LoveHeaven at MAHED 2025: Text-based Hate and Hope Speech Classification Using AraBERT-Twitter Ensemble

Nguyen Thien Bao, Dang Van Thin
University of Information Technology-VNUHCM,
Vietnam National University, Ho Chi Minh City, Vietnam
23520127@gm.uit.edu.vn
thindv@uit.edu.vn

Abstract

This paper presents our system for Sub-task 1 in MAHED 2025 (Zaghouani et al., 2025) shared task: Text-based Hate and Hope Speech Classification. We propose a robust pipeline built upon the bert-base-arabertv02-twitter model, leveraging domain-specific preprocessing, hyperparameter optimization with Optuna, and a K-Fold ensemble strategy. This system ranked 4 th among all participating teams on the leaderboard. We discuss technical design choices, the results of ablation studies, and the impact of preprocessing and model selection on final performance.

1 Introduction

Social media in the Arabic-speaking world exhibits a dynamic interplay between hateful and hopeful expressions, often entangled with rich dialectal diversity, code-switching, and informal orthography that complicate automatic detection. Beyond text, hateful content is increasingly conveyed via multimodal artifacts such as memes (Alam et al., 2024), motivating systems capable of analyzing both textual and visual modalities. Within this context, MAHED 2025 (Zaghouani et al., 2025) is organized as a shared task at ArabicNLP 2025 (co-located with EMNLP 2025), covering hope/hate and emotion detection in single-task, multi-task, and multimodal settings.

This paper presents a text-only system for Subtask 1, where the input is Arabic text (MSA or dialect) and the output is one of three labels: *hate*, *hope*, or *not_applicable*. The task evaluates systems by macro-averaged F1, a metric robust under class imbalance.

2 Related Work

Pre-trained transformer models for Arabic, notably AraBERT (Antoun et al., 2020), have established strong baselines on sentiment, dialect identification, and harmful content detection. AraBERTv0.2-Twitter (Antoun et al., 2020) extends this by further pretraining on a large corpus of tweets to better handle dialectal and informal Arabic. Recent datasets for harmful, offensive, and hopeful Arabic speech (Zaghouani and Biswas, 2025a; Zaghouani et al., 2024; Zaghouani and Biswas, 2025b) highlight the need for balanced evaluation metrics like macro-F1. For multimodal hateful content, studies such as (Alam et al., 2024) show the value of multimodal fusion techniques.

3 Background

3.1 Task Setup

Sub-task 1 requires a three-way classification: *hate*, *hope*, and *not_applicable*, for short Arabic text. The evaluation uses macro-F1 to handle class imbalance. In particular, the validation and test labels are concealed from participants. Predictions must be submitted to the official leaderboard to obtain macro-F1 scores, promoting strong generalization and preventing tuning on these datasets.

3.2 SubTask1 and its dataset

Sub-task 1 is a three-way classification problem: *hate*, *hope*, *not_applicable*. Input is short Arabic text in MSA or dialect. The dataset (Zaghouani et al., 2024) includes contributions from multiple platforms, with annotations performed manually by native speakers. Training set: 6,890 labeled samples; validation set: 1,476 unlabeled. Evaluation uses macro-F1 as the primary metric.

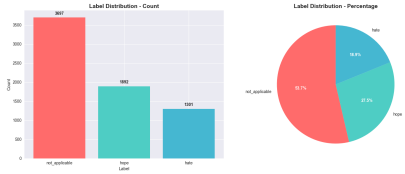

Figure 1: Label Distribution in training data

Proceedings of The Third Arabic Natural Language Processing Conference, pages 595–598

Key dataset observations:

- Quite imbalanced label distribution but can be acceptable (Figure 1).
- Short, noisy social-media texts — suitable for 128–256 BERT token length. After evaluating both configurations on the validation and test data, we found that a token length of 256 is more suitable for our pipeline in this task, providing better performance and results.(Figure 2)

