CLTL at ArAIEval Shared Task: Multimodal Propagandistic Memes Classification Using Transformer Models

Yeshan Wang
CLTL, Vrije Universiteit Amsterdam
Amsterdam, The Netherlands
y.wang11@student.vu.nl

Ilia Markov
CLTL, Vrije Universiteit Amsterdam
Amsterdam, The Netherlands
i.markov@vu.nl

Abstract

We present the CLTL system designed for the ArAIEval Shared Task 2024 on multimodal propagandistic memes classification in Arabic. The challenge was divided into three subtasks: identifying propagandistic content from textual modality of memes (subtask 2A), from visual modality of memes (subtask 2B), and in a multimodal scenario when both modalities are combined (subtask 2C). We explored various unimodal transformer models for Arabic language processing (subtask 2A), visual models for image processing (subtask 2B), and concatenated text and image embeddings using the Multilayer Perceptron fusion module for multimodal propagandistic memes classification (subtask 2C). Our system achieved 77.96% for subtask 2A, 71.04% for subtask 2B, and 79.80% for subtask 2C, ranking 2nd, 1st, and 3rd on the leaderboard.

1 Introduction

In the digital age, memes have gained immense popularity across various age groups through social media platforms such as Facebook and Twitter. In recent years, however, they have been increasingly employed for propagandistic purposes (G. De Leon and Ballesteros-Lintao, 2021; Kingdon, 2021). These memes often deploy a range of propagandistic techniques, including logical fallacies, causal oversimplification, and exaggeration, paired with images with strong positive/negative emotional implications to subtly sway audience opinions and propagate misinformation (Nieubuurt, 2021).

Given the multimodal nature and powerful impact of these memes, it is crucial to develop classification systems capable of detecting propagandistic memes effectively. Previous studies in propaganda detection primarily focused on textual content (Barrón-Cedeño et al., 2019; Alam et al., 2022; Hasanain et al., 2023, 2024a) or fine-grained persuasion techniques analysis (Dimitrov et al., 2024). To the best of our knowledge, the ArAIEval 2024 Shared Task (Hasanain et al., 2024b) is the first shared task on propaganda detection in multimodal scenarios, specifically within Arabic memes. The task is divided into three subtasks: subtasks 2A and 2B focus on classifying the textual and visual modalities of a given meme separately to detect whether it is propagandistic or not, whereas subtask 2C integrates both modalities for multimodal classification to detect whether a meme is propagandistic.

We carried out various experiments using state-of-the-art Arabic language processing models and vision models, and employed the Multilayer Perceptron (MLP) fusion module (Shi et al., 2021) with a prediction layer on top for multimodal classification. Without the need for text preprocessing and feature engineering, our system secured 2nd place in subtask 2A, 1st place in subtask 2B, and 3rd place in subtask 2C.

2 Data

The dataset used in the ArAIEval 2024 Shared Task comprises a collection of Arabic memes sourced from various social media platforms: Facebook, Twitter, Instagram, and Pinterest (Alam et al., 2024). Each meme is labeled as either 'propagandistic' or 'non-propagandistic' and includes both the original image file and its corresponding extracted textual content. The single dataset was used for the three subtasks covered in the competition.

The statistics of the dataset, in terms of the number of memes per class as well as the class distribution, are provided in Table 1. It can be observed that the dataset is imbalanced in terms of the represented classes, with the propagandistic content constituting the minority class with 28.14%, 28.21%, and 28.17% of the training, development, and test data, respectively.

Proceedings of The Second Arabic Natural Language Processing Conference, pages 501–506
August 16, 2024 ©2024 Association for Computational Linguistics

Table 1: Dataset statistics in terms of the number of memes per class (# Num) and class distribution (%).

Label	**Train**		**Dev**		**Test**	
	# Num	**%**	**# Num**	**%**	**# Num**	**%**
propaganda	603	28.14	88	28.21	171	28.17
not_propaganda	1,540	71.86	224	71.79	436	71.83
Total	**2,143**	**100**	**312**	**100**	**607**	**100**

3 Methodology

We conducted comprehensive experiments to evaluate the effectiveness of several state-of-the-art transformer models for Arabic language and image processing to address subtasks 2A and 2B, respectively. For subtask 2C, we used the Multilayer Perceptron (MLP) fusion module, followed by a prediction layer for multimodal classification. We did not apply preprocessing steps and fed the input representations provided with the dataset into the transformer models. All the models were fine-tuned on the training data and evaluated on the development set provided by the shared task organizers.

