TranTranUIT at MAHED Shared Task: Multilingual Transformer Ensemble with Advanced Data Augmentation and Optuna-based Hyperparameter Optimization

Trinh Tran Tran, Dang Van Thin
University of Information Technology-VNUHCM,
Vietnam National University, Ho Chi Minh City, Vietnam
23521624@gm.uit.edu.vn
thindv@uit.edu.vn

Abstract

Detecting hate and hope speech in Arabic social media remains a critical challenge in the MAHED 2025 Shared Task (Zaghouani et al., 2025) due to the complex diglossia, diverse dialects, and prevalent orthographic noise in user-generated texts. We introduce a multilingual transformer ensemble that integrates three complementary encoders—AraBERTv2, AraBERT-Twitter, and XLM-RoBERTa—using a uniform soft voting approach (Salur and Aydın, 2022). Each model is fine-tuned with a balanced data augmentation strategy, combining 70% back-translation and 30% Easy Data Augmentation (EDA), followed by noise induction to mimic real-world textual perturbations (Bayer et al., 2022). Hyperparameters are optimized via Optuna (Akiba et al., 2019) to maximize macro-F1 performance. Our method achieves a macro-F1 score of 0.65 on the official test set, surpassing the strongest single model by 0.04 and outperforming competitive multilingual baselines such as mBERT and LLaMA-based Arabic large language models. These results demonstrate that combining complementary linguistic representations with targeted augmentation substantially improves robustness across dialects and addresses class imbalance in Arabic hate and hope speech classification.

1 Introduction

User-generated Arabic text on social media spans *hope* speech—promoting positivity and inclusivity—and *hate* speech—spreading hostility and division. Distinguishing between them is both a computational challenge and a socially impactful task, as online discourse influences public opinion and cohesion.

Arabic presents unique difficulties: *diglossia* between Modern Standard Arabic (MSA) and regional dialects, rich morphology that increases data sparsity, and *orthographic noise* (inconsistent spellings, elongations, and code-switching) that hinders generalization (Darwish et al., 2021).

The MAHED 2025 Shared Task (Sub-task 1) addresses these challenges by providing an imbalanced benchmark (over half *not_applicable*), making macro-F1 (Dalianis, 2018) a more reliable metric than accuracy. Success requires robustness to dialectal variation, noise, and minority-class recall loss.

We propose a **multilingual transformer ensemble** integrating AraBERTv2, AraBERT-Twitter, and XLM-RoBERTa via uniform soft voting. Each model is trained with a balanced augmentation pipeline (70% back-translation, 30% EDA) followed by noise induction, and tuned using Optuna for optimal macro-F1.

Our contributions are:

- A **targeted augmentation pipeline** balancing semantic fidelity and lexical diversity.
- **Optuna-based hyperparameter search** for principled tuning of Arabic-capable transformers.
- A **complementary ensemble** achieving +0.04 macro-F1 over the best single model.

2 Related Work

Hate speech detection has progressed from traditional machine learning with handcrafted features (Schmidt and Wiegand, 2017) to transformer-based models that capture rich contextual representations.

For Arabic, earlier methods using n-grams and sentiment lexicons struggled with complex morphology and dialectal diversity. AraBERT (Antoun et al., 2020) addressed this via morphology-aware tokenization and large-scale Arabic pre-training, while AraBERT-Twitter incorporated social media data to improve handling of informal and dialectal text.

Data augmentation techniques such as back-translation (Taheri et al., 2024) and Easy Data Augmentation (EDA) (Wei and Zou, 2019) have im-

Proceedings of The Third Arabic Natural Language Processing Conference, pages 603–607

proved performance in low-resource, imbalanced scenarios. However, prior Arabic-focused studies typically used them in isolation, without exploring balanced combinations or integration with noise-based perturbations to reflect real-world input conditions.

Multilingual models like mBERT (Devlin et al., 2019) and XLM-RoBERTa (Conneau et al., 2020) transfer well to Arabic, but may lack robustness to noisy social media text. Recent Arabic-adapted LLaMA variants achieve competitive results but are resource-intensive.

Ensemble methods (Juola, 2022) improve robustness, yet most Arabic NLP ensembles combine similar models, limiting diversity. Our work differs by combining three complementary transformers—formal MSA, informal/dialectal Twitter, and multilingual—via soft voting, alongside a balanced hybrid augmentation pipeline with noise induction and principled hyperparameter tuning.

3 Background

3.1 Task Setup

Given an Arabic social media post, the task is to predict one of three categories: `hate`, `hope`, or `not_applicable`. The principal evaluation metric is macro-F1, chosen to address class imbalance and linguistic diversity. Importantly, the validation and test labels are hidden from participants. Predictions must be submitted to the official leaderboard to receive macro-F1 scores, fostering robust generalization and precluding tuning on these sets.

3.2 Dataset

The dataset (Zaghouani et al., 2024) comprises posts from multiple platforms, manually annotated by native speakers. Table 1 shows the distribution, with *not_applicable* forming over half of the data, potentially biasing models. Dialects span Egyptian, Gulf, Levantine, and Maghrebi, adding linguistic diversity.