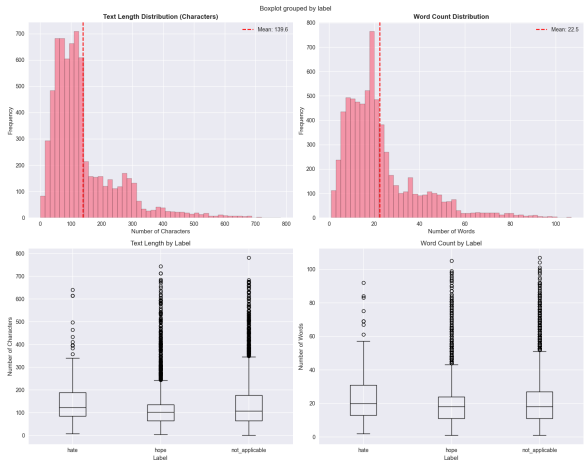

Figure 2: Text length distribution

4 System Overview

4.1 Duplicate handling

The dataset is of high quality with no missing values but has around 320 duplicate entries found in the training set.
Duplicate handling in the training dataset:

- Same text, conflicting labels → remove all.
- Same text, same label → keep one text.

The final label distribution after handling duplicate (Figure 3)

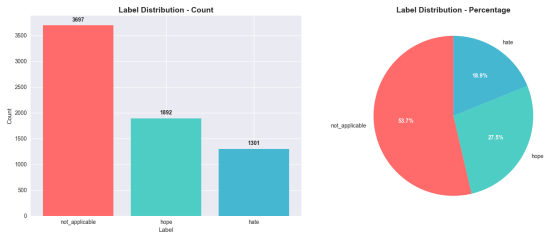

Figure 3: Label distribution after handling duplicate training set

4.2 Preprocessing

We import ArabertPreprocessor from arabert.preprocess for automatically handles (Antoun et al., 2020):

- Text normalization.
- Remove non-Arabic characters, URLs, mentions.
- Tokenize via HuggingFace AutoTokenizer (max_len=256).

4.3 Model and System

Our approach uses aubmindlab/bert-base-arabertv02-twitter (Antoun et al., 2020), pretrained on ∼60M tweets, alongside Arabic-aware preprocessing and Optuna-driven hyperparameter tuning. A 4-fold ensemble is used for robustness.(Figure 4)

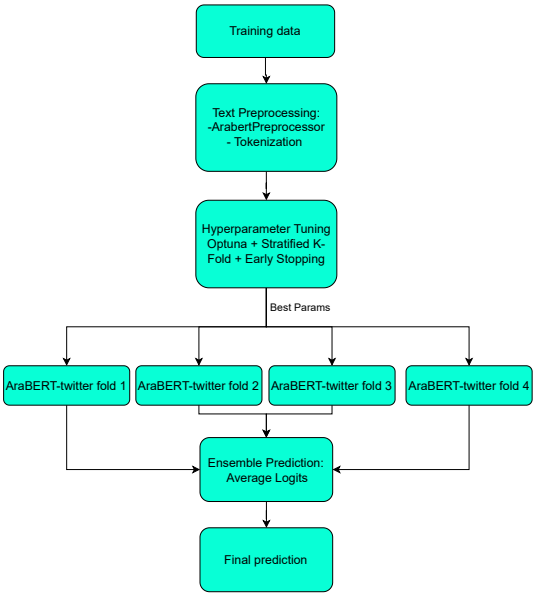

Figure 4: Pipeline of Technique in subtask 1

After having Best parameters from Optuna, the dataset was split into 4 folds using StratifiedK-Fold. We trained 4 separate AraBERT models from scratch, one for each fold, using the best parameters found by Optuna. The inference is based on the average logits from all folds.

5 Experimental Setup

5.1 Resources

We trained and evaluated all models on Kaggle Notebooks (free tier) with a single NVIDIA Tesla

P100 16GB GPU, 2 vCPUs, and approximately 13GB RAM. The environment used PyTorch 2.6.0, Transformers 4.52.4, and Optuna 4.4.0 on the default Kaggle Linux image (Python 3.11). Training sessions were constrained by free-tier time limits; each fold completed within a single session.

5.2 Hyperparameter Search

The Optuna experiment was run with 18 trials with tokenizer max_length = 256 and 30 trials with max_length = 128, both using early stopping (patience = 3). Due to resource constraints, the experiment completed all trials before the early stopping criteria were met. After evaluating on the validation and test data, we chose max_length = 256 with Optuna (n_trials = 18) as the better choice for our pipeline.
Loss: cross-entropy.
Optimizer: AdamW.
Scheduler: linear warmup-decay.
Hyperparameter tuning via Optuna:

- Learning Rate $\in [1 \times 10^{-6}, 1 \times 10^{-5}]$
- Batch size $\in \{8, 16, 32\}$
- Epochs $\in [2, 5]$

Best parameters: Learning Rate $\approx 9.74 \times 10^{-6}$, batch=8, epochs=4.

