ANLPers at MAHED Shared Task: From Hate to Hope: Boosting Arabic Text Classification

Yasser Alhabashi[1] **Serry Sibaee**[1*] **Omer Nacar**[2] **Adel Ammar**[1] **Wadii Boulila**[1]
[1]Prince Sultan University, Riyadh, Saudi Arabia
[2]Tuwaiq Academy – Tuwaiq Research and Development Center
{yalhabashi, ssibaee, aammar , wboulila}@psu.edu.sa
{o.najar}@tuwaiq.edu.sa
[*]Corresponding author: ssibaee@psu.edu.sa

Abstract

The detection of harmful online content, including hate speech and propaganda, is particularly challenging in multimodal and multilingual contexts such as Arabic social media. This work addresses **Sub-task 1: Text-based Hate and Hope Speech Classification** in the MAHED2025 (Zaghouani et al., 2025) challenge, where the goal is to classify Arabic text into *hate*, *hope*, or *not_applicable*. We develop a system based on pre-trained Arabic BERT models with three fine-tuning strategies, combined with a custom preprocessing pipeline for noise removal, normalization, and diacritic stripping. To address class imbalance and lexical sparsity, we augment the training data with synthetically generated paraphrases via the OpenAI API. Experimental results on the official test set demonstrate that our best configuration, **BERT-base-AraBERTv02 + NN** with cleaning and generated data, achieves a macro-F1 score of **0.6747** F1. Error analysis reveals that mislabeled training instances significantly limit model performance, suggesting that future improvements may be achieved through systematic dataset refinement. Our approach highlights the importance of preprocessing, augmentation, and careful architectural choices for robust Arabic text classification.

1 Introduction

The rapid growth of social media has transformed the way information is produced, shared, and consumed, enabling unprecedented reach and immediacy. However, this openness has also facilitated the large-scale dissemination of harmful content such as hate speech, propaganda, and other forms of toxic communication. While such material may appear in text, images, or videos, multimodal formats like memes present a unique challenge for automated detection due to their combination of linguistic and visual cues, cultural references, and implicit meanings (Alam et al., 2024). These challenges are further compounded when hateful or propagandistic elements are intertwined, requiring models to capture subtle contextual overlaps between intent, emotion, and target.

Existing research has made significant progress in detecting harmful content across various modalities, languages, and levels of granularity. For instance, several studies have focused on annotating and analyzing large datasets for hate speech, offensive language, and related emotional attributes, particularly in underrepresented languages such as Arabic [(Zaghouani et al., 2024a),(Zaghouani and Biswas, 2025b)]. Others have highlighted the need to move beyond binary classification toward multi-label frameworks that capture target type, severity, and overlapping categories (Alam et al., 2024). Despite these advances, a number of persistent issues hinder progress: small and heterogeneous datasets, low inter-annotator agreement, inconsistent evaluation methodologies, and model performance drops when applied across domains or languages (Bäumler et al., 2025).

Moreover, most prior work treats modalities in isolation—either text-only or image-only—leaving limited exploration of their intersection, especially in contexts where textual and visual signals work jointly to convey harmful messages (Zaghouani et al., 2024a). Multilingual and cross-linguistic challenges remain especially acute, with the scarcity of high-quality annotated datasets further complicating model development (Zaghouani and Biswas, 2025b). Additionally, while transformer-based models such as BERT have shown strong performance in single-modality tasks (Bäumler et al., 2025), their application in complex, multimodal, multi-label scenarios remains underexplored.

Our work addresses these challenges through a novel approach that integrates advanced NLP pre-processing techniques with BERT-based model training, enabling more accurate and nuanced detection of harmful multimodal content. By leveraging fine-tuned linguistic preprocessing to normalize

Proceedings of The Third Arabic Natural Language Processing Conference, pages 590–594

and enrich textual data before BERT training, we improve the model's ability to capture subtle semantic and contextual cues that are often missed in raw text. This combination not only enhances classification accuracy in multi-label settings but also facilitates better generalization across different domains and linguistic varieties. In doing so, our approach bridges critical gaps identified in the literature and provides a scalable pathway toward more robust and context-aware harmful content detection systems.

