ANLP-UniSo at MAHED Shared Task: Detection of Hate and Hope Speech in Arabic Social Media based on XLM-RoBERTa and Logistic Regression

Yasmine El Abed[1], **Mariem Ben Arbia**[1], **Saoussen Ben Chaabene**[1], **Omar Trigui**[1,2]
[1]University of Sousse, Sousse 4000, Tunisia
[2]MIRACL Laboratory, Sfax, Tunisia
yasmine.elabed@isgs.u-sousse.tn, mariem.benarbia@isgs.u-sousse.tn,
saoussenbenchabane@isgs.u-sousse.tn, Omar.Trigui@isgs.u-sousse.tn

Abstract

In this paper, we present our system for Subtask 1 of the MAHED 2025 shared task, which involves classifying Arabic text into three categories: Hate, Hope, and not_applicable. Our methodology integrates XLM-RoBERTa embeddings with supervised ML and deep learning techniques. After applying Arabic-specific preprocessing, we extract contextual embeddings and mitigate class imbalance using SMOTE . We then train LR and LSTM classifiers on the augmented features space, supplemented by a similarity calculation with Zero-Shot for prediction validation. The system was evaluated in two phases: using the initial validation set, and the official updated datasets. Results show competitive performance, particularly in boosting recall for minority classes with a macro score of 0.60.

Keywords: Arabic text classification; Hate speech; Hope speech; XLM-RoBERTa; SMOTE; LR; LSTM.

1 Introduction

Hate speech and hope speech are increasingly important phenomena in online discourse, with significant implications for social harmony, policy-making, and content moderation (Ahmad et al., 2024; Charfi et al., 2024). Detecting such speech in Arabic presents unique challenges due to the language's morphological complexity, dialectal diversity, and scarcity of labeled datasets (Haidar, 2021). These challenges are further compounded by class imbalance, where certain categories are underrepresented, leading to biased models and reduced generalization (Haj Ahmed et al., 2024).

Recent works on hope speech detection have shown that transformer-based models combined with transfer learning can achieve competitive results, but most research has focused on English datasets, leaving Arabic-specific challenges under-explored.

The MAHED 2025 shared task (Zaghouani et al., 2025) provides an opportunity to address this gap by focusing on a three-way classification task for Arabic language texts into hate, hope, and not_applicable categories. This task is critical in combating online hostility while also promoting positive and constructive discourse online. However, existing literature and models often focus exclusively on identifying offensive content, neglecting the importance of detecting encouraging or hopeful messages (Chakravarthi et al., 2022). Moreover, many approaches suffer from language imbalance, inadequate handling of class imbalance, and poor generalization across tasks involving morphologically rich languages such as Arabic (Antoun et al., 2020).

To tackle these challenges, we developed a classification pipeline specifically tailored for Arabic. Our pipeline integrates XLM-RoBERTa embeddings, followed by class balancing (SMOTE), and classification using a logistic regression model. We also explored a comparative approach using a Bidirectional LSTM neural network (Conneau et al., 2020).

Our system was submitted to Subtask 1 of MAHED 2025. On the official leaderboard, our final submission achieved a macro Score of 0.60.

2 Background

The MAHED 2025 shared task (Zaghouani et al., 2025) focuses on the multimodal detection of hope and hate emotions in Arabic online content. In particular, Subtask 1, which we participated in, targets a three-way text classification problem. Given short Arabic text segments collected from social media platforms, the system must predict one of three categories: *Hate*, *Hope*, or *Not_Applicable*.

Proceedings of The Third Arabic Natural Language Processing Conference, pages 639–644

These texts are written in either Modern Standard Arabic or dialectal variations commonly used on social media. They often include informal spelling, emojis, and punctuation typical of online discourse.

- **Hate**: The text contains hate speech or promotes harmful or hostile rhetoric. Example: أي جاهل غبي متخلف مضحوك عليه يقول الشيعه مسلمين

- **Hope**: The text expresses positive sentiment, optimism, or encourages unity and peaceful resolution. Example: آلاتنسوا اهل حلب من دعواتكم اللهم فرج عنهم وارحمهم وغفر لموتاهم اللهم اميين

- **Not_Applicable**: The text is neutral or unrelated to the conflict context, and does not express hate or hope. Example: آو لانني احب الاشياء الراقية

Several datasets support this shared task, including *So Hateful!* (Zaghouani et al., 2024), *EmoHopeSpeech* (Zaghouani and Biswas, 2025b), and an *Annotated Corpus of Arabic Tweets for Hate Speech* (Zaghouani and Biswas, 2025a). These resources provide the foundation for the MAHED dataset and highlight the novelty of our approach.

