HTU at MAHED Shared Task: Ensemble-Based Classification of Arabic Hate and Hope Speech Using Pre-trained Dialectal Arabic Models

Abdallah Saleh
Al Hussein Technical University
abdullahhsalehds@gmail.com

Mariam Biltawi
Al Hussein Technical University
Mariam.Biltawi@htu.edu.jo

Abstract

Emotional contagion, the phenomenon where emotions spread between individuals, shows the importance of detecting both hope and hate speech in digital communications. This emotional transmission can amplify positive sentiments that foster community resilience or propagate harmful content that divides societies. While hate speech detection in Arabic has been extensively studied, hope speech detection has received comparatively limited attention, creating an imbalance in the understanding of emotional influence online. To address this gap, MAHED 2025 sub-task 1 introduced the task of detecting both hope and hate speech using a substantial dataset designed for developing robust classification models. This paper presents an ensemble approach combining three Transformer-based encoder models with soft voting and weighted loss functions to address class imbalance issues. Those models, ArabicDeBERTa-DA, BERT-DA, and MARBERTV2, have been continually pre-trained on different domains of Arabic, showing the benefits of continual pre-training both on downstream performance and computational efficiency. The proposed ensemble model achieved the highest performance in the competition with an F1 macro score of 72.3% using an ensemble voting of the best-performing variants.

1 Introduction

Arabic NLP researchers have extensively investigated the problem of hate speech in Arabic, particularly through the construction of datasets and the development of detection systems. In contrast, hope speech has received considerably less attention, with only a few datasets available and, consequently, fewer detection systems. sub-task 1 of MAHED 2025 (Zaghouani et al., 2025) seeks to address this gap by introducing a dataset that incorporates both hate speech and hope speech.

The dataset encompasses two varieties of Arabic: Modern Standard Arabic (MSA) and Dialectal Arabic (DA). This diversity introduces significant challenges in developing robust detection systems, as the linguistic variation between these forms of Arabic is substantial. Most pre-trained Arabic language models have been trained almost exclusively on a single domain, which limits their ability to generalize effectively to this dataset. Given the high computational cost of pre-training from scratch, it is often impractical to train a new model entirely for an unseen domain. A practical alternative is continual training, whereby a pre-trained model is further adapted to a new domain through additional pre-training, not only gaining the ability to generalize to a new domain but also retaining performance on the original domain, if done correctly.

Only a limited number of Arabic language models have undergone continual pre-training, with notable exceptions such as MARBERTV2(Abdul-Mageed et al., 2021) and AraBERTv0.2 Twitter(Antoun et al., 2020). MARBERTV2 is an enhanced version of the original MARBERT model, specifically designed to better capture the multilingual and multi-dialectal nature of Arabic text. It was continually pre-trained on diverse Arabic corpora thereby improving its robustness for cross-dialectal tasks. Similarly, AraBERTv0.2 Twitter was continually pre-trained on Twitter data (specifically, approximately 60 million tweets), which enables it to better handle the DA semantic and syntactic language patterns prevalent on social media. This specialized pre-training makes it particularly well-suited for social media text classification tasks, such as the detection of hate and hope speech in informal Arabic discourse.

Ensemble learning in deep learning represents a powerful paradigm that combines predictions from multiple models to achieve superior performance compared to any individual model. This approach leverages the principle of diversity among models, where different models, architectures, training procedures, or data representations can capture

Proceedings of The Third Arabic Natural Language Processing Conference, pages 651–657

complementary/independent patterns and their corresponding errors in the data(Goodfellow et al., 2016). If models are diverse enough, which might translate to their errors being independent, ensemble models will perform significantly better than their members/submodels (Goodfellow et al., 2016). In the context of transformer-based models like BERT variants, ensemble methods can combine models that have been trained on different domains, use different tokenization algorithms or sizes, or have been fine-tuned with a range of hyperparameters. The ensemble typically aggregates predictions through techniques such as majority voting for classification tasks, weighted averaging of probability distributions, or more sophisticated methods like stacking, where a meta-learner is trained to optimally combine the base models' outputs.

The main contribution of this paper is the introduction of two continually pre-trained models, BERT-DA and ArabicDeBERTa-DA, which have been continually pre-trained on DA data. These models, when combined with MARBERTV2 in an ensemble of pre-trained language models, achieved state-of-the-art performance in sub-task 1 of MAHED 2025. The remainder of this paper is organized as follows: Section 2 reviews related work, Section 3 describes an overview of the proposed system and methodology, Section 4 presents the experimental setup, Section 5 presents the results, and Section 6 provides a discussion and conclusion.

