# AraMinds at MAHED 2025: Leveraging Vision-Language Models and Contrastive Multi-task Learning for Multimodal Hate Speech Detection

**Mohamed Zaytoon, Ahmed Salem, Ahmed Sakr, Hossam Elkordi**
Department of Computer and Systems Engineering
Alexandria University, Egypt
{mohamed.zaytoon24,es-AhmedMahmod2022,
es-ahmedsakr20,es-hossam.elkordi2018}@alexu.edu.eg

## Abstract

Detecting hate speech in social media content is essential to provide a safe space for people to connect. Memes have been used lately to sarcastically express one's opinion, and they can be used to hide harmful intentions and spread hateful speech. In this work, we build our system that detects hateful speech in memes by combining visual and textual features and merging them using different techniques to detect the inherent meaning and overcome the challenge of vast dialectal differences and the variety of topics discussed. To improve our system's robustness, we combine different techniques, such as multi-tasking, contrastive learning, and vision language modeling in a final ensemble model that secured us the third place in the MAHED 2025 shared-task leaderboard with a macro-f1 score of 0.74, showing strong performance on the evaluation set.

## 1 Introduction

The social media content of the Arabic-speaking world is a complex footprint of social and political expression due to the diverse topics discussed on it, the different points of view introduced, and the different narratives they are presented in. People tend to reflect their hopeful and hateful sentiments on social media platforms, projecting them in different formats of content, such as memes, videos, and textual blog posts (Al-Saqqa et al., 2024; Mulki et al., 2019). The increase of the meme culture over the last few years provided an abundance of multimodal data that introduced nuanced techniques to hide complex and harmful messages through humour and irony (Kiela et al., 2020; Alam et al., 2024a). This necessitates the need for a means of automatic detection of such hateful content to enable safer online platforms for people to express their opinion (Chen and Pan, 2022).

We address the need for robust hateful digital content detection by focusing on the multimodal of Arabic memes (Arya et al., 2024). Such systems must have the capacity to analyze both the visual and textual components of a meme to produce a binary classification of "hate" or "no-hate", thereby mitigating the spread of harmful online content.

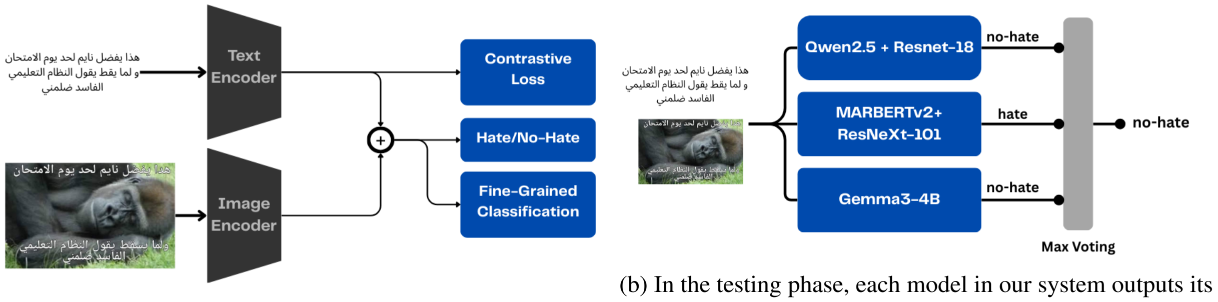
One of the challenges in this task can be the linguistic diversity of Arabic, including vast dialectal variations (Habash, 2010) and cultural expressions with double meanings that can obscure intent (Elkordi et al., 2024). Additional hurdles include the wide range of viral social and political topics, such as politics, religion, and gender, and data-specific issues like the scarcity of clean annotations and a significant imbalance where non-hateful memes are more common (Mulki et al., 2019).

To overcome these obstacles, our system must handle the complexities of the Arabic language by analyzing the interaction between visual and textual features to uncover the actual intent behind sarcastic content. A key requirement is the ability to generalize from limited data across diverse topics and expressions, enabling the model to differentiate between benign cultural commentary and genuinely hateful projections.