Sentiment and emotion analysis in Arabic has gained increasing attention due to the complexity and richness of the language. A number of shared tasks and benchmarks have emerged to foster research in this area.

Recently, the ArAIEval 2022 shared task (Hasanain et al., 2023) addressed Arabic implicit emotion detection using tweets. The participating systems primarily employed transformer-based models such as AraBERT and multilingual BERT, achieving notable performance. In a related context, Daouadi et al (Daouadi et al., 2024) have shown that applying data augmentation alongside fine-tuning transformer models (e.g., ensemble of pre-trained models) can effectively mitigate class imbalance and significantly improve F1 scores in Arabic hate speech detection tasks.

These studies demonstrate the importance of robust pre-processing, balanced datasets, and fine-tuned multilingual models in Arabic text classification tasks.

3 System Overview

To address the challenge of Arabic text classification into *hate*, *hope*, and *not_applicable* categories, we propose a system that combines a multilingual transformer-based model with machine learning techniques, data balancing, and data augmentation strategies. Our system is composed of six steps as follows:

3.1 Data Preprocessing and Tokenization

Arabic sentences were cleaned by removing diacritics, URLs, emojis, elongations, stop words, and rare tokens. The resulting text was normalized and tokenized at the sentence level using XLMRoberta-Tokenizer, to meet transformer input specifications.

3.2 Data Augmentation via Back-Translation

To address class imbalance in the hope category, back-translation was applied. Sentences were translated from Arabic to English and then back to Arabic using automated translation APIs, producing semantically equivalent yet lexically diverse samples.

3.3 Feature Representation with XLM-R

XLM-RoBERTa (XLM-R) from Hugging Face Transformers was used as the feature extractor due to its effectiveness in multilingual and low-resource settings. Trained on 100 languages, including Arabic, it captures both syntactic and semantic nuances, making it suitable for Arabic social media text containing dialectal variations.

3.4 Label Encoding

To prepare the text data for machine learning, the categorical labels (hate, hope, not_applicable) were converted into numerical format using integer label encoding. This transformation is essential for compatibility with scikit-learn and deep learning frameworks, which require numerical inputs for both training and evaluation. The mapping preserved the original class distribution while enabling efficient optimization of loss functions (for Logistic Regression and LSTM).

3.5 Data Balancing with SMOTE

To mitigate the class imbalance problem, we employed SMOTE (Synthetic Minority Oversampling Technique) on the training set. SMOTE generates synthetic samples of the minority classes (hope and not_applicable) in the embedding space, thus helping the classifier learn more balanced decision boundaries 1.

3.6 Classification Models

To perform the classification task, we trained and evaluated a diverse set of models. For traditional ML approaches, we employed Logistic Regression. While for neural network architectures, we explored LSTM models

4 Experimental Setup

Table 1 presents the class distribution of the training set, which contains 6,890 Arabic sentences distributed across the three categories.

Table 1: Class distribution in the MAHED 2025 training set

Class	Samples
Not Applicable	3697
Hope	1892
Hate	1301

The dataset was released in two phases. In the First phase, we received a training set and an initial validation set. In the Second phase, the organizers updated the validation set and provided an unseen test set.

4.1 Development Phase

4.1.1 Logistic Regression Model Performance in the Developpement Phase

The LR model demonstrated competitive performance in the ternary classification task, achieving a macro F1-score of 0.57 which calculated as the average of class-wise F1-scores: 0.47, 0.51, 0.72. Class-wise metrics reveal nuanced behavior: the model exhibited strong performance for the majority class "Not_applicable" (precision = 0.64, recall = 0.81, F1 $\bar{0}$.72), suggesting effective handling of prevalent patterns. However, minority classes like "Hate" (precision $\bar{0}$.58, recall $\bar{0}$.40) and "Hope" (precision = 0.63, recall = 0.43) showed lower recall, indicating challenges in capturing these instances (see Appendix, Table 2). The confusion matrix further highlights this imbalance, with notable misclassifications of "Hate" and "Hope" samples as "Not_applicable" (see Appendix, Figure 2).

