CUET_Zahra_Duo@Mahed 2025: Hate and Hope Speech Detection in Arabic Social Media Content using Transformer

Walisa Alam, Mehreen Rahman, Shawly Ahsan and Mohammed Moshiul Hoque
Department of Computer Science and Engineering
Chittagong University of Engineering and Technology, Bangladesh
{u2004015, u2004033, 22mcse105}@student.cuet.ac.bd, moshiul_240@cuet.ac.bd

Abstract

In recent years, online social life has become an integral part of the global landscape, with social media platforms enabling users to express a wide range of emotions and opinions. In the Arabic-speaking world, navigating the dual nature of content—encompassing both hate and hope speech—remains challenging due to linguistic and cultural complexities. The MAHED 2025 shared task at ArabicNLP 2025 addressed this by focusing on detecting both hate and hope speech in Arabic social media. This paper describes our approach for subtask 1, utilizing various machine learning, deep learning, and transformer models for classification. AraBERT-large-v2 yielded the highest macro f_1-score of 0.698, earning 8^{th} place on the leaderboard.

1 Introduction

Social media platforms, such as Facebook and Twitter, enable widespread communication but also accelerate the spread of hate speech, which can fuel hostility and deepen social divides. The hateful content often spreads farther and faster than non-hateful material, mainly due to closely connected online communities (Mathew et al., 2019). In Arabic-speaking contexts, detecting hate speech is challenging due to dialectal diversity, frequent code-switching, rich morphological structures, orthographic variation, and cultural nuances (Elmadany et al., 2024).

Transformer architectures have significantly advanced hate speech detection. Recent research has shifted from traditional and deep learning models to transformer-based approaches, including BERT and its multilingual variants. Although these models achieve state-of-the-art performance, they also introduce higher computational costs, algorithmic biases, data scarcity, and inconsistent evaluation practices. The **MAHED 2025** shared task (Zaghouani et al., 2025) focuses on detecting hope and hate emotions in Arabic content. This study addresses subtask 1, which involves classifying hate and hope speech in Arabic texts. The primary contributions of this work are as follows:

- Investigated the efficacy of various machine learning models (Logistic Regression, Decision Tree, Random Forest, Naive Bayes, MNB, KNN, and XGBoost), deep learning models (CNN, BiLSTM, and CNN-BiLSTM), and transformer-based models (MARBERT, AraBERT-base, and AraBERT-large) in detecting both hate and hope speech in Arabic texts.
- Presented a transformer-based approach using **AraBERT-large** to classify Arabic social media texts into *hate*, *hope*, and *not_applicable* categories.

2 Related Work

Extensive research has been conducted on hate and hope speech detection, ranging from classical ML to DL and from transformer models to large language models (LLMs). Roy et al. (2022) applied classical ML models, using Logistic Regression and TF-IDF features, Random Forest, and XGBoost. Their best-performing model was Random Forest, with an F1 score of 0.96. Yang et al. (2023) used several LLMs like GPT-3.5-turbo-0613, Flan-T5, T5-large, GPT-2-large, and two variants of the HARE framework, Fr-HARE and Co-HARE, to improve accuracy. Among the models tested, Co-HARE with Flan-T5 (large) achieved the highest accuracy. Usman et al. (2025) addresses multilingual hate speech detection in English, Urdu, and Spanish using a trilingual dataset of 10,193 tweets. The evaluated models include LLMs (GPT-3.5 Turbo, Qwen 2.5 72B), transformers (BERT, RoBERTa), and SVM with TF-IDF features. Qwen 2.5 72B achieved the best performance overall, especially in the joint multilingual setting.

Proceedings of The Third Arabic Natural Language Processing Conference, pages 700–705

Recent research has seen significant advancements in the detection of Arabic hate and hope speech. Zaghouani et al. (2024) evaluated LR, RF, Gradient Boosting, SVM, Decision Tree, and AraBERT for this task. AraBERT was the best-performing model, with an accuracy of 0.83. Charfi et al. (2024) introduced the ADHAR dataset covering various dialects. Using AraBERT, they achieved high performance in hate speech detection (F1 score of 0.95). Alghamdi et al. (2024) presented AraTar, where AraBERTv0.2 (base) achieved the best performance. Yagci et al. (2024) worked on Turkish and Arabic hate speech detection in the HSD-2Lang shared task, using AraBERTv02-Twitter fine-tuned with the AdamW optimizer. Their best-performing configuration achieved 0.89 accuracy and 0.74 F1 score for Arabic texts. AlDahoul and Zaki (2025) addressed Arabic hate and hope speech detection, where an ensemble of GPT-4o-mini, Gemini Flash 2.5, and Google text embedding with SVM, combined with a fine-tuned GPT-4o-mini hope/not classifier, achieved the best performance (macro-F1 score of 72.1%).