2 Background

(Ke et al., 2023) have shown that continual pre-training of a general-domain language model to a specific domain increases the performance of the further pre-trained model on the target domain. They have proposed the continual domain-adaptive pre-training, continual DAP for short, methodology to allow for general-domain language models to continue training on domain specific data, ensuring both learning from the new domain, while avoiding catastrophic forgetting of the model's general knowledge.

(Lee et al., 2023) have shown, at least within the domain of computer vision, that continual learning doesn't always bring an increase in performance, especially in strong pre-trained models like CLIP. It was also shown that the algorithm used to training doesn't always have the performance boost when used on continued training phase.

(Ibrahim et al., 2024) have noticed that while adapting language models to new domain knowledge is more data efficient, the process of continued pre-training is challenging. As a sub-optimal continued pre-training can lead to either catastrophic forgetting of previously trained general knowledge of the model or poor performance on the new domain. Hence, different learning rate schedulers have been proposed to allow the model to learn without forgetting its past knowledge. The authors have shown one of the most efficient and simplest way to continually pre-train is to use a simple learning rate scheduler.

Continual pre-training remains a relatively underutilized strategy in Arabic NLP. While models such as AraBERT (Antoun et al., 2020) and MARBERTV2 (Abdul-Mageed et al., 2021) exemplify its application, explicit discussion of continual pre-training is scarce in the broader literature. AraBERT was extended via continual pre-training to adapt to domain-specific nuances such as social media, as in the AraBERTv02-Twitter variant, while MARBERTV2 further refines the original MARBERT by additional pre-training on MSA corpus with longer sequence lengths.

(Sarkar et al., 2022) have also investigated parameter and data efficient continual pre-training approaches for the Arabic language, showing that further adaptation of pre-trained multilingual models (mBERT) with DA can have robust performance. Beyond these instances, continual pre-training in Arabic remains largely unexplored.

(Anezi, 2022) addressed the need for an intelligent system that can detect hate speech in Arabic, as the author highlighted its importance for national security and combating issues like cyberbullying. The author introduced a new dataset of 4,203 Arabic social media comments, which are classified into seven distinct categories: content against religion, racist content, content against gender equality, violent content, insulting/bullying content, normal positive comments, and normal negative comments. This dataset is notably larger and more granular in its classification than most existing Arabic hate speech datasets. The core of the study is a proposed deep recurrent neural network (RNN) model, called DRNN-2. The model's performance was evaluated on three different classification tasks: binary (positive vs. negative), three-class (positive, negative, and hate speech), and the full seven-class classification. The DRNN-2 model achieved a training accuracy of 99.73% for binary classification, 95.38% for the three-class task, and 84.14% for the

seven-class task. These results are reported to be higher than those of similar methods in the current literature, demonstrating the model's effectiveness in tackling the complexities of Arabic text and providing a potentially valuable tool for monitoring online content.

(Almaliki et al., 2023) were among the first researchers to pre-train language models and fine-tune them for the task of hate speech detection. They introduced Arabic BERT-Mini Model (ABMM), a smaller BERT variant with a reduced hidden dimension compared to BERT. After pre-training, it was fine-tuned on a newly released dataset of around 9,500 labeled hate speech documents. Despite its modest size estimated to be approximately 11 million parameters, it outperformed much larger models such as AraBERT.

(Gandhi et al., 2024) also delved into the importance of detecting hate speech, while discussing literature about the topic since 2020. Alongside the comprehensive literature review, they introduced a methodology to tackle multi-label and multi-class hate speech in Indonesian. With the LSTM model attaining the highest accuracy compared to Logistic Regression models.

(Chakravarthi, 2022) proposed a different alternative to suppression of hate speech, namely the promotion and assistance of hope speech. In their study, the authors proposed the first multilingual hope speech datasets collected from YouTube along with a novel deep learning architecture to train on this dataset and detect hope speech. The dataset included English, Tamil, and Malayalam text. The multi-annotator annotation process yielded consistent results across annotators, proved by the fairly high inter-annotator agreement of 0.6+ score on Krippendorff's Alpha metric and reaching as high as a near-perfect agreement of 0.85 on labeling Malayalam text. After dataset collection and annotation, they tested it on a suite of machine learning algorithms, ranging from traditional SVM and Logistic Regression models, among others, to their proposed deep learning-based system with a CNN model with T5-sentence embedding and IndicBERT. The proposed model outperformed all other models with an F1 score of 0.75, 0.62, and 0.67 on English, Tamil, and Malayalam, respectively.