2 Background

We use the Arabic hate/hope speech dataset introduced by (Zaghouani et al., 2024b; Zaghouani and Biswas, 2025c,a) as part of the **Sub-task 1: Text-based Hate and Hope Speech Classification** in the MAHED2025 shared task (Zaghouani et al., 2025). The goal is to classify Arabic text—either Modern Standard Arabic (MSA) or dialectal—into one of three categories:

- **Hate**: Hostile, offensive, or discriminatory content.
- **Hope**: Optimistic, encouraging, or positive sentiment.
- **Not Applicable**: Neutral text without hate or hope signals.

The input is an Arabic sentence, and the output is a label from the set {hate, hope, not_applicable}. Below (in Table 1) are example instances from the dataset, with their corresponding labels:

	Text	Transleted text	Label
1	@Kiatuh السلق قايمين على فيصل طلال عشانه مايعرف الارقوز 😂😂😂	@Kiatuh, the Saluq are after Faisal Talal because he doesn't know the clown 😂😂😂.	not_applicable
2	ولذبحنا أموالهم هم يدفعون عرب ولكن من يهود يرضعون ,https://t.co/8wA9kg03CX	And to slaughter our wealth, they pay Arabs, but from Jews they suckle. https://t.co/8wA9kg03CX	hate
3	Alamer : يامرحبامرحب يخص المير في ارض العرب ذي صحبته مكسب ولقيانه بشاره غاليهيا من كتب في القلب عنوانه ورقمه والل...	Alamer, welcome, welcome! He is dear to the Emir in the land of the Arabs. His friendship is a treasure, and meeting him is precious good news. O one whose name and number are written in the heart...	hope
4	لقد كان لدي شعور مميز للغاية بأني أستطيع الذهاب إلى أي مكان	I had a very special feeling that I could go anywhere.	hope

Table 1: Dataset instances.

The original training set contains **6,890** instances with columns `text` and `label`. A class distribution analysis reveals a moderate imbalance toward the *not_applicable* class, with a majority/minority ratio of approximately **2.84**. Table 2 shows the distribution.

Label	Count	Percent
hate	1,301	18.88%
hope	1,892	27.46%
not_applicable	3,697	53.66%
Total	6,890	100%

Table 2: Class distribution in the original dataset.

Texts in the dataset average **22.48** words (median **18**; 95th percentile **54**) and **139.64** characters (median **109.5**; 95th percentile **357**). These figures indicate that most inputs are relatively short, but there is a long tail of longer utterances. The observed imbalance motivates the use of macro-averaged metrics for evaluation and, during training, class-aware strategies such as re-weighting or targeted augmentation to improve model robustness across all categories.

3 System Overview

Our system is built upon pre-trained transformer-based Arabic language models, with multiple fine-tuning strategies. We explored three main architectures:

Variant A: BERT as Frozen Embeddings + Neural Network. We freeze the BERT encoder (Sibaee et al., 2024), compute average-pooled sentence embeddings, and train a feed-forward neural network. Two configurations were tested: one with 8 layers. All hidden layers use GELU activations and optional batch normalization.

Variant B: Fine-tuning BERT End-to-End. We fine-tune the BERT model directly for the classification task by updating all encoder parameters during training. A single linear classification head is applied on top of the pooled sentence representation.

Variant C: Fine-tuning BERT + Additional Fully Connected Layers. We fine-tune the BERT encoder and append two additional fully connected layers before the classification layer. These layers use GELU activations and optional batch normalization to capture higher-level abstractions.

All models incorporate our cleaning pipeline, which removes Latin characters, symbols, emojis, and Arabic diacritics, normalizes Unicode, and collapses extra spaces.

4 Experimental Setup

4.1 Data

We evaluate our models under three data preparation settings:

1. **Without Cleaning:** raw text as provided in the original dataset.
2. **With Cleaning:** applying the custom preprocessing function described in Section 4.2.
3. **With Cleaning + Generated Data:** combining cleaned text with synthetically generated paraphrases to increase lexical diversity.

Without augmentation, the training set contains 1,000 samples per class (3,000 total) and the validation set contains 250 samples per class (750 total). With generated data, the training set grows to 4,000 samples per class (12,000 total) and the validation set includes 300 samples per class (900 total). The official test set is provided by the task organizers.

