AAA at MAHED Shared task: A Systematic Encoder Evaluation for Arabic Hope and Hate Speech Classification

Ahmed Elzainy, Hazem Abdelsalam, Ahmed Samir*,
Mohamed Amin*
Alexandria university, Egypt
{es-ahmedkhalil2025, es-Hazemabdelsalam2024, es-AhmedAbdelMaksoud2025,
es-MohamedE.Amin2025}@alexu.edu.eg

Abstract

Arabic hate speech detection presents unique challenges due to the language's morphological complexity, dialectal diversity, and the subtle nature of emotional expressions in social media. In this paper, we present our submission to the MAHED shared task for Arabic hate speech classification, which aims to classify Arabic text into three categories: hope, hate, and not_applicable. This task is crucial for building safer online communities and has applications in content moderation, social media analysis, and digital wellbeing initiatives. We systematically evaluate six transformer-based encoders, comparing Arabic-specific models (MARBERT, AraBERT, ALCALM) against multilingual alternatives (XLM-RoBERTa, LaBSE, BGE). Our approach demonstrates that specialized Arabic models specially encoders trained on more than one dialect like marber significantly outperform their multilingual counterparts, with MARBERT achieving the best overall performance. Using our proposed methodology, we achieved competitive results on the MAHED shared task with a macro-F1 score of 0.707 on the test split, securing a strong position in the final competition rankings.

1 Introduction

This paper details the system we developed for the MAHED (Multimodal Detection of Hope and Hate Emotions in Arabic Content) shared task, hosted at the Arabic Natural Language Processing Conference (ArabicNLP 2025)(Zaghouani et al., 2025). Our work addresses the critical challenge of Arabic hate speech classification, a multi-class problem designed to distinguish between hope, hate, and neutral expressions in Arabic social media content.

The importance of this task has grown substantially with the increasing prevalence of Arabic content online and the urgent need for effective content moderation systems. Robust hate speech detection systems have critical real-world applications in social media platforms for automated content filtering, in digital wellbeing initiatives for protecting vulnerable users, in research contexts for understanding online discourse patterns, and in policy-making for developing evidence-based regulations around online hate speech.

The challenge of emotion classification in Arabic is particularly acute due to the language's intrinsic complexities. Arabic is characterized by its rich morphological system, where words can be derived from trilateral or quadrilateral roots through complex patterns, making surface-level features less reliable. Furthermore, the phenomenon of diglossia—the coexistence of Modern Standard Arabic (MSA) with numerous regional dialects—means that emotional expressions often carry dialectal nuances that may not be immediately apparent to standard language models. Additionally, the subtlety of hate speech and sarcastic expressions in Arabic social media creates challenges for automated detection systems that must distinguish between explicit and implicit emotional content.

To address these challenges, we conducted a systematic evaluation of six transformer-based encoders, comparing their effectiveness on Arabic emotion classification. Our methodology focuses on fine-tuning individual models with careful hyperparameter optimization rather than ensemble approaches, allowing us to identify the most capable single-model solution for deployment scenarios where computational efficiency is crucial.

The key contributions and findings of our work can be summarized as follows:

- We demonstrate the systematic evaluation of six diverse transformer encoders on Arabic emotion classification, providing comprehensive performance comparisons across Arabic-specific and multilingual models.

*Equal contribution

Proceedings of The Third Arabic Natural Language Processing Conference, pages 615–619

- We show that Arabic-specific models (MARBERT(Abdul-Mageed et al., 2021), ALCLAM(Murtadha et al., 2024), AraBERT(Antoun et al.)) significantly outperform multilingual alternatives, with MARBERT achieving the best balance of performance and robustness.

2 Background

Arabic hate speech detection has evolved from traditional machine learning approaches using handcrafted features to modern transformer-based methods. Early work in Arabic sentiment analysis relied on lexicon-based approaches and statistical features, but these methods struggled with the morphological richness and dialectal variation of Arabic text.

The introduction of pre-trained language models revolutionized Arabic NLP, with BERT-based models like AraBERT (Antoun et al.) and MARBERT (Abdul-Mageed et al., 2021) demonstrating significant improvements over previous approaches. These models leverage large-scale Arabic corpora to learn contextual representations that better capture the nuances of Arabic text.

Multilingual models such as XLM-RoBERTa (Conneau et al., 2019) have shown competitive performance across multiple languages, but their effectiveness on Arabic-specific tasks remains a subject of investigation. Recent work has suggested that language-specific pre-training often provides advantages for morphologically rich languages like Arabic (Abdul-Mageed et al., 2021)(Antoun et al.).

