DesCartes-HOPE at MAHED Shared task 2025: Integrating Pragmatic Features for Arabic Hope and Hate Speech Detection

Leila Moudjari[1], Mélissa Hacene Cherkaski[1], and Farah Benamara[1,2]
[1] IRIT, Université de Toulouse, ANITI, Toulouse, France
[2] IPAL, CNRS-NUS-A*STAR, Singapore
firstname.lastname@irit.fr

Abstract

This work presents our system for MAHED 2025 Task 1, which focuses on classifying Arabic text into Hope Speech, Hate Speech, or Not Applicable. Our approach combines dialect-aware contextual embeddings with pragmatic features—including speech acts, irony detection, and emotion cues—to capture the nuanced ways in which hope and hate are expressed across diverse Arabic varieties. We also employ targeted data augmentation to improve robustness in underrepresented categories. Experimental results show that incorporating speech act and emotion information significantly enhances detection performance. This approach allowed us to secure the fifth place in the official ranking[1] out of 60 participants, 25 of whom appeared on the final leaderboard, with a macro-F1 score of 0.7010. Our results are promising and mark a first step towards speech-act-aware hope/hate detection for Arabic social media.

1 Introduction

Online hate speech poses serious challenges globally, eroding social cohesion and enabling marginalization. In the Arabic-speaking world, these challenges are exacerbated by the rich tapestry of dialects—such as Egyptian, Levantine, Gulf, Maghrebi—and the frequent use of figurative language, such as irony, that makes automated detection especially complex.

While Arabic hate speech detection has received considerable attention—with resources such as the ADHAR multi-dialect corpus providing richly annotated datasets across both Modern Standard Arabic and major regional dialects, facilitating high-performance classification systems (Charfi et al., 2024)—research on Arabic hope speech detection remains limited compared to other languages. Prior research on hope speech has explored a range of perspectives, from peace-oriented discourse (Palakodety et al., 2019) to multilingual detection for promoting inclusion (Chakravarthi, 2020). Other works have examined expressions of regret and past-oriented hope (Balouchzahi et al., 2023), and the expression of wish in products reviews and political discussions (Goldberg et al., 2009). Building on these prior works, the *CDB model* (Da Silva et al., 2025) introduces a more fine-grained and linguistically grounded classification of hope through counterfactual, desire and belief.

More recently, the EmoHopeSpeech dataset was introduced, a bilingual resource annotated for both emotions and hope speech in English and Arabic, offering fine-grained emotional labels and linguistic variety for deeper analysis (Zaghouani and Biswas, 2025). Additionally, the emergence of innovative computational frameworks has further advanced the study of hope speech: for instance, recent methods leverage emotion-aware modeling to better distinguish hope expressions from neutrality or negativity, However, these approaches remain relatively unexplored in the context of Arabic hope/hate speech detection, and have largely focused on emotions alone (Badawi, 2025).

Building on these insights, the MAHED 2025 Shared Sub-Task1 (Zaghouani et al., 2025) presents an opportunity to jointly study hope and hate speech within a unified framework that accounts for Arabic linguistic variation and rich emotional subtleties. In this work, we extend prior emotion-based approaches by newly incorporating additional pragmatic features—most notably speech acts and irony—alongside emotion categories (e.g., anger, sadness, joy, love). Our system combines dialect-aware transformer embeddings with these pragmatic and affective cues, supported by targeted data augmentation to improve robustness for underrepresented dialects and classes.

Our model achieves a strong F1 score of 0.7010,

[1]Team: IRIT_HOPE.

Proceedings of The Third Arabic Natural Language Processing Conference, pages 632–638

	hope	hate	not applicable	Total
Train-set	1,892	1,301	3,697	6,890
Validation-set	409	261	806	1,476
Test-set	422	287	768	1,477

Table 1: Distribution of the MAHED dataset across training, validation, and test splits, showing the balance of classes for Subtask 1 (hope, hate, and not applicable).