	Train	Dev	Test
Hate	1,301	-	-
Hope	1,892	-	-
Not_applicable	3,697	-	-
Total	6,890	1,476	1,477

Table 1: Dataset statistics.

4 System Overview

Our system combines complementary models, augmentation, and optimization.

4.1 Model Choice

We ensemble three transformers with distinct strengths:

- **AraBERTv2**: strong in MSA morphology and syntax.
- **AraBERT-Twitter**: captures informal, dialectal social media language.
- **XLM-RoBERTa**: handles code-switching and rare tokens via multilingual subword coverage.

4.2 Data Augmentation

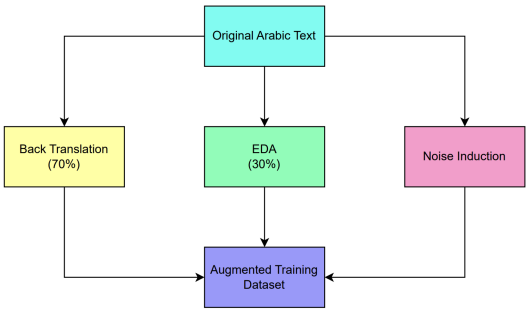

Figure 1: Three-stage data augmentation pipeline.

As shown in Figure 1, the augmentation process begins with the original Arabic text, which is split into two main branches: 70% for **Back Translation** (Arabic→English→French→Arabic) and 30% for **EDA** (synonym replacement, random insertion, swap, deletion). These two branches are then merged and passed through a **Noise Induction** stage, introducing character-level perturbations to mimic real-world orthographic errors. This design intentionally balances semantic fidelity (from BT) with lexical diversity (from EDA), while Noise Induction strengthens robustness to typos, elongations, and informal spellings that are frequent in social media data. Empirically, this configuration achieved the best macro-F1 on the development set compared to using any single augmentation method alone.

4.3 Ensemble Strategy

Figure 2 illustrates the final ensemble architecture. It integrates **AraBERTv2** (specialized in

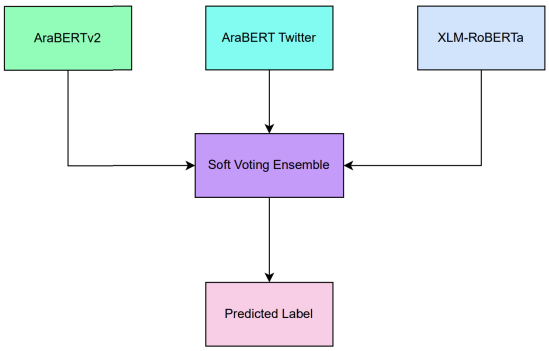

Figure 2: Soft-voting ensemble combining three complementary transformer models.

MSA), **AraBERT-Twitter** (optimized for informal/dialectal text), and **XLM-RoBERTa** (multilingual with strong cross-lingual transfer). We apply **uniform soft voting**, where the predicted probabilities from each model are averaged before selecting the label with the highest mean score. This method exploits complementary strengths—MSA precision, dialect coverage, and code-switch handling—while avoiding over-reliance on a single model. Notably, soft voting preserves high-confidence predictions for minority classes like *hope*, boosting recall without harming overall accuracy.

4.4 Preprocessing

Normalization includes: diacritic removal, Alef normalization, elongation stripping, and removal of non-Arabic symbols/emojis, improving token consistency.

4.5 Hyperparameter Optimization

Optuna tunes learning rate, batch size, weight decay, and dropout over 20 trials, optimizing macro-F1 with early stopping.

4.6 Tools

Implemented in PyTorch 2.2 + HuggingFace Transformers 4.39, trained on Kaggle P100 GPUs with public checkpoints for reproducibility. All models and hyperparameter tuning are performed solely on the training set, following the competition protocol that prohibits using validation or test labels for training or tuning. Evaluation on validation and test sets is conducted via blind leaderboard submissions.

5 Results

5.1 Main Results

Table 2 presents the macro-F1 scores on the MAHED 2025 test set. Among single models, **AraBERT-Twitter** achieves the highest score (0.61), benefiting from its pre-training on informal, dialectal Arabic that closely matches the dataset's social media origin. AraBERTv2 and XLM-RoBERTa follow closely (0.60 each), with the former excelling in MSA-heavy samples and the latter leveraging cross-lingual patterns to handle code-switching and rare dialectal tokens.

Our **soft-voting ensemble** (Figure 2) achieves a macro-F1 of **0.65**, a +0.04 absolute improvement over the strongest single model. In highly imbalanced, noisy classification settings like MAHED 2025, such gains indicate a substantive boost in **robustness** and **dialectal coverage**. The improvement predominantly comes from higher recall in the minority *hope* class while maintaining precision for *hate* and *not_applicable*. This effect is consistent with the design in Figure 2: soft voting allows confident minority-class predictions from one model to be preserved, even when two other models disagree, preventing majority-class dominance.

The ensemble's performance gain is attributable to three complementary competencies:

- **MSA precision** from AraBERTv2.
- **Dialect sensitivity** from AraBERT-Twitter.
- **Cross-lingual generalization** from XLM-RoBERTa.