Our main contributions can be summarized in the following points:

- Applied multi-tasking technique to benefit from fine-grained classed and contrastive learning to extract meaningful features that can group samples of the same class closer in the embedding space.
- Used a pretrained vision language model in our classification task to benefit from it generalization and multilingual abilities.
- Combined the different techniques we used in a maximum voting ensemble that is robust in multimodal hate speech detection and secured the third place in the shared task leaderboard (Zaghouani et al., 2025).

*Proceedings of The Third Arabic Natural Language Processing Conference*, pages 626–631
November 8-9, 2025 ©2025 Association for Computational Linguistics



(a) Our dual-encoder training setup in the multi-tasking setup. Both input modalities are encoded separately, then the embeddings are concatenated, and the binary cross-entropy loss is applied on the merged embedding and the text embedding while the contrastive loss is applied only on the textual embedding.

(b) In the testing phase, each model in our system outputs its prediction, then a maximum voting is applied to produce the final prediction. Our ensemble benefit from the different encoding and merging approaches we used with both modalities, including VLMs and encoder/decoder-only models for text embedding.

Figure 1: Illustration of our training in the multi-tasking setup and the maximum voting ensemble in the testing phase

## 2 Background

Prior research has explored hate speech detection through different classical and deep learning techniques in both unimodal and multimodal settings. The main focus was on the textual content only (Chhabra and Vishwakarma, 2023), relying on classical techniques such as Bag-of-Words (Husain and Uzuner, 2022), TF-IDF (Kumar and Varalakshmi, 2021), Word Embedding e.g. Word2Vec, GloVe, and FastText (Plaza-del Arco et al., 2021), and hybrid methods that combine CNN and GRU or integrate attention mechanisms for improved performance (Zhang et al., 2018).

Then, the focus is switched to deep learning techniques that rely on contextual representation using recurrent networks such as RNNs and LSTMs or BERT-based models (Devlin et al., 2019) that rely on the self-attention technique. More recent work has used both images and text for better contextual representation and accurate results (Kiela et al., 2020). In this setup, multiple techniques have been explored, such as early fusion, late fusion (Lippe et al., 2020), and pre-trained vision language models (Chen and Pan, 2022).

Considering the Arabic language, this is the first use of multimodal memes for hateful speech detection. A previous task explored the use of such a setup for propaganda detection from memes (Hasanain et al., 2024). Participants of this task explored different techniques to integrate visual and textual features to produce the final prediction, such as using multi-agent LLMs to detect the propaganda (Alam et al., 2024a) or using contrastive learning with a multi-objective function (Zaytoon et al., 2024).

## 3 System Overview

In this section, we present different components of our system. First, we show the backbones and the fusion technique we used in a dual-encoder architecture. Then, we present how we benefited from the instruction capabilities of pretrained vision language models (VLMs) (Bordes et al., 2024). Then, we showcase how we improved the classification performance using a multi-task approach. Finally, we present our system as an ensemble of all the above components.

### 3.1 Dual-Encoders

In this component, we employed a separate encoder for each modality. For the text modality, we relied on pretrained language models and tested two different approaches. First, we used an encoder-only model, `MARBERTv2` (Abdul-Mageed et al., 2021), which is known for its robustness against dialectal Arabic. We used a version of it that is trained for hate speech detection in the Egyptian dialect (Ahmed et al., 2022). Second, we used a pretrained decoder-only LLM, `Qwen2.5-1.5B` (Team, 2024) to provide a more general representation of the textual input. For the image modality, we used convolutional neural network backbones, specifically `ResNet-101` (He et al., 2016) and `ResNeXt-101` (Xie et al., 2017) models, to capture both global and fine-grained visual features.

After the extraction of both visual and textual features, they are concatenated to form a single multimodal representation. Next, we apply binary cross-entropy on this final representation and the textual embedding, along with a contrastive loss on the textual features only, using in-batch sampling

| Split | Number of Samples |
|---|---|
| Total Training | 2,452 |
| Validation | 606 |
| Test | 500 |

Table 1: Number of samples in each split of the dataset.

to select positive and negative samples as shown in 1.

