CUET-823 at MAHED 2025 Shared Task: Large Language Model-Based Framework for Emotion, Offensive, and Hate Detection in Arabic

Ratnajit Dhar, Arpita Mallik
Department of Computer Science and Engineering
Chittagong University of Engineering and Technology, Bangladesh
{u2004008, u2004023}@student.cuet.ac.bd

Abstract

This paper presents our system for Subtask-2: Emotion, Offensive Language, and Hate Detection in the MAHED 2025 Shared Task at ArabicNLP 2025. We address the challenge of multi-label classification in Arabic social media text using a two-stage, prompt-based framework with large language models. In the first stage, our system classifies emotions into 12 distinct categories; in the second stage, it detects offensive messages and, when relevant, further identifies the presence of hate speech. Both stages leverage the Meta-Llama-3.1-8B model, fine-tuned to capture the diverse linguistic and dialectal characteristics of Arabic. Our approach achieved a macro F1-score of 0.518 on the official test set, placing 4th in Subtask 2. The results demonstrate the effectiveness of prompt-based modeling for complex Arabic text classification and contribute a practical, LLM-based solution for emotion and hate speech detection in low-resource scenarios.

1 Introduction

The growing influence of social media platforms has fundamentally transformed how individuals express emotions and communicate across digital spaces, with Arabic-speaking communities having represented one of the fastest-growing user bases globally (Ali and Aleqabie, 2024; Alqahtani and Alothaim, 2022). According to Statista, the internet user population in the United Arab Emirates (UAE) has peaked in 2025 and has increased by almost a thousand users compared to the previous year. This rapid growth has generated vast amounts of user-generated content in diverse Arabic dialects. Consequently, there has been a growing need for robust NLP tools to understand and moderate Arabic content, especially given the mix of emotions and potential for offensive or hateful language. Yet, existing moderation systems have often struggled with Arabics morphological complexity, dialect diversity (Center for Democracy and Technology, 2023), and limited training data and cultural awareness (AL-Sarayreh et al., 2023).

To address these challenges, we have participated in Subtask-2 of the MAHED 2025 Shared Task on Multimodal Detection of Hope and Hate Emotions in Arabic Content (Zaghouani et al., 2025). The purposive focus of this task has been to identify the emotion expressed in Arabic social media text, determine whether the text is offensive, and, if offensive, further assess whether it contains hate content.

To achieve our goal, we have employed a large language model (LLM), specifically a lightweight Meta-Llama-3.1-8B, as the core of our system. This powerful LLM has been fine-tuned using the Unsloth framework, allowing us to efficiently adapt it to the challenging Arabic emotion, offensive language, and hate speech detection tasks. The use of a large language model has enabled us to build an effective system that has achieved competitive performance (macro F1-score: 0.518), ranking 4th among all submissions. The main contributions of this work have been:

- Proposed a two-stage prompt-based framework linking emotion and hate speech detection.
- Applied lightweight LLM fine-tuning for efficient Arabic multi-label classification.
- Showed conditional prompting outperforms flat multi-label methods in low-resource Arabic NLP.

Further implementation details can be accessed via the GitHub repository.[1]

[1] https://github.com/ratnajit-dhar/MAHED

Proceedings of The Third Arabic Natural Language Processing Conference, pages 620–625
November 8-9, 2025 ©2025 Association for Computational Linguistics

2 Background

The detection of hate speech and offensive language in Arabic has remained challenging due to its complex morphology and dialectal variation. The MAHED 2025 Shared Task (Zaghouani et al., 2025) has required systems to analyze short Arabic texts (mostly tweets) and assign multiple labels. The dataset (Zaghouani et al., 2024; Zaghouani and Biswas, 2025b,a) used in this work was introduced at the ArabicNLP 2025 workshop under the MAHED 2025 Shared Task. Examples of input texts and their corresponding output labels are provided in Appendix A.

