CIC-NLP at MAHED 2025 TASK 1:Assessing the Role of Bigram Augmentation in Multiclass Arabic Hate and Hope Speech Classification

Obiadoh A. E., Abiola O.J, Ogunleye T.D., Tewodros B.A.[1], Abiola T.O[1],
[1]Instituto Politecnico Nacional, Centro de Investigacion en Computacion, CDMX, Mexico.,
Correspondence: tabiola2025@cic.ipn.mx

Abstract

This study investigates the impact of bigram-based data augmentation on the joint classification of hate speech, hope speech, and neutral content in multilingual social media contexts, with a particular focus on Arabic. While previous research has shown the benefits of augmentation in text classification, its effectiveness in nuanced domains such as hate and hope speech remains underexplored. Using the annotated MAHED dataset, we compare three scenarios: a baseline without augmentation, global bigram augmentation, and classwise bigram augmentation. The baseline achieved 68.25% accuracy (macro-F1 = 0.6729) on the test set. Global bigram augmentation slightly reduced accuracy to 63.0% (macro-F1 = 0.62), showing no improvement over the baseline. Classwise augmentation achieved 93% accuracy on the validation set but dropped sharply to 59.65% accuracy (macro-F1 = 0.4726) on the test set, indicating severe overfitting. These results suggest that bigram-based methods are sensitive to class imbalance and may harm generalisation when applied unevenly across classes. We conclude by highlighting the need for more balanced, context-aware augmentation strategies in socially impactful NLP tasks.

1 Introduction

Hate speech and hope speech represent two critical yet contrasting forms of online expression. Hate speech fosters hostility, discrimination, and division (Alshahrani et al., 2025; **?**), while hope speech promotes unity, resilience, and positive social change (**?**). With the rapid growth of social media platforms, especially in multilingual and dialect-rich contexts such as Arabic, the automatic detection of these speech forms has become a pressing challenge. Although hate speech detection has received significant research attention (Al-Sukhani et al., 2025; Gasmi et al., 2025), hope speech detection remains comparatively underexplored, and the combined classification of both introduces unique complexities. These challenges include linguistic diversity, scarcity of high-quality annotated datasets, and the nuanced cultural and contextual variations in language use (Alrasheed et al., 2025).

Data augmentation has emerged as a promising strategy to improve the robustness and generalisation of natural language processing models, particularly in low-resource scenarios. Among these, bigram-based augmentation methods have shown success in enhancing text classification performance by enriching contextual co-occurrence patterns. However, their efficacy in nuanced, multiclass problems—such as joint hate and hope speech classification—remains uncertain. In this study, we investigate the impact of different bigram augmentation strategies, namely global and classwise augmentation, in comparison with a non-augmented baseline. Through a comprehensive empirical evaluation, we identify scenarios where augmentation may fail to deliver expected gains and discuss the implications for future work in socially impactful NLP applications.

2 Background

Recent advances in text classification have been driven by the adoption of Large Language Models (LLMs) across diverse domains. Early transformer-based approaches showed strong performance on complex linguistic tasks (Kolesnikova and Gelbukh, 2020; Adebanji et al., 2022), while more recent studies have explored fine-tuning and prompt-based methods for low-resource and multilingual contexts (Abiola et al., 2025c,b). Shared tasks and benchmarks (Ojo et al., 2023; Achamaleh et al., 2025) have further tested LLM robustness in noisy, real-world settings, and other works (Oladepo et al., 2025; Abiola et al., 2025a) have integrated contextual cues to improve predictive performance.

In the context of Arabic hate and hope speech

Proceedings of The Third Arabic Natural Language Processing Conference, pages 599–602

detection, challenges arise from dialectal diversity, morphological richness, and scarcity of annotated resources. The MAHED shared task (Zaghouani et al., 2025) addresses this by providing a labelled dataset with three categories: *hate*, *hope*, and *not_applicable*, encouraging participants to explore robust, generalisable classification approaches. Our submission focuses on a MARBERT-based pipeline with hybrid lexical–contextual augmentation via bigrams.

3 System Overview

Our system combines light preprocessing, a transformer encoder (MARBERT), and three bigram augmentation strategies. We use MARBERT (`UBC-NLP/MARBERT`) to capture deep contextual semantics and append frequent bigrams as explicit lexical cues. This design addresses two key challenges: (1) dialectal variation, by using MARBERT's pretraining coverage, and (2) sparse surface features, by injecting high-frequency n-grams into the input.

3.1 Preprocessing

We normalise Arabic text with the `ArabertPreprocessor` (AraElectra profile), preserving emojis to retain affective cues. No morphological segmentation is applied.

