REGLAT at MAHED Shared Task: A Hybrid Ensemble-Based System for Arabic Hate Speech Detection

Nsrin Ashraf[1,2], Mariam Labib[1,3], Tarek Elshishtawy[4], Hamada Nayel[2,5]

[1]Computer Engineering, Elsewedy University of Technology, Cairo, Egypt
[2]Department of Computer Science, Faculty of Computers and Artificial Intelligence, Benha University, Egypt
[3]Department of Electronics and Communications Engineering, Faculty of Engineering, Mansoura University, Egypt
[4]Department of Information Systems, Faculty of Computers and Artificial Intelligence, Benha University, Egypt
[5]Department of Computer Engineering and Information, College of Engineering, Wadi Ad Dwaser, Prince Sattam Bin Abdulaziz University, Al-Kharj 16273, Saudi Arabia

Correspondence: hamada.ali@fci.bu.edu.eg

Abstract

Hope and hate speech detection in natural language processing addresses the challenge of identifying social media content within the fast-paced environment of online platforms. Hopeful speech that promotes supportive and inclusive language plays a crucial role in counteracting online toxicity, whereas hate speech poses threats and challenges to society. This paper focuses on text-based Arabic hate and hope speech detection, demonstrating the system submitted by the REGLAT team to the MAHED shared task held in conjunction with ArabicNLP 2025. The proposed system employs an ensemble-based model that combines a TF-IDF + Logistic Regression classifier with a fine-tuned AraBERTv2 model as baselines. A majority voting approach is then applied to aggregate the predictions. The proposed model reported an F1 score of 0.58. These promising results are notable given the simplicity of the system's architecture, and they highlight the potential of our approach for improving the performance of this task.

1 Introduction

Hope speech detection in Natural Language Processing (NLP) is a multifaceted research area positioned at the intersection of computational linguistics and artificial intelligence. With the rapid growth of digital communication and social media platforms, the volume and influence of online speech, particularly language that supports mental health and social harmony, have increased significantly (Balouchzahi et al., 2023). Hope speech refers to positive, supportive, and inclusive language that can counteract online toxicity, promote mental health, and offer solidarity to marginalized or vulnerable communities. Detecting and amplifying such speech can play a vital role in mitigating conflict, encouraging resilience during crises, and encouraging inclusive digital spaces (Sharma et al., 2025).

The core objective of hope and hate speech detection is to automatically identify, classify and respond to emotionally charged or socially impactful content within fast-moving streams of user-generated data (Ashraf et al., 2022). Given the scale and speed of online discourse, manual annotation is no longer practical. As a result, modern systems rely on NLP and learning techniques ranging from traditional keywords and lexicon-based methods to more approaches involving machine learning, deep learning, and transformers such as BERT and GPT (Ahmad et al., 2024; ArunaDevi and Bharathi, 2024). These models enable a deeper contextual understanding of language, which is essential for accurately distinguishing between supportive and harmful expressions in diverse linguistic and cultural settings.

The main contribution of this work lies in the development and evaluation of a hybrid approach for Arabic hate speech detection submitted to MAHED shared task (Zaghouani et al., 2025). Arabic is one of the six official languages of the United Nations and has more than 400 million native speakers. Arabic NLP poses unique challenges compared to other languages, due to the complexity of the morphological structure, rich inflection, and diverse dialects (

Proceedings of The Third Arabic Natural Language Processing Conference, pages 645–650

such as Egyptian, Algerian, Tunisian, Gulf Region, Levant, Iraqi etc.) (AbuElAtta et al., 2023; Sobhy et al., 2025). The proposed work combines traditional machine learning techniques with modern transformer-based models. We demonstrate that despite the growing dominance of deep learning, lightweight models such TF-IDF with Logistic Regression remain highly competitive, particularly when complemented by contextual embeddings from models such as AraBERTv2. Furthermore, we propose an ensemble strategy using majority voting to find predictions from both approaches, which yielded the best overall performance in our experiments.

