NYUAD at MAHED Shared Task: Detecting Hope, Hate, and Emotion in Arabic Textual Speech and Multi-modal Memes Using Large Language Models

Nouar AlDahoul
Computer Science Department
New York University
Abu Dhabi, UAE
nouar.aldahoul@nyu.edu

Yasir Zaki
Computer Science Department
New York University
Abu Dhabi, UAE
yasir.zaki@nyu.edu

Abstract

The rise of social media and online communication platforms has led to the spread of Arabic textual posts and memes as a key form of digital expression. While these contents can be humorous and informative, they are also increasingly being used to spread offensive language and hate speech. Consequently, there is a growing demand for precise analysis of content in Arabic text and meme. This paper explores the potential of large language models to effectively identify hope, hate speech, offensive language, and emotional expressions within such content. We evaluate the performance of base LLMs, fine-tuned LLMs, and pre-trained embedding models . The evaluation is conducted using a dataset of Arabic textual speech and memes proposed in the ArabicNLP MAHED 2025 challenge. The results underscore the capacity of LLMs such as GPT-4o-mini, fine-tuned with Arabic textual speech, and Gemini Flash 2.5, fine-tuned with Arabic memes, to deliver the superior performance. They achieve up to 72.1%, 57.8%, and 79.6% macro F1 scores for task 1, 2, and 3, respectively and secure first place overall in the challenge[1] (Zaghouani et al., 2025). The proposed solutions offer a more nuanced understanding of both text and memes for accurate and efficient Arabic content moderation systems.

1 Introduction

AI content moderation refers to the use of artificial intelligence to monitor, evaluate, and manage content across digital platforms[2]. By ensuring that posts comply with community standards and legal regulations, it helps create safer, more respectful, and law-abiding online environments. Its role has become increasingly vital as the volume and complexity of online content continue to grow. Despite growing efforts, Arabic content moderation still lags behind. Challenges such as dialect diversity, limited training data, and under-resourced tools make it difficult to ensure effective moderation across Arabic-speaking regions[3],[4].

Although Arabic is spoken by around 380 million people, it is far from being a uniform language[5]. It consists of six major regional dialect groups, so for classifiers to work effectively, they must be trained across all these dialects. The rise of social media and online communication platforms has led to the spread of Arabic textual posts and memes as a key form of digital expression. There is a growing need to develop methods for detecting hateful text and memes, as they can perpetuate harmful stereotypes and contribute to the spread of offensive language and hate speech in digital spaces (Zaghouani et al., 2024; Zaghouani and Biswas, 2025a; AlDahoul et al., 2024a).

To have a full understanding of the emotional landscape of online communication, recognition of emotional expression can provide deeper insight into user sentiment and foster empathy. Additionally, emotional expression classification has valuable applications such as monitoring mental health and tailoring personalized recommendations (Zaghouani and Biswas, 2025b).

Memes are especially widespread and can be potent tools for spreading propaganda, inciting hate, or conveying humor. LLMs have been shown to have superior performance in various domains and applications (AlDahoul et al., 2025, 2024b). For meme understanding, having textual and visual inputs, LLMs can analyze both the linguistic content and the underlying visual elements of a meme.

[1] https://marsadlab.github.io/mahed2025/#

[2] https://verpex.com/blog/website-tips/ai-content-moderation

[3] https://techglobalinstitute.com/announcements/blog/content-moderation-arabic-hebrew-in-under-resourced-regions/

[4] https://www.mei.edu/publications/content-moderation-trends-mena-region-censorship-discrimination-design-and-linguistic

[5] https://techglobalinstitute.com/announcements/blog/content-moderation-arabic-hebrew-in-under-resourced-regions/

Proceedings of The Third Arabic Natural Language Processing Conference, pages 575–584

Our analyses and experiments center around the following research questions: **RQ1**: Can a pre-trained embedding model, combined with trained SVM or DNN classifiers, effectively detect hate and hope speech in Arabic text and memes? **RQ2**: Are existing safety classification and content moderation solutions capable of detecting hate speech in Arabic memes? **RQ3**: To what extent do state-of-the-art base LLMs excel in detecting hate speech in Arabic memes? **RQ4**: Can fine-tuned LLMs detect emotion, hope, hate, and offensive content in Arabic text with high accuracy? **RQ5**: Can fine-tuned LLMs detect hateful Arabic memes with high accuracy?

2 Related Work

Several studies have investigated hate speech and offensive language in Arabic text (Mohaouchane et al., 2019; Kaddoura et al., 2023; Mubarak et al., 2023; Shapiro et al., 2022; Albadi et al., 2018; Bennessir et al., 2022). They utilized Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM), CNN-LSTM (Mohaouchane et al., 2019; Kaddoura et al., 2023), word embeddings with simple Recurrent Neural Networks (RNN) (Albadi et al., 2018) and MARBERT (Shapiro et al., 2022; Bennessir et al., 2022). The datasets used for analysis contain social posts and tweets.