Early pioneering work has developed foundational methods for Arabic emotion detection utilizing Twitter data in the context of the Egyptian revolution and has found that it was possible to automatically detect emotions from Arabic tweets after appropriate preprocessing (Rabie and Sturm, 2014). Further studies have provided sizable progress through a variety of approaches. A study to collect Arabic dialect datasets by scraping tweets through Olympic hashtags has derived an accuracy of 68.12% with Complement Naive Bayes classifiers (Al-Khatib and El-Beltagy, 2017). Another one has built large-scale COVID-19 datasets with 5.5 million tweets and has achieved an 83% F1-score for emotion classification using LSTM models (Al-Laith and Alenezi, 2021). The advent of transformer-based models has transformed Arabic emotion detection. Research has shown transformer-based models, such as AraBERT, have been superior to traditional machine learning methods (Qaddoumi, 2022). Arabic hate speech and offensive language detection, an equally challenging space, has had progress on large-scale datasets with special tags for vulgarity and hate speech where researchers have achieved F1-scores of 83.2% using state of the art techniques (Mubarak et al., 2020). The integration of emotional knowledge with hate speech detection through multi-task learning frameworks has shown promising results, with studies demonstrating approximately 3% improvement when combining emotional analysis with hate speech detection tasks (Mnassri et al., 2023).

Building on prior work, our system employs a two-stage prompt-based approach using large language models to efficiently address the problem of hierarchical emotion and hate speech detection with limited data and different dialects.

3 System Overview

Our system has been built upon a two-stage, prompt-based classification framework, using Meta-Llama-3.1-8B as the backbone large language model. The main design choice has been to separate emotion classification from detecting hate and offensive speech, providing appropriate prompts and fine-tuning approaches for each stage. An overview of the architecture is illustrated in Figure 1.

In the first stage, we have fine-tuned the model to detect emotion using the prompts that have explicitly highlighted the 12 emotion categories. An example of a prompt that has been used during training and during the inferencing phase follows:

> The following text is an Arabic text. Your task is to classify the emotion expressed in the text into one of the following categories:
> 1. anger, 2. disgust, 3. neutral, 4. love, 5. joy, 6. anticipation, 7. optimism, 8. sadness, 9. confidence, 10. pessimism, 11. surprise, 12. fear
>
> **Text:** {text}
>
> **Response:** {response}

In the second stage, we employed a dedicated prompt asking the model to create a response in two steps, strictly following the task instructions:

> You are given an Arabic text. Your task is to classify whether the text is offensive or not offensive.
> - If the text is offensive, respond with: offensive
> - Then, further classify the text as either:
> - hate (if it expresses hate speech)
> - not_hate (if it does not express hate speech)
> - If the text is not offensive, respond with: not offensive
> - Then, respond with: `not_applicable` for the second classification.
>
> **Text:** {text}
>
> **Response:**
> 1. {offensive}
> 2. {hate}

Based on the hierarchical nature of the task labels, this two-stage design has treated emotion classification as a foundational step, whereas hate speech classification has only been taken into consideration if the text has already been deemed offensive. Such a conditional setup has not only reduced label confusion but also allowed the model to focus on distinct linguistic patterns at each stage. This approach has aligned with prior work showing that hierarchical approaches have had better performance than flat multi-label classification for multi-label tasks as it reduces a complex task into

