AlexUNLP-MZ at ArAIEval Shared Task: Contrastive Learning, LLM Features Extraction and Multi-Objective Optimization for Arabic Multi-Modal Meme Propaganda Detection

Mohamed Zaytoon, Nagwa ElMakky, Marwan Torki
Department of Computer and Systems Engineering
Alexandria University, Egypt
{mohamed.zaytoon24, nagwamakky, mtorki}@alexu.edu.eg

Abstract

The rise of memes as a tool for spreading propaganda presents a significant challenge in the current digital environment. In this paper, we outline our work for the ArAIEval Shared Task2 in ArabicNLP 2024. This study introduces a method for identifying propaganda in Arabic memes using a multimodal system that combines textual and visual indicators to enhance the result. Our approach achieves the first place in text classification with Macro-F1 of 78.69%, the third place in image classification with Macro-F1 of 65.92%, and the first place in multimodal classification with Macro-F1 of 80.51%.

1 Introduction

In today's digital era, memes have emerged as a potent instrument for disseminating propaganda. It is a method employed to influence individuals to adopt specific thoughts or behaviors. While this practice has existed for an extended period, social media has become a prevalent platform for its dissemination, especially considering its role as a primary news source. Those who engage in social media propaganda may have intentions to trick or control people.

Memes combining humor or visuals with specific messages, can bypass traditional filters and influence opinions surprisingly well (Nieubuurt, 2020) (Alam et al., 2024). Their effectiveness stems from their ability to be understood across languages and cultures. This makes them ideal for political groups, extremists, and even advertisers, who can leverage their virality to spread ideas or promote agendas. Memes can be persuasive, using humor, sarcasm, or fear to either reinforce existing beliefs or stir up conflict.

Spotting propaganda in memes is tough. Machine learning models need tons of knowledge across many areas (politics, history, etc.) to sort real content from propaganda.

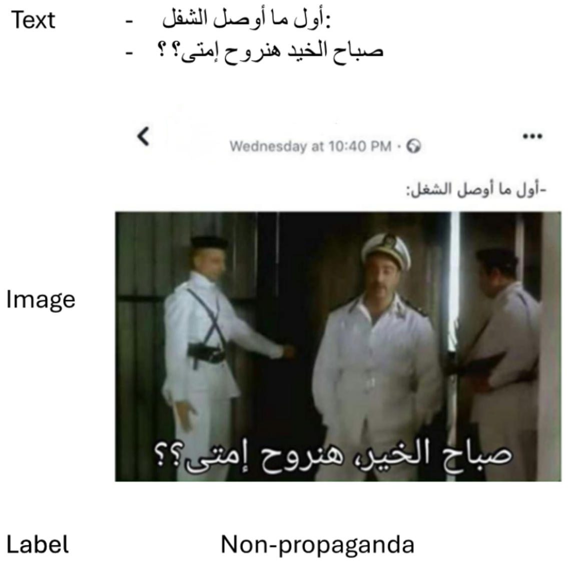

Figure 1: Sample from the dataset

Main Contributions

- We apply four objective functions, three objective for each modality and one contrastive objective to enhance the whole system.
- Using LLMs as a feature extractor to extract some common feature to improve our models learning

2 Related Work

Traditional methods for detecting propaganda have focused on analyzing textual content. Researchers have identified recurring linguistic features in propaganda (Hasanain et al., 2024a) (Hasanain et al., 2023a).

Although detecting propaganda in text is common, truly understanding bias in memes requires analyzing both text and image elements. A recent

Proceedings of The Second Arabic Natural Language Processing Conference, pages 512–517
August 16, 2024 ©2024 Association for Computational Linguistics

