful content (Lippe et al., 2020). From another perspective, some focus on processing text only that extracted from memes (Boinepelli et al., 2020), while others view it as a multi-modal task (Du et al., 2020). Despite the diversity of tasks, the ultimate goal remains the same: understanding the impact of memes on society. Our team participated in the ArAIEval Shared Task, specifically focusing on Task 2C, which involves analyzing multimodal content (Hasanain et al., 2024b). This content includes textual data extracted from memes and their corresponding images to identify propagandistic elements. We developed a methodology that combines the best-performing text and image pre-trained models, using MARBERT for textual analysis and ResNet50 for visual analysis.

2 Related Work

Recent research has focused on classifying memes and extracting features using advanced computational methods. For example, one study (Suryawanshi et al., 2023) addresses the classification of multimodal offensive memes by employing Natural Language Inference (NLI) techniques and fine-tuning RoBERTa models to refine the task into a unimodal offensive text classification challenge. Another significant study (Dimitrov et al., 2021) proposes a multi-label multimodal approach for detecting propaganda techniques in memes. This method involves annotating a corpus of 950 memes with 22 different propaganda techniques found in text, images, or both, utilizing a multilabel setup for the annotations.

A different work (Ouaari et al., 2022) proposes a feature extraction method for multimodal meme classification using Deep Learning approaches. They aim to classify the sentiment in memes using Multi-Embedding, Residual Network, and Bimodal Autoencoder techniques, highlighting the importance of sentiment analysis for monitoring and managing negative content online. Addition-

Proceedings of The Second Arabic Natural Language Processing Conference, pages 473–477

ally, a study (Deng et al., 2023) introduces a Meme-Integrated Deep Learning (MIDL) framework for meme classification and analysis, employing BERT and ResNet architectures. They discuss the limitations of traditional methods and the potential of the proposed framework to advance research in online culture analysis. Furthermore, surveys address challenges in meme classification and understanding harmful memes, discussing a typology of harmful memes and the lack of suitable datasets for studying certain types, such as those featuring self-harm and extremism. They emphasize the need for cutting-edge solutions and fine-grained analysis in this area. Works (Sharma et al., 2022), and (Afridi et al., 2021) collectively aim to enhance meme classification, sentiment analysis, and understanding of harmful content on social media platforms, contributing to research in online culture analysis and combating misinformation.

3 Methodology

This section outlines our methodology for detecting propagandistic content in Arabic memes. It details the dataset, pre-trained models for text and image processing, system architecture, and training procedures, collectively enabling our system to analyze multimodal data effectively.

3.1 Dataset

The dataset compiles memes from various social media platforms including Facebook, Twitter, Instagram, and Pinterest (Hasanain et al., 2023, 2024a). Each meme is labeled as either propagandistic or not propagandistic. The dataset is structured as a JSON file containing the following fields: "id": a unique identifier assigned to each meme; "img_path": the file path or URL pointing to the image associated with the meme; "text": the textual content of the meme, often presenting a dialogue or interaction between characters; and "class_label": indicates whether the meme is classified as "not propaganda" or "propaganda". As depicted in Figure 1, all subsets (training, development, and test) exhibit a similar distribution pattern, with a significantly higher number of non-propaganda memes compared to propaganda memes. This consistent imbalance across the sets challenges developing robust models that effectively recognize less represented propagandistic content.

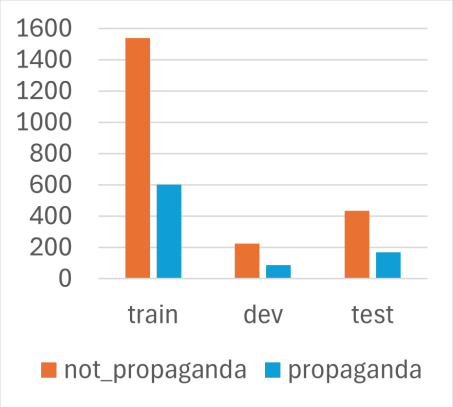

Figure 1: Distribution of memes in the dataset by Classification. This bar chart shows the number of memes classes within the training, dev, and test subsets.