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{BCE}}^{\text{fusion}} + \mathcal{L}_{\text{BCE}}^{\text{text}} + \mathcal{L}_{\text{CL}}^{\text{text}} \tag{1}$$

### 3.2 Vision Language Models (VLMs)

In this component, we explored the instruction-following capabilities of pretrained VLMs and applied supervised fine-tuning to the recent `Gemma3-4B` model (Team et al., 2025) on our task. The model is provided with an instruction that explains the task and both modalities, text and image. The model is trained in the next token prediction setup, and the cross-entropy loss is applied on the model's output tokens only that include either **"hate"** or **"no-hate"** only.

### 3.3 Multi-Task

The shared task dataset had another fine-grained classification for each sample, including nine distinct categories. We observed a strong correlation between those categories and the original binary classes. Hence, we decided to benefit from this correlation as an additional supervision by employing a multi-classification objective using our dual-encoder component with two classification heads instead of one.

### 3.4 Ensemble

Finally, we combined all three components in a maximum voting ensemble as shown in figure 1b, where each of the three models is treated equally and the class with the highest vote is picked as our final classification.

## 4 Experimental Setup

### 4.1 Dataset

The dataset is collected from different social media platforms such as Facebook, Instagram, and Pinterest. Then, the dataset went through multiple filteration stages filtered including de-duplication, text extraction, and memes identification. Then, the dataset is randomly sampled from the original 6k samples (Alam et al., 2024b) and gets annotated.

| Method | Multi-Tasking | macro-F1 |
|---|---|---|
| `Qwen2.5-1.5B + ResNet-18` | ✗ | 0.689 |
| `Qwen2.5-1.5B + ResNet-18` | ✓ | 0.702 |
| `Qwen2.5-0.5B + ResNet-18` | ✓ | 0.663 |
| `Qwen2.5-1.5B + ResNeXt-101` | ✓ | 0.699 |
| `MARBERTv2 + ResNet-18` | ✓ | 0.705 |
| `MARBERTv2 + ResNeXt-101` | ✓ | 0.703 |

Table 2: Macro-F1 results on the validation split using the dual-encoder approach for comparison between single and multi-tasking, as well as the size of both textual and visual encoders and the architecture of the textual encoder.

Table 1 indicates the distribution of the provided dataset. It includes 2,452 samples for training, 606 samples for validation, and 500 samples for final evaluation of the model's performance.

### 4.2 Training Setup

For the dual-encoders components, we trained the model for 150 epochs with a learning rate of 0.001 and used AdamW optimizer (Loshchilov and Hutter, 2017). We used different batch sizes based on the backbone sizes. We used a batch size of 2 for `Qwen2.5-0.5B` and 16 for `Qwen2.5-1.5B` and `MARBERT`. For the `Gemma3-4B`, we trained the model for 10 epochs with a learning rate of 5e-6 with a cosine scheduler and a batch size of 2. We evaluated our system during training on the validation set using the macro-F1 score, which treats each class equally. All training was done on a single NVIDIA RTX-3090 GPU.

## 5 Results

In this section, we present a detailed overview of our experiments. All results are reported on the validation set using the macro-F1 score, and finally, we report our test set results on the leaderboard.

### 5.1 Effect of Multi-Tasking

We conducted our first experiment to assess the effect of the fine-grained categories on the main classification task. We used `Qwen2.5-1.5B` and `ResNet-18` for this experiment. We can see in the first two rows of table 2 that the multi-tasking improved the classification performance by 0.013.

### 5.2 Size of the Text Encoder

We tested the effect of the text encoder size without changing the image encoder. This experiment was done using the multi-tasking setup. Results

| Method | Contrastive Embedding | macro-F1 |
|---|---|---|
| `Qwen2.5-1.5B + ResNet-18` | text | 0.702 |
| `Qwen2.5-1.5B + ResNet-18` | fused | 0.666 |
| `Qwen2.5-1.5B + ResNet-18` | text + fused | 0.687 |
| `VLM - Gemma3-4B` | - | 0.692 |

Table 3: Comparison between the dual-encoder approach and fine-tuning a pre-trained VLM approach. In the dual-encoder approach, different embeddings were used for the contrastive objective during training.

show that the size increase improved our system performance by 0.039.