Although many hate speech datasets exist, including in Arabic, few researchers have developed a corpus with hope speech, let alone one with Arabic text in it. (Zaghouani and Biswas, 2025b) were among few Arabic NLP researchers to have collected text and created a hope speech dataset. It is a bilingual Arabic-English dataset with around 38,000 data points, collected from social media sites. Not only does it annotate text on emotion labels, but also for intensity, complexity, and cause, categorizing hope speech with both binary and more granular labels. Its reliability can be attributed to its high inter-rater agreement, with a very good - near perfect - agreement, ranging from 0.75 - 0.85 Fleiss' Kappa.

(ArunaDevi and Bharathi, 2024) went beyond just simply creating hope speech datasets, acknowledging the importance of automated and intelligent systems in detection of hope speech, especially in social media platforms due to the "snowball effect" of speech in social media in general Where, according to a Facebook (now Meta) study, emotional states found in comments can directly influence the emotional state of users reading those comments (Kramer et al., 2014). Hence, having systems that can detect hope speech can be of immense usefulness to foster a positive environment where certain types of speech are promoted, further influencing a positive emotional state of their users. The authors proposed different intelligent systems to detect hope speech, ranging from traditional machine models like Multinomial Naive Bayes classifier and Support Vector Machines to the usage of BERT. Unsurprisingly, the BERT model outperformed traditional machine learning models while Multinomial Naive Bayes had a respectable performance.

3 System Overview

The MAHED sub-task 1 (Zaghouani et al., 2024; Zaghouani and Biswas, 2025b,a) dataset comprises MSA and DA instances, split into training, validation, and test sets by the task's organizers. Each text is accompanied by a target label, with possible labels being either "hate", "hope", or "not_applicable". The dataset exhibited somewhat of a class imbalance, with the training dataset having 3,697 text instances labeled as "not_applicable", 1,892 text instances labeled as "hope", and only 1,301 text instances labeled as "hate". Text preprocessing involved the removal of English, emoticons, symbols, and Arabic diacritics. The clean text was then tokenized with a sequence length of 128 tokens, with sentences shorter or longer being padded or truncated, respectively. The tokenized data was stored for later usage.

Pre-trained language models ArabicDeBERTa-DA, BERT-DA, and MARBERTV2 were used. ArabicDeBERTa and BERT-MSA were first pre-trained on a large corpus of approximately 2.2 billion MSA tokens, using Masked Language Modeling (MLM) task, with cross entropy loss as the loss function. The pre-training's aim is for the model to gain general language understanding, which can be leveraged later on downstream tasks, increasing performance. After pre-training exclusively on MSA data, the models have gained knowledge about the syntax, morphology, and semantics of MSA. To facilitate understanding in DA tasks, ArabicDeBERTa-DA and BERT-DA were continually pre-trained on DA text obtained from (Al-Fetyani et al., 2023). While the previous models were first trained on MSA data then DA data, MARBERTV2 was first trained on a large corpus of DA data, then continually pre-trained on MSA data. MARBERTV2 first pre-training phase included around 15.6 billion tokens. Its continual pre-training phase with MARBERTV2 also used MLM as the task type, with cross entropy loss as the loss function.

A custom ensemble model was designed to integrate three transformer-based models: BERT-DA, ArabicDeBERTa-DA, and MARBERTV2, each equipped with a classification head consisting of a linear layer, Tanh activation, and a final linear layer mapping to three output classes. The ensemble combined their outputs through soft voting by averaging logits to produce the final prediction.

4 Experimental Setup

The model was trained for 5 epochs with a batch size of 16, using the AdamW optimizer with a learning rate of 7×10^{-6}. Cross-entropy loss with class weights was applied to address class imbalance. Validation was performed periodically during training.

Performance was evaluated using macro-averaged precision, recall, F1-score, and accuracy on the validation and test sets, demonstrating the effectiveness of the ensemble in multi-class classification.

5 Results

The ensemble model performed at its best within ¾ of the first epoch, achieving a 72.2% F1 score. To achieve the best performance across different ensemble models, different seeds for initialization were set for each ensemble model and then the best performing ensemble models runs were hard voted for submission, achieving 72.3%. The Macro F1 score achieved by this model gained it its first place position in the competition. Each model achieved around 67% F1 score on test dataset, hence their combined performance boosted the score by 5-6%. The addition of a weighted loss function with 1.02 for hate, 0.98 for hope and 1 for not applicable allowed for higher weighting of error for misclassification of the minority class. The different weight initialization allowed for better diversity of heads. The proposed system has close to 0.5 billion parameters. Appendix A includes a confusion matrix of models performance on test set along with error analysis of 3 text instances.

