NguyenTriet at MAHED Shared Task: Ensemble of Arabic BERT Models with Hierarchical Prediction and Soft Voting for Text-Based Hope and Hate Detection

Nguyen Minh Triet and Dang Van Thin
University of Information Technology-VNUHCM
Quarter 34, Linh Xuan Ward, Ho Chi Minh City, Vietnam
Vietnam National University, Ho Chi Minh City, Vietnam
23521652@gm.uit.edu.vn and thindv@uit.edu.vn

Abstract

We present the NguyenTriet system for the MAHED 2025 shared task on multimodal detection of hope and hate emotions in Arabic content (Zaghouani et al., 2025). The challenge was divided into three subtasks: text-based hate and hope speech classification in Arabic text; multitask emotion, offensive language, and hate detection in Arabic text with a hierarchical structure; and detecting hateful memes from multimodal text-image pairs. Our participation focused on Subtasks 1 and 2. For Subtask 1, we employed an ensemble of Arabic BERT models for multi-class classification. In Subtask 2, we implemented a hierarchical classification framework utilizing a similar ensemble methodology, where emotion predictions are leveraged through a cascaded pipeline architecture to inform downstream hate and offensive detection tasks. Our approach achieved macro-F_1 scores of 0.707 (3rd place) on Subtask 1 and 0.553 (2nd place) on Subtask 2.

1 Introduction

The detection of hope and hate emotions in multimodal Arabic content has become increasingly critical in the era of social media, where memes and text-based posts can rapidly disseminate polarizing messages (Zaghouani et al., 2024a; Alam et al., 2024b). The MAHED 2025 Shared Task addresses this challenge through three subtasks: (1) text-based hate and hope speech classification in Arabic text, (2) multitask emotion, offensive, and hate detection in Arabic text with a hierarchical structure encompassing emotion classification, offensiveness detection, and hate speech identification, and (3) multimodal hateful meme detection combining Arabic text and images. This task is particularly important for Arabic, a language with diverse dialects and cultural nuances, where automated detection can aid in moderating harmful content while promoting positive discourse (Zaghouani et al., 2025).

Our system employs transformer-based models fine-tuned on the provided datasets, leveraging ensemble techniques and emotion-aware inputs to handle the hierarchical nature of the subtasks. For Subtask 1, we focus on multi-class classification of memes into hate, hope, or not_applicable categories using soft voting ensembles of Arabic-specific BERT variants. For Subtask 2, we adopt a cascaded pipeline that first predicts emotions, then incorporates these predictions into offensiveness and hate detection models. Key findings include the effectiveness of emotion integration in improving downstream tasks and the robustness of ensembles in handling class imbalances. Experiments demonstrate that our ensemble approach enhances performance on imbalanced datasets, with final scores of 0.707 (ranking 3rd) on Subtask 1 and 0.553 (ranking 2nd) on Subtask 2. Our approach achieved competitive rankings, highlighting challenges such as label imbalance, dialectal variations, disambiguating subtle emotions like pessimism due to limited examples, and dialectal ambiguity.

2 Background

2.1 Data

The dataset of MAHED 2025 includes Modern Standard Arabic (MSA) and various dialects, with genres primarily from social media content such as tweets and memes. All content was collected from public social media, anonymized, and annotated by native speakers. The task setup involves three subtasks:

Subtask 1 (Text-based Hate and Hope Speech Classification): Classifying Arabic text into three categories: 'hate' (content propagating hostility or prejudice), 'hope' (content inspiring positivity or optimism), or 'not_applicable' (neutral or unrelated content). The dataset (Zaghouani et al., 2024b; Zaghouani and Biswas, 2025b) contains 9,843 instances with notable class imbalance:

Proceedings of The Third Arabic Natural Language Processing Conference, pages 585–589

'not_applicable' dominates at 53.36%, followed by 'hope' (27.65%) and 'hate' (18.97%).

Subtask 2 (Emotion, Offensive, and Hate Detection - Multitask): Hierarchical classification framework with three sequential stages: (1) emotion classification among 12 categories (neutral, anger, anticipation, disgust, fear, joy, love, optimism, pessimism, sadness, surprise, trust), (2) binary offensiveness detection, and (3) conditional hate classification applied only to offensive content. The dataset (Zaghouani et al., 2024b; Zaghouani and Biswas, 2025a) comprises 8,515 instances with significant imbalances across all levels: 'anger' dominates emotions (25.94%) while 'fear' represents only 0.91%; offensiveness skews toward 'no' (70.79%); hate labels favor 'not_hate' (82.40% among offensive samples).