Train Dataset	hope	hate	not applicable	Total
MAHED	1,892	1,301	3,697	6,890
MAHED+subtasks2	1,892	1,604	3,697	7,193
MAHED+MLMA	1,892	2,730	3,697	8,319
MAHED+Synthetic	3,226	1,301	3,697	8,224
MAHED+MLMA+Synthetic	3,226	2,730	3,697	9,653

Table 2: Statistics of the original MAHED train dataset and its augmented variants across the *hope*, *hate*, and *not applicable* classes.

securing fifth place in the official MAHED ranking. In the following sections, we detail our methodology, highlight the impact of speech-act and emotion features, and discuss avenues for advancing hope/hate detection in Arabic social media.

2 Task Overview

MAHED 2025 Task 1 focuses on classifying Arabic social media posts into three categories: hope, hate, or not applicable. The input consists of raw Arabic tweets, encompassing both Modern Standard Arabic (MSA) and various dialects. The output is a single categorical label. For instance:

1. "معاً يمكننا بناء مستقبل أفضل لأطفالنا" (Together, we can build a better future for our children.) → hope
2. "كل المهاجرين لصوص ومجرمون يجب طردهم فوراً" (All immigrants are thieves and criminals who should be expelled immediately.) → hate
3. "اليوم هو يوم مشمس وجميل" (Today is a sunny and beautiful day.) → not applicable

Dataset Details. The MAHED 2025 dataset for sub-task 1 comprises 9,843 annotated instances, divided into training, development, and test sets as shown in table 1.

All instances were collected from public social media platforms, anonymized, and annotated by native speakers, ensuring a Cohen's Kappa agreement greater than 0.85.

3 System Description

Our system for the MAHED2025 shared task builds upon our previous work on exploiting language models for Arabic text classification (Moudjari et al., 2021; Moudjari and Benamara, 2025). In this section, we detail the preprocessing pipeline, data augmentation strategies, and feature integration methods used in our approach.

3.1 Data Augmentation

The MAHED train dataset exhibits a notable class imbalance (see Table 2), with both the *hate* and *hope* categories significantly underrepresented. To mitigate this imbalance and improve model robustness, we implemented targeted data augmentation strategies for these classes.

Hate Class Augmentation. We augmented the hate speech data by incorporating additional annotated instances from two sources:

- **MAHED+subtasks2**: 303 hate-labeled examples were extracted from the second sub-task MAHED: Emotion, Offensive Language, and Hate Detection.
- **MAHED+MLMA**: 1,428 samples annotated with direct offensive and hateful sentiment labels were retrieved from the MLMA dataset (Ousidhoum et al., 2019).

Hope Class Augmentation. Due to the scarcity of publicly available Arabic hope speech datasets, we generated synthetic hope speech data using the ChatGPT-4o language model. By providing in-context examples from the MAHED dataset, we generated 1,334 additional hope-labeled instances designed to preserve domain relevance and linguistic characteristics. The newly augmented dataset is hereafter referred to as **MAHED+Synthetic**.
We instructed the model to generate several hundred Arabic texts covering a wide range of dialects—including Gulf, Egyptian, Maghrebi, and Levantine—and supplemented this with dedicated runs producing several hundred instances for each individual dialect to ensure balanced representation. For each run, we provided dialect-specific examples to guide generation (Figure 1 illustrates the prompts used).

We further combined the newly added inputs from **MAHED+MLMA** and **MAHED+Synthetic** to create **MAHED+MLMA+Synthetic**.

3.2 Enriching Datasets with Pragmatic Features

To provide richer input representations, we automatically augment the original MAHED dataset

You are a language model that generates **Arabic/Golf/Maghrebi/Egyptian/Levantine/diverse dialectal** social media comments expressing *hope*, *optimism*, and *encouragement*. Below are several tweets examples labeled as "hope" from the MAHED dataset. Please create new, original texts that are similar in style, tone, and dialect diversity. Avoid copying exact sentences. Focus on messages that inspire positivity, peace, and solidarity.

Examples:

1) يا مَن عَلِقتُ بِحَبلِ حُبِّه وَأَماطَ عَنّي ظِلَّ قُربِه لِمَ لا تَرِقُّ لِعاشِقٍ مُتَعَثِّرٍ بِذُيولِ كَربِه لا تُنبِتَن زَهرَ الصَبا بَةِ بِالقَلى في أَرضِ قَلبِه وَخَفِ العُقوبَةَ فَالقَتى لا بُدَّ أَن يُجزى بِذَنبِهِ

2) متضامنه

3) لا معايير تربطنا بالحب وتجعل ذلگ الحب اگثر برگهّ ان تعيشه دون ان تُدرگ انگ وقعت في شباگه ثغرشتاء ❄️🍃 قروب سفير للدعم …

Generate 10 new hope speech messages in Arabic.