In our setup, both the input and output are text: the input is a sentence and the output is a label from the set {hope, hate, not_applicable}. For example, the input ابتهاج محمد اول امريكيه تشارك في الاوليمبياد بحجاب ، بدأت النهارده is classified as hope. We evaluate on a dataset of Arabic social media posts (Zaghouani et al., 2025) (train: 6890, validation: 1476, test: 1477), which provides a realistic benchmark for emotion and hate speech detection. This task goes beyond sentiment analysis by requiring fine-grained distinctions between emotional states while handling the informal nature of online discourse.

3 System Overview

3.1 Model Architecture

Our approach employs a standard fine-tuning methodology using transformer-based encoders with task-specific classification heads. The general architecture consists of four main components:

1. **Input Processing**: Text tokenization using model-specific tokenizers optimized for Arabic text handling.
2. **Encoder Layer**: Pre-trained transformer encoder providing contextualized representations of input sequences.
3. **Classification Head**: Linear transformation layer mapping encoder outputs to class probabilities.
4. **Loss Function**: Cross-entropy loss with class weighting to address dataset imbalance.

3.2 Evaluated Models

We systematically evaluate six transformer-based encoders representing different pre-training approaches and language coverage:

Arabic-Specific Models:

- **MARBERT** (Abdul-Mageed et al., 2021): Bidirectional encoder pre-trained specifically on Arabic social media content, optimized for informal Arabic text processing.
- **AraBERT-Twitter** (Antoun et al.): Large-scale Arabic BERT model with Twitter-specific pre-training, designed for social media content understanding.
- **ALCLAM** (Murtadha et al., 2024): Contemplative language model designed for deeper Arabic text understanding and reasoning tasks.

Multilingual Models:

- **XLM-RoBERTa** (Conneau et al., 2019): Cross-lingual encoder supporting 100+ languages, including Arabic, trained on diverse multilingual corpora.
- **LaBSE** (Feng et al., 2022): Language-agnostic sentence encoder designed for cross-lingual text representation and similarity tasks.

- **BGE-m3** (Chen et al., 2024): Bidirectional and generative encoder optimized for text embedding and representation learning.

4 Experimental Setup

4.1 Data Split

	Train	Validation	Test
Number of samples	6890	1476	1477

Table 1: Dataset split for Arabic social media posts.

4.2 Training Setup

Key training parameters for our best-performing model (MARBERT) include:

- Learning rate: 1×10^{-6} (optimized through systematic search)
- Batch size: 64 (training), 128 (evaluation)
- Training epochs: 3 with early stopping
- Warmup ratio: 0.1 for learning rate scheduling
- Weight decay: 0.01 for regularization
- Maximum sequence length: 170 tokens

4.3 Evaluation Framework

Following the shared task guidelines, we employ comprehensive evaluation metrics:

- **Primary Metric**: Macro-averaged F1 score for balanced evaluation across all classes
- **Secondary Metrics**: Accuracy, macro-averaged precision, and recall

5 Results

5.1 Overall Performance

Table 2 presents comprehensive performance comparisons for all models evaluated in the test set. The results demonstrate clear performance advantages for Arabic-specific models over their multilingual counterparts other than alclam model, which we were surprised with the performance of the model.

Model	F1-M	Acc.	Prec.	Rec.
XLM-RoBERTa	0.645	0.653	0.645	0.647
MARBERT	**0.707**	**0.712**	**0.705**	**0.710**
AraBERT-Twitter	0.630	0.658	0.669	0.616
BGE	0.588	0.617	0.631	0.585
ALCLAM Base v2	0.404	0.563	0.684	0.442
LaBSE	0.627	0.654	0.655	0.611

Table 2: Performance comparison across evaluated models on the test set.

5.2 Key Findings

The experimental results reveal several important insights:

Arabic-Specific Model Superiority: ALCALM, MARBERT, and AraBERT-Twitter achieved the highest performance scores, demonstrating the critical importance of language-specific pre-training for Arabic emotion classification tasks.

Multilingual Model Limitations: Bge-m3 showed substantially lower performance despite its broad language coverage, suggesting that multilingual models may not effectively capture the subtle linguistic nuances required for Arabic emotion classification, also XLM-RoBERTa, although higher than both ALCLAM and AraBERT-Twitter, still lower than the MARBERT which is trained on more than one arabic dialect.

Performance-Robustness Trade-offs: MARBERT achieved the best balance across all evaluation metrics, making it the most reliable choice for deployment scenarios requiring consistent performance.

5.3 Error Analysis

Detailed analysis of model predictions on the validation set reveals several patterns in classification errors:

1. **Implicit Hate Expression**: Subtle hate speech often misclassified as neutral content. Example: مايفعله ميسي في الملعب شكله اخذ منشطات لازم يفحصونه (What Messi does on the field looks like he took stimulants They need to test him) True: not_applicable, but contains subtle accusatory language that could be misinterpreted.

2. **Dialectal Variation Impact**: Regional dialects and informal expressions create clas-

sification challenges. Example: كسم امتحان انهـــارده (Damn today's exam) True: not_applicable, but vulgar dialectal expressions can be difficult to classify accurately.