To study the proportion of hate speech and offensive language in Arabic tweets, AraBERT was utilized (Zaghouani et al., 2024). They found that 15% of tweets contained offensive language, while 6% included hate speech. Additionally, their annotated tweet dataset provided a valuable contribution to the limited availability of Arabic data related to hate speech and offensive language (Zaghouani et al., 2024). It was found that AraBERT outperformed conventional machine learning classifiers (Zaghouani et al., 2024).

Even though there are several English emotion datasets, there is still a shortage of comprehensive Arabic datasets that support the analysis of both emotions and hope speech. (Zaghouani and Biswas, 2025b) proposed an Arabic dataset, fostering better cross-linguistic analysis of emotions and hope speech. They fine-tuned the AraBERT model (Antoun et al., 2020) for the hate-hope classification task.

Building on previous research, numerous studies broadened the scope to tackle the challenge of detecting Arabic content across multiple modalities. In the context of Arabic propaganda identification (Alam et al., 2024b; Hasanain et al., 2024), separate feature extractors were employed for text and images. Moving from propaganda to hate, a multi-modal analysis of Arabic memes was done to further detect hate in memes. They used a fusion of features extracted from AraBERT for text and ConvNxT for images (Alam et al., 2024a).

3 Materials and Methods

3.1 Dataset Overview

Here we describe the datasets proposed in the ArabicNLP MAHED 2025 challenge (Zaghouani et al., 2025) that we utilized to run our experiments. The **first dataset** is text-based speech that includes 9,843 examples for training, 1,476 for validation, and 1,477 for testing. The goal of using this data is to classify the speech text into one of three categories: hope, hate, or not_applicable.

The **second dataset** is a text-based multi-task set that contains 8,515 examples (5,960 for training, 1,277 for validation, and 1,278 for testing) and supports three types of sub-tasks. The first sub-task aims to classify each text into one of twelve emotions: neutral, anger, anticipation, disgust, fear, joy, love, optimism, pessimism, sadness, surprise, or trust. The second sub-task aims to detect offensive language in the text, labeling it as either yes or no. When offensive language is detected, the third sub-task classifies the text as either hate or not_hate.

The **third dataset** targets multi-modal hateful meme detection. It has 4,500 examples (2,143 for training, 312 for validation, and 606 for testing) annotated with two labels: hateful and not hateful. Each meme example includes an image and its extracted Arabic text.

3.2 Methods

Detection of Hope and Hate in Arabic Speech: For this task, first, we fine-tuned 2 LLMs such as GPT-4o-mini[6] (namely LLM 1 in Table 1), and Gemini Flash 2.5[7] (Team et al., 2023)(namely LLM 2 in Table 1) using the training and validation sets from the **first dataset**. Secondly, we utilized Google text embedding[8]+ SVM (Hearst et al.,

[6] https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/

[7] https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/#gemini-2-5-thinking

[8] https://developers.googleblog.com/en/gemini-embedding-text-model-now-available-gemini-api

1998) (namely LLM 3 in Table 1). To improve the accuracy, we used ensemble learning (namely Ensemble in Table 1) that used majority voting among previous 3 models. We found that many hope samples were predicted as not_applicable. So we added hope/not_applicable fine-tuned GPT-4o-mini to address this issue. we named our solution in Table 1. We reported the results of inference on a testing set.

Multi-task Detection of Emotional Expressions, Offensive Language, and Hate Speech: For this task, three GPT-4o-mini models were fine-tuned using the training and validation sets from the **second dataset** for three epochs with a learning rate multiplier of 1.8. We reported the results of inference on a testing set.

To address the class imbalance in hate/not-hate sub-task, we over-sampled the minority 'hate' text by a factor of five to achieve a more balanced distribution between the 'hate' and 'not-hate' classes.

Multi-modal Detection of Arabic Hateful Memes: For this task, we have evaluated several methods, including base LLMs, fine-tuned LLMs, and embedding models, to find the best solution. We tested all solutions using the testing data of 606 Arabic memes.

First, we started with assessing the performance of **embedding models** that can combine their outputs with traditional classifiers for hate/not-hate classification. We used the Google multi-modal pre-trained embedding model (multimodalembedding@001)[9] to generate embedding vectors for each text and image in each meme. The embedding vector has 512 dimensions. Later, we aggregated the two embedding vectors of text and image by computing their element-wise average first and then by concatenating the two vectors. Finally, we added a support vector machine (SVM) (Hearst et al., 1998) to classify the resulting embedding vector into two classes: hate and not-hate. We assessed four scenarios: text embedding vector only, image embedding vector only, average of text and image embedding vectors, and concatenation of text and image embeddings. We fine-tuned hyper-parameters of SVM to get the highest F1 and F2 scores. We found that regularization parameter C = 0.1, kernel = radial basis function (rbf), gamma = scale, and balanced class weighted loss function are the optimal hyperparameters for the three scenarios except the text-only scenario, where C = 1 is optimal. Additionally, we replaced SVM with a deep neural network (DNN) (LeCun et al., 2015) whose architecture was optimized to get the optimal one with the highest F1 and F2 scores.