Previous research on Arabic hate and hope speech detection has been constrained by limited data. In this study, we overcame these constraints by implementing improved data cleaning and augmentation strategies.

3 Task and Dataset Distribution

The MAHED 2025 shared task aims to advance research on detecting hate speech, hope speech, and emotional expressions in Arabic content (Zaghouani et al., 2024; Zaghouani and Biswas, 2025b,a). We participated in subtask 1, which involved classifying Arabic texts into three categories:

- **Hate:** Text expressing hostility, bias, or discrimination against certain individuals or groups.
- **Hope:** Text that communicates positivity, encouragement, or supportive messages.
- **Not Applicable:** Text that does not contain elements of hate or hope speech.

The dataset consists of text samples collected from Arabic social media platforms and is divided into a training set (T_{trn}), validation set (T_{val}) and test set (T_{tst}). Table 1 shows the statistics of the dataset.

Table 1: Dataset statistics, where T_w, T_{uw}, T_{mw}, and T_{aw} indicate the number of total words, unique words, maximum words per text, and average words per text in the training set, respectively.

Attributes	Hate	Hope	N /A	Total
T_{trn}	1,301	1,892	3,697	6,890
T_{val}	261	409	806	1,476
T_{tst}	287	422	768	1,477
T_w	30,855	41,317	82,700	1,54,872
T_{uw}	15,606	22,499	42,212	68,126
T_{mw}	92	105	107	–
T_{aw}	23.0	21.0	22.0	–

4 Methodology

This study explores several ML, DL, and transformer-based architectures. As shown in Figure 1, the adopted model follows a multi-stage design.

4.1 Data Preprocessing

The text preprocessing pipeline systematically cleans Arabic tweets to improve the model performance. It removes punctuation (including Arabic symbols), numbers, Latin letters, emojis, and extra whitespace, converts the text to lowercase, and normalizes the Arabic script by removing Tashkeel. Additionally, tweet-specific elements, such as URLs, mentions, and hashtags, were handled, and informal text was converted to standard Arabic. This preprocessing ensures that the input data is normalized and noise-free, making it suitable for ML, DL, and transformer-based models.

4.2 Data Augmentation

We employed contextual word embedding–based augmentation using the Arabic BERT model (Antoun et al., 2020) via the *nlpaug*[1] library, where selected words were replaced with contextually similar alternatives predicted by the model. This approach preserves the semantic meaning of the original text while introducing lexical and structural variations, ensuring that the augmented samples retain their original classification labels.

4.3 Overview of the Adopted Model

We adopted ML, DL, and transformer-based classifiers for Arabic hate and hope speech detection.

4.3.1 ML Models

For feature representation, we employed the TF-IDF scheme to represent the textual data. Using the

[1] https://github.com/makcedward/nlpaug

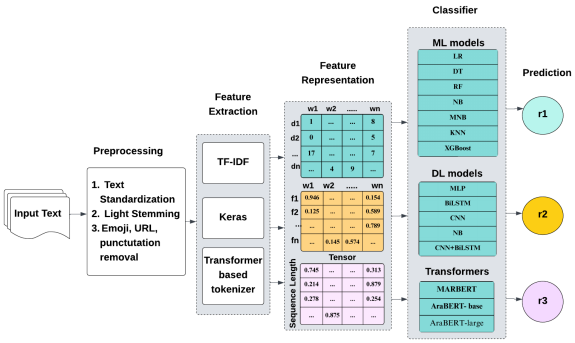

Figure 1: Overview of the adopted model

TF-IDF features, we evaluated several ML classifiers, including Logistic Regression (LR), Decision Tree (DT), Support Vector Machine (SVM), Multinomial Naive Bayes (MNB), Random Forest (RF), K-Nearest Neighbors (KNN), and XGBoost (XGB). LR was trained with a maximum of 1200 iterations, while DT was trained with its default configuration. RF was optimized with tuned estimators and depth, and SVM was employed with a linear kernel. For XGBoost, we applied '*multi:softprob*' objective, 100 rounds for boosting, multiclass logloss for evaluation, and '*gpu_hist*' for the tree construction algorithm. The KNN model was trained with 12 neighbours.