Figure 1: Prompt design for data generation. The prompts guided the model to produce diverse, dialect-rich hope speech texts, ensuring both stylistic variation and balanced representation.

and its augmented variants with emotion, irony, and speech act labels. The emotion and irony annotations follow the configuration described in our previous work (Moudjari and Benamara, 2025) and are detailed in this section, while the speech act annotations follow Benamara et al. (2024). Appendix A provides further details on the datasets used for each annotation type (emotion, irony, and speech acts). Throughout the remainder of this paper, we denote each dataset d (either MAHED or one of its augmented variants) enriched with emotion, irony, and speech act features as d_{emo}, d_{irony}, and d_{SAct}, respectively.

Emotion Detection. We fine-tuned the AceGPT model (Huang et al., 2024) on the **Sem18$_{\text{MSA+Mixed}}$** dataset (Mohammad et al., 2018), which consists of Arabic tweets annotated for eleven emotions (anger, disgust, fear, joy, sadness, etc.). Although the prompting was done in MSA, the dataset itself contains both MSA and various dialectal forms, offering a rich and diverse training resource for emotion classification.

Irony Detection. For irony detection, we fine-tuned the same model on the **IDAT$_{\text{MSA+Mixed}}$** dataset (Ghanem et al., 2019), which consists of Arabic tweets labeled for binary irony classification (ironic vs. non-ironic). Similar to the emotion dataset, it includes a mix of MSA and dialectal varieties, enabling robust evaluation across linguistic registers.

Speech Acts. Following our previous work (Benamara et al., 2024), we employed the `arabertv02-twitter` model (Antoun et al., 2020), fine-tuned on the ArSAS$_{MSA+Mixed}$ dataset (Elmadany et al., 2018), to predict the underlying communicative function of each tweet. The model classifies speech acts into four categories: Subjective, Assertive, Interrogative, and Jussive — corresponding in Arabic to: موضوعي, تأكيدي, استفهامي, and أمري, respectively.

3.3 Model Architecture

Our final architecture builds on `arabertv02-twitter`,[2] a BERT-based model pretrained on a large corpus of Arabic tweets and adapted to the challenges of social media text, including dialectal variation, orthographic inconsistency, and noisy user-generated content. We fine-tuned this model for multi-class classification on the MAHED dataset and its augmented variants (see Table 2). The input text is tokenized and fed into the base model, and class probabilities are produced through a softmax output layer. Training is performed using weighted cross entropy, with early stopping based on the development set F1 score. We train the model for three epochs with a learning rate of $2e-5$, employing the Adam optimizer with an epsilon value of $1e-8$. The batch size is fixed at 16 for training and 128 for validation.

[2]It is worth noting that during the development phase, we submitted several runs using alternative embedding models, including `CAMeL-Lab/bert-base-arabic-camelbert-msa`, `CAMeL-Lab/bert-base-arabic-camelbert-mix`, `SI2M-Lab/DarijaBERT`, and `SI2M-Lab/DarijaBERT-mix`, as well as Arabic-centric large language models such as `FreedomIntelligence/AceGPT-v2-8B` and `FreedomIntelligence/AceGPT-v2-8B-Chat`. Notably, the AceGPT models gave results similar to `bert-base-arabertv02-twitter`. Nevertheless, `bert-base-arabertv02-twitter` proved to be the most effective model in our experiments, and thus we focus our reported results on this model.

3.4 Feature Integration

We explored several strategies for integrating pragmatic cues into the model:

Token-level Augmentation: The most effective approach was to append the predicted emotion, irony, and speech act labels directly to the raw input text prior to tokenization. This method consistently yielded the best performance across our experiments.

Separate Embedding Channels: We also experimented with multi-channel architectures, processing the original text and the additional cues in parallel before merging their representations. However, this approach did not lead to performance gains; in fact, it resulted in a $\sim 2\%$ drop in validation accuracy compared to token-level augmentation.