In the **second experiment**, we assessed the capacity of multi-modal pre-trained **safety classifiers** for hate detection in memes.

Llama Guard 4[10],[11] (Chi et al., 2024) is a multi-modal safety classifier with 12 billion parameters, trained jointly on both text and images. It uses a dense architecture derived from the Llama 4 Scout pre-trained model, which has been pruned and fine-tuned specifically for content safety classification. In this work, our focus is on the 'hate' category, which refers to text that demeans or dehumanizes individuals based on sensitive personal characteristics. We focus on all examples that have been flagged under the hate category only.

Omni-moderation-latest[12] is a moderation endpoint used to check whether text or images are potentially harmful. Its output includes several categories and their confidence values. The moderator sets the flag to true if it classifies the content as harmful. The limitation of this moderator is that for categories such as 'hate' or 'hate/threatening,' it supports only text. We consider all examples that have triggered the safety flag.

In the **third experiment**, we ran **Gemini Flash 2.5**, a base model with a system prompt (Prompt 1). We also ran the **GPT-4o-mini** base model with Prompts 1, 2, and 3 (available in the Appendix).

To improve the detection performance, we fine-tuned several LLMs in a supervised learning setting. We started by tuning Gemini Flash 2.5 using Prompt 3. To address the class imbalance, we over-sampled the minority 'hate' memes by a factor of nine to achieve a more balanced distribution between the 'hate' and 'no_hate' classes. The hyper-parameters used are three epochs, learning_rate_multiplier of 0.5, an adapter size of 2, an off threshold in safety_settings, and disabled thinking. Additionally, we also fine-tuned **Llama 3.2-11B**[13] (Dubey et al., 2024) using both text and image inputs from the training data. We used

/

[9] https://cloud.google.com/vertex-ai/generative-ai/docs/embeddings/get-multimodal-embeddings

[10] https://www.llama.com/docs/model-cards-and-prompt-formats/llama-guard-4/

[11] https://huggingface.co/meta-llama/Llama-Guard-4-12B

[12] https://platform.openai.com/docs/guides/moderation

[13] https://huggingface.co/meta-llama/Llama-3.2-11B-Vision-Instruct

Low-Rank Adaptation (LoRA) (Hu et al., 2022) as the Parameter-Efficient Fine-Tuning (PEFT) (Xu et al., 2023) method for fine-tuning utilizing the unsloth framework. The fine-tuned Llama 3.2-11B model was uploaded to Huggingface: https://huggingface.co/NYUAD-ComNets/Llama3.2-MultiModal-Hate_Detector_Memes

Finally, we fine-tuned **Paligemma2**[14] (Steiner et al., 2024) namely "google/paligemma2-3b-pt-224". The parameters of the vision tower and the language model are frozen, while only the parameters of the multi-modal projector are set to be trainable.

In the previous fine-tuning experiments, we used OpenAI[15] for tuning each GPT-4o-mini. Additionally, we used Google AI vertex studio[16] for tuning Gemini Flash 2.5.

4 Results and Discussion

Hate/Hope Detection in textual speech: In this task, the ensemble method of majority voting among the three LLMs improved the performance as shown in Table 1. Moreover, adding the hope/not classifier to better distinguish real hope samples from those predicted as not_applicable achieved the best performance metrics and ranked second in the leaderboard (Zaghouani et al., 2025) which addresses **RQ4**. It is also worth mentioning that embedding model + SVM (LLM3) shows good performance which answers **RQ1**.

Task	Accuracy %	Macro Precision %	Macro Recall %	Macro F1 Score %
LLM 1	70.6	70.6	69	69.7
LLM 2	69.7	68.6	72.2	69.9
LLM 3	70.6	71.6	67.2	68.9
Ensemble	71.9	71.7	71.2	71.4
Our Solution	**72.3**	71.6	**72.9**	**72.1**

Table 1: Performance metrics for Task 1 (hop/hate/not_applicable)

Multi-task Detection: The three fine-tuned GPT-4o-mini for multi-task (emotion, offensive, hate) achieved the best performance compared to other methods in the leaderboard (Zaghouani et al., 2025) evaluated on a testing set which addresses **RQ4**. More details in Table 2. The model achieved a Macro F1-score of 57.8%, an accuracy of 75.0%, a precision of 61.2%, and a recall of 57.8% over all three sub-tasks.