### 5.3 Size of the Image Encoder

Also, we tested the effect of increasing the image encoder size. In this experiment, we compared `ResNet-18` and `ResNeXt-101`. Our results show that the size of the image encoder didn't have a huge improvement on the classification performance.

### 5.4 Encoder vs. Decoder Models

We tested changing the text encoder architecture and tried using a bi-directional encoder. We used `MARBERTv2` model, and tested it with both `ResNet-18` and `ResNeXt-101`. Using an encoder-only model improved the performance when using different sizes of the image encoder. Also, increasing the size of the image encoder doesn't improve the model's performance.

### 5.5 Dual-Encoders vs. VLM

Finally, we compared fine-tuning a pre-trained vision language model with the dual-encoder setup with the multi-tasking objective. In the dual-encoder setup, we tested the application of the contrastive objective on different feature vectors: the text embeddings, the image embeddings, and the fused embeddings. We can see that applying the contrastive loss on the text embedding only was the best performing. Also, fine-tuning VLM had very close results and was better than other dual-encoder setups.

### 5.6 Test Set Results

In this section, we report the results of different systems we submitted and the final maximum voting ensemble we made from them in table 4. We chose the submitted models based on their experimental results shown in tables 2 and 3. Our max-voting ensemble achieved a macro-f1 score of 0.74 and secured our third place in the leaderboard.

| Method | macro-F1 |
|---|---|
| `Qwen2.5-1.5B + ResNet-18` | 0.71 |
| `MARBERTv2 + ResNeXt-101` | 0.72 |
| `VLM - Gemma3-4B` | 0.72 |
| Ensemble | **0.74** |

Table 4: Macro-F1 scores on the test set submitted on the shared task leaderboard.

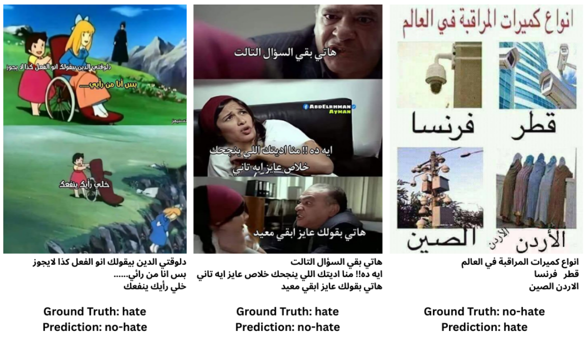


Figure 2: Examples of failure cases from our system.

### 5.7 Qualitative and Error Analysis

Figure 2 shows cases that our system failed to correctly predict. In the first sample, our system prediction was misguided by the cartoonish scene and failed to identify its hateful stance when discussing religious opinions. In the second image, the obviously sarcastic text over-shadowed the hateful and offensive scene displayed in the image. In the final image, our system identifies the sample as hateful due to its mocking and stereotyping nature.

## 6 Conclusion

This paper investigated our work in subtask-3 of the MAHED 2025 shared task for hate speech detection in memes. We used contrastive learning and multi-tasking techniques in our dual-enconder component with late embedding fusion. We also fine-tuned a vision language model to benefit from its instruction-following, multi-lingual, and generalization capabilities in the classification task that covers multiple Arabic dialects and different topics. Lastly, we combine different models we built in a robust maximum voting ensemble that secured us the third place in the competition leaderboard with macro-f1 score of 0.74.

## 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*. Association for Computational Linguistics.

Ibrahim Ahmed, Mostafa Abbas, Rany Hatem, Andrew Ihab, and Mohamed Waleed Fahkr. 2022. Fine-tuning arabic pre-trained transformer models for egyptian-arabic dialect offensive language and hate speech detection and classification. In *2022 20th International Conference on Language Engineering (ESOLEC)*.

Samar Al-Saqqa, Arafat Awajan, and Bassam Hammo. 2024. A survey of hate speech detection for arabic social media: Methods and datasets. *Procedia Computer Science*. 15th International Conference on Emerging Ubiquitous Systems and Pervasive Networks / 14th International Conference on Current and Future Trends of Information and Communication Technologies in Healthcare EUSPN/ICTH 2024.