6 Results

We experimented with two state-of-the-art Arabic BERT models from the aubmindlab repository using the same pipeline.

Metric	AraBERT-Twitter	AraBERT
F1	0.6563	0.6403
Accuracy	0.6775	0.6511
Precision	0.6600	0.6343
Recall	0.6533	0.6477

Table 1: Comparison of AraBERT and AraBERT-Twitter on Validation data

Due to its higher F1-score on the validation data (0.66 compared to 0.64), the arabertv02-twitter model was selected for the final pipeline. Its specialization in social media text is particularly relevant to the dialectal and informal nature of the dataset. Moreover, the arabertv02-twitter model also outperformed the other model on the test data.

Metric	AraBERT-Twitter	AraBERT
F1	0.7030	0.7017
Accuracy	0.7130	0.7109
Precision	0.7100	0.7061
Recall	0.6990	0.6982

Table 2: Comparison of AraBERT and AraBERT-Twitter on Test data

To evaluate the impact of the ensemble approach, we compared our 4-fold StratifiedKFold ensemble against training a single `aubmindlab/bert-base-arabertv02-twitter` model on the full training set using the same best hyperparameters found via Optuna. The single-model setup slightly underperforms in the validation and test dataset compared to the ensemble, suggesting that ensembling mitigates variance and improves robustness, particularly under class imbalance conditions. This aligns with our observation that different folds capture complementary patterns in the training data.

6.1 Leaderboard

Our system ranked 4th on the official competition Leaderboard, with a Macro F1 score just 0.02 behind the top-ranked team. Our Accuracy and Precision placed us in the top 3, while our competitive recall (0.699) secured a position in the top 4. This result showcases a quite strong overall performance.

Rank	Team Name	Macro F1-score (Leaderboard)	Accuracy	Precision	Recall
1	HTU	0.723	0.725	0.717	0.730
2	NYUAD	0.721	0.723	0.716	0.729
3	AAA	0.707	0.712	0.705	0.710
3	NguyenTriet	0.707	0.705	0.692	0.737
4	LoveHeaven	0.703	0.713	0.710	0.699

Figure 5: The Final Leaderboard by macro-F1

7 Conclusion

We presented a competitive text-only system for MAHED 2025 Sub-task 1, ranking 4th by macro-F1 on the Leaderboard. In conclusion, our proposed AraBERT-based ensemble framework, optimized with Stratified K-Fold and Optuna for macro-F1, demonstrates significant effectiveness

in classifying Arabic text into hate, hope, and not_applicable categories, highlighting the potential of transformer-based models combined with ensemble learning for nuanced emotion detection in low-resource languages.

7.1 Limitations

This work uses text-only inputs; multimodal cues from images/memes are not modeled. Dialectal diversity and code-switching can reduce recall on minority or subtle cases, especially *hope* vs *not_applicable*. Label subjectivity around borderline cases can introduce noise across folds. Resource constraints (free-tier Kaggle Notebooks) limited the breadth of hyperparameter exploration.

7.2 Ethical Considerations

Misclassifying harmful content as benign can cause user harm and under-enforcement; human-in-the-loop moderation is recommended in high-stakes deployments. Data derived from social media may contain sensitive content and PII; usage should respect licensing, privacy, and minimize potential disparate impacts on dialect communities.

7.3 Future work

Future work includes expanding to multimodal inputs (images/memes), stronger dialect handling, and uncertainty-aware inference.

Acknowledgements

This research was supported by The VNUHCM-University of Information Technology's Scientific Research Support Fund.

References

Firoj Alam, Md Rafiul Biswas, Uzair Shah, Wajdi Zaghouani, and Georgios Mikros. 2024. 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.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. Arabert: Transformer-based model for arabic language understanding. *arXiv preprint arXiv:2003.00104*.

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

Full hyperparameters and code are available at: https://github.com/Limdim1604/MAHED2025