Task	Accuracy %	Macro Precision %	Macro Recall %	Macro F1 score %
Emotion	59.9	57.2	49.9	51.7
Offensive/ Not	85.4	82.0	84.8	83.1
Hate/Not	63.8	-	-	-

Table 2: performance metrics for multi-task (task 2). Hate/Not detection is influenced by the offensive detection step, and some evaluation metrics cannot be computed because samples predicted as non-offensive yield NaN values and are excluded from the Hate/Not detector.

Hate Detection in Memes: Table 3 presents performance metrics for a variety of models. A pre-trained multi-modal embedding model was found to effectively detect hate speech in Arabic memes using either SVM or DNN, answering **RQ1**. Both LLaMA 4 Guard and OpenAI content moderator show lower recall and F1-scores, especially OpenAI one, suggesting limitations in the existing safety classification solutions on this task, which addresses **RQ2**. Among the base LLMs, GPT-4o demonstrated stronger performance compared to Gemini Flash 2.5, answering **RQ3**.

Fine-tuned Gemini Flash 2.5 demonstrates superior performance across all metrics. Similarly, fine-tuned Llama 3.2 11B consistently ranks second. The results indicate that fine-tuning significantly boosts models' capabilities, which addresses **RQ5**. On the other hand, fine-tuned PaliGemma2 underperforms compared to other models.

Table 4 shows Google's multi-modal embedding model results with SVM for different input modalities. The findings indicate that the average embedding vector outperforms slightly the image-only embedding. This suggests that adding text embeddings does not provide an advantage for classification. One explanation is that Google's embedding model processes the text within the meme's image. The performance of text-only embeddings is the lowests. We also ran GPT-4o-mini with the three prompts as shown in Table 5. Even though Prompt 3 produced the highest accuracy and macro F1 score, Prompt 1 gave the highest macro F2 score, suggesting a better prompt to detect the hate class specifically.

Flash Flash 2.5 achieved the best performance in

[14] https://huggingface.co/google/paligemma-3b-pt-224

[15] https://platform.openai.com/finetune/

[16] https://console.cloud.google.com/vertex-ai/studio/

LLM	Accuracy %	Macro Precision %	Macro Recall %	Macro F1 score %	Macro F2 score %
embedding + SVM	77.56	70.48	70.83	70.65	70.76
embedding + DNN	77.56	70.32	69.97	70.14	70.04
OpenAI content moderator	72.77	57.27	52.85	51.18	51.75
Llama 4 Guard	71.45	63.36	64.38	63.77	64.11
GPT-4o-mini	79.21	72.49	71.29	71.84	71.50
Gemini Flash 2.5	64.19	62.47	66.36	61.09	63.20
Fine-tuned Gemini Flash 2.5	**83.33**	**78.84**	**74.91**	**76.49**	**75.46**
Fine-tuned Llama 3.2 11B	80.36	74.09	73.14	73.58	73.31
Fine-tuned Paligemma2	76.73	68.95	67.49	68.12	67.72

Table 3: Performance of base and fine-tuned LLMs for task 3.

	Accuracy %	Macro Precision %	Macro Recall %	Macro F1 score %	Macro F2 score %
image embedding + SVM	77.23	70.10	70.61	70.34	70.50
text embedding + SVM	66.50	57.15	57.64	57.33	57.50
Avg. embedding of image&text + SVM	**77.56**	**70.48**	**70.83**	**70.65**	**70.76**
Concatenate embedding of image&text + SVM	76.07	68.53	68.76	68.64	68.71

Table 4: Performance of different input modalities combined with evaluated on validation set in task 3.

	Accuracy %	Macro Precision %	Macro Recall %	Macro F1 score %	Macro F2 score %
GPT-4o-mini Prompt 1	74.75	70.70	**76.23**	71.34	**73.62**
GPT-4o-mini Prompt 2	79.21	72.49	71.29	71.84	71.50
GPT-4o-mini Prompt 3	**82.51**	**80.86**	69.44	**72.29**	70.14

Table 5: GPT-4o-mini base model performance under different prompts evaluated on validation set in task 3.

the leaderboard (Zaghouani et al., 2025) evaluated on a testing set of 500 memes. The model achieved a Macro F1-score of 79.6%, an accuracy of 80.0%, a precision of 79.4%, and a recall of 80.4%.

Limitations

One limitation of this work is the subjective nature of the annotations poses challenges, as different annotators may interpret and label content differently. This introduces potential inconsistencies in the training data, which could affect the model's performance.

Another key limitation is the models' ability to understand and process different Arabic dialects.

References

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*, pages 380–390. 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*.

Nuha Albadi, Maram Kurdi, and Shivakant Mishra. 2018. Are they our brothers? analysis and detection of religious hate speech in the arabic twittersphere. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 69–76. IEEE.

Nouar AlDahoul, Myles Joshua Toledo Tan, Harishwar Reddy Kasireddy, and Yasir Zaki. 2024a. Advancing content moderation: Evaluating large language models for detecting sensitive content across text, images, and videos. *arXiv preprint arXiv:2411.17123*.