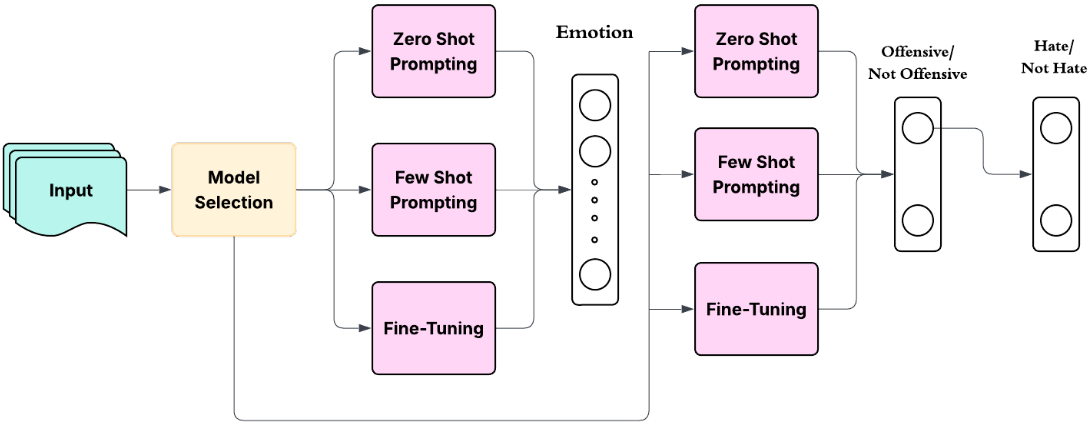

Figure 1: Two-Stage LLM Framework for Arabic Emotion, Offensive, and Hate Speech Detection.

simpler sub-tasks (Galea et al., 2017; Yang et al., 2023). To ensure deterministic outputs during evaluation, we have set the generation temperature to 0.0 for all inference stages.

All fine-tuning and evaluation have used only the official MAHED 2025 dataset; no external data, manual features, or class balancing techniques have been applied.

4 Experimental Setup

4.1 Dataset

We have used the MAHED 2025 dataset from the shared task on Arabic emotion, offensive language, and hate speech classification. Each post is labeled with one of 12 emotions, an offensive label (yes or no), and, if offensive, a hate label (hate or not hate). The training set contained 5,960 posts (1,744 offensive, 303 hate), the validation set 1,277 posts (363 offensive, 68 hate), and the test set consisted of 1,278 unlabeled posts, used solely for final shared task evaluation. The detailed distribution of labels in the training set is shown in Table 1.

4.2 Model and Hyperparameters

We have fine-tuned the Meta-Llama-3.1-8B model using 4-bit quantization and LoRA (Hu et al., 2021) with rank 16 for 3 epochs. Emotion and offensive/hate models have been trained for 1000 and 500 steps, respectively, using AdamW optimizer (Loshchilov and Hutter, 2019) with a learning rate of 2e-4, weight decay 0.01, and a linear scheduler with 5 warmup steps. We have employed gradient accumulation and capped input sequences at 2048 tokens. Gradient checkpointing has been enabled to reduce memory usage.

Label	Category	Count
Emotion	Anger	1551
	Disgust	777
	Neutral	661
	Love	593
	Joy	533
	Anticipation	491
	Optimism	419
	Sadness	335
	Confidence	210
	Pessimism	194
	Surprise	143
	Fear	53
	Total	**5960**
Offensive	No	4216
	Yes	1744
Hate (if offensive)	Not Hate	1441
	Hate	303
	Total	**1744**

Table 1: Distribution of emotion, offensive, and hate speech labels in the MAHED 2025 training set.

4.3 Libraries and Frameworks

We have used the Unsloth framework (v2024.8) for fine-tuning and the Hugging Face Transformers library (v4.44.0) for model loading/inference. Fine-tuning was configured with TRL, and all models were quantized to 4-bit using BitsAndBytes (v0.43.1) to reduce memory usage.

4.4 Evaluation Metrics

We have used macro-averaged F1-score as the primary evaluation metric. Additionally, we have reported per-class precision and recall to analyze how the model has handled both common and underrepresented emotion and hate categories.

5 Results

In this section, we present the results of our Arabic emotion, offensive language, and hate speech classification task by comparing different prompting strategies as well as models in order to demonstrate their abilities in tackling the problems of this multi-label task.