2 Background

Machine learning has played a pivotal role in the advancement of NLP, allowing computers to learn patterns from textual data for tasks such as sentiment analysis, machine translation and text classification (Kamal et al., 2024). Traditional machine learning models, such as Naive Bayes and Support Vector Machines, are heavily based on hand-crafted features and vector representations such as TF-IDF and word embeddings (Khairy et al., 2024).

The rise of deep learning models such as Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM), and Convolutional Neural Networks (CNNs) brought significant improvements by automatically learning features from large amount of data. Doghmash and Saad (2025) presented a dual approach to combating hate speech in Arabic social media content. The first part focuses on hate speech detection, where the authors evaluated several deep learning models (RNN, CNN, and CNN-RNN) trained from scratch using AraVec word embeddings (Soliman et al., 2017). Among these, CNN outperforms all other models reported a macro F1-score and accuracy of 0.51 and 0.80 respectively. In contrast, transformer-based models (e.g., QARiB (Abdelali et al., 2021), MARBERT (Abdul-Mageed et al., 2021), AraBERT (Antoun et al., 2020)) significantly outperformed traditional deep learning models. Among these, the QARiB model combined with AraBERT preprocessing achieved the best results, obtaining a 0.92, 0.95 macro F1-score and accuracy respectively.

More recently, transformer-based models such as BERT, MARBERT, and RoBERTa (Liu et al., 2019) have revolutionized NLP by capturing complex contextual relationships through self-attention mechanisms. Abdelsamie et al. (2026) proposed a multi-task learning (MTL) approach using transformer models (AraBERT, MARBERT, MARBERTv2) to improve hate speech detection across Arabic dialects. Each dialect is treated as a separate task to address semantic ambiguity caused by dialectal differences. The AraBERT model learns both shared and dialect-specific features, achieving higher F1-scores than traditional single-task models (up to 0.98 for Egyptian) and 0.85 for MTL combined dialects. It also generalizes well to unseen datasets, proving more effective in detecting hate speech across diverse Arabic dialects.

Daouadi et al. (2024) conducted extensive experiments to optimize hyperparameters and evaluate the effectiveness of transformer-based models for Arabic hate speech detection. Initially, three pre-trained models were fine-tuned using varying parameters. The authors implemented ensemble learning using majority and average voting. These methods further improved performance, with majority voting reaching a weighted F1-score of 0.86. Furthermore, applying data augmentation using external datasets and semi-supervised learning boosted the F1-score to 0.86 and outperformed prior methods across multiple hate speech categories. These models achieve state-of-the-art performance on a wide range of NLP tasks and are now widely adopted in both academic and industrial applications.

3 System Overview

In this study, We experimented with a range of machine learning and deep learning models for text classification. Traditional approaches such as Support Vector Machines (SVMs) have proven effective in handling high-dimensional data, making them well-suited for text classification tasks. Similarly, Logistic Regression (LR) offers a simple yet interpretable linear model that estimates class membership probabilities. Moving beyond these classical methods, we explored Deep Neural Networks (DNNs), which employ multi-layer architectures with ReLU activation functions and dropout regularization to mitigate overfitting. Finally, we fine-tuned transformer-based models, including AraBERTv2 (Antoun et al., 2020) and

CAMeL-BERT (Inoue et al., 2021), both pretrained on large-scale Arabic corpora and shown to be highly effective for Arabic NLP.

3.1 Dataset

The MAHED shared task focusing on the detection of hate speech, hope speech, and emotional expression in Arabic content through three sub-tasks, our team participated in subtask 1 (Zaghouani and Biswas, 2025; Zaghouani et al., 2024). The proposed dataset focuses on classifying Text-based Hate and Hope Speech in Arabic dialects and Modern Standard Arabic (MSA). It consists of manually annotated data collected from social media posts. Each text instance in the dataset is labeled for **`Hate`**, **`Hope`** and **`Not Applicable`**. A general statistics of the class distribution over the dataset is shown in Figure 1. Data were split into training (70%), validation (15%), and testing (15%) sets by the shared task organizers, ensuring consistency and fairness across all participating systems.