We carried out our experiments on the Google Colaboratory platform with an NVIDIA L4 GPU, utilizing the PyTorch framework and AutoGluon library (Shi et al., 2021). We set uniform hyperparameter settings for all the examined models: a base learning rate of 1e-4, decay rate of 0.9 using cosine decay scheduling, batch size of 8, maximum training epochs of 10, and optimization via the AdamW optimizer. Each model was trained for around 12 minutes, regardless of the subtask.

3.1 Unimodal experiments: textual modality

Our language model selection strategy was inspired by previous studies that showed that language-specific models pre-trained with larger vocabulary and bigger language-specific datasets usually outperform multilingual models such as mBERT (Virtanen et al., 2019). Consequently, our experiments focused on comparing several state-of-the-art Arabic language models:

- **MARBERT** (Abdul-Mageed et al., 2021): a transformer-based model pre-trained on a large corpus of 1 billion Arabic tweets, which predominantly includes modern standard Arabic and various dialectal forms, thus enhancing its language understanding capability across various Arabic dialects. The model was pre-trained using the same architectural framework as BERT (Devlin et al., 2019), focusing exclusively on the masked language modeling (MLM) objective but omitting the next sentence prediction (NSP) component. The large amount of tweets in the training data makes the model particularly suited for processing short social media texts.

- **CAMeLBERT** (Inoue et al., 2021): a collection of BERT models pre-trained on Arabic texts with different sizes and variants, including modern standard Arabic, dialectal Arabic, and classical Arabic. Recently, the authors proposed a new version called CAMeLBERT-MIX-SA, which is pre-trained on a mix of 167GB of texts in the aforementioned three Arabic variants and fine-tuned on the ASTD (Nabil et al., 2015), ArSAS (AbdelRahim Elmadany and Magdy, 2018), and SemEval datasets (Rosenthal et al., 2017). We used the CAMeLBERT-MIX-SA version in our experiments.

- **GigaBERT** (Lan et al., 2020): a customized bilingual BERT model designed for zero-shot transfer learning from English to Arabic. The model is pre-trained on a multilingual corpus that includes 6.1 billion tokens in English and 4.3 billion in Arabic, sourced from the Gigaword corpora (Parker et al., 2011), Wikipedia, and the OSCAR corpus (Suárez et al., 2019). The authors customized the vocabulary and augmented their data with code-switched samples to improve cross-lingual performance. These modifications enabled GigaBERT to achieve superior performance compared to other multilingual BERT-based models on various Arabic natural language processing tasks, such as named entity recognition and part-of-speech tagging.

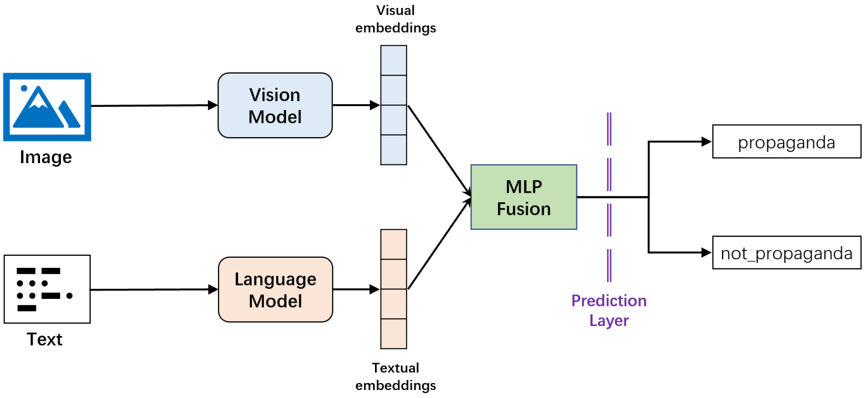

Figure 1: An overview of the multimodal classification system.

3.2 Unimodal experiments: visual modality

We examined two visual models for classifying the visual modality, i.e., meme's image only:

- **EVA** (Fang et al., 2023): a vision-centric foundation model designed to explore the limits of visual representation at scale using only publicly accessible data. It introduces masked image modeling (MIM) pre-training task that reconstructs masked-out image-text aligned vision features based on visible image patches. This enabled EVA to efficiently scale up to one billion parameters and achieve state-of-the-art performance across various downstream visual tasks (e.g., image and video classification, object detection) without extensive supervised training.

- **CAFormer** (Yu et al., 2024): a vision model that incorporates a mix of depthwise separable convolutions and vanilla self-attention mechanisms within a MetaFormer architecture, which allows the model to effectively manage computational complexity while capturing long-range dependencies. The model achieved a top-1 accuracy of 85.5% on ImageNet-1K (Russakovsky et al., 2015), demonstrating its effectiveness under supervised training conditions without the need for external data or distillation.