3.2 Models

To effectively handle the multimodal data, we compared three pre-trained models for both text and image processing to determine the most efficient method for detecting propagandistic content.

3.2.1 Pre-trained Text Models

MARBERT (Abdul-Mageed et al., 2021) is a bidirectional transformer-based model specifically designed for Arabic language processing. This model was pre-trained on extensive and varied datasets (about 6 Billion tweets) to enable effective transfer learning for Arabic dialects. MARBERT focuses on the diverse Arabic dialects, many of which have not been extensively studied due to the scarcity of resources. AraBERT (Antoun et al.) is another prominent model in Arabic NLP, fine-tuned to understand Modern Standard Arabic (MSA) and Arabic dialects. It was trained using a large corpus that included 200 million sentences, totalling 77 billion words. QARiB (Abdelali et al., 2021), the Arabic and Dialectal BERT (QARiB) model, was trained on a vast dataset that included 180 million text phrases and nearly collected 420 million tweets.

3.2.2 Pre-trained Images Models

ResNet50 (He et al., 2016) is a Convolutional Neural Network (CNN) architecture comprising 50 layers. It is trained on large-scale image datasets, such as ImageNet, to learn hierarchical features from images. ResNet50 is widely used for tasks like image classification, object detection, and image segmentation. Its architecture features residual connections, which help mitigate the vanishing gradient problem during training. Swin Transformer (Swin-V2) (Liu et al., 2021) is a transformer-based

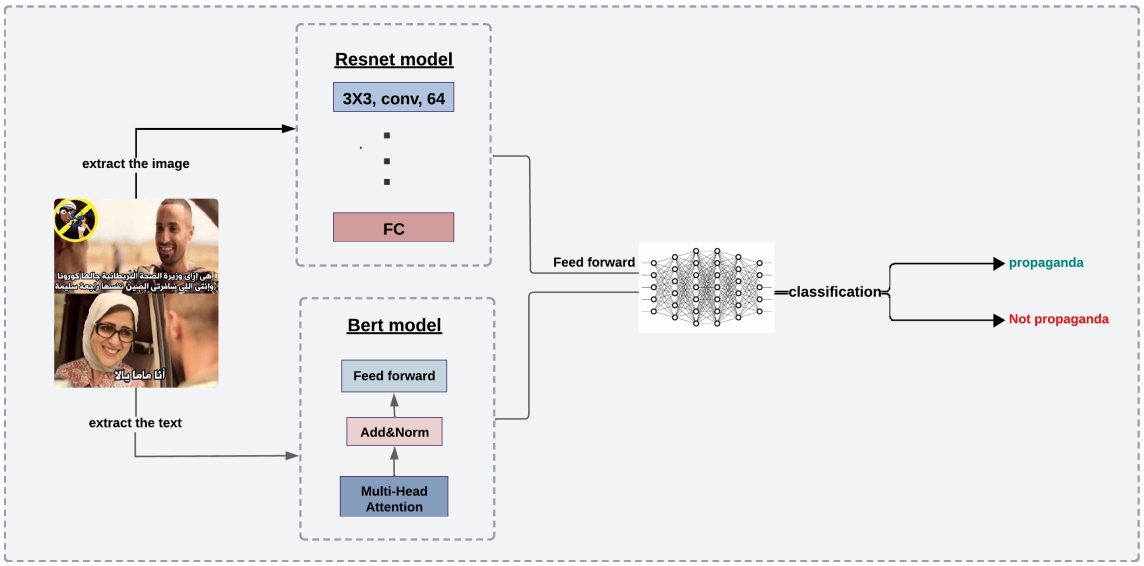

Figure 2: Architecture of our proposed multimodal classification system for propaganda detection.

architecture designed for vision tasks. It employs a hierarchical design with shifted windows to capture local and global information effectively. Swin-V2 is a specific variant of the Swin Transformer, characterized by its unique configuration of layers and parameters. This model is trained on large-scale image datasets for image classification, object detection, and semantic segmentation tasks. The Swin Transformer is notable for its strong performance and scalability in handling large images. MaxViT (Tu et al., 2022) is a variation of the Vision Transformer (ViT) architecture based on the transformer model. MaxViT introduces modifications to the transformer architecture to enhance performance and efficiency. It uses the self-attention mechanism to capture global dependencies in images. MaxViT is trained on image datasets like other image models for tasks such as image classification, object detection, and image generation.