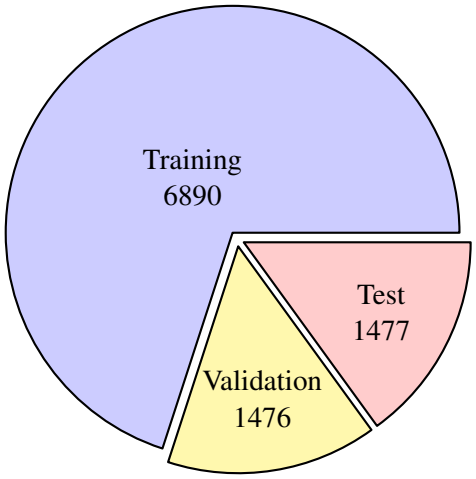

Figure 1: Dataset statistics across training, validation, and test splits.

4 Experimental Setup

In this section, experimental setup, model configurations, dataset preprocessing, evaluation metrics and a detailed analysis of the results have been presented. We have conducted a series of experiments to assess the effectiveness of various pre-trained transformer models and machine learning baselines for Arabic hate speech detection. The experiments are organized into preprocessing, future extraction, hyper-parameter tuning, and evaluation of individual and ensemble models.

4.1 Dataset Preprocessing

A text cleaning strategy using regular expression, implemented using the **`Regex`**, and **`NLTK`** packages. Text cleaning was applied according to the following steps:-

- Remove URLs, mentions, whitespace, punctuation, symbols and emojis
- Normalize Arabic letters
- Remove English letters and numbers

4.2 Model Implementation

This study employed a variety of learning techniques to compare their performance on the same task and determine the impact of each technique on the classification results. These models were selected on the basis of their diversity in mathematical structure and complexity. These models include traditional machine learning, deep learning, and transformer-based models.

For machine learning technique, SVM and LR have been implemented using TF-IDF as a text representation technique over *unigrams*, *bigrams* and *trigrams*. SVM is particularly effective for handling high-dimensional textual data, while LR was used to compute the class probabilities, offering a simple yet robust baseline for text classification tasks. To further enhance the performance of these ML models, a fine-tuning technique was applied using a range of transformer-based models, including `AraBERTv2` and `CAMeLBERT-finetuned-2e-5`.

Deep learning model was applied using multiple hidden layers, incorporating the ReLU activation function and dropout regularization to reduce overfitting. In addition, an ensemble model combining Bi-LSTM and CNN architectures was examined to capture both sequential and local features of the text. This ensemble was tested with two types of word embeddings: a specifically designed for Arabic `AraVec` ($dim = 300$) and a widely used general-purpose embedding `GloVe` ($dim = 200$). These configurations aimed to improve the model's ability to understand semantic and syntactic nuances in Arabic text.

Transformer-based models were applied using pre-trained Arabic language models `AraBERTv2` and `CAMelBERT`. Both models were fine-tuned for sequence classification with

three output labels, utilizing the HuggingFace `AutoTokenizer` for tokenization and `AutoModelForSequenceClassification` for model initialization. The training was performed on GPU with parameter alignment through the ignore mismatched sizes option to ensure compatibility. These configurations leveraged the rich contextual representations of Arabic text captured by the transformer models, aiming to enhance classification performance across the **`Hate`**, **`Hope`** and **`Not Applicable`** categories.

Among all the models evaluated, the hybrid technique combining TF-IDF-based Logistic Regression (LR) and transformer-based AraBERT classification, followed by majority voting, achieved the best overall performance. This approach effectively leverages the strengths of both traditional machine learning and deep contextual representations. The LR-based model captured key lexical features, especially effective in high-dimensional sparse text data, while the AraBERT classifier provided deep semantic understanding through pre-trained language representations. To ensure optimal performance, we experimented with different hyperparameter configurations for both models and selected the best-performing settings based on validation results Table 1. In addition, we evaluated several Arabic transformer models and identified AraBERT as the most effective.