3.3 Multimodal experiments

We employed a multimodal architecture that integrates language and visual models, serving as text and image encoders to extract contextualized embeddings from textual and visual inputs. The resulting embeddings are concatenated using the Multilayer Perceptron (MLP) fusion module (Shi et al., 2021), where the top vector representations from the different models are combined into a single vector. A prediction layer is subsequently added to classify each instance into one of the predefined categories: propaganda or not_propaganda. Figure 1 illustrates the details of this multimodal architecture, which has proven effective for detecting other types of harmful multimodal content, such as multimodal hate speech (Wang and Markov, 2024b) and fine-grained types of hateful memes (Wang and Markov, 2024a).

4 Results

We report the results obtained on the development and test sets in terms of the official evaluation metric: macro-averaged F1 score.

The results for subtask 2A are provided in Table 2. The MARBERT model outperformed the other examined transformer models on the development set by 1–2 F1 points. On the test set, this model showed a drop of about 2.5 F1 points, achieving an F1 score of 77.96%, which is marginally (0.73 F1 points) lower than the best-performing system in this subtask.

Table 2: Results for subtask 2A on the dev and test sets.

Set	Language model	macro-F1
Dev	MARBERT	**80.48**
	CAMeLBERT	79.32
	GigaBERT	78.66
Test	MARBERT	77.96

The results for subtask 2B are shown in Table 3. It can be observed that the EVA model showed higher performance than CAFormer on the development set. We once again observe a drop in performance when the model is evaluated on the test set (ca. 6 F1 points). Nonetheless, the results obtained on the test set for this subtask are substantially higher than those of the second-runner, with a difference of 4.7 F1 points.

Table 3: Results for subtask 2B on the dev and test sets.

Set	Vision model	macro-F1
Dev	EVA	**76.89**
	CAFormer	75.23
Test	EVA	71.04

The performance comparison results for subtask 2C are detailed in Table 4. Although MARBERT was the best-performing language model in isolation, GigaBERT combined with EVA showed higher performance after multimodal fusion. We submitted the multimodal model that showed the best results on the development set for further evaluation on the test set and achieved an F1 score of 79.80, which is very close to the second place (79.87 F1 score) and only 0.71 F1 points away from the first place (80.51 F1 score).

Table 4: Results for subtask 2C on the dev and test sets.

Set	Multimodal model	macro-F1
Dev	GigaBERT + EVA	**82.60**
	MARBERT + EVA	81.59
	CAMeLBERT + EVA	80.79
	MARBERT + CAFormer	80.53
	GigaBERT + CAFormer	80.26
	CAMeLBERT + CAFormer	77.52
Test	GigaBERT + EVA	79.80

The obtained results indicate that the models trained on texts extracted from memes show higher performance than the models trained on the visual modality, implying the importance of textual modality for detecting propagandistic memes. When both modalities are combined, there is a further increase in performance by around 2 F1 points.

We provide confusion matrices for the submitted models on the test set for all three subtasks in Figures 2, 3, and 4. The multimodal model for subtask 2C exhibited the best performance and the lowest false positive and false negative rates in propagandistic memes classification. Conversely, the vision model for subtask 2B performs the worst among the three, with the lowest number of true positives (98) and the highest number of false negatives (73).

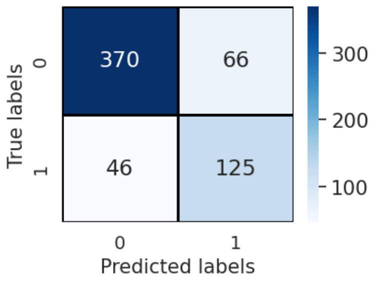

Figure 2: Confusion matrix for the MARBERT model on the test set for subtask 2A (0 = not_propaganda, 1= propaganda).

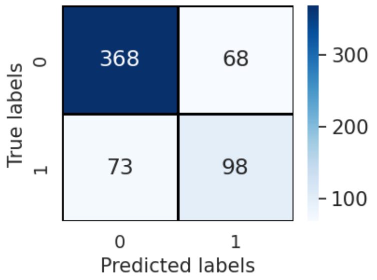

Figure 3: Confusion matrix for the EVA model on the test set for subtask 2B (0 = not_propaganda, 1= propaganda).

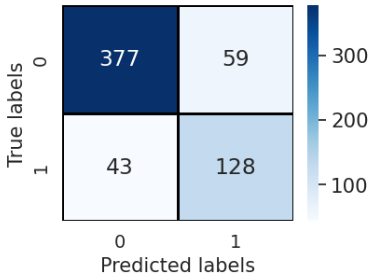

Figure 4: Confusion matrix for the multimodal model on the test set for subtask 2C (0 = not_propaganda, 1= propaganda).