Nouar AlDahoul, Myles Joshua Toledo Tan, Harishwar Reddy Kasireddy, and Yasir Zaki. 2024b. Exploring vision language models for facial attribute recognition: Emotion, race, gender, and age. *arXiv preprint arXiv:2410.24148*.

Nouar AlDahoul, Myles Joshua Toledo Tan, Raghava Reddy Tera, Hezerul Abdul Karim, Chee How Lim, Manish Kumar Mishra, and Yasir Zaki. 2025. Multitasking vision language models for vehicle plate recognition with vehiclepaligemma. *Scientific Reports*, 15(1):1–15.

Wissam Antoun, Fady Baly, and Hazem Hajj.

2020. Arabert: Transformer-based model for arabic language understanding. *arXiv preprint arXiv:2003.00104*.

Mohamed Aziz Bennessir, Malek Rhouma, Hatem Haddad, and Chayma Fourati. 2022. icompass at arabic hate speech 2022: Detect hate speech using qrnn and transformers. 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 176–180.

Jianfeng Chi, Ujjwal Karn, Hongyuan Zhan, Eric Smith, Javier Rando, Yiming Zhang, Kate Plawiak, Zacharie Delpierre Coudert, Kartikeya Upasani, and Mahesh Pasupuleti. 2024. Llama guard 3 vision: Safeguarding human-ai image understanding conversations.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

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*.

Marti A. Hearst, Susan T Dumais, Edgar Osuna, John Platt, and Bernhard Scholkopf. 1998. Support vector machines. *IEEE Intelligent Systems and their applications*, 13(4):18–28.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Sanaa Kaddoura, Suja A Alex, Maher Itani, Safaa Henno, Asma AlNashash, and D Jude Hemanth. 2023. Arabic spam tweets classification using deep learning. *Neural Computing and Applications*, 35(23):17233–17246.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *nature*, 521(7553):436–444.

Hanane Mohaouchane, Asmaa Mourhir, and Nikola S Nikolov. 2019. Detecting offensive language on arabic social media using deep learning. In *2019 sixth international conference on social networks analysis, management and security (SNAMS)*, pages 466–471. IEEE.

Hamdy Mubarak, Sabit Hassan, and Shammur Absar Chowdhury. 2023. Emojis as anchors to detect arabic offensive language and hate speech. *Natural Language Engineering*, 29(6):1436–1457.

Ahmad Shapiro, Ayman Khalafallah, and Marwan Torki. 2022. Alexu-aic at arabic hate speech 2022: Contrast to classify. *arXiv preprint arXiv:2207.08557*.

Andreas Steiner, André Susano Pinto, Michael Tschannen, Daniel Keysers, Xiao Wang, Yonatan Bitton, Alexey Gritsenko, Matthias Minderer, Anthony Sherbondy, Shangbang Long, et al. 2024. Paligemma 2: A family of versatile vlms for transfer. *arXiv preprint arXiv:2412.03555*.

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*.

Lingling Xu, Haoran Xie, Si-Zhao Joe Qin, Xiaohui Tao, and Fu Lee Wang. 2023. Parameter-efficient fine-tuning methods for pretrained language models: A critical review and assessment. *arXiv preprint arXiv:2312.12148*.

Wajdi Zaghouani and Md Rafiul Biswas. 2025a. An annotated corpus of arabic tweets for hate speech analysis. *arXiv preprint arXiv:2505.11969*.

Wajdi Zaghouani and Md Rafiul Biswas. 2025b. Emohopespeech: An annotated dataset of emotions and hope speech in english and arabic. *arXiv preprint arXiv:2505.11959*.

Wajdi Zaghouani, Md Rafiul Biswas, Mabrouka Bessghaier, Shimaa Ibrahim, Georgios Mikros, Abul Hasnat, and Firoj Alam. 2025. MAHED shared task: Multimodal detection of hope and hate emotions in arabic content. In *Proceedings of the Third Arabic Natural Language Processing Conference (Arabic-NLP 2025)*, Suzhou, China. Association for Computational Linguistics.

Wajdi Zaghouani, Hamdy Mubarak, and Md Rafiul Biswas. 2024. So hateful! building a multi-label hate speech annotated arabic dataset. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15044–15055.

A Appendix

A.1 Prompts used

The following prompts were used in this work for the three tasks. The prompt in orange color aims to classify text into hate, hope, and not_applicable. The three prompts in cyan color aim to classify text into:

- offensive language or not.
- hate or not.
- one of the emotions in a predefined set of emotions.

The prompts in red, blue, and green color are three different prompts used to classify memes into hate or not.