6 Conclusion

This work highlights the importance of detecting both hope and hate speech in Arabic digital communication, as emotional contagion can either foster resilience and harmony or amplify societal divides. To address this challenge, this paper proposed an ensemble model integrating three continually pre-trained transformer-based encoder models: BERT-DA, ArabicDeBERTa-DA, and MARBERTV2. By leveraging continual pre-training on dialectal Arabic and combining models through soft voting with weighted loss functions, the proposed system achieved state-of-the-art results in sub-task 1 of MAHED 2025, with the the ensemble of best-performing models obtained a macro F1-score of 72.3%.

The results demonstrate the effectiveness of ensemble learning in handling linguistic diversity across Arabic varieties along with the performance boost and computational efficiency of continually pre-trained models. Beyond competition performance, the contribution of this work lies in introducing dialect-aware pre-trained models that can be extended to a wide range of downstream tasks. Future research can build on this by exploring larger-scale pre-training and continual pre-training efficacy and transferability between different Arabic dialects along with their downstream performance on tasks such as MAHED 2025 sub-task 1.

Limitations

A key limitation of this study is its proposed model's large size, having a nearly half a billion parameter model makes it unfeasible to run in

most edge devices and CPU environments without significant delay in response. This makes the model impractical to run in production and resource-constrained environments.

Acknowledgments

The authors express gratitude to Al Hussein Technical University for its support and for fostering a research environment committed to innovation.

Research supported with Cloud TPUs from Google's TPU Research Cloud (TRC).

References

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021. ARBERT & MARBERT: Deep bidirectional transformers for Arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7088–7105, Online. Association for Computational Linguistics.

Mohammad Al-Fetyani, Muhammad Al-Barham, Gheith Abandah, Adham Alsharkawi, and Maha Dawas. 2023. Masc: Massive arabic speech corpus. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 1006–1013. IEEE.

Malik Almaliki, Abdulqader M Almars, Ibrahim Gad, and El-Sayed Atlam. 2023. Abmm: Arabic bert-mini model for hate-speech detection on social media. *Electronics*, 12(4):1048.

Faisal Yousif Al Anezi. 2022. Arabic hate speech detection using deep recurrent neural networks. *Applied Sciences*, 12(12):6010.

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

S ArunaDevi and B Bharathi. 2024. Machine learning based approach for hope speech detection. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2024), co-located with the 40th Conference of the Spanish Society for Natural Language Processing (SEPLN 2024), CEUR-WS. org*.

Bharathi Raja Chakravarthi. 2022. Hope speech detection in youtube comments. *Social Network Analysis and Mining*, 12(1):75.

Ankita Gandhi, Param Ahir, Kinjal Adhvaryu, Pooja Shah, Ritika Lohiya, Erik Cambria, Soujanya Poria, and Amir Hussain. 2024. Hate speech detection: A comprehensive review of recent works. *Expert Systems*, 41(8):e13562.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press. http://www.deeplearningbook.org.

Adam Ibrahim, Benjamin Thérien, Kshitij Gupta, Mats L Richter, Quentin Anthony, Timothée Lesort, Eugene Belilovsky, and Irina Rish. 2024. Simple and scalable strategies to continually pre-train large language models. *arXiv preprint arXiv:2403.08763*.

Zixuan Ke, Yijia Shao, Haowei Lin, Tatsuya Konishi, Gyuhak Kim, and Bing Liu. 2023. Continual pre-training of language models. *arXiv preprint arXiv:2302.03241*.

Adam DI Kramer, Jamie E Guillory, and Jeffrey T Hancock. 2014. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, 111(24):8788–8790.

Kuan-Ying Lee, Yuanyi Zhong, and Yu-Xiong Wang. 2023. Do pre-trained models benefit equally in continual learning? In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6485–6493.

Soumajyoti Sarkar, Kaixiang Lin, Sailik Sengupta, Leonard Lausen, Sheng Zha, and Saab Mansour. 2022. Parameter and data efficient continual pre-training for robustness to dialectal variance in arabic. *arXiv preprint arXiv:2211.03966*.

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 Results analysis

Figure 1 shows the performance of the best results against the test dataset. Hate speech has been rarely predicted as hope speech, and vice versa.

Table 1 shows examples of text along with their actual and predicted labels, as Error Analysis. Examples of hate speech and overtly vulgar sentiment were avoided. The examples call attention to the ambiguous nature of the text, highlighting the challenging nature of the task along with annotating it.