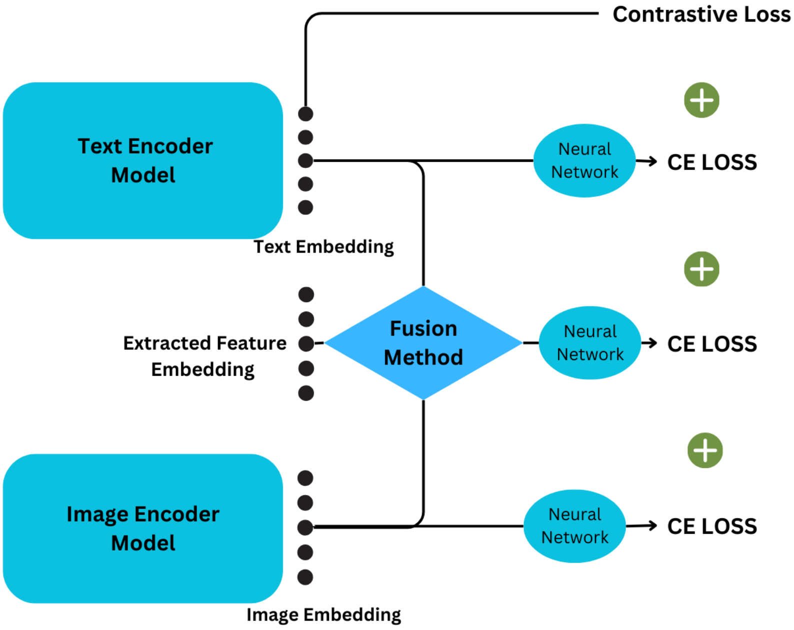

Figure 2: Our system for the multimodality model

study explored ways to label memes to identify political bias, showing how important it is to consider both the image, text (Dimitrov et al., 2021).

Furthermore, the survey in (Alam et al., 2022a) and Arabic shared tasks specifically for propaganda detection in (Alam et al., 2022b) , (Hasanain et al., 2023b) and (Hasanain et al., 2024b)) highlight the growing importance of addressing this issue in a multilmodal context.

3 Data

The data set for meme propaganda detection collected from different social media (e.g., Facebook, Twitter, Instagram and Pinterest). It comprises a collection of images, captions (mesme text content) and the corresponding labels as shown in Figure 1 stored in JSON file.

The data is split into training, validation, and testing sets. The statistics of the training and validation sets are given in Table 1. The testing set consists of 607 instances to check the robustness of the propaganda detection models developed.

Additionally, we processed our data by eliminating English words, stripping diacritics, and removing any HTML elements if present. Furthermore, we incorporated the extracted features from the Language Model Models (LLMs), as illustrated in Figure 3, into both our training and validation datasets in binary form (0 and 1).

4 System

Our system relies on two encoders components for the two modalities and a fusion module along with three objective functions incorporating contrastive loss, as depicted in Figure 2. We incorporate a range of techniques, with a strong focus on the multimodal approach, which yields the best performance.

4.1 Weighted Loss

Machine learning struggles with imbalanced data where one class has many more examples than others. This bias in training data leads to poorly performing models, especially for the minority class. Traditional methods don't handle this imbalance well.

Weighted cross entropy loss tackles imbalanced data by giving more weight to minority classes during training. This adjustment in the loss function improves model performance, especially for the under-represented classes, making it a valuable technique in various fields (Rezaei-Dastjerdehei et al., 2020) (Phan and Yamamoto, 2020).

4.2 Features Extraction

Large Language Models (LLMs) like ChatGPT and Gemini (Team et al., 2023) can handle tasks with little to no examples (zero-shot learning) thanks to their vast training data. One way to use LLMs

	Training set	Validation set
not propaganda	1540	224
propaganda	603	88
Total	2143	312

Table 1: Statistics of the Training and Validation sets

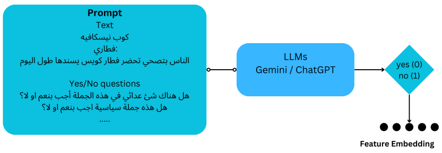
Prompt
Text
كوب نيسكافيه
فطاري:
الناس بتصحي تحضر فطار كويس يسندها طول اليوم
Yes/No questions
هل هناك شئ عدائي في هذه الجملة أجب بنعم او لا؟
هل هذه جملة سياسية اجب بنعم او لا؟
.....
LLMs
Gemini / ChatGPT
yes (0)
no (1)
Feature Embedding

Figure 3: Extracting extra features using LLMs (ChatGPT / Gemini) to embed with our models

Model	Method	macro-F1
	w/o EF+TIL+CL	0.799
bloomz-1b1	w/o EF	0.800
+	w/o TIL	0.808
resnext101_32x8d	w/o CL	0.814
	w EF+TIL+CL	**0.825**

Table 2: A comparison for multi-modality bloomz-1b1 + resenext101_32x8d model without using any criteria, without extracted features (EF), without text and image losses (TIL), without contrastive loss, and with all the criteria together.

is to extract features in text. We can use their understanding of language to identify features useful for NLP tasks.The LLM is given text data and asked yes/no questions about specific details (e.g., whether a given statement is offensive, political, etc.) as shown in Figure 3. These features are based on the model's embeddings, which are numerical representations of words capturing meaning and relationships within the text.