Synthetic Data Generation. To address class imbalance and enhance linguistic diversity, we expanded the training set with **synthetically generated paraphrases** of existing samples. Paraphrases were produced using the **GPT4-mini** (OpenAI et al., 2024), guided by prompts designed to generate semantically equivalent Arabic sentences while preserving the original class labels. The generation process introduced lexical, structural, and stylistic variations without altering the underlying meaning, enabling the model to better generalize to unseen expressions.

table 3 presents examples of generated sentences alongside their corresponding labels.

	Generated data	Translated generated data	Label
1	أشعر بالكراهية تجاه الظلم.	I feel hatred toward injustice.	hate
2	استلمت الراتب اليوم وقررت ألغي خطتي الشرائية وأجلس في البيت أنظم ميزانيتي للأيام الجاية بهدوء.	I received my salary today and decided to cancel my shopping plans and stay home to calmly organize my budget for the coming days.	not_applicable
3	الشخص الناجح لا يستسلم بل يحاول	A successful person does not give up but tries many times.	hope

Table 3: Examples of synthetically generated Arabic data with corresponding labels.

4.2 Preprocessing

The custom text cleaning pipeline performs the following steps:

- Remove non-Arabic letters.
- Remove punctuation symbols.
- Remove emojis and pictographs.
- Remove Arabic diacritics.
- Remove diacritics from other languages via Unicode normalization.

Finally, multiple spaces are collapsed into a single space, preserving the core Arabic words.

4.3 Training Details

We use the `AdamW` optimizer with a linear decay learning rate schedule and warmup. Learning rates tested across experiments include 1×10^{-4}, 2×10^{-5}, 1×10^{-5}, and 1×10^{-6}. The batch size is fixed at 32. Early stopping is applied with a patience of 10 to prevent overfitting; no fixed epoch count is used. For most experiments, we use a dropout rate of 0.3, while for **Variant C** we additionally test a higher dropout rate of 0.7.

5 Results

Results are reported using the official evaluation metric (average macro-F1-score). Table 4 presents the validation and test **average macro-F1-score** for all model variants under the three data preparation settings. Our best **test** result is **0.6747**, achieved with **BERT-base-AraBERTv02 (Antoun et al., 2021) + NN** using cleaning, generated data, and a learning rate of 2×10^{-5}.

6 Error Analysis

Despite achieving competitive macro-F1 scores, our models' performance is limited by annotation quality. A manual review of a subset of the training data revealed a substantial proportion of mislabeled instances, which can mislead the learning process and reduce model generalization.

6.1 Quantitative Error Breakdown

We manually evaluated a random sample of 100 training examples. Out of these, 78 samples were correctly labeled, while 22 (22%) were found to be mislabeled. The dataset is heavily skewed toward the *not_applicable* class, followed by *hope* and *hate*, as shown in Figure 1. Figure 2 illustrates the number of correct vs. mislabeled samples for each class.

Model	LR	Clean	Gen. Data	Dropout	Val Avg. Macro-F1	Test Avg. Macro-F1
BERT-base-AraBERTv02	1e-5	Yes	Yes	0.3	0.6281	0.6504
BERT-base-AraBERTv02 + NN	2e-5	Yes	No	0.7	0.6386	0.6736
BERT-base-AraBERTv02 + NN	2e-5	Yes	Yes	0.3	0.6394	**0.6747**
BERT-base-AraBERTv02 embd + 8 layers	2e-5	Yes	Yes	0.3	0.5863	0.6235

Table 4: Comparison of experimental settings and corresponding validation/test macro-F1 scores. Settings are shown first for clearer interpretability.