4.3.2 DL Models

For the DL models, the text was first tokenized using the Keras library[2] with a maximum vocabulary size of 10,000 words, and sequences were padded or truncated to a fixed length of 150 tokens. Multiple neural network architectures were implemented, including Multi-Layer Perceptron (MLP) with TF-IDF inputs, BiLSTM, CNN, and CNN+BiLSTM. MLP was configured with 3 layers (512, 256, 256), ReLU activation, softmax output, and trained for 150 epochs with a batch size of 64 and early stopping. The BiLSTM model consisted of 512 units, a dropout rate of 0.2, and a dense layer of 256. It was trained for 150 epochs with a batch size of 64 and learning rate decay (0.96, 1000). The CNN architecture applied convolution layers with 512 and 256 filters, a kernel size of 5, vocabulary size of 10,000, maximum input length of 150, dropout rate of 0.2, and early stopping. The hybrid CNN+BiLSTM combined a 512-filter convolution layer with a 256-unit BiLSTM layer, followed by a dense layer of 128 units and a dropout rate of 0.2. The CNN+BiLSTM model was trained for 100 epochs with a batch size of 64 and early stopping. All models employed embedding layers and a softmax function for multiclass classification.

[2] https://keras.io/

4.3.3 Transformer-Based Models

For each transformer-based model, the texts were tokenized and padded using their respective tokenizers from the HuggingFace library. We employed several transformer-based models for Arabic text classification, including AraBERT-base (Safaya et al., 2020), MARBERT (Abdul-Mageed et al., 2021), and AraBERT-large (Antoun et al., 2020). Each transformer comprises multiple encoder layers with multi-head self-attention, feed-forward networks, residual connections, and layer normalization, with dropout applied to the hidden states and attention weights to prevent overfitting. The contextual representation of the [CLS] token was fed into a fully connected layer for classification into three categories (hope, hate, and not applicable). AraBERT-base and MARBERT were trained with a learning rate of 1×10^{-5} for 20 epochs with a batch size of 128, while AraBERT-large was trained with varying learning rates (3.5×10^{-6} to 2×10^{-5}), epochs, and batch sizes (128 and 256), with or without augmentation.

5 Result Analysis

All experiments were conducted on Kaggle using two NVIDIA Tesla T4 GPUs with 16 GB of GPU memory each and 32 GB system RAM. We evaluated the performance of the models using several metrics, including precision, recall, and macro f_1 score (MF1). MF1 was chosen as the primary metric to ensure a balanced performance evaluation of the models. Table 2 presents a comparative analysis of the performance achieved by ML, DL, and transformer-based models for Arabic text-based classification of hope and hate speech.

Among the ML classifiers, Naive Bayes performed best, likely because its probabilistic nature enhanced its ability to predict positive outcomes, boosting Recall and thus MF1, which is especially suitable in cases where positive instances are harder to capture. In the DL category, CNN + BiLSTM performed best, with an MF1 score of 0.619, because it effectively integrated CNN's local feature extraction with BiLSTM's sequential context modeling, resulting in a stronger precision–recall balance. However, AraBERT-large was the standout performer among the AraBERT family. Outper-

Table 2: Performance comparison of ML, DL, and transformer-based models for Arabic hate and hope speech classification.

Model	Precision	Recall	MF1
LR	0.652	0.519	0.542
DT	0.516	0.498	0.506
RF	0.652	0.492	0.509
NB	0.638	0.541	0.563
MNB	0.789	0.381	0.325
KNN	0.597	0.494	0.513
XGBoost	0.633	0.517	0.538
MLP	0.617	0.554	0.581
BiLSTM	0.634	0.565	0.582
CNN	0.598	0.594	0.607
CNN + BiLSTM	0.601	0.623	0.619
MARBERT	0.640	0.640	0.640
AraBERT-base	0.600	0.600	0.600
AraBERT-large	**0.694**	**0.697**	**0.695**

forming MARBERT and AraBERT-base in Precision, Recall, and MF1 scores, AraBERT-large emerged as the top variant with the highest scores across all metrics, achieving an MF1 score of 0.695. This superior performance can be attributed to AraBERT-large's broader contextual coverage and stronger capacity to capture the morphological richness of Arabic, which enabled it to outperform the smaller models.

5.1 Ablation Study

The results of the ablation study on the classification of hatred and hope discourse in Arabic are shown in Table 3, with distinct reports for the development and testing stages of the models. The best-performing model was trained for a maximum of 20 epochs, incorporating early stopping with a patience of 4. The model converged after 12 epochs. In the development phase, the batch size, learning rate, and sequence length had a clear effect. A batch size of 8 and a learning rate of 1×10^{-5} led to stable training, whereas increasing the rate to 2×10^{-5} slightly reduced performance. Preprocessing combined with augmentation helped AraBERT-large and MARBERT achieve the highest MF1 score of 0.64, whereas raw data or preprocessing alone yielded lower scores.