3.2 Bigram Augmentation

We explore:

- **Global-top**: top-K bigrams across the corpus, appended to all samples.
- **Class-specific**: top-K bigrams per class, appended based on ground-truth labels.
- **Unsupervised test-time**: predicted dominant class bigrams appended using overlap heuristics.

3.3 Training Setup

We compare:

1. **Baseline**: MARBERT with no augmentation (10 epochs).
2. **Hybrid**: MARBERT with bigram-augmented text (4 epochs).

Training uses AdamW (HuggingFace defaults), batch size $= 16$, maximum sequence length $= 128$, and model selection by validation macro-F1.

Class	Precision	Recall	F1	Support
0	0.59	0.63	0.61	238
1	0.62	0.55	0.58	359
2	0.69	0.71	0.70	729

Table 1: Validation metrics — Baseline.

Class	Precision	Recall	F1	Support
0	0.53	0.69	0.60	238
1	0.62	0.57	0.59	359
2	0.69	0.65	0.67	729

Table 2: Validation metrics — Global bigram augmentation.

4 Experimental Setup

The MAHED dataset is split into `train`, `val`, and `test` as per organisers. Labels are encoded via `LabelEncoder` for consistency. Evaluation metric: macro-F1 (primary), along with accuracy, precision, and recall.

5 Results

5.1 Validation Performance

The baseline achieved macro-F1 $= 0.63$ (accuracy $= 0.65$), with the majority class performing best. Global bigrams improved minority-class recall but reduced majority-class accuracy. Classwise bigrams yielded extremely high validation performance (macro-F1 $= 0.92$) but failed to generalise.

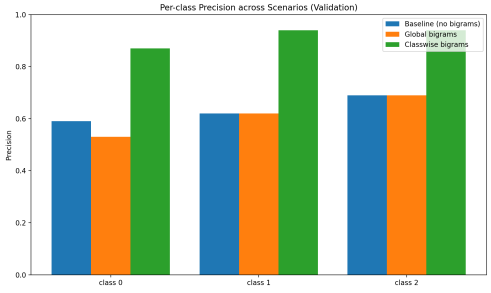

Figure 1: Per-class precision (validation).

5.2 Test Performance and Generalisation

The baseline maintained macro-F1 $= 0.6729$ on test data, while classwise bigrams dropped sharply to 0.4726 due to overfitting.

5.3 Error Analysis

Global bigrams: Provided minor recall gains for minority classes but reduced precision for the majority class.

Class	Precision	Recall	F1	Support
0	0.87	0.87	0.87	238
1	0.94	0.95	0.95	359
2	0.94	0.94	0.94	729

Table 3: Validation metrics — Classwise bigram augmentation.

Scenario	Accuracy	Precision	Recall	Macro-F1
Baseline (test)	0.6825	0.6742	0.6733	0.6729
Classwise bigrams (test)	0.5965	0.6802	0.4660	0.4726

Table 4: Test metrics: Baseline vs. Classwise bigrams.

Classwise bigrams: Boosted validation scores artificially by memorising label-specific tokens, which became noise in test scenarios.

Other factors: Token truncation and domain shift likely reduced augmentation benefits.

6 Conclusion

Global bigram augmentation offered only small gains, while classwise augmentation inflated validation results but failed in generalisation. This underscores the risk of label-tied augmentation in imbalanced, nuanced datasets and points to the need for label-agnostic, domain-robust augmentation strategies.

Acknowledgments

The work was done with partial support from the Mexican Government through the grant A1-S-47854 of CONACYT, Mexico, grants 20241816, 20241819, and 20240951 of the Secretaría de Investigación y Posgrado of the Instituto Politécnico Nacional, Mexico. The authors thank the CONACYT for the computing resources brought to them through the Plataforma de Aprendizaje Profundo para Tecnologías del Lenguaje of the Laboratorio de Supercómputo of the INAOE, Mexico and acknowledge the support of Microsoft through the Microsoft Latin America PhD Award.

7 Limitations

The small, imbalanced dataset may have skewed augmentation effects, with classwise augmentation risking overfitting for rare classes. We only tested bigram-based methods, leaving other strategies (e.g., paraphrasing, back-translation, contextual augmentation) unexplored. Evaluation was confined to in-domain data, so cross-domain and cross-dialect generalisation is uncertain. Finally,

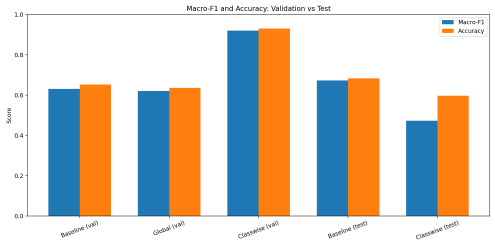

Figure 2: Macro-F1 and accuracy for validation and test.

we did not assess interpretability, which is important to prevent augmentation-induced bias.