Subtask 3 (Multimodal Hateful Meme Detection): Binary classification of Arabic memes requiring analysis of both visual content and embedded Arabic text to determine 'hateful' or 'non-hateful' labels. The dataset (Alam et al., 2024b) contains 3,561 instances with class imbalance favoring 'non-hateful' memes (75.31%).

2.2 Related Work

Related work encompasses several key efforts in Arabic NLP for hate speech, emotion detection, and multimodal analysis.

Prior studies have explored multi-label classification of hate speech from social media tweets and focused analyses of harmful content, providing baselines for binary or multi-class detection (Zaghouani et al., 2024a; Biswas and Zaghouani, 2025a).

Bilingual approaches to emotions and hope speech have advanced positive discourse identification through paired language modeling (Biswas and Zaghouani, 2025b).

In the multimodal domain, investigations into propagandistic content in Arabic memes have established baselines for detecting harmful visual-textual combinations (Alam et al., 2024b), with extensions employing multi-agent large language models for nuanced propaganda analysis (Alam et al., 2024a).

Furthermore, propaganda span annotation has utilized large language models for fine-grained identification in news articles and memes (Hasanain et al., 2024a,b), demonstrating the efficacy of LLMs in capturing subtle spans while often neglecting hierarchical emotion integration.

Participating in subtask 1 and 2, our contribution's novelty lies in combining soft-voting ensembles of Arabic-specific BERT models with a cascaded emotion-integrated pipeline for hierarchical detection. This approach enhances robustness against class imbalances and dialectal variations, outperforming prior single-model methods or non-cascaded ensembles by explicitly leveraging predicted emotions to inform offensiveness and hate predictions in a structured manner.

3 System Overview

3.1 Approach

Our system comprises key components, including text preprocessing, classifiers formed through soft voting ensembles of Arabic-specific BERT models, and a hierarchical structure designed for Subtask 2 to address the task's inherent hierarchical nature.

3.2 Text Preprocessing

A critical component of our system is the text preprocessing pipeline, which addresses challenges such as dialectal variations, noisy social media content (e.g., emojis, URLs, and mentions), and orthographic inconsistencies in Arabic script. The preprocessing function is implemented as follows:

- Demojize emojis to their Arabic descriptions using the `emoji` library.
- Strip tashkeel (diacritics), tatweel (elongation), and normalize ligatures with `pyarabic.araby`.
- Normalize alef maksura and teh marbuta using `camel_tools.utils.normalize`.
- Remove URLs, mentions, hashtags, and non-alphanumeric characters (except punctuation like !?.) via regular expressions.
- Remove Arabic stopwords from NLTK's Arabic stopwords list.

This pipeline reduces text length and noise, improving model focus on semantic content.

3.3 Pre-trained Models

We employed two Arabic-specific BERT models, both pre-trained on extensive Arabic social media corpora, to capitalize on their robust understanding of dialectal variations and informal language patterns characteristic of tweets and social media content:

- **MARBERTv2** (Abdul-Mageed et al., 2021): A comprehensive model designed to handle both Dialectal Arabic (DA) and Modern Standard Arabic (MSA). Pre-trained using masked language modeling (MLM) on a substantial corpus of approximately 1 billion Arabic tweets, this model demonstrates exceptional performance on social media-related NLP tasks across diverse Arabic linguistic varieties and regional dialects.

- **AraBERTv0.2-Twitter** (Antoun et al.): A specialized variant optimized specifically for Arabic dialectal content and Twitter-style communications, built upon the BERT-Base architecture. Through continued pre-training via MLM on approximately 60 million curated Arabic tweets, this model incorporates an extensive vocabulary of dialectal expressions and colloquialisms, making it exceptionally well-suited for processing noisy, abbreviated social media text with informal linguistic structures.

3.4 Systems Details

The systems for Subtasks 1 and 2 are built upon classifiers structured as follows.

Subtask 1: The architecture consists of a single multi-class classifier that receives processed text and performs classification into three labels: hate, hope, or not_applicable.