3.3 System Architecture

Our system architecture is strategically designed to effectively process and integrate the outputs from the aforementioned text and image models, as illustrated in Figure 2. Each text model generates a high-dimensional embedding of 768 dimensions, capturing the nuances of the Arabic language, while each image model outputs feature vectors—2048 dimensions from ResNet50, 1024 from Swin Transformer V2, and 768 from MaxViT—that encapsulate both the local and global contextual details of the images. These outputs are concatenated to form a comprehensive feature vector whose dimensionality depends on the specific models used. For instance, combining MARBERT with ResNet50 yields a vector of 2816 dimensions.

This feature vector is processed through a fusion layer, which often reduces the dimensionality to a more manageable size (512 dimensions). This is achieved by passing the concatenated vector through a dense layer that integrates the textual and visual information into a unified representation. This combined representation is then fed into a classification head, consisting of a fully connected layer with a softmax activation function, which classifies the content into "propagandistic" and "not propagandistic" categories. This integration of text and image embeddings ensures a robust analysis, enhancing our capability to detect nuanced propagandistic content in multimodal data such as Arabic memes.

3.4 Experimental Setup

We employ the Cross-Entropy Loss function for our binary classification tasks and the Adam optimizer. It is renowned for efficiently handling sparse gradients and adaptive learning rate adjustments. Training begins with a learning rate of 2e-5, a standard for fine-tuning BERT-based models, with a batch size of 16 to balance computational efficiency and memory constraints. Each model is trained for up to three epochs, as preliminary results indicated that more extended training periods did not significantly enhance performance and could lead to overfitting. To prevent overfitting, a dropout rate of 0.3 is implemented in both the text

Model	precision	recall	F1-score
MARBERT	0.778	0.739	0.754
ARABERT	0.752	0.696	0.713
QARiB	0.728	0.694	0.706

Table 1: Comparative performance metrics of pre-trained text models on development set.

model and the fusion layers to promote robustness by randomly deactivating neurons during training. We used the official metric proposed by the shared tasks organisers (macro-average F1-score) to assess model effectiveness.

4 Results and Discussion

4.1 Text-Based Models

To evaluate our adopted pre-trained models namely MARBERT, ARABERT, and QARiB, we adopted the F1-score as a key metric to assess their effectiveness on the development set in detecting propagandistic content. MARBERT emerged as the leading model, achieving a macro-average F1-score of 0.754. This model displayed robust capabilities in discriminating between propagandistic and non-propagandistic content, evidenced by its balanced performance across the classes. Table 1 shows the detailed results of this evaluation.

ARABERT recorded a macro-average F1-score of 0.713. While it showed proficiency in recognizing non-propagandistic content, its performance in identifying it was less effective, impacting its overall score. QARiB, with the lowest macro-average F1-score of 0.706, demonstrated challenges in consistently identifying more nuanced propagandistic elements, indicating a need for further refinement. Based on these results, MARBERT was selected for further experiments.

4.2 Integrated Text and Image Models (Multimodel)

In our investigation, we combined our top-performing text model, MARBERT, with different image models to find the most effective approach for detecting propagandistic content in Arabic memes. The evaluation conducted on the development set showed that integrating MARBERT with MaxViT demonstrated superior performance, achieving an F1-score of 0.791, with precision at 0.838 and recall at 0.842. This configuration outperformed both Swin-V2, which posted an F1-score of 0.786, and ResNet50, which recorded a score of 0.753. Due to its outstanding results, MaxViT was selected for further assessment on the test set. Table 2 shows the detailed results of this evaluation.

Model	Precision	Recall	F1-score
Development set			
ResNet50	0.803	0.810	0.7531
Swin-V2	0.829	0.833	0.7866
MaxViT	0.838	0.842	0.791
Test set			
MaxViT	0.790	0.809	0.798

Table 2: Comparative performance metrics of multimodels on development and test sets. The text-based model used alongside the image model is MARBERT.