5 Conclusion

In this paper, we presented the CLTL system developed for the ArAIEval 2024 Shared Task on multimodal propagandistic memes classification. Our approach involved leveraging state-of-the-art transformer models for both textual and visual modalities and employing the Multilayer Perceptron fusion module to combine text and image representations for multimodal classification. Our system secured 2nd place with a 77.96% macro-F1 score for subtask 2A, 1st place with 71.04% for subtask 2B, and 3rd place with 79.8% for subtask 2C.

References

Hamdy Mubarak AbdelRahim Elmadany and Walid Magdy. 2018. ArSAS: An Arabic speech-act and sentiment corpus of tweets. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA).

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. *arXiv preprint 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.

Alberto Barrón-Cedeño, Israa Jaradat, Giovanni Da San Martino, and Preslav Nakov. 2019. Proppy: Organizing the news based on their propagandistic content. *Information Processing & Management*, 56(5):1849–1864.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Dimitar Dimitrov, Firoj Alam, Maram Hasanain, Abul Hasnat, Fabrizio Silvestri, Preslav Nakov, and Giovanni Da San Martino. 2024. SemEval-2024 task 4: Multilingual detection of persuasion techniques in memes. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 2009–2026, Mexico City, Mexico. Association for Computational Linguistics.

Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. 2023. EVA: Exploring the limits of masked visual representation learning at scale. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19358–19369.

Faye Margarette G. De Leon and Rachelle Ballesteros-Lintao. 2021. The rise of meme culture: Internet political memes as tools for analysing philippine propaganda. *Journal of Critical Studies in Language and Literature*, 2(4):1–13.

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)*, pages 2724–2744, Torino, Italia. ELRA and ICCL.

Maram Hasanain, Firoj Alam, Hamdy Mubarak, Samir Abdaljalil, Wajdi Zaghouani, Preslav Nakov, Giovanni Da San Martino, and Abed 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)*, pages 483–493, 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.

Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2021. The interplay of variant, size, and task type in Arabic pre-trained language models. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 92–104, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Ashton Kingdon. 2021. *The Meme Is the Method: Examining the Power of the Image Within Extremist Propaganda*, pages 301–322. Springer International Publishing, Cham.

Wuwei Lan, Yang Chen, Wei Xu, and Alan Ritter. 2020. An empirical study of pre-trained transformers for Arabic information extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4727–4734, Online. Association for Computational Linguistics.

Mahmoud Nabil, Mohamed Aly, and Amir Atiya. 2015. ASTD: Arabic sentiment tweets dataset. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2515–2519, Lisbon, Portugal. Association for Computational Linguistics.

Joshua Troy Nieubuurt. 2021. Internet memes: Leaflet propaganda of the digital age. *Frontiers in Communication*, 5.

Robert Parker, David Graff, Ke Chen, Junbo Kong, and Kazuaki Maeda. 2011. Arabic gigaword fifth edition. *Philadelphia: Linguistic Data Consortium*.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. SemEval-2017 task 4: Sentiment analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 502–518, Vancouver, Canada. Association for Computational Linguistics.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252.

Xingjian Shi, Jonas Mueller, Nick Erickson, Mu Li, and Alex Smola. 2021. Multimodal AutoML on structured tables with text fields. In *8th ICML Workshop on Automated Machine Learning (AutoML)*.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures. In *7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7)*. Leibniz-Institut für Deutsche Sprache.

Antti Virtanen, Jenna Kanerva, Rami Ilo, Jouni Luoma, Juhani Luotolahti, Tapio Salakoski, Filip Ginter, and Sampo Pyysalo. 2019. Multilingual is not enough: BERT for Finnish. *CoRR*, abs/1912.07076.

Yeshan Wang and Ilia Markov. 2024a. CLTL at DIMEMEX shared task: Fine-grained detection of hate speech in memes. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2024), co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2024)*.

Yeshan Wang and Ilia Markov. 2024b. CLTL@Multimodal hate speech event detection 2024: The winning approach to detecting multimodal hate speech and its targets. In *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024)*, pages 73–78, St. Julians, Malta. Association for Computational Linguistics.

Weihao Yu, Chenyang Si, Pan Zhou, Mi Luo, Yichen Zhou, Jiashi Feng, Shuicheng Yan, and Xinchao Wang. 2024. Metaformer baselines for vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(2):896–912.