3. **Context-Dependent Statements**: Expressions requiring broader context for accurate interpretation. Example: هذا يقطع الطريق على عصاباتهم المتحالفة معهم (This cuts the road on their allied gangs) True: not_applicable, but political references require contextual understanding for proper classification.

6 Conclusion

The experimental results provide valuable insights into the effectiveness of different architectural approaches for Arabic emotion classification. The consistent superiority of Arabic-specific models reinforces the importance of language-specialized pre-training for morphologically complex languages.

Model Architecture Insights: The performance gap between Arabic-specific and multilingual models suggests that the linguistic complexity of Arabic—including its rich morphology, dialectal variations, and unique emotional expression patterns—requires specialized model architectures trained on Arabic-specific corpora.

Task-Specific Challenges: Our error analysis reveals that the primary difficulties lie in detecting implicit emotional content rather than explicit expressions. This finding has important implications for system deployment, suggesting that additional context or multi-turn analysis might improve performance on ambiguous cases.

Practical Deployment Considerations: MARBERT's balanced performance across all metrics makes it the most suitable choice for production deployment, where consistent reliability is more important than peak performance on specific metrics.

The main challenges identified in our analysis include:

- **Class Imbalance Effects**: The dominance of neutral content in the shared task dataset continues to pose challenges for balanced classification performance.
- **Implicit Expression Detection**: Subtle emotional expressions, particularly sarcasm and implicit hate speech, remain difficult to accurately classify.
- **Dialectal Variation Impact**: Different Arabic dialects introduce additional complexity that current models handle with varying degrees of success.
- **Context Dependency**: Many emotional expressions require a wider conversational or situational context for an accurate interpretation.

We presented a systematic evaluation of transformer-based encoders for Arabic emotion classification in the MAHED shared task. The results show that Arabic-specific models, especially MARBERT, outperform multilingual alternatives in capturing morphological and dialectal nuances. Key challenges remain to detect implicit and sarcastic expressions, handle class imbalance, and address dialectal variation.

Future work will explore lightweight ensembles of Arabic-specific models, advanced training strategies (e.g., curriculum learning), and incorporating broader context or multimodal cues to improve subtle emotion detection.

Limitations: We focused on single-model fine-tuning with limited hyperparameter exploration, which may cap performance.

Acknowledgments

The authors thank the MAHED shared task organizers for providing the dataset(Zaghouani and Biswas, 2025b)(Zaghouani and Biswas, 2025a)(Zaghouani et al., 2024) and the evaluation framework. We also acknowledge the open-source community for the pre-trained models used in this work.

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 (Volume 1: Long Papers)*, pages 7088–7105, Online. Association for Computational Linguistics.

Wissam Antoun, Fady Baly, and Hazem Hajj. Arabert: Transformer-based model for arabic language understanding. In *LREC 2020 Workshop Language Resources and Evaluation Conference 11–16 May 2020*, page 9.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *Preprint*, arXiv:2402.03216.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Ahmed Murtadha, Alfasly Saghir, Bo Wen, Qasem Jamaal, Ahmed Mohammed, and Yunfeng Liu. 2024. Alclam: Arabic dialectal language model. *Arabic NLP 2024*.

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 Detailed Model Specifications

MARBERT Configuration:

```
TrainingArguments(
    output_dir="./checkpoints_marbert",
    eval_strategy="epoch",
    save_strategy="epoch",
    per_device_train_batch_size=64,
    per_device_eval_batch_size=128,
    num_train_epochs=3,
    learning_rate=1e-6,
    warmup_ratio=0.1,
    weight_decay=0.01,
    logging_strategy="epoch",
    save_total_limit=2,
    load_best_model_at_end=True,
    metric_for_best_model="f1_macro",
    greater_is_better=True,
    dataloader_num_workers=2
)
```

Model Architecture Details:

- **MARBERT**: 12 layers, 768 hidden dimensions, 12 attention heads
- **AraBERT-Twitter**: 24 layers, 1024 hidden dimensions, 16 attention heads
- **ALCALM**: 12 layers, 768 hidden dimensions, 12 attention heads
- **XLM-RoBERTa**: 12 layers, 768 hidden dimensions, 12 attention heads
- **LaBSE**: 12 layers, 768 hidden dimensions, 12 attention heads
- **BGE**: Variable architecture depending on specific variant

B Hardware and Runtime Details

All experiments were conducted on GPU-accelerated hardware with the following specifications:

- GPU: NVIDIA Tesla V100 with 32GB memory
- Training time: Approximately 2-4 hours per model for 3 epochs
- Framework: PyTorch 1.12+ with Hugging Face Transformers 4.21+
- Additional libraries: scikit-learn, numpy, pandas