Our official (fine-tuning LLaMA-3.1-8B) system has had a macro-average F1-score of **0.518** on the test set, placing us in **4th** place out of all of the participating systems.

To gain insight into the impact of different model architectures and prompting strategies, we have compared a number of LLMs using zero-shot, few-shot, and fine-tuned setups. The results, evaluated on the development set, are summarized in Table 2.

In addition to our overall performance, our team (CUET_823) achieved the **highest precision (0.617)**, showing strong effectiveness at minimizing false positives despite slightly lower F1-scores.

Model Name	Prompting Strategy	Macro F1
LLaMA-3.1 8B	Fine-tuning	0.554
	Zero-shot	0.484
	Few-shot	0.477
Mistral 7B v0.3	Zero-shot	0.412
	Few-shot	0.420
	Fine-tuning	0.435
Qwen-2 7B	Zero-shot	0.458
	Few-shot	0.407
	Fine-tuning	0.415
CodeGemma 7B	Zero-shot	0.388
	Few-shot	0.397
	Fine-tuning	0.421
Zephyr 7B	Zero-shot	0.374
	Few-shot	0.386
	Fine-tuning	0.410
Gemma 3 4B	Zero-shot	0.395
	Few-shot	0.402
	Fine-tuning	0.428

Table 2: Macro F1 performance of different models on the validation set.

Fine-tuning has clearly outperformed both zero-shot and few-shot prompting strategies across all models. The LLaMA-3.1-8B model has consistently achieved the highest scores, validating our choice of model and training strategy.

5.1 Error Analysis

While this system has performed very well overall, it has struggled in borderline cases of offensive/the hate speech classification decision. In many cases, the system has flagged text as offensive but has failed to escalate to hate, especially where hatefulness was implied or culturally coded. Below are a few representative errors from the validation set:

Text: شكرا لك وجزاك الله خيرا تحياتي @ZADXII
True: love, no, -
Predicted: neutral, no, -
A positive, polite thank-you message was predicted as neutral rather than 'love' (politeness vs. affection confusion).

Text: العبيد زود انك كريه تسوي ذا الحركات يامريض
True: disgust, yes, hate
Predicted: disgust, yes, not_hate
Racist slur went undetected as hate, showing model's hesitation to escalate from 'offensive' to 'hate' without explicit group targeting.

6 Conclusion

In this work, we have presented a hierarchical two-stage system for Arabic emotion, offensive language, and hate speech detection. The experiments have shown that detecting emotions first and then applying conditional offensive and hate speech classification has been effective due to the strong correlation between these tasks. The hierarchical approach has also been useful in addressing dialectal diversity while being resource-efficient.

Although the system has performed competitively, there have been some limitations. Generalizability has been limited as we have employed only a single dataset and architecture, and static prompting may struggle with evolving language. Not exploring Arabic-specific transformer models may also have limited performance. Future work could explore further architectures, new ensemble methods, dynamic prompting rather than static, wider dialect coverage, and multimodal features for better robustness and contextual understanding.

Acknowledgments

We thank the organizers of the MAHED 2025 Shared Task for providing the dataset and evaluation framework. We are grateful to the anonymous reviewers for their constructive feedback that helped improve this paper.

References

Amr Al-Khatib and Samhaa R El-Beltagy. 2017. Emotional tone detection in arabic tweets. In *International Conference on Computational Linguistics and Intelligent Text Processing*, pages 105--114. Springer.

Ali Al-Laith and Mamdouh Alenezi. 2021. Monitoring peoples emotions and symptoms from arabic tweets during the covid-19 pandemic. *Information*, 12(2):86.

Sallam AL-Sarayreh, Azza Mohamed, and Khaled Shaalan. 2023. Challenges and solutions for arabic natural language processing in social media. In *International conference on Variability of the Sun and sun-like stars: from asteroseismology to space weather*, pages 293--302. Springer.