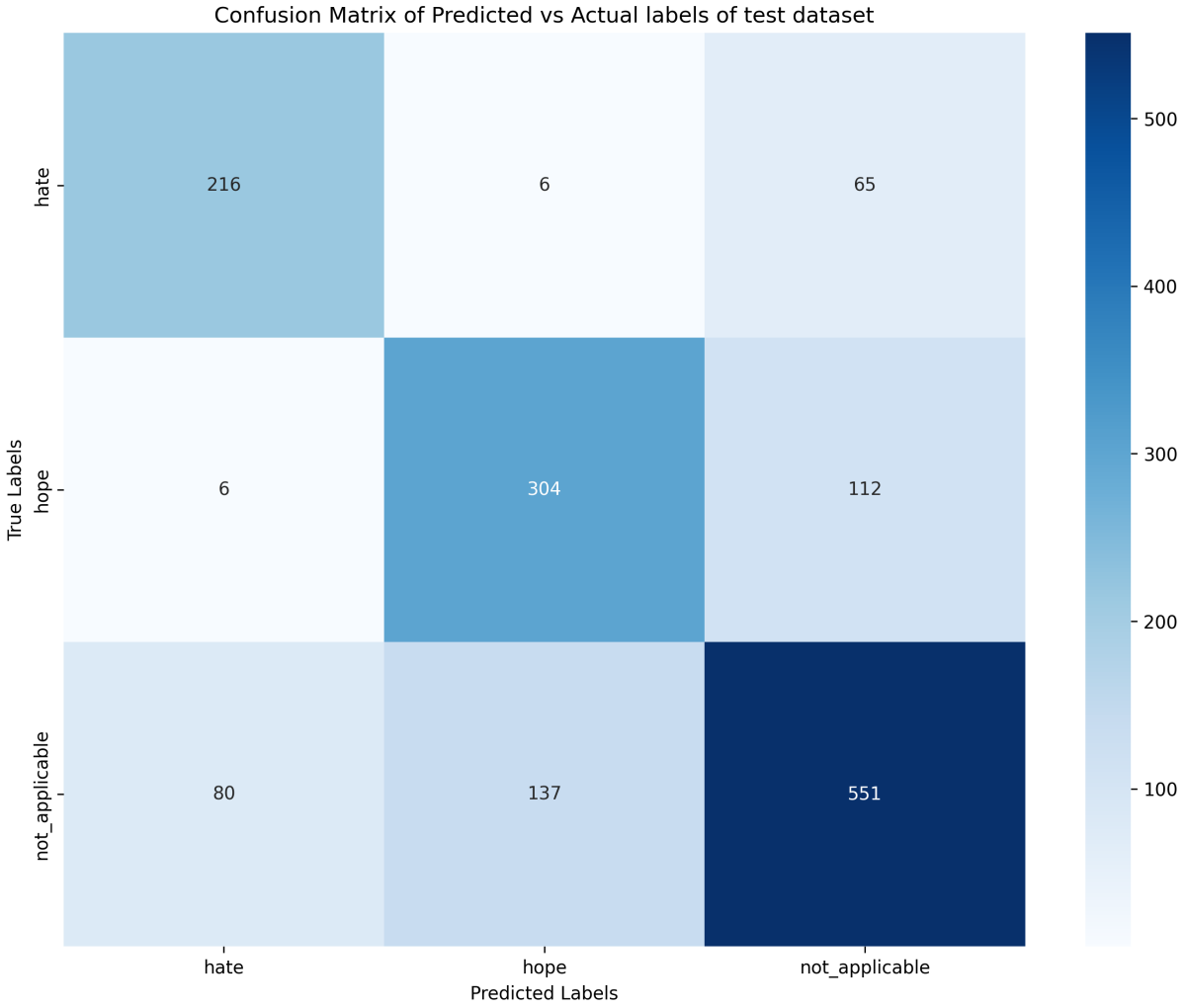

Figure 1: Confusion Matrix displaying performance of best predicted results vs actual labels by class/

Ground Truth	**Predicted**	**Translated Text**	**Arabic Type**
Hope	Not applicable	Oppression Oh God, I seek refuge in You from the oppression of men	DA
Not applicable	Hope	It is not just words of love and flirtation, but rather it is caring and taking care of the one you love and staying with him for a lifetime without changing your feelings towards him, friendship.	MSA
Not applicable	Hope	I put my heart between your hands and swore not to care.	MSA

Table 1: Examples of ground truth vs predictions with Arabic text. The Arabic text was translated using Google Translate.