4.3 Contrastive Loss

Contrastive loss train models are created by comparing data points, such as grouping similar ones and separating dissimilar ones (Wang and Liu, 2021). This helps the model understand the key differences between data points. It offers advantages over traditional classification losses (Shapiro et al., 2022). Specifically, we used NT-Xent contrastive loss (Sohn, 2016) from the PyTorch Metric Learning library (Musgrave et al., 2020) to learn informative representations from the data .

4.4 Models

To understand Arabic text and images, we applied transfer learning, a technique that reuses knowledge from pre-trained models for new tasks. This increases performance compared to training from scratch (Ibrahim et al., 2020).

4.4.1 Uni-Modality Models

The dataset provided contains both image and text data, offering a unique opportunity to move beyond the training of individual unimodality models for text and image classification.

Text Classification We build text encoder models using pre-trained models from Huggingface. For Arabic dialects, first we use bert-base-arabic-camelbert-da trained on various dialects (Inoue et al., 2021) and bert-base-arabertv02-twitter specifically for tweets (Antoun et al., 2020).

Then, instead of Arabic dialects pre-trained models, we used BLOOM and BLOOMZ(Muennighoff et al., 2022) from Big Science. These handle a broader range of tasks due to their training on huge, general-domain datasets. Here, text is classified based on hidden-state embeddings from an LLM fed into a neural network. To enhance training, the model combines contrastive loss with LLM-extracted features, resulting in a stronger and more informative text classification system

Imaged Classification Unlike typical image tasks, meme propaganda requires understanding both the image and its caption, making it a challenge for standard image classification. In short, image and text work together in memes, creating a complex classification problem.

For this challenge, we used pre-trained CNN architectures from PyTorch library (maintainers and contributors, 2016) the used architectures like ResNet, ResNeXt, and DenseNet. For this task we apply weighted cross entropy loss.

4.4.2 Multi-Modality Models

Our system analyzes both text and images in memes (multimodal) using separate branches for each data type and LLMs for feature extraction. Finally, we combine these elements using a fusion module. We tried different fusion techniques e.g. concatenation, bilinear fusion, gated, and attention fusion to create a single, comprehensive representation.

We use different objective functions to effectively train our model. This includes separate weighted cross-entropy losses for text, image, and combined data, plus a contrastive function that im-

Type	Model Name	macro-F1
Text Classification	bert-base-arabertv02-twitter	0.780
	bert-base-arabic-camelbert-da	0.761
	bloom-560m	0.791
	bloom-1b1	0.789
	bloomz-560m	0.769
	bloomz-1b1	**0.801**
	bloomz-3b1	0.764
Image Classification	densenet201	0.671
	resnet18	0.728
	resnet101	0.723
	resnext101_32x8d	**0.742**
	resnext101_64x4d	0.719
Multi-Modal Classification	bloomz-1b1 + resnext101_32x8d (gated fusion)	0.800
	bloomz-1b1 + resnext101_32x8d (bilinear fusion)	0.793
	bloomz-1b1 + resnext101_32x8d (concatenation fusion)	0.825
	bloomz-1b1 + resnext101_32x8d (attention fusion)	0.824
	bloomz-1b1 + resnet101 (concatenation fusion)	**0.831**

Table 3: The models performance on macro-f1 score across the validation dataset.

proves the models understanding of text information. Our network utilizes separate encoders for both text and image modalities. These encoders extract high-level features that capture the semantic content within each data type. Subsequently, the extracted features are combined together as shown on Figure 2. This combo helps the model learn well from both text and images, boosting its overall performance.

4.5 Training

In the training process, across the three models we trained with varying batch sizes of 128, 64, and 32 on a V100 GPU. Using the Adam optimizer with learning rates of 0.01, 0.001, and 0.0001, and weight decay values of 0.01 and 0.001. Additionally, we tried to fine-tune different layers and freezing parts of the model.

5 Results

In our submission to the official test set, we used Bloomz-1b1 for text classification which achieves the first place with a Macro-F1 of 78.69%, ResNeXt101_32x8d, for image classification which achieves the third place with a Macro-F1 of 65.92%, and a combination of these models with attention fusion for the multimodal task and we achieved the first place with a Macro-F1 of 80.51%.

To assess the impact of our system design on the Bloomz-1b1 model with a ResNeXt-101_32x backbones, we compared the performance of three training approaches on the validation set using extracted features, contrastive loss, and employing both text and image losses. Our findings demonstrate that the best results are achieved by combining all three methods as shown in Table 2.