In the testing phase, AraBERT-large with preprocessing achieved the best MF1 score (0.69). Using a smaller learning rate of 3.5×10^{-6} and a longer sequence length of 256 improved generalization, highlighting the importance of careful hyperparameter tuning along with preprocessing.

5.2 Error Analysis

A detailed error analysis was carried out to understand the performance of the best-performing model (i.e., AraBERT-large).

5.2.1 Quantitative Error Analysis

The results highlight the strong performance of the AraBERT-large model in classifying Arabic social media texts into hate and hope categories. The confusion matrix shown in Figure 2 provides a quantitative breakdown of the predictions.

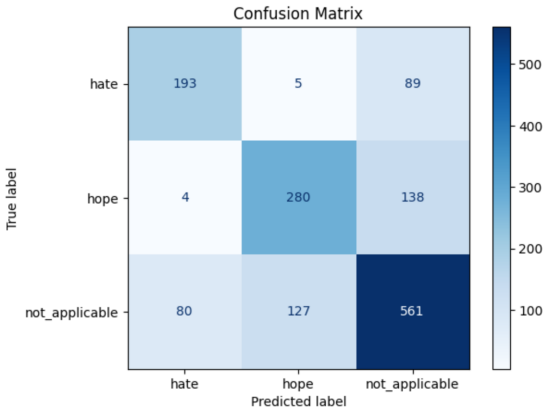

Figure 2: Confusion matrix of the AraBERT-large model for the test set

The analysis showed that the model successfully identified 193 hate samples, 280 hope samples, and 561 not_applicable samples. However, there were a few misclassifications, with 5 hate instances incorrectly labeled as hope and 4 hope instances misclassified as hate. A larger source of error comes from confusion with the not_applicable class, where 89 hate and 138 hope samples are wrongly predicted as not_applicable, whereas 80 not_applicable samples are labeled as hate and 127 as hope. These errors can be attributed to overlapping linguistic cues across categories and the class imbalance.

5.2.2 Qualitative Error Analysis

Table 4 demonstrates some sample predictions made by the AraBERT-large model. Here, samples 2 and 3 were correctly classified, whereas samples 1, 4, and 5 were misclassified. Sample 1 was mislabelled as not_applicable when it was hate due to sarcasm diluting the hateful signals. Sample 4 was predicted as hate instead of not_applicable because of the strong offensive words that the model associates with hate, and Sample 5 was inaccurately classified as hope instead of not_applicable because the positive and uplifting tone resembles

Table 3: Ablation study on the impact of hyperparameters on the performance of the transformer-based models.

Model	Method	Batch	LR	Epochs	MaxLen	MF1
Development Phase						
AraBERT-base	Preproc + Aug	8	1×10^{-5}	20	128	0.60
MARBERT	Preproc + Aug	8	1×10^{-5}	20	128	0.64
AraBERT-large	Raw data	8	1×10^{-5}	20	128	0.61
AraBERT-large	Preprocessed	8	1×10^{-5}	20	128	0.62
AraBERT-large	Preproc + Aug	8	1×10^{-5}	20	128	0.64
AraBERT-large	Preproc + Aug	8	2×10^{-5}	3	128	0.63
Testing Phase						
AraBERT-large	Raw data	8	1×10^{-5}	20	128	0.67
AraBERT-large	Preproc + Aug	8	3.5×10^{-6}	20	256	0.69
AraBERT-large	**Preprocessed**	**8**	3.5×10^{-6}	**20**	**256**	**0.69**

the hope class. These nuances highlight the importance of qualitative analysis in understanding the model performance in specific cases. Moreover, we observed that dialectal words introduce challenges in the detection of hate and hope speech. Sample 1 includes the Gulf/Yemeni dialect expression "ba'aysh" ("with what") and sample 5 contains the Egyptian/Levantine colloquial word "teslamy" ("thank you" / "bless you"). The presence of dialectal expressions in these samples underscores the complexity of accurately classifying texts in diverse Arabic dialects.

Table 4: Sample output predictions by the AraBERT-large model, where Arabic texts were translated using Google Translate.