4.1.2 LSTM Model Performance in the Developpement Phase

The LSTM model achieved a macro F1-score of 0.56, reflecting a trade-off between recall and precision across classes. It showed strong recall for minority classes ("Hate": 0.69, "Hope": 0.70), outperforming Logistic Regression in capturing these instances, but with lower precision (0.47, 0.48), indicating higher false positives. Conversely, the majority class "Not_applicable" had high precision (0.71) but low recall (0.44), suggesting conservative predictions and misclassifications to other classes (see Appendix, Table 3). The confusion matrix confirmed these trends, with notable true positives for minority classes but elevated false positives (see Appendix, Figure 3).

4.2 Test Phase

4.2.1 Logistic Regression Model Performance in the Test Phase

The LR model achieved a macro F1-score of 0.57 during testing, demonstrating varied performance across classes. For minority classes, a trade-off between precision and recall was observed: "Hate" showed high recall (0.68) but moderate precision (0.45), while "Hope" had more balanced metrics (0.53 precision, 0.65 recall). The majority class, "Not_applicable", exhibited strong precision (0.70) but lower recall (0.51), indicating conservative predictions(see Appendix, Table 4). The confusion matrix revealed challenges in distinguishing emotional content ("Hate"/"Hope") from neutral cases, with frequent misclassifications favoring the majority class(see Appendix, Figure 4).

4.2.2 LSTM Model Performance in the Test Phase

The LSTM model achieved a macro F1-score of 0.56, reflecting a trade-off between recall and precision across classes. It showed strong recall for minority classes ("Hate": 0.74, "Hope": 0.70), outperforming Logistic Regression in capturing these instances, but with lower precision (0.43, 0.51 respectively), indicating higher false positives (47 and 78 misclassified as "Not_applicable"). Conversely, the majority class ("Not_applicable") had high precision (0.73) but low recall (0.42), suggest-

ing conservative predictions and frequent misclassifications to other classes (216 as "Hate", 252 as "Hope") (see Appendix, Table 5). The confusion matrix confirmed these trends, with 192 true positives for "Hate" and 288 for "Hope", but elevated false positives that reveal the model's tendency to default to neutral classifications (see Appendix, Figure 5).

5 Results

Our system einvolves two distinct approaches for classifying Arabic texts into three categories: a traditional machine learning model (Logistic Regression) and a deep learning architecture (LSTM). The comparative analysis reveals important insights about their respective strengths and weaknesses. For the official phase developpement, we have obtained the following results:

- Macro F1-score for LR: **0.5658**
- Macro F1-score for LSTM:**0.5561**

For the official phase test, we have obtained the following results:

- Macro F1-score for LR: **0.5740**
- Macro F1-score for LSTM: **0.5551**

The logistic regression model offers more transparent decision-making processes compared to the black-box nature of LSTM (see Appendix, Figure 1).

5.1 Error Analysis

The logistic regression classification model achieved a score of 0.6, meaning that 40% of the data was misclassified. To improve performance, it would be beneficial to explore other models like few-shot or one-shot learning, which can better understand the meaning of words and perform classification with minimal training data.
Although our system achieved competitive results on the MAHED 2025 shared task, several systematic misclassifications were observed. A recurrent error pattern was the confusion between *Hate* and *Hope*. For instance, the sentence ألله يلعن حيطتكم الغبيه صغار الشرقيه وستبغون صغار was annotated as *Hate*, but the model predicted *Hope*. This indicates that the presence of certain positive lexical cues can mislead the classifier, even when the overall semantic orientation is hostile. Similarly, the text آحينما انظر الي مثل هؤلاء السعداء في الامم .. آ was misclassified as *Hope* although it was labeled as *Hate*, reflecting the difficulty of capturing sarcasm and figurative expressions. Another frequent source of error was the misclassification of hateful or politically charged content as *Not_Applicable*. Conversely, the system sometimes failed to recognize hopeful messages. For example, آما جمعه الاسلام لن تفرقه السياسه الجزائر المغربْ, which conveys unity and optimism, was annotated as *Hope* but predicted as *Hate*.

5.2 Discussion

Our system was built by combining contextual word embeddings from XLM-RoBERTa with a logistic regression classifier. A comparison with an LSTM classifier did not reveal a substantial improvement, and logistic regression proved to be more stable and consistent across evaluation settings. These results suggest that while multilingual transformers such as XLM-RoBERTa can provide a strong baseline for Arabic text classification.