Upon testing, the MaxViT-MARBERT combination maintained strong performance, achieving a precision of 0.7904 and a recall of 0.809, resulting in an F1-score of 0.798. These results solidified the model's effectiveness in accurately identifying propagandistic content, confirming the robustness of our selection based on earlier evaluations. These findings highlight the critical role of integrating advanced image processing techniques with sophisticated text analysis to tackle the complex challenge of multimodal propaganda detection.

5 Conclusion

Our participation in the ArAIEval Shared Task 2024 led to the development of a robust system that integrates advanced text and image processing techniques to detect propagandistic content in Arabic memes. We used MARBERT for textual analysis and MaxViT for visual analysis. Our approach placed second with an F1-socre of 0.798, showing its effectiveness in handling complex multimodal data. These results validate our methodology and highlight the importance of combining different modalities to improve content classification systems. In the future, we plan to refine these techniques, explore additional model integrations, and expand our dataset to enhance the system's performance and adaptability across various multimedia platforms.

Acknowledgments

We thank RIOTU labs at Prince Sultan University in Saudi Arabia for allowing us to use the lab server.

References

Ahmed Abdelali, Sabit Hassan, Hamdy Mubarak, Kareem Darwish, and Younes Samih. 2021. Pre-training bert on arabic tweets: Practical considerations.

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.

Tariq Habib Afridi, Aftab Alam, Muhammad Numan Khan, Jawad Khan, and Young-Koo Lee. 2021. A multimodal memes classification: A survey and open research issues. In *Innovations in Smart Cities Applications Volume 4: The Proceedings of the 5th International Conference on Smart City Applications*, pages 1451–1466. Springer.

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.

Sravani Boinepelli, Manish Shrivastava, and Vasudeva Varma. 2020. Sis@ iiith at semeval-2020 task 8: An overview of simple text classification methods for meme analysis. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1190–1194.

Xuxiang Deng, Yifan Liu, and Qihao Yan. 2023. Meme-integrated deep learning: A multimodal classification fusion framework to fuse meme culture into deep learning. In *2023 International Conference on Image, Algorithms and Artificial Intelligence (ICIAAI 2023)*, pages 130–145. Atlantis Press.

Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021. Detecting propaganda techniques in memes. *arXiv preprint arXiv:2109.08013*.

Yuhao Du, Muhammad Aamir Masood, and Kenneth Joseph. 2020. Understanding visual memes: An empirical analysis of text superimposed on memes shared on twitter. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pages 153–164.

Marc J Dupuis and Andrew Williams. 2019. The spread of disinformation on the web: An examination of memes on social networking. In *2019 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, pages 1412–1418. IEEE.

Maram Hasanain, Fatema Ahmed, and Firoj Alam. 2023. 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, 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. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Phillip Lippe, Nithin Holla, Shantanu Chandra, Santhosh Rajamanickam, Georgios Antoniou, Ekaterina Shutova, and Helen Yannakoudakis. 2020. A multimodal framework for the detection of hateful memes. *arXiv preprint arXiv:2012.12871*.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022.

Sofiane Ouaari, Tsegaye Misikir Tashu, and Tomáš Horváth. 2022. Multimodal feature extraction for memes sentiment classification. In *2022 IEEE 2nd Conference on Information Technology and Data Science (CITDS)*, pages 285–290. IEEE.

Shivam Sharma, Firoj Alam, Md Shad Akhtar, Dimitar Dimitrov, Giovanni Da San Martino, Hamed Firooz, Alon Halevy, Fabrizio Silvestri, Preslav Nakov, and Tanmoy Chakraborty. 2022. Detecting and understanding harmful memes: A survey. *arXiv preprint arXiv:2205.04274*.

Shardul Suryawanshi, Mihael Arcan, Suzanne Little, and Paul Buitelaar. 2023. Multimodal offensive meme classification with natural language inference. In *Proceedings of the 4th Conference on Language, Data and Knowledge*, pages 134–145.

Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. 2022. Maxvit: Multi-axis vision transformer. In *European conference on computer vision*, pages 459–479. Springer.