Subtask 2: The architecture employs a hierarchical structure comprising three classifiers: (1) an emotion classifier for 12 emotion categories, (2) a binary offensiveness classifier that incorporates the predicted emotion as additional context, and (3) a binary hate classifier applied only to samples predicted as offensive, similarly augmented with the emotion label. In the training phase, the three classifiers are trained sequentially: first the emotion classifier, followed by the offensiveness classifier, and finally the hate classifier. In this process, the predicted emotion from the emotion classifier is replaced with the true emotion label to ensure accurate context augmentation for the downstream offensiveness and hate classifiers.

A classifier comprises two models (MARBERTv2 and AraBERTv0.2-Twitter) that perform tokenization and prediction independently. We applied the simple soft voting technique to merge the predictions of the two models, in which we sum up the probability output of the two classifiers and choose the sentiment class with the highest probability as the final prediction.

The hierarchical architecture for Subtask 2 and a classifier architecture are illustrated in Figure 1.

4 Experimental Setup

Data split usage: We utilized the provided train, validation, and test sets for both subtasks. The training set was used exclusively for model training, the validation set for hyperparameter tuning and evaluation during development, and the test set for final evaluation. No data augmentation or splitting was applied beyond the provided sets.

Configuration Settings: All experiments were conducted using a P100 GPU on the Kaggle platform. For both subtasks, the hyperparameters selected to train the two models included a learning rate of 3e-5, weight decay of 0.1, batch size of 32 for MARBERTv2 and 16 for arabert-twitter, over 2 epochs. The loss function employed was a class-weighted CrossEntropyLoss to effectively handle class imbalances during training. Optimization was performed using AdamW with a cosine annealing learning rate scheduler.

Evaluation Metrics: Task evaluation metrics are summarized as macro-averaged F1-score for all subtasks, as per the official guidelines, emphasizing balanced performance across imbalanced classes.

External Tools and Libraries: transformers (v4.20.0), torch (v2.0.0), pandas (v2.0.0), numpy (v1.24.0), scikit-learn (v1.2.0), pyarabic (v0.6.14), emoji (v2.0.0), camel_tools (v1.2.0), nltk (v3.8.0), and scipy (v1.10.0).

5 Results

5.1 Official Results

The official evaluation was conducted on the test set using macro-averaged F1-score. Our ensemble system achieved a macro-F1 of 0.707, ranking 3rd on Subtask 1. For Subtask 2, the system obtained a macro-F1 of 0.553, ranking 2nd. These results represent the official submission scores. The top 3 teams' results in subtask 1 and 2 are demonstrated in Table 1 and 2.

Ranking	Team	Macro-F1
Top 1	HTU	0.723
Top 2	NYUAD	0.721
Top 3 (Ours)	NguyenTriet	0.707

Table 1: Top 3 rankings for Subtask 1 on the test set.

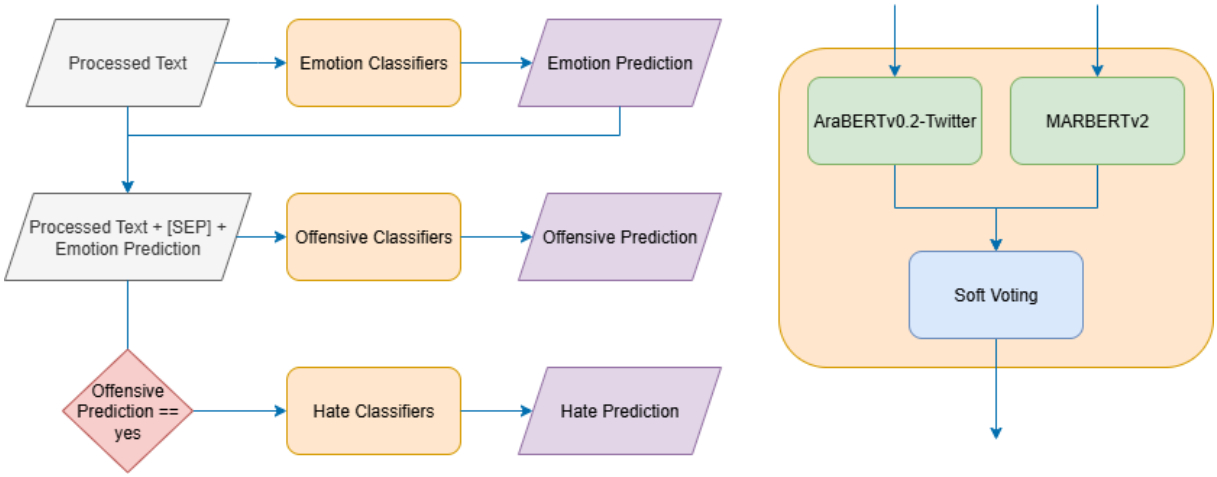

Figure 1: Hierarchical architecture for Subtask 2 (left) and a classifier architecture (right).