6 Conclusion

In this work, we have described our participation in MAHED 2025 sub-task 1. We have developed a system which classifies sentences extracted from Arabic social media in three categories : Hate, Hope and not_applicable. Our system is based on a combination of XLM-RoBERTa embeddings with Logistic Regression and LSTM classifiers, augmented by SMOTE for class imbalance and back-translation. For future work, we plan to experiment with Arabic-specific pretrained models such as MARBERT, as well as few-shot and one-shot learning methods to better capture semantic nuances.

References

Ahmad, M., Usman, S., Farid, H., Ameer, I., Muzammil, M., Hamza, A., Sidorov, G., and Batyrshin, I. (2024). Hope speech detection using social media discourse

(posi-vox-2024): A transfer learning approach. *Journal of Language and Education*, 10(4):31–43.

Antoun, W., Baly, F., and Hajj, H. (2020). AraBERT: Transformer-based model for arabic language understanding. *arXiv preprint arXiv:2003.00104*.

Chakravarthi, B. R., Muralidaran, V., Hande, A., Philip, J., McCrae, J. P., Buitelaar, P., Ponnusamy, R., Suryawanshi, S., Sherly, E., and Bandyopadhyay, S. (2022). Hope speech detection for equality, diversity and inclusion. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion (LT-EDI 2022)*, pages 378–388, Dublin, Ireland. Association for Computational Linguistics.

Charfi, A., Besghaier, M., Akasheh, R., Atalla, A., and Zaghouani, W. (2024). Hate speech detection with adhar: a multi-dialectal hate speech corpus in arabic. *Frontiers in Artificial Intelligence*, 7.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Daouadi, K. E., Boualleg, Y., and Haouaouchi, K. E. (2024). Ensemble of pre-trained language models and data augmentation for hate speech detection from arabic tweets. *CoRR*, abs/2407.02448.

Haidar, B. (2021). A survey on hate speech detection in arabic social media. *Journal of King Saud University - Computer and Information Sciences*.

Haj Ahmed, A., Yew, R.-J., Minocher, X., and Venkatasubramanian, S. (2024). Navigating dialectal bias and ethical complexities in levantine arabic hate speech detection. *arXiv preprint arXiv:2412.10991*.

Hasanain, M., Alam, F., Mubarak, H., Abdaljalil, S., Zaghouani, W., Nakov, P., Da San Martino, G., and Freihat, A. A. (2023). Araieval shared task: Persuasion techniques and disinformation detection in arabic text. *CoRR*, abs/2311.03179.

Zaghouani, W. and Biswas, M. R. (2025a). An annotated corpus of arabic tweets for hate speech analysis. *arXiv preprint arXiv:2505.11969*.

Zaghouani, W. and Biswas, M. R. (2025b). Emohopespeech: An annotated dataset of emotions and hope speech in english and arabic. *arXiv preprint arXiv:2505.11959*.

Zaghouani, W., Biswas, M. R., Bessghaier, M., Ibrahim, S., Mikros, G., Hasnat, A., and Alam, F. (2025). MAHED shared task: Multimodal detection of hope and hate emotions in arabic content. In *Proceedings of the Third Arabic Natural Language Processing Conference (ArabicNLP 2025)*, Suzhou, China. Association for Computational Linguistics.

Zaghouani, W., Mubarak, H., and Biswas, M. R. (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 Appendix

A.1 Tables

Table 2: Classification report of development phase for Logistic Regression

Class	Precision	Recall	F1-score
Hate	0.58	0.40	0.47
Hope	0.63	0.43	0.51
Not_applicable	0.64	0.81	0.72

Table 3: Classification report of development phase for LSTM

Class	Precision	Recall	F1-score
Hate	0.47	0.69	0.56
Hope	0.48	0.70	0.57
Not_applicable	0.71	0.44	0.54

Table 4: Class-wise performance metrics

Class	Precision	Recall	F1-score
Hate	0.45	0.68	0.54
Hope	0.53	0.65	0.58
Not_applicable	0.70	0.51	0.59

Table 5: Class-wise performance metrics

Class	Precision	Recall	F1-score
Hate	0.43	0.74	0.54
Hope	0.51	0.70	0.59
Not_applicable	0.73	0.42	0.53