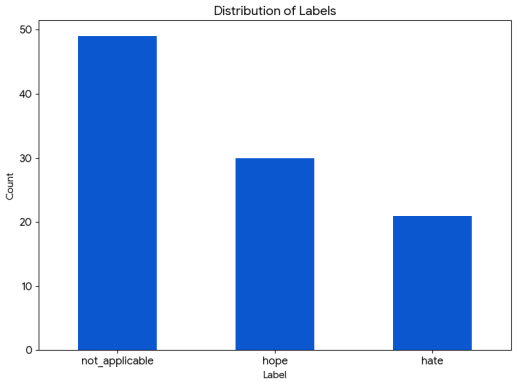

Figure 1: Distribution of studied samples

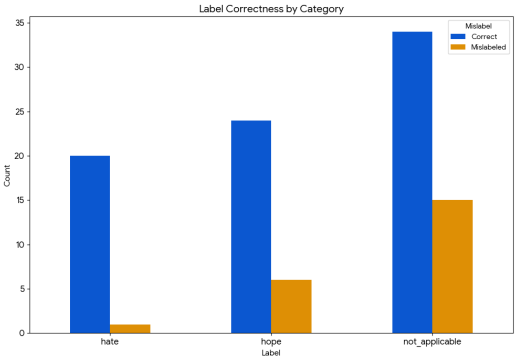

Figure 2: Label correctness by category, showing the number of correctly labeled vs. mislabeled samples per class.

6.2 Label Quality Summary

Figure 2 summarizes the distribution of correctly labeled vs. mislabeled samples by true class. While all three categories are affected by labeling errors, 'not_applicable' exhibits the highest mislabel rate relative to its class size (8 of 29 samples, ∼27.6%). Nearly a quarter of the reviewed data was mislabeled, highlighting that annotation noise is a major bottleneck. These findings suggest that systematic dataset relabeling or consensus-based annotation is crucial to improving model robustness (Sibaee et al., 2025), showing in Table 5 after correcting the labels for each category.

Labels	Correctly Labeled	After Correction	Total
Hate	21	8	29
Hope	30	7	37
Not_applicable	49	7	56
Total	78	22	100

Table 5: Breakdown of correctly labeled and mislabeled samples per true class in the manually reviewed subset, and after correcting each category.

7 Conclusion

In this paper, we introduced a BERT-based Arabic text classification system developed for the MAHED2025: Task-1 challenge, integrating tailored preprocessing, synthetic data generation, and multiple fine-tuning strategies. Our best configuration, combining AraBERT embeddings with additional neural network layers and generated data, achieved a macro-F1 score of 0.6747, demonstrating the effectiveness of our approach. However, manual error analysis revealed a considerable proportion of mislabeled instances in the dataset, which limits performance even with advanced models. Future work will focus on improving annotation quality through re-labeling or consensus-based methods, as well as exploring domain adaptation, cross-lingual transfer, and multimodal extensions to build more accurate, robust, and context-aware systems for harmful content detection in underrepresented languages like Arabic.

Acknowledgments

We would like to thank Prince Sultan University for their generous support in enabling this research.

References

Firoj Alam, Md. Rafiul Biswas, Uzair Shah, Wajdi Zaghouani, and Georgios Mikros. 2024. Propaganda tonbsp;hate: A multimodal analysis ofnbsp;arabic memes withnbsp;multi-agent llms. In *Web Information Systems Engineering – WISE 2024: 25th International Conference, Doha, Qatar, December 2–5,*

2024, Proceedings, Part V, page 380–390, Berlin, Heidelberg. Springer-Verlag.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2021. Arabert: Transformer-based model for arabic language understanding. *Preprint*, arXiv:2003.00104.

Julian Bäumler, Louis Blöcher, Lars-Joel Frey, Xian Chen, Markus Bayer, and Christian Reuter. 2025. A survey of machine learning models and datasets for the multi-label classification of textual hate speech in english. *Preprint*, arXiv:2504.08609.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Serry Sibaee, Abdullah Alharbi, Samar Ahmad, Omer Nacar, Anis Koubaa, and Lahouari Ghouti. 2024. ASOS at KSAA-CAD 2024: One embedding is all you need for your dictionary. In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 697–703, Bangkok, Thailand. Association for Computational Linguistics.

Serry Sibaee, Omer Nacar, Adel Ammar, Yasser Al-Habashi, Abdulrahman Al-Batati, and Wadii Boulila. 2025. From guidelines to practice: A new paradigm for arabic language model evaluation. *Preprint*, arXiv:2506.01920.

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. *Preprint*, arXiv:2505.11959.

Wajdi Zaghouani and Md Rafiul Biswas. 2025c. 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.