Zakaria H Ali and Hiba J Aleqabie. 2024. Emotion detection in arabic text in social media: A brief survey. *Al-Furat Journal of Innovations in Electronics and Computer Engineering*, 3(2):412--421.

Ghadah Alqahtani and Abdulrahman Alothaim. 2022. Emotion analysis of arabic tweets: Language models and available resources. *Frontiers in Artificial Intelligence*, 5:843038.

Center for Democracy and Technology. 2023. Moderating maghrebi arabic content on social media. https://cdt.org/insights/moderating-maghrebi-arabic-content-on-social-media/. Accessed: 2025-08-03.

Dieter Galea, Paolo Inglese, Lidia Cammack, Nicole Strittmatter, Monica Rebec, Reza Mirnezami, Ivan Laponogov, James Kinross, Jeremy Nicholson, Zoltan Takats, and 1 others. 2017. Translational utility of a hierarchical classification strategy in biomolecular data analytics. *Scientific Reports*, 7(1):14981.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Khouloud Mnassri, Praboda Rajapaksha, Reza Farahbakhsh, and Noel Crespi. 2023. Hate speech and offensive language detection using an emotion-aware shared encoder. In *ICC 2023-IEEE International Conference on Communications*, pages 2852--2857. IEEE.

Hamdy Mubarak, Ammar Rashed, Kareem Darwish, Younes Samih, and Ahmed Abdelali. 2020. Arabic offensive language on twitter: Analysis and experiments. *arXiv preprint arXiv:2004.02192*.

Abdelrahim Qaddoumi. 2022. Arabic sentiment ensemble nadi shared task 2. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP 2022). Association for Computational Linguistics*.

Omneya Rabie and Christian Sturm. 2014. Feel the heat: Emotion detection in arabic social media content. In *The International Conference on Data Mining, Internet Computing, and Big Data (BigData2014)*, pages 37--49. Kuala Lumpur Citeseer.

Statista. United arab emirates: Number of internet users from 2010 to 2025. https://www.statista.com/statistics/1389944/uae-number-of-internet-users/. Accessed: 2025-08-03.

Youpeng Yang, Qiuhong Zeng, Gaotong Liu, Shiyao Zheng, Tianyang Luo, Yibin Guo, Jia Tang, and Yi Huang. 2023. Hierarchical classification-based pan-cancer methylation analysis to classify primary cancer. *BMC bioinformatics*, 24(1):465.

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 (ArabicNLP 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 Input and Output Examples

The example below shows a sample input and the corresponding output labels from the train dataset:

Example 1

Input: أحد التجار الشباب العمانيين يقول للاسف لما يكون عندهم
كاش يروحوا هايبرماركت ولما يريدوا صبر يتسوقوا من عندي!!<LF>متى
سندرك أن تسوقنا من تاجر عماني فتح لبيت عماني ودعما لاقتصاد الوطن
، واذا اردتم التأكد فسألوا موظفي البنوك كم من آلاف الريالات يحولها
التجار الأجانب إلى الخارج يوميا . https://t.co/tBeNnETQ4z

Output:
Emotion: neutral
Offensive: no
Hate: not applicable

Example 2

Input: RT @tlbakhsh: @AddadRuh مخصصة للاجانب فقط
والسعودي نكه!! اشياء عجيبة غريبة مانشوفها غير في السعودية!! المشكلة
الاجانب نفسهم في دولهم مايعمل...

Output:
Emotion: anger
Offensive: yes
Hate: yes

Example 3

Input: مسيرات جمعة غضب القدس تتواصل في مختلف مناطق
البحرين تنديداً بخطوة النظام في التطبيع مع العدو الاسرائيلي - ١٨
سبتمبر البحرين#<LF><LF>٢٠٢٠ #التطبيع $< LF > \# <$
$LF > \#\#\#Bahrainhttps : //t.co/kBhiuqdJfN$

Output:
Emotion: anger
Offensive: yes
Hate: not_hate