Component	**Hyperparameters**
TF-IDF Vectorizer	max_df = 0.9 min_df = 5 max_features = 50000 ngram_range = (1,3) sublinear_tf = True norm = 'l2' lowercase = True stop_words = None
Logistic Regression (LR)	class_weight = balanced max_iter = 1000
AraBERT (Transformer)	num_labels = 3 Optimizer: AdamW Learning rate = $2e^{-5}$ Batch size = 16 Epochs = 3 Tokenizer: AraBERT pretrained vocabulary

Table 1: Hyperparameters used in the ensemble model

By applying majority voting between the optimized LR and AraBERT predictions, the system achieved superior classification accuracy and robustness. This ensemble not only mitigated the weaknesses of individual models, but also significantly outperformed standalone deep learning and transformer models, making it the most effective method in our study.

The model was experimented with various configurations to determine the optimal settings for our models. Given that our datasets are imbalanced, as the shared task organizers selected macro F1-Score metric to ensure weight balancing for each label. The final parameters and evaluation metrics are summarized in Table 1.

All experiments were conducted on Google Colab using an NVIDIA T4 GPU with 16 GB of VRAM and 25 GB of system RAM. This setup ensured efficient training and fine-tuning of the transformer-based models while maintaining reproducibility of results.

5 Results and Discussions

This section reports and analyzes the results of our experiments across various configurations and model architectures. We evaluated the performance of traditional machine learning, deep learning, and transformer-based models for Arabic hate speech detection. The results include the impact of hyperparameter tuning and the effectiveness of ensemble learning technique. Our results are very competitive compared to the other teams as shown in Table 2. Among traditional models, SVM and LR achieved macro F1-scores of 0.42 and 0.40 respectively. The CNN-BiLSTM deep learning model underperformed, with a low macro F1-score of 0.31, likely due to limited data or lack of contextual embeddings. However, incorporating transformer-based embeddings significantly improved the results. When combined with **`AraBERTv2`** and **`CAMeLBERT`**, both SVM and LR models showed noticeable performance gains. In particular, **`LR-AraBERTv2`** achieved the highest macro F1-score of 0.58, followed by **`LR-CAMeLBERT`** at 0.54. These results highlight the importance of contextualized transformer embeddings, especially when paired with lightweight classifiers such as LR, to enhance classification performance in Arabic hate speech detection.

Model	Word Representation	Macro F1-score
SVM	TF-IDF	0.42
LR	TF-IDF	0.40
CAMeLBERT	Transformer (AutoTokenizer)	0.36
AraBERTv2	Transformer (AutoTokenizer)	0.46
CNN-BiLSTM	GloVe (200)	0.12
CNN-BiLSTM	AraVec (300)	0.31
SVM-AraBERTv2	TF-IDF + Transformer	0.47
SVM-CAMeLBERT	TF-IDF + Transformer	0.44
LR-AraBERTv2	TF-IDF + Transformer	**0.58**
LR-CAMeLBERT	TF-IDF + Transformer	0.54

Table 2: System performance results on MAHED dataset

6 Limitations

The performance of the models in this study was constrained by several key factors. First, the dataset was imbalanced, which limited the ability of the models to generalize effectively across all classes. Second, the lack of large-scale, domain-specific pre-trained language models for Arabic hate speech reduced the effectiveness of transformer-based approaches, as they struggled to capture the nuanced and context-dependent expressions of hate across dialects. Finally, existing resources for Arabic NLP remain limited compared to high-resource languages, which restricts the range of architectures and embeddings that can be effectively applied.

Future work could address these limitations by developing larger and more balanced datasets, creating domain-specific pre-trained models tailored for hate speech detection, and incorporating dialect-aware modeling. Data augmentation, transfer learning, and multi-task learning also represent promising directions for overcoming data scarcity and improving robustness.

7 Acknowledgment

The ***REGLAT*** team appreciates the efforts that have been offered by Benha University, Elsweedy University of Technology, and Prince Sattam Bin Abdulazizi University.

References

Ahmed Abdelali, Sabit Hassan, Hamdy Mubarak, Kareem Darwish, and Younes Samih. 2021. Pre-training bert on arabic tweets: Practical considerations. *Preprint*, arXiv:2102.10684.