Sample Text	Actual	Predicted
@shatat_20 @zainabera نعوضكم يالعراكي بس انقلع قم يعني بأيش (@shatat_20 @zainabera How can we compensate you? Just get out of here, you brat)	hate	N/A
@Mp_M_Alhazmi هو كما الشيعه دين والخرافه الشرك (@Mp_M_Alhazmi The Shiite religion is polytheism and superstition.)	hate	hate
الشعب... ايها معكم ..مواصلون الخير مساء عمانتل 51 من الجماعيه الهجره (Good evening.. We continue with you, dear people... The mass migration from Omantel 51)	hope	hope
وجوهكم تتبدل خنازير منكم اوسخ يوجد لا والنظام داعش مقولاتكم اين البرق بسرعه اجل فيها تستنجد ليش واحده لعمله وجهان الان؟ (There is no one dirtier than you pigs, your faces change at lightning speed. Where are your sayings? ISIS and the regime are two sides of the same coin. Why are you seeking their help now?)	N/A	hate
..... الاوليمبياد في مصر فخر سمير ساره تسلمي (Sarah Samir, Egypt's pride in the Olympics... Thank you)	N/A	hope

6 Conclusion

This work evaluated multiple ML, DL, and transformer-based models for detecting hate and hope speech in Arabic. The AraBERT-large model demonstrated the highest performance, achieving a macro f_1 score of 0.69 and surpassing all other models tested, benefiting from its broader contextual coverage and stronger ability to capture the morphological richness of the Arabic language. However, the system is limited by class imbalance, challenges in capturing nuanced or context-dependent meanings, and its dependence on the quality of the training data. Future work should focus on augmenting the dataset to mitigate class imbalance, integrating multilingual or cross-domain data, and investigating hybrid-model architectures to enhance predictive accuracy.

Limitations

The developed model has several limitations. It relies solely on textual input and cannot leverage multimodal signals, such as images or videos that often accompany social media posts. Its performance is also sensitive to preprocessing and augmentation strategies, which may not generalize well to unseen data. Moreover, training on a single dataset introduces the risk of bias and limits the model's adaptability to other dialects, domains, and code-switched texts.

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.

Nouar AlDahoul and Yasir Zaki. 2025. Detecting hope, hate, and emotion in arabic textual speech and multimodal memes using large language models. *Preprint*, arXiv:2508.15810.

Seham Alghamdi, Youcef Benkhedda, Basma Alharbi, and Riza Batista-Navarro. 2024. AraTar: A corpus to support the fine-grained detection of hate speech targets in the Arabic language. In *Proceedings of the 6th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT) with Shared Tasks on Arabic LLMs Hallucination and Dialect to MSA Machine Translation @ LREC-COLING 2024*, pages 1–12, Torino, Italia. ELRA and ICCL.

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.

Anis Charfi, Mabrouka Bessghaier, 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.

Ahmed Elmadany, Wajdi Zaghouani, and Nizar Habash. 2024. A survey on arabic natural language processing: Challenges and applications. *ACM Computing Surveys*, 57(2):1–40.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2019. Spread of hate speech in online social media. In *Proceedings of the 10th ACM Conference on Web Science (WebSci '19)*, pages 173–182, Boston, MA, USA. ACM.

Pradeep Roy, Snehaan Bhawal, Abhinav Kumar, and Bharathi Raja Chakravarthi. 2022. IIITSurat@LT-EDI-ACL2022: Hope speech detection using machine learning. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 120–126, Dublin, Ireland. Association for Computational Linguistics.

Ali Safaya, Moutasem Abdullatif, and Deniz Yuret. 2020. KUISAIL at SemEval-2020 task 12: BERT-CNN for offensive speech identification in social media. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 2054–2059, Barcelona (online). International Committee for Computational Linguistics.

Muhammad Usman, Muhammad Ahmad, M. Shahiki Tash, Irina Gelbukh, Rolando Quintero Tellez, and Grigori Sidorov. 2025. Multilingual hate speech detection in social media using translation-based approaches with large language models. *Preprint*, arXiv:2506.08147.

Utku Yagci, Egemen Iscan, and Ahmet Kolcak. 2024. ReBERT at HSD-2Lang 2024: Fine-tuning BERT with AdamW for hate speech detection in Arabic and Turkish. In *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024)*, pages 195–198, St. Julians, Malta. Association for Computational Linguistics.

Yongjin Yang, Joonkee Kim, Yujin Kim, Namgyu Ho, James Thorne, and Se young Yun. 2023. Hare: Explainable hate speech detection with step-by-step reasoning. *Preprint*, arXiv:2311.00321.

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.