Normalised Log Feature Scaling: Since the direct insertion of categorical features into text was the most effective, we also experimented with a numeric encoding pipeline for these cues — first normalising label values (z-score), then scaling to $[0, 1]$, and finally applying a log transformation (`normalLog`). While this representation was numerically well-behaved and closer to direct token insertion in terms of accuracy, it did not outperform the plain token-level augmentation approach.

4 Results and Discussion

For all experiments, we used the official MAHED 2025 training set and its augmented version for model fitting and the development set for hyperparameter tuning and model selection. The results reported in Table 3 correspond to performance on the test set as evaluated on the Codabench platform. The final system submitted to the shared task was chosen based on its macro F-score during the development phase, then retrained on the full training data and evaluated by the organizers on the official test set to produce the leaderboard score.

Table 3 presents the experimental results obtained on the MAHED dataset and its augmented variants. Overall, the results show that augmentations incorporating emotion cues (`_emo`) and speech acts (`_SActs`) generally improve performance over the baseline. The best-performing configuration, MAHED+MLMA+Synthetic$_{\text{SAct}}$, reached a macro F-score of 0.7014, outperforming both our MAHED baseline (0.6400) and the official BERT baseline (0.5300). In contrast, adding irony features did not yield consistent gains—either in isolation or in combination with other features—suggesting possible redundancy or the introduction of noise in certain configurations. This outcome can be attributed to the fact that, upon inspection, we found that over 95% of the inputs in the file were labeled as non-ironic.

Test Dataset	F-score
ShardTask baseline	0.5300
Our baseline	0.6400
MAHED$_{\text{emo}}$	0.7000
MAHED$_{\text{irony}}$	0.6900
MAHED$_{\text{SAct}}$	**0.7010***
MAHED$_{\text{emo+irony}}$	0.6800
MAHED$_{\text{SAct+emo+irony}}$	0.6900
MAHED+subtasks2	0.6200
MAHED+MLMA	0.6900
MAHED+Synthetic	0.6800
MAHED+MLMA+Synthetic	0.6900
MAHED+MLMA+Synthetic$_{\text{emo}}$	**0.7007**
MAHED+MLMA+Synthetic$_{\text{irony}}$	0.6934
MAHED+MLMA+Synthetic$_{\text{SAct}}$	**0.7014**

Table 3: Macro F-scores of `bert-base-arabertv02-twitter` fine-tuned on the MAHED dataset and its augmented variants. The score marked with * corresponds to the official leaderboard submission, for which full precision is available. Bolded scores indicate newly obtained runs.

Dataset	hate	hope	not applicable
Our baseline	0.6643	0.5392	0.7081
MAHED$_{\text{emo}}$	0.703	0.6611	0.7237
MAHED$_{\text{SAct}}$	0.7038	0.669	0.7297
MAHED+MLMA+Synthetic$_{\text{emo}}$	0.7078	0.6757	0.7187
MAHED+MLMA+Synthetic$_{\text{SAct}}$	0.7094	0.6635	0.7314

Table 4: Class-wise macro F-scores for the baseline and top-performing augmented configurations on the official MAHED 2025 test set. Scores are reported for each class (*hate*, *hope*, and *not applicable*) along with the overall macro F-score.

Table 4 presents class-wise macro F-scores for the best-performing configurations, alongside our baseline system. In addition to the overall macro F-score, we report separate scores for the *hate*, *hope*, and *not applicable* classes. This breakdown allows us to assess whether specific augmentations, such as emotion or speech act features, offer balanced improvements across all categories or disproportionately benefit particular classes. Emotion features seem especially beneficial for improving the hope class, while speech acts give more balanced

improvements across classes, particularly boosting hate and not applicable. To assess whether the observed performance differences between models are statistically meaningful, we conducted McNemar's tests on paired classification outputs. Results revealed no significant differences among most top configurations, except for $\text{MAHED}_{\text{SAct}}$ over MAHED+MLMA+$\text{Synthetic}_{\text{SAct}}$ ($p = 0.0322$ using McNemar's test), see Appendix Section B for more detailed on these tests.