Hate/Hope Detection Prompt

Classify each text into one of the three categories 'hope','hate',or 'not_applicable'

Offensive Language Detection Prompt

You are an expert in offensive language detection.
Carefully read the input text and determine whether it is offensive.
Respond with only one word, either "yes" if the text is offensive, or "no" if it is not.
Text: "<your text here>"
Answer:

Hate/not-Hate Detection Prompt

Classify each text into one of the two categories 'hate', or 'not_hate'

Emotion Detection Prompt

You are an expert in fine-grained emotion classification.
Given a short piece of text, your task is to classify the **dominant emotion** expressed in the text.
Use **only one word** from the following list of predefined emotion categories:

[anger, disgust, neutral, love, joy, anticipation, optimism, sadness, confidence, pessimism, surprise, fear]
Carefully read the input and select the most appropriate emotion based on the **overall tone, context, and word choice**.
If multiple emotions are expressed, choose the one that is **most central** to the message.
Respond with only one word, exactly as listed.
Text: "<your text here>"
Answer:

Hate Detection in Memes Prompt 1

You're an expert Arabic social media hate content analyst. Your goal is to flag any content that is derogatory, disparaging, promotes negative stereotypes, incites any form of harm (even humorously), or contributes to a hostile environment.
Provide the following for analysis:
Meme Description (Visuals): Detailed text description of all visual elements (characters, symbols, actions, etc.).
Meme Text (Arabic & English): Exact Arabic text and its literal English translation. Include any emojis/symbols.
Context: Explain who is targeted (individuals, groups, professions, etc.), relevant current events, cultural/historical references, and any known slang/derogatory terms.
User Intent (if known): Any information on the creator/sharer's likely purpose.
Analysis: Your detailed reasoning, explaining how visuals, text, and context contribute to hate.
Focus on dehumanization, negative stereotypes, incitement (even if satirical), hostile environment creation, contempt, ridicule, disgust, targeting based on role, disparaging language (ableism, body shaming, etc.), and normalization of problematic behavior.
Final Answer: hate/no hate

Hate Detection in Memes Prompt 2

You are a hate speech analyst evaluating Arabic social media memes.
Your goal is to flag hate speech, including implicit hate masked as humor or political discourse.
Hate includes:
Stereotyping or mocking a group based on ethnicity, nationality, gender, religion, or identity Jokes or memes that promote harmful generalizations
Political content that blames or dehumanizes a people, not just a government or policy
However, you must not flag:
Political critique focused on leadership, government, or policy (not people)
Cultural observations or satire without negative generalization
Your analysis must:
Describe the image and interpret the Arabic text
Explain whether it includes group-based bias or stereotypes
End with:
Analysis: your reasoning
Final Answer: 'hate' or 'no hate'

Hate Detection in Memes Prompt 3

You are a hate speech analyst evaluating Arabic social media memes.
Your goal is to classify meme into hate or no_hate

A.2 Hyper-parameters for various models

The following are Hyper-parameters used for training DNN, and fine-tuning PaliGemma2, and Llama 3.2- 11B. Table 6 describes the DNN's architecture.

Hyper-parameters for DNN

- Adam optimizer,
- weighted class binary cross-entropy loss fuction
- 100 epochs
- 128 batch size
- early stopping with patience = 3.

Training Configuration of PaliGemma2

- number of training epochs: 3
- per-device training batch size: 2
- gradient accumulation steps: 8
- warm-up steps: 2
- learning rate: 2e-5
- weight decay: 1e-6
- Adam optimizer beta2 value: 0.999
- optimizer type: Adamw_hf
- early stopping callback with patience=2.

Fine-tuning Configurations of Llama 3.2-11B

- the training batch size per device is set to 4.
- gradients are accumulated over 4 steps.
- the learning rate warm-up lasts for 5 steps.
- the total number of training steps is 150.
- the learning rate is set to 0.0002.
- the optimizer used is 8-bit AdamW
- weight decay is set to 0.01.
- a linear learning rate scheduler is used.

Layer Type	Output Shape	Activation	Description
Input Layer	(512,)	–	Input vector representing image embedding
Dense Layer	(256,)	ReLU	Fully connected layer on image input
Dropout	(256,)	Dropout 0.5	Regularization
Dense Layer	(128,)	ReLU	Further transformation of image embedding
Dropout	(128,)	Dropout 0.5	Regularization
Dense Layer	(64,)	ReLU	Compressed feature representation
Dropout	(64,)	Dropout 0.5	Regularization
Input Layer	(512,)	–	Input vector representing text embedding
Dense Layer	(256,)	ReLU	Fully connected layer on text input
Dropout	(256,)	Dropout 0.5	Regularization
Dense Layer	(128,)	ReLU	Intermediate transformation
Dropout	(128,)	Dropout 0.5	Regularization
Dense Layer	(64,)	ReLU	Compressed feature representation
Dropout	(64,)	Dropout 0.5	Regularization
Concatenate	(128,)	–	Merge image and text features (64 + 64)
Dense Layer	(128,)	ReLU	Combined representation processing
Dropout	(128,)	Dropout 0.5	Regularization
Dense Layer	(1024,)	ReLU	High-capacity layer for rich interaction
Dropout	(1024,)	Dropout 0.5	Regularization
Dense Layer	(1,)	Sigmoid	Final prediction for binary classification

Table 6: Architecture of the dual-branch DNN model for image and text fusion.