Acknowledgments

References

Tolulope Abiola, Olumide Ebenezer Ojo, Grigori Sidorov, Olga Kolesnikova, and Hiram Calvo. 2025a. CIC-IPN at SemEval-2025 task 11: Transformer-based approach to multi-class emotion detection. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 1609–1615, Vienna, Austria. Association for Computational Linguistics.

Tolulope O. Abiola, Tewodros A. Bizuneh, Oluwatobi J. Abiola, Temitope O. Oladepo, Olumide E. Ojo, Adebanji O. O., Grigori Sidorov, and Olga Kolesnikova. 2025b. Cic-nlp at genai detection task 1: Leveraging distilbert for detecting machine-generated text in english. In *Proceedings of the 31st International Conference on Computational Linguistics (COLING)*, Abu Dhabi, UAE. Association for Computational Linguistics.

Tolulope O. Abiola, Tewodros A. Bizuneh, Fatima Uroosa, Nida Hafeez, Grigori Sidorov, Olga Kolesnikova, and Olumide E. Ojo. 2025c. Cic-nlp at genai detection task 1: Advancing multilingual machine-generated text detection. In *Proceedings of the 31st International Conference on Computational Linguistics (COLING)*, Abu Dhabi, UAE. Association for Computational Linguistics.

Tewodros Achamaleh, Tolulope Olalekan Abiola, Lemlem Eyob Kawo, Mikiyas Mebraihtu, and Grigori Sidorov. 2025. CIC-NLP@DravidianLangTech 2025: Detecting AI-generated product reviews in Dravidian languages. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 502–507, Acoma, The Albuquerque Convention Center, Albuquerque, New Mexico. Association for Computational Linguistics.

Olaronke Oluwayemisi Adebanji, Irina Gelbukh, Hiram Calvo, and Olumide Ebenezer Ojo. 2022. Sequential models for sentiment analysis: A comparative study. In *Advances in Computational Intelligence, 21st Mexican International Conference on Artificial*

Intelligence, MICAI 2022, Proceedings, Part II, Monterrey, Mexico. Springer.

Hassan Al-Sukhani, Qusay Bsoul, Abdelrahman H. Elhawary, Ziad M. Nasr, Ahmed E. Mansour, Radwan M. Batyha, Basma S. Alqadi, Jehad Saad Alqurni, Hayat Alfagham, and Magda M. Madbouly. 2025. Multilingual hate speech detection: Innovations in optimized deep learning for english and arabic hate speech detection. *SN Computer Science*, 6(205).

Sadeem Alrasheed, Suliman Aladhadh, and Abdulatif Alabdulatif. 2025. Protecting intellectual security through hate speech detection using an artificial intelligence approach. *Algorithms*, 18(4):179.

Eman S. Alshahrani, Mehmet S. Aksoy, and Ahmed Emam. 2025. Detection of hate speech and offensive language in arabic text: A systematic literature review. *Applied Computational Intelligence and Soft Computing*. First published: 13 April 2025.

Karim Gasmi, Ibtihel Ben Ltaifa, Alameen Eltoum Abdalrahman, Omer Hamid, Mohamed Othman Altaieb, and Shahzad Ali. 2025. Hybrid feature and optimized deep learning model fusion for detecting hateful arabic content. *IEEE Access*, 13.

O. Kolesnikova and A. Gelbukh. 2020. A study of lexical function detection with word2vec and supervised machine learning. *J. Intell. Fuzzy Syst.*, 39.

Olumide E. Ojo, Olaronke O. Adebanji, Hiram Calvo, Damian O. Dieke, Olumuyiwa E. Ojo, Seye E. Akinsanya, Tolulope O. Abiola, and Anna Feldman. 2023. Legend at araieval shared task: Persuasion technique detection using a language-agnostic text representation model. *Preprint*, arXiv:2310.09661.

Temitope Oladepo, Oluwatobi Abiola, Tolulope Abiola, Abdullah , Usman Muhammad, and Babatunde Abiola. 2025. Predicting emotion intensity in text using transformer-based models. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 1677–1682, Vienna, Austria. Association for Computational Linguistics.

Wajdi Zaghouani, Md Rafiul Biswas, Mabrouka Bessghaier, Shimaa Ibrahim, Georgio Mikros, Abul Hasnat, and Firoj Alam. 2025. Overview of mahed shared task: Multimodal detection of hope and hate emotions in arabic content. In *The Third Arabic Natural Language Processing Conference (ArabicNLP 2025)*, Suzhou. Association for Computational Linguistics.

A Example Appendix

This is an appendix.