Ranking	Team	Macro-F1
Top 1	NYUAD	0.578
Top 2 (Ours)	NguyenTriet	0.553
Top 3	HTU	0.535

Table 2: Top 3 rankings for Subtask 2 on the test set.

5.2 Analysis

We first compare performance across different settings on the test set for Subtask 1, including individual models (MARBERTv2, AraBERTv0.2-Twitter) and the ensemble setting (combining MARBERTv2, AraBERTv0.2-Twitter using soft-voting). Table 3 summarizes the performance for Subtask 1.

Configuration	Macro-F1
MARBERTv2	0.692
AraBERTv0.2-Twitter	0.698
Ensemble	0.707

Table 3: Performance comparison for Subtask 1 across configurations on test set.

Next, for Subtask 2, we compare settings on the test set, distinguishing multiclass (non-hierarchical) and hierarchical configurations for individual models (multiclass MARBERTv2, multiclass arabert-twitter-large, hierarchical MARBERTv2, hierarchical arabert-twitter-large) and ensembles (multiclass Ensemble, hierarchical Ensemble). Table 4 summarizes the performance for Subtask 2.

These comparisons demonstrate the effectiveness of the ensemble architecture, which consistently outperforms individual models by 1-2% across both subtasks on test set, highlighting its role in enhancing robustness and reducing variance. Additionally, the hierarchical (cascaded) structure in Subtask 2 proves superior to multiclass approaches, improving macro-F1 by 4-5%, as it better captures dependencies between emotion, offensiveness, and hate predictions through contextual augmentation.

Configuration	Macro-F1
Multiclass MARBERTv2	0.483
Multiclass AraBERTv0.2-Twitter	0.490
Multiclass Ensemble	0.510
Hierarchical MARBERTv2	0.538
Hierarchical AraBERTv0.2-Twitter	0.547
Hierarchical Ensemble	0.553

Table 4: Performance comparison for Subtask 2 across configurations on test set.

6 Conclusion

In this paper, we presented our system for the MAHED 2025 Shared Task, which leverages Arabic-specific BERT ensembles with soft voting and a hierarchical cascaded pipeline for Subtask 2 to detect hope and hate emotions in Arabic content. Our approach achieved competitive results, demonstrating the effectiveness of ensemble methods and emotion augmentation in handling class imbalances and hierarchical dependencies.

Several promising directions emerge for enhancing system performance: implementing targeted data augmentation strategies for underrepresented classes, incorporating large language models (LLMs) to leverage their contextual understanding capabilities to more effectively address class imbalances.

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.

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 Ahmad, Md. Arid Hasan, and Maram Hasanain. 2024b. ArMeme: Propagandistic content in Arabic memes. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21071–21090, Miami, Florida, USA. 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.

Md. Rafiul Biswas and Wajdi Zaghouani. 2025a. An annotated corpus of arabic tweets for hate speech analysis. *CoRR*, abs/2505.11969.

Md. Rafiul Biswas and Wajdi Zaghouani. 2025b. Emohopespeech: An annotated dataset of emotions and hope speech in english and arabic. *CoRR*, abs/2505.11959.

Maram Hasanain, Fatema Ahmad, and Firoj Alam. 2024a. Large language models for propaganda span annotation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14522–14532, Miami, Florida, USA. Association for Computational Linguistics.

Maram Hasanain, Fatema Ahmed, and Firoj Alam. 2024b. Can gpt-4 identify propaganda? annotation and detection of propaganda spans in news articles. In *Proceedings of the 2024 Joint International Conference On Computational Linguistics, Language Resources And Evaluation*, LREC-COLING 2024, Torino, Italy.

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. 2024a. 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, Torino, Italia. ELRA and ICCL.

Wajdi Zaghouani, Hamdy Mubarak, and Md Rafiul Biswas. 2024b. 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.