A.3 Confusion matrices for various models

The following are confusion matrices presenting the models' performance in terms of False Positives, False Negatives, True Positives, and True Negatives.

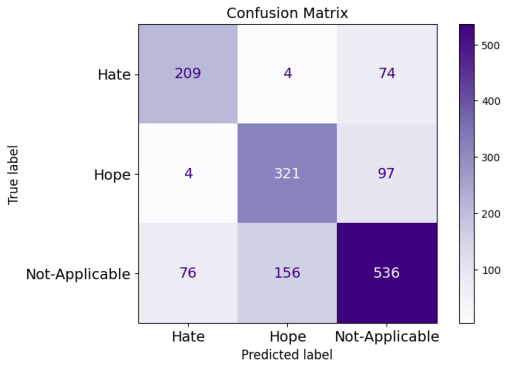

Figure 1: Confusion matrix of testing set in task 1 for hope/hate/not_applicable classification in text using ensemble of 3 fine-tuned LLMs (gpt-4o-mini, Gemini Flash 2.5, and Google text embedding + SVM) + fine-tuned gpt-4o-mini for hope/not

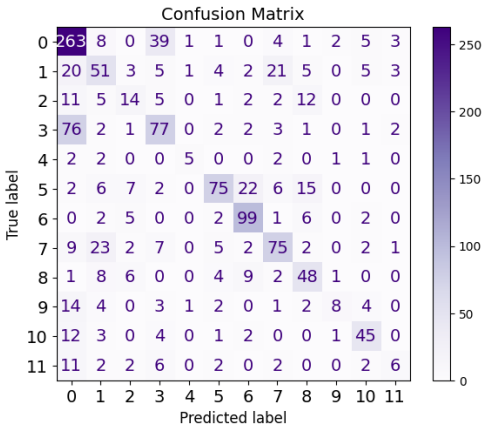

Figure 2: Confusion matrix of testing set in task 2 for emotion classification in text using Fine-tuned GPT-4o-mini. class 0: Anger, class 1: Anticipation, class 2: Confidence, class 3: Disgust, class 4: Fear, class 5: Joy, class 6: Love, class 7: Neutral, class 8: Optimism, class 9: Pessimism, class 10: Sadness , class 11: Surprise

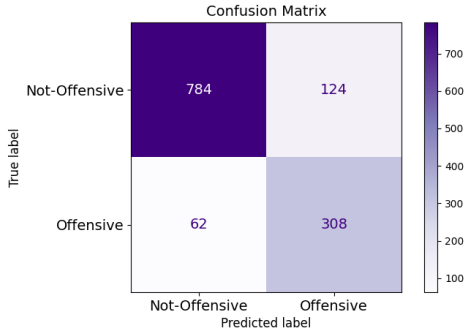

Figure 3: Confusion matrix of testing set in task 2 for offensive detection in text using Fine-tuned GPT-4o-mini.

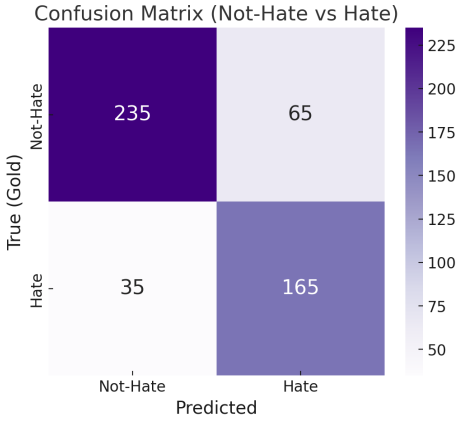

Figure 4: Confusion matrix of testing set for hate detection in memes using Fine-tuned Gemini Flash 2.5

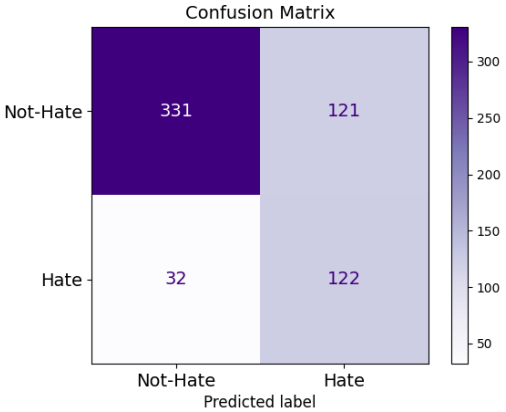

Figure 5: Confusion matrix of validation set for hate detection in memes using GPT-4o-mini with Prompt 1