These findings underscore the importance of pragmatic and affective cues—particularly emotion and speech act information—in detecting hope and hate speech in Arabic social media.

5 Conclusion

We presented an approach to hope and hate speech detection for Arabic social media, leveraging dialect-aware contextual embeddings, pragmatic features (emotion, irony and speech act), and targeted data augmentation. Our results show that dialect sensitivity and augmentation substantially improve performance across Arabic varieties, and that incorporating affective and pragmatic cues—especially speech acts—yields further gains. These findings underscore the importance of modeling both linguistic diversity and communicative intent in fine-grained content moderation. Future work will explore contrastive learning to better disentangle hope and hate in the embedding space, as well as cross-task transfer from sentiment and stance datasets to enrich affective representations and enhance generalization.

Acknowledgments

This work has been supported by DesCartes: the National Research Foundation, Prime Minister's Office, Singapore under its Campus for Research Excellence and Technological Enterprise (CREATE) program. It has also benefited from the AI Interdisciplinary Institute ANITI under the Grant agreement n°ANR-23-IACL-0002.

Limitations

While our system demonstrates improved performance in detecting hope and hate speech across Arabic dialects, several limitations remain. First, the reliance on synthetic data—particularly for the under-represented hope class—introduces a risk of distributional mismatch between generated and naturally occurring texts. Second, our augmentation process covered only four major dialect families; smaller regional varieties remain underexplored. Third, pragmatic features such as irony and speech acts were derived from automatically predicted labels, which may propagate upstream errors into the final classification. Finally, our experiments were limited to the MAHED dataset, and generalizability to other genres (e.g., spoken discourse, formal writing) remains to be validated. Future work will address these issues by expanding dialectal coverage, improving the robustness of feature extraction, and testing cross-domain applicability.

References

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

Soran Badawi. 2025. Hopedetect: a multicomponent deep learning framework for hope detection in kurdish language. *The Computer Journal*.

Fazlourrahman Balouchzahi, Sabur Butt, Grigori Sidorov, and Alexander Gelbukh. 2023. Reddit: Regret detection and domain identification from text. *Expert Systems with Applications*, 225:120099.

Farah Benamara, Alda Mari, Romain Meunier, Véronique Moriceau, Leila Moudjari, and Valentin Tinarrage. 2024. Digging communicative intentions: The case of crises events. *Dialogue & Discourse*, 15(1):1–44.

Bharathi Raja Chakravarthi. 2020. Hopeedi: A multilingual hope speech detection dataset for equality, diversity, and inclusion. In *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, pages 41–53.

Anis Charfi, Mabrouka Besghaier, Raghda Akasheh, Andria Atalla, and Wajdi Zaghouani. 2024. Hate speech detection with adhar: a multi-dialectal hate speech corpus in arabic. *Frontiers in Artificial Intelligence*, 7:1391472.

Tulio Ferreira Leite Da Silva, Gonzalo Freijedo Aduna, Farah Benamara, Alda Mari, Zongmin Li, Li Yue, and Jian Su. 2025. Cdb: A unified framework for hope speech detection through counterfactual, desire and belief. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4448–4463.

A Elmadany, Hamdy Mubarak, and Walid Magdy. 2018. Arsas: An arabic speech-act and sentiment corpus of tweets. *Osact*, 3:20.

Bilal Ghanem, Jihen Karoui, Farah Benamara, Véronique Moriceau, and Paolo Rosso. 2019. Idat@fire2019: Overview of the track on irony detection in arabic tweets. In *Mehta P., Rosso P., Majumder P., Mitra M. (Eds.) Working Notes of the Forum for Information Retrieval Evaluation (FIRE 2019). CEUR Workshop Proceedings. In: CEUR-WS.org, Kolkata, India, December 12-15*, pages 10–13.

Andrew B Goldberg, Nathanael Fillmore, David Andrzejewski, Zhiting Xu, Bryan Gibson, and Xiaojin Zhu. 2009. May all your wishes come true: A study of wishes and how to recognize them. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 263–271.