A.2 Figures

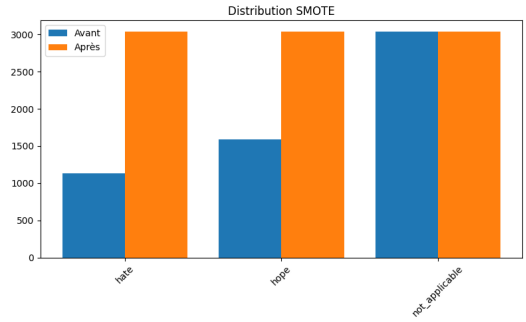

Figure 1: Visualizing Data Balancing with SMOTE

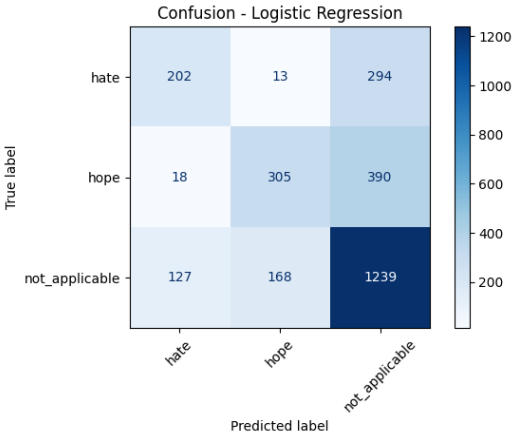

Figure 2: Confusion Matrix of development phase for Logistic Regression

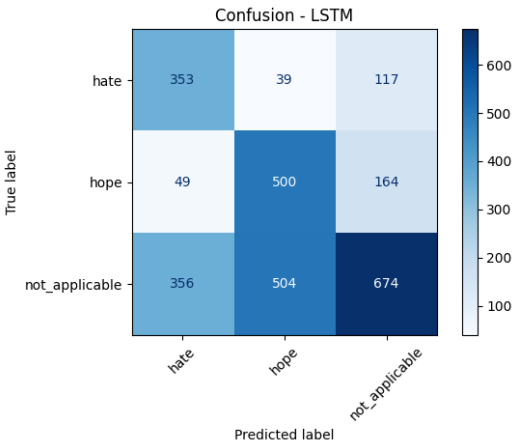

Figure 3: Confusion Matrix of development phase for LSTM

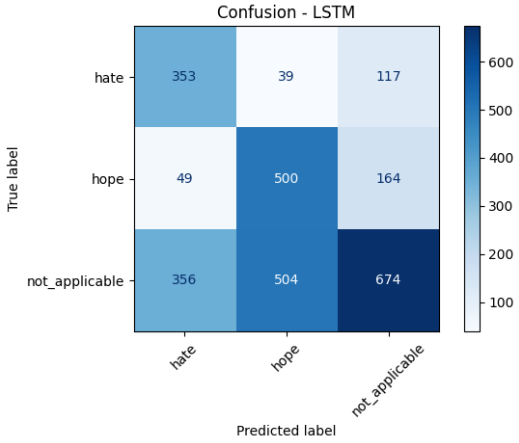

Figure 4: Confusion Matrix of test phase for Logistic Regression

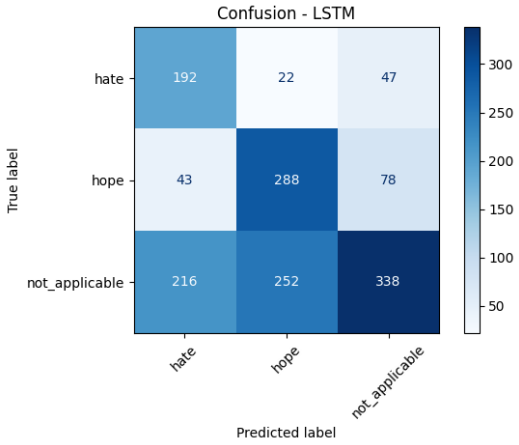

Figure 5: Confusion Matrix of test phase for LSTM

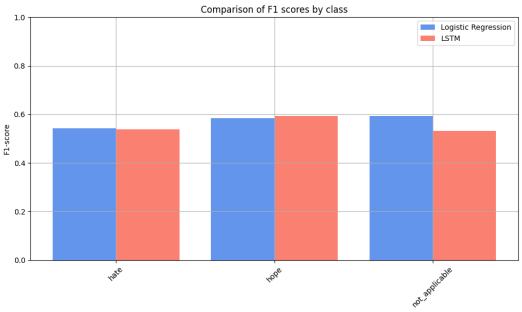

Figure 6: Comparative Analysis of Logistic Regression and LSTM Models