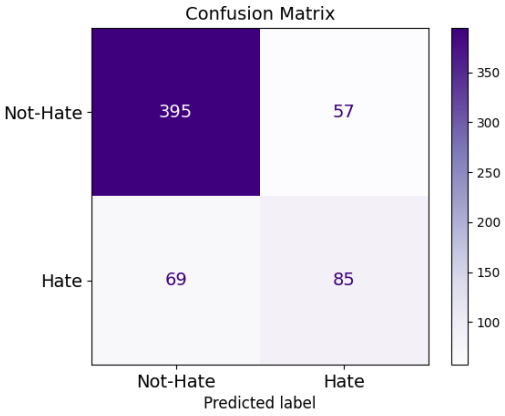

Figure 6: Confusion matrix of validation set for hate detection in memes using GPT-4o-mini with Prompt 2

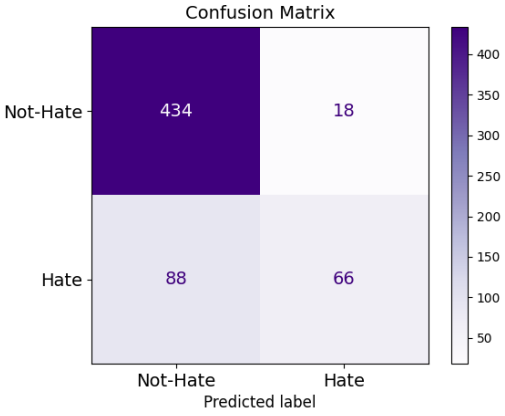

Figure 7: Confusion matrix of validation set for hate detection in memes using GPT-4o-mini with Prompt 3

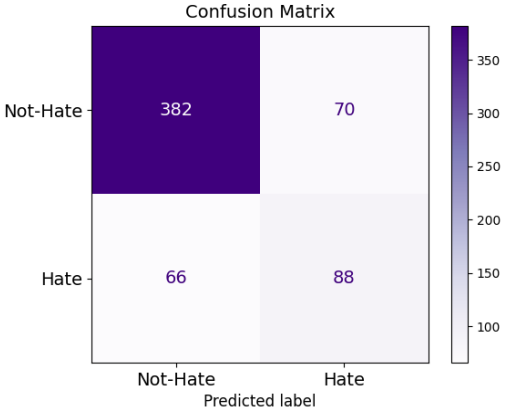

Figure 8: Confusion matrix of validation set for hate detection in memes using average embeddings of image and text + SVM

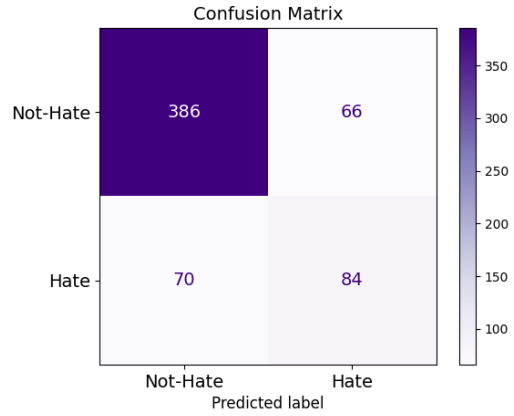

Figure 9: Confusion matrix of validation set for hate detection in memes using average embeddings of image and text + DNN

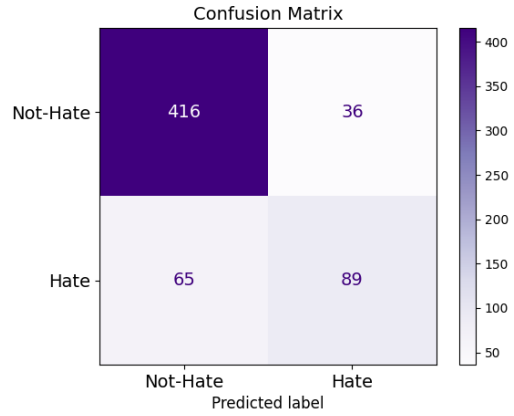

Figure 10: Confusion matrix of validation set for hate detection in memes using Fine-tuned Gemini Flash 2.5

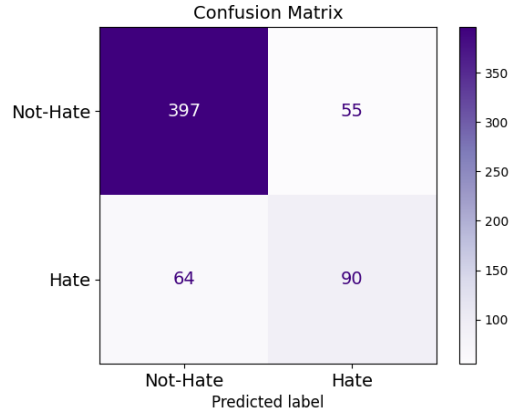

Figure 11: Confusion matrix of validation set for hate detection in memes using Fine-tuned Llama 3.2 11B