Mahmoud Mohamed Abdelsamie, Shahira Shaaban Azab, and Hesham A. Hefny. 2026. The dialects gap: A multi-task learning approach for enhancing hate speech detection in arabic dialects. *Expert Systems with Applications*, 295:128584.

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.

Ahmed H. AbuElAtta, Mahmoud Sobhy, Ahmed A. El-Sawy, and Hamada Nayel. 2023. Arabic regional dialect identification (ardi) using pair of continuous bag-of-words and data augmentation. *International Journal of Advanced Computer Science and Applications*, 14(11).

Muhammad Ahmad, Sardar Usman, Humaira Farid, Iqra Ameer, Muhammad Muzammil, Ameer Hamza, Grigori Sidorov, and Ildar Batyrshin. 2024. Hope speech detection using social media discourse (posivox-2024): A transfer learning approach. *Journal of Language and Education*, 10(4):31–43.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based model for Arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15, Marseille, France. European Language Resource Association.

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 Conference of the Spanish Society for Natural Language Processing (SEPLN 2024), Valladolid, Spain, September 24, 2024*, volume 3756 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Nsrin Ashraf, Mohamed Taha, Ahmed Abd Elfattah, and Hamada Nayel. 2022. NAYEL @LT-EDI-ACL2022: Homophobia/transphobia detection for equality, diversity, and inclusion using SVM. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 287–290, Dublin, Ireland. Association for Computational Linguistics.

Fazlourrahman Balouchzahi, Grigori Sidorov, and Alexander Gelbukh. 2023. Polyhope: Two-level hope speech detection from tweets. *Expert Systems with Applications*, 225:120078.

Kheir Eddine Daouadi, Yaakoub Boualleg, and Kheir Eddine Haouaouchi. 2024. Ensemble of pre-trained language models and data augmentation for hate speech detection from arabic tweets. *arXiv preprint arXiv:2407.02448*.

Salam Thabet Doghmash and Motaz Saad. 2025. Arabic hate speech identification and masking in social media using deep learning models and pre-trained models fine-tuning. *Preprint*, arXiv:2507.23661.

Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2021. The interplay of variant, size, and task type in Arabic pre-trained language models. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, Kyiv, Ukraine (Online). Association for Computational Linguistics.

Sammer Kamal, Hamada Nayel, and Ahmed Shalaby. 2024. Enhancing hadith classification using statistical and semantic feature fusion and dimension reduction. In *2024 12th International Japan-Africa Conference on Electronics, Communications, and Computations (JAC-ECC)*, pages 160–163.

Marwa Khairy, Tarek M Mahmoud, Ahmed Omar, and Tarek Abd El-Hafeez. 2024. Comparative performance of ensemble machine learning for arabic cyberbullying and offensive language detection. *Language Resources and Evaluation*, 58(2):695–712.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Deepawali Sharma, Vedika Gupta, Vivek Kumar Singh, and Bharathi Raja Chakravarthi. 2025. Stop the hate, spread the hope: An ensemble model for hope speech detection in english and dravidian languages. *ACM Transaction Asian Low-Resource Language Information Processing*.

Mahmoud Sobhy, Ahmed H AbuElAtta, Ahmed A El-Sawy, and Hamada Nayel. 2025. Swarm intelligence for handling out-of-vocabulary in Arabic Dialect Identification with different representations. *Neural Computing and Applications*, pages 1–27.

Abu Bakr Soliman, Kareem Eissa, and Samhaa R. El-Beltagy. 2017. Aravec: A set of arabic word embedding models for use in arabic nlp. *Procedia Computer Science*, 117:256–265. Arabic Computational Linguistics.

Wajdi Zaghouani and Md. Rafiul Biswas. 2025. An annotated corpus of arabic tweets for hate speech analysis. *Preprint*, arXiv:2505.11969.

Wajdi Zaghouani, Md Rafiul Biswas, Mabrouka Bessghaier, Shimaa Ibrahim, Georgio 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, Torino, Italia. ELRA and ICCL.