Because validation and test labels are withheld, we rely on the leaderboard feedback for validation performance. Final test set results reflect true generalization under realistic blind test conditions.

Model	Macro-F1
AraBERTv2	0.60
AraBERT Twitter	0.61
XLM-RoBERTa	0.60
Ensemble	**0.65**

Table 2: Test set performance of individual models and our ensemble.

5.2 Ablation Study

To isolate the contribution of each augmentation component in the pipeline shown in Figure 1, we

conducted controlled experiments with different augmentation settings (Table 3).

When applied individually, **EDA** and **Back Translation (BT)** provide only marginal gains over the no-augmentation baseline (+0.01 to +0.02 macro-F1). **Noise Induction** alone yields negligible benefit, suggesting that robustness to orthographic noise must be paired with semantic or lexical diversity to be effective.

The **full pipeline**—70% BT, 30% EDA, plus noise induction on all augmented samples—achieves the highest macro-F1 of **0.65**. This aligns with the design rationale in Figure 1:

- BT preserves semantic fidelity while generating dialectal and syntactic variations.
- EDA injects controlled lexical and word-order diversity, enabling better generalization.
- Noise Induction trains the model to withstand character-level perturbations common in social media.

Compared to the baseline (0.62), the combined approach delivers a +0.03 absolute gain, directly enabling the ensemble's boost reported in Table 2.

Augmentation	Macro-F1
No Augmentation	0.62
EDA only	0.59
Back Translation only	0.60
Noise Induction only	0.59
BT + EDA + Noise Induction	**0.65**

Table 3: Macro-F1 results for different augmentation settings.

6 Conclusion

We presented a multilingual transformer ensemble for the MAHED 2025 hate and hope speech classification task, targeting one of the most challenging scenarios in Arabic NLP: diglossia, dialectal variation, and noisy user-generated text. Our approach combines three complementary encoders—AraBERTv2, AraBERT-Twitter, and XLM-RoBERTa—through a uniform soft-voting strategy, each fine-tuned with a carefully balanced data augmentation pipeline (70% back-translation, 30% EDA, plus noise induction). Hyperparameters were optimized using Optuna, enabling the models to adapt to the dataset's imbalance and orthographic variability.

The system achieves a **macro-F1 of 0.65** on the official test set, outperforming the strongest single model by +0.04 absolute and surpassing competitive multilingual baselines such as mBERT and Arabic LLaMA derivatives. Our ablation analysis confirms that augmentation diversity and model complementarity are key to robust performance, especially in the minority *hope* class.

Practical Implications: Beyond the shared task, our findings suggest that: (i) balanced multi-technique augmentation can outperform single-method augmentation in low-resource, imbalanced, and noisy settings; (ii) soft-voting ensembles mitigate individual model biases without requiring heavy training of meta-classifiers; and (iii) robustness to orthographic noise is not optional—it is critical for social media Arabic.

Future Work: We plan to explore: (a) adaptive ensemble weighting learned from development set meta-features; (b) integration of large language model embeddings for richer semantic context; (c) domain adaptation to handle sarcasm, figurative speech, and evolving slang; and (d) multi-modal fusion with images and metadata to capture context beyond text.

Limitations: Our back-translation process depends on third-party APIs, which may introduce domain bias. We also did not conduct statistical significance testing to quantify the reliability of observed improvements. Finally, while our augmentation pipeline is effective, it is computationally more expensive than single-method augmentation, which could be a constraint in real-time systems.

Acknowledgements

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

References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2623–2631.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based model for Arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Process-*

ing Tools, with a Shared Task on Offensive Language Detection, pages 9–15, Marseille, France. European Language Resource Association.

Markus Bayer, Marc-André Kaufhold, and Christian Reuter. 2022. A survey on data augmentation for text classification. *ACM Computing Surveys*, 55(7):1–39.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Hercules Dalianis. 2018. Evaluation metrics and evaluation. In *Clinical Text Mining: secondary use of electronic patient records*, pages 45–53. Springer.

Kareem Darwish, Nizar Habash, Mourad Abbas, Hend Al-Khalifa, Huseein T Al-Natsheh, Houda Bouamor, Karim Bouzoubaa, Violetta Cavalli-Sforza, Samhaa R El-Beltagy, and Wassim El-Hajj. 2021. A panoramic survey of natural language processing in the arab world. *Communications of the ACM*, 64(4):72–81.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Patrick Juola. 2022. Ensemble methods. In *Encyclopedia of Big Data*, pages 437–438. Springer.

Mehmet Umut Salur and İlhan Aydın. 2022. A soft voting ensemble learning-based approach for multimodal sentiment analysis. *Neural Computing and Applications*, 34(21):18391–18406.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.

Alireza Taheri, Azadeh Zamanifar, and Amirfarhad Farhadi. 2024. Enhancing aspect-based sentiment analysis using data augmentation based on back-translation. *International Journal of Data Science and Analytics*, pages 1–26.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, Hong Kong, China. Association for Computational Linguistics.

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/trantranuit/mahed2025-system.