Firoj Alam, Md Rafiul Biswas, Uzair Shah, Wajdi Zaghouani, and Georgios Mikros. 2024a. Propaganda to hate: A multimodal analysis of arabic memes with multi-agent llms. In *International Conference on Web Information Systems Engineering*. Springer.

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

Greeshma Arya, Mohammad Kamrul Hasan, Ashish Bagwari, Nurhizam Safie, Shayla Islam, Fatima Rayan Awad Ahmed, Aaishani De, Muhammad Attique Khan, and Taher M Ghazal. 2024. Multimodal hate speech detection in memes using contrastive language-image pre-training. *IEEe Access*.

Florian Bordes, Richard Yuanzhe Pang, Anurag Ajay, Alexander C. Li, Adrien Bardes, Suzanne Petryk, Oscar Mañas, Zhiqiu Lin, Anas Mahmoud, Bargav Jayaraman, Mark Ibrahim, Melissa Hall, Yunyang Xiong, Jonathan Lebensold, Candace Ross, Srihari Jayakumar, Chuan Guo, Diane Bouchacourt, Haider Al-Tahan, and 22 others. 2024. An introduction to vision-language modeling. *ArXiv*.

Yuyang Chen and Feng Pan. 2022. Multimodal detection of hateful memes by applying a vision-language pre-training model. *Plos one*.

Anusha Chhabra and Dinesh Kumar Vishwakarma. 2023. A literature survey on multimodal and multilingual automatic hate speech identification. *Multimedia Systems*.

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*.

Hossam Elkordi, Ahmed Sakr, Marwan Torki, and Nagwa El-Makky. 2024. AlexuNLP24 at AraFinNLP2024: Multi-dialect Arabic intent detection with contrastive learning in banking domain. In *Proceedings of the Second Arabic Natural Language Processing Conference*, Bangkok, Thailand. Association for Computational Linguistics.

Nizar Y Habash. 2010. *Introduction to Arabic natural language processing*. Morgan & Claypool Publishers.

Maram Hasanain, Md Arid Hasan, Fatema Ahmed, Reem Suwaileh, Md Rafiul Biswas, Wajdi Zaghouani, and Firoj Alam. 2024. Araieval shared task: propagandistic techniques detection in unimodal and multimodal arabic content. *arXiv preprint arXiv:2407.04247*.

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*.

Fatemah Husain and Ozlem Uzuner. 2022. Investigating the effect of preprocessing arabic text on offensive language and hate speech detection. *Transactions on Asian and Low-Resource Language Information Processing*.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in neural information processing systems*.

PM Ashok Kumar and K Varalakshmi. 2021. Hate speech detection using text and image tweets based on bi-directional long short-term memory. In *2021 International conference on disruptive technologies for multi-disciplinary research and applications (CENTCON)*. IEEE.

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*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Hala Mulki, Hatem Haddad, Chedi Bechikh Ali, and Halima Alshabani. 2019. L-hsab: A levantine twitter dataset for hate speech and abusive language. In *Proceedings of the third workshop on abusive language online*.

Flor Miriam Plaza-del Arco, M Dolores Molina-González, L Alfonso Urena-López, and M Teresa Martín-Valdivia. 2021. Comparing pre-trained language models for spanish hate speech detection. *Expert Systems with Applications*.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Qwen Team. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. 2017. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Wajdi Zaghouani, Md Rafiul Biswas, Mabrouka Bessghaier, Shimaa Ibrahim, Georgio Mikros, Abul Hasnat, and Firoj Alam. 2025. Overview of mahed shared task: Multimodal detection of hope and hate emotions in arabic content. In *The Third Arabic Natural Language Processing Conference (ArabicNLP 2025)*. Association for Computational Linguistics.

Mohamed Zaytoon, Nagwa M El-Makky, and Marwan Torki. 2024. Alexunlp-mz at araieval shared task: contrastive learning, llm features extraction and multi-objective optimization for arabic multi-modal meme propaganda detection. In *Proceedings of The Second Arabic Natural Language Processing Conference*.

Ziqi Zhang, David Robinson, and Jonathan Tepper. 2018. Detecting hate speech on twitter using a convolution-gru based deep neural network. In *European semantic web conference*. Springer.