6 Discussion

Table 3 shows the validation set results which indicate that Bloomz-1b1 performed best for text data, while ResNeXt101_32x8d achieved the best results for image data. For multimodal tasks, the best performing model combined Bloomz-1b1 for text with Resnet101 using concatenation fusion.

Our system that analyzes both text and images (multi-modal) significantly beats analyzing just text or images alone (uni-modal). This shows the value of combining different data types for better understanding. Features extracted from large language models (LLMs) are key factors, as they provide a strong foundation for the model understanding of the data. In future work, we aim to leverage ensemble methods by combining our most effective models. We will also explore incorporating specialized propaganda detection models and leverage the power of large vision language models for multimodal analysis. By extracting richer features from

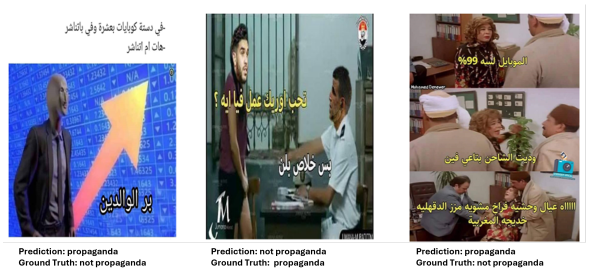

Figure 4: Failure examples from the multi-modality model.

the data we anticipate significant improvements in propaganda detection accuracy.

Figure 4 reveals some weaknesses in the model for predicting the correct class. It demonstrates the limitations of the current model ability to accurately identify public figures within images, also it indicates that mentioning countries or names may mislead the model.

7 Conclusion

This research investigated the role of memes as propaganda tools within the ever-expanding landscape of social media. We developed a system capable of identifying propaganda in memes by analyzing their multimodal elements text, image, and the combination of both. Our findings highlight the critical role of textual content in propaganda detection within memes. Analyzing the language proved to be a key factor, as words often carry more explicit messaging compared to visuals alone.

Furthermore, the system demonstrated the effectiveness of utilizing multiple objective optimization techniques. This, combined with the inclusion of strategically extracted features, enhanced the models overall accuracy in identifying propagandistic memes.

References

Firoj Alam, Stefano Cresci, Tanmoy Chakraborty, Fabrizio Silvestri, Dimiter Dimitrov, Giovanni Da San Martino, Shaden Shaar, Hamed Firooz, and Preslav Nakov. 2022a. A survey on multimodal disinformation detection. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6625–6643, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

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. 2022b. 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. 2020. AraBERT: Transformer-based model for Arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15, Marseille, France. European Language Resource Association.

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*.

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.

Mai Ibrahim, Marwan Torki, and Nagwa El-Makky. 2020. AlexU-BackTranslation-TL at SemEval-2020 task 12: Improving offensive language detection using data augmentation and transfer learning. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1881–1890, Barcelona (online). International Committee 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*, Kyiv, Ukraine (Online). Association for Computational Linguistics.

TorchVision maintainers and contributors. 2016. Torchvision: Pytorch's computer vision library. https://github.com/pytorch/vision.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.

Kevin Musgrave, Serge J. Belongie, and Ser-Nam Lim. 2020. Pytorch metric learning. *ArXiv*, abs/2008.09164.

Joris T. Nieubuurt. 2020. Internet memes: Leaflet propaganda of the digital age. *Frontiers in Psychology*, 11:1709.

Trong Huy Phan and Kazuma Yamamoto. 2020. Resolving class imbalance in object detection with weighted cross entropy losses. *arXiv preprint arXiv:2006.01413*.

Mohammad Reza Rezaei-Dastjerdehei, Amirmohammad Mijani, and Emad Fatemizadeh. 2020. Addressing imbalance in multi-label classification using weighted cross entropy loss function. In *2020 27th National and 5th International Iranian Conference on Biomedical Engineering (ICBME)*, pages 333–338. IEEE.

Ahmad Shapiro, Ayman Khalafallah, and Marwan Torki. 2022. AlexU-AIC at Arabic hate speech 2022: Contrast to classify. In *Proceedinsg of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools with Shared Tasks on Qur'an QA and Fine-Grained Hate Speech Detection*, pages 200–208, Marseille, France. European Language Resources Association.

Kihyuk Sohn. 2016. Improved deep metric learning with multi-class n-pair loss objective. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Feng Wang and Huaping Liu. 2021. Understanding the behaviour of contrastive loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2495–2504.