Huang Huang, Fei Yu, Jianqing Zhu, Xuening Sun, Hao Cheng, Song Dingjie, Zhihong Chen, Mosen Alharthi, Bang An, Juncai He, Ziche Liu, Junying Chen, Jianquan Li, Benyou Wang, Lian Zhang, Ruoyu Sun, Xiang Wan, Haizhou Li, and Jinchao Xu. 2024. AceGPT, localizing large language models in Arabic. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8139–8163, Mexico City, Mexico. Association for Computational Linguistics.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. SemEval-2018 task 1: Affect in tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 1–17, New Orleans, Louisiana. Association for Computational Linguistics.

Leila Moudjari and Farah Benamara. 2025. Are dialects better prompters? a case study on arabic subjective text classification. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 17356–17371.

Leila Moudjari, Farah Benamara, and Karima Akli-Astouati. 2021. Multi-level embeddings for processing arabic social media contents. *Computer Speech & Language*, 70:101240.

Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019. Multilingual and multi-aspect hate speech analysis. In *Proceedings of EMNLP*. Association for Computational Linguistics.

Shriphani Palakodety, Ashiqur R KhudaBukhsh, and Jaime G Carbonell. 2019. Hope speech detection: A computational analysis of the voice of peace. *arXiv preprint arXiv:1909.12940*.

Wajdi Zaghouani and Md Rafiul Biswas. 2025. 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.

A Datasets

Sem18$_{MSA+Mixed}$ (Mohammad et al., 2018): We use the Emotion Classification (E-C) subset from the SemEval-2018 Task 1 "Affect in Tweets" challenge[3]. This dataset contains tweets collected in 2017 and manually annotated into 11 emotion categories: *Anger, Anticipation, Disgust, Fear, Joy, Love, Optimism, Pessimism, Sadness, Surprise*, and *Trust*.

IDAT$_{MSA+Mixed}$ (Ghanem et al., 2019): This dataset comprises tweets on various political issues and events in the Middle East from 2011 to 2018. The tweets are written in Modern Standard Arabic (MSA) as well as Egyptian, Gulf, Levantine, and Maghrebi dialects, and each tweet is manually labeled as Ironic or Not-Ironic.

ArSAS$_{MSA+Mixed}$ (Arabic Speech-Act and Sentiment Corpus of Tweets): is a manually annotated dataset comprising over 21,000 Arabic tweets drawn from diverse dialects and topics. Each tweet is labeled with one of six speech-act categories—Assertion, Expression, Recommendation, Question, Request, and Miscellaneous—as well as one of four sentiment labels: Positive, Negative, Neutral, or Mixed.

B Statistical Significance Analysis

Comparison	p-value
$\text{MAHED}_{\text{emo}}$ vs $\text{MAHED}_{\text{SAct}}$	0.3173
$\text{MAHED+MLMA+Synthetic}_{\text{emo}}$ vs $\text{MAHED+MLMA+Synthetic}_{\text{SAct}}$	0.1416
$\text{MAHED}_{\text{emo}}$ vs $\text{MAHED+MLMA+Synthetic}_{\text{emo}}$	0.3173
$\text{MAHED}_{\text{SAct}}$ vs $\text{MAHED+MLMA+Synthetic}_{\text{SAct}}$	**0.0322**

Table 5: McNemar's test p-values comparing top-performing configurations. Statistically significant results ($p < 0.05$) are in bold.

To assess whether the observed differences were statistically significant, we conducted McNemar's tests between the top configurations. The comparisons between $\text{MAHED}_{\text{emo}}$ and $\text{MAHED}_{\text{SAct}}$ ($p = 0.3173$), as well as between their augmented counterparts $\text{MAHED+MLMA+Synthetic}_{\text{emo}}$ and

[3] https://huggingface.co/datasets/SemEvalWorkshop/sem_eval_2018_task_1

MAHED+MLMA+Synthetic$_{\text{SAct}}$ ($p = 0.1416$), did not yield significant differences, indicating comparable performance. Similarly, the difference between MAHED$_{\text{emo}}$ and MAHED+MLMA+Synthetic$_{\text{emo}}$ was not significant ($p = 0.3173$). However, the comparison between MAHED$_{\text{SAct}}$ and MAHED+MLMA+Synthetic$_{\text{SAct}}$ showed a statistically significant improvement ($p = 0.0322$), suggesting that dataset augmentation benefits speech act–enriched models more consistently.