SmolLab_SEU at MAHED Shared Task: Do Arabic-Native Encoders Surpass Multilingual Models in Detecting the Nuances of Hope, Hate, and Emotion?

Md. Abdur Rahman[1] **Md Sabbir Dewan**[2*] **Md. Tofael Ahmed Bhuiyan**[1*]
Md. Ashiqur Rahman[3]
[1]Department of Computer Science and Engineering, Southeast University, Bangladesh
[2]School of IT, Murdoch University, Australia
[3]School of Computer and Cyber Sciences, Augusta University, Georgia, USA
2021200000025@seu.edu.bd, 35254922@student.murdoch.edu.au,
tofael1104@gmail.com, mdrahman@augusta.edu

Abstract

The dynamic interplay of hope and hate speech on Arabic social media presents a critical challenge for content moderation and digital discourse analysis. This paper presents our systems for the MAHED 2025 shared task on Multimodal Detection of Hope and Hate Emotions in Arabic Content, addressing the two text-based subtasks. Our approach centers on a systematic, empirical comparison of Arabic-native versus large-scale multilingual Transformer encoders to determine the optimal pre-training strategy for this nuanced domain. Comprehensive evaluations demonstrate the clear superiority of Arabic-native models, with our ARBERTv2-based system achieving the highest performance. We secured 11th place in Subtask 1 with a macro F1-score of 0.682 and 5th place in Subtask 2 with a macro F1-score of 0.514. Error analysis reveals persistent challenges in interpreting implicit language and overcoming severe class imbalance, particularly in distinguishing targeted hate from general offensiveness. This work contributes a robust benchmark for this comparison and underscores the importance of language-specific pre-training for nuanced affective computing in Arabic.

1 Introduction

The proliferation of social media has transformed the Arabic-speaking world into a complex information ecosystem where constructive and destructive narratives compete. This duality is starkly represented by the concurrent rise of hate speech and hope speech, making their automatic detection paramount for content moderation and understanding online discourse (Mubarak et al., 2017). While early Arabic NLP efforts focused on general sentiment, the community has shifted towards more nuanced, high-impact tasks like hate speech detection.

The advent of large pre-trained Transformers (Devlin et al., 2019) has revolutionized this field, becoming the de facto standard. However, a fundamental architectural question remains for Arabic: do exclusively pre-trained Arabic-native models offer a performance advantage over large-scale multilingual models like XLM-RoBERTa (Conneau et al., 2020)? The latter may offer broader linguistic generalization, while the former might better capture language-specific nuances, dialects, and cultural contexts.

The MAHED 2025 shared task at ArabicNLP 2025 (Zaghouani et al., 2025) provides an ideal testbed to investigate this question. Its focus on the duality of hope and hate speech, alongside a complex emotion classification challenge, pushes beyond simple toxicity detection. In this paper, we present our systems for Subtask 1 and 2, systematically evaluating a diverse suite of Arabic-native and multilingual Transformer models to empirically answer this question. Our implementation is made publicly available to ensure reproducibility.[1]

The main contributions of our work:

- We present a systematic empirical comparison of seven distinct Transformer architectures, investigating the performance trade-offs between Arabic-native and multilingual encoders for nuanced affective computing.
- We developed robust systems for both subtasks, including a cascaded pipeline for Subtask 2 that explicitly models the hierarchical dependencies between offensive and hate speech detection, allowing for specialized classifier optimization.
- We establish a strong benchmark demonstrating the clear superiority of Arabic-native models, with our ARBERTv2-based system

[*]Authors contributed equally to this work.

[1]https://github.com/borhanitrash/ArabicNLP-EMNLP

Proceedings of The Third Arabic Natural Language Processing Conference, pages 658–664
November 8-9, 2025

achieving competitive performance. Our detailed error analysis further illuminates the specific challenges posed by semantic ambiguity and class imbalance in this domain.

2 Related Works

The automatic detection of nuanced affective states, including hate and hope speech, is a critical area of research in Arabic Natural Language Processing (NLP). Our work builds upon recent advancements in deep learning for sentiment and emotion analysis, particularly those leveraging Transformer-based architectures.

Recent efforts in Arabic affective computing highlight the success of pre-trained models. For instance, Cherrat et al. (2024) demonstrated the efficacy of AraBERT-based models for sentiment analysis across Standard Arabic and Moroccan dialect, showcasing their ability to capture complex linguistic features. Similarly, for Arabic tweet classification, Al-Onazi et al. (2023) developed a framework combining Deep Belief Networks with advanced hyperparameter optimization, while Elfaik et al. (2023) engineered a feature-fusion model using hybrid RNN-CNN architectures to tackle multi-label affect analysis. These studies affirm the power of deep learning for Arabic text but often focus on general sentiment or a broad spectrum of emotions.

This trend of applying sophisticated deep learning models extends to other languages and related tasks. Researchers have employed CNNs for detecting violent incitement in Urdu (Khan et al., 2024), hierarchical attention networks for depression detection from English tweets (Khafaga et al., 2023), and various hybrid architectures for emotion classification in Afan Oromo (Abdella and Sori, 2024). Furthermore, the field is advancing towards more complex methodologies, such as the tri-modal (text, audio, visual) graph neural networks for emotion recognition proposed by Al-Saadawi and Das (2024).

While these studies establish the effectiveness of Transformer models, a critical gap remains in the direct, empirical comparison of Arabic-native versus multilingual pre-training strategies for the complex, concurrent detection of hope, hate, and fine-grained emotions. Our work addresses this gap by leveraging the MAHED 2025 shared task as a rigorous testbed to provide a robust benchmark and a detailed analysis of model performance on this challenging domain.

Split	**Instances**	**Unique Words**	**Total Words**
Train	6,890	62,744	147,285
Validation	1,476	17,553	30,731
Test	1,477	17,891	31,492

Table 1: Dataset statistics for Subtask 1.

Split	**Instances**	**Unique Words**	**Total Words**
Train	5,960	45,015	115,279
Validation	1,277	13,726	25,346
Test	1,278	13,339	24,596

Table 2: Dataset statistics for Subtask 2.

3 Task and Dataset Description

We participated in the two text-based tracks of the MAHED 2025 shared task (Zaghouani et al., 2025), which provides a standardized framework to evaluate systems on challenging affective computing tasks in Arabic. We formalize the subtasks as follows:

Subtask 1: Hate and Hope Speech Classification. A three-way classification problem where the input is an Arabic text and the output is a single label from the set {`hate`, `hope`, `not_applicable`}. For example, a text translating to "All immigrants are thieves and criminals, they must be deported immediately" is labeled as `hate`.

Subtask 2: Emotion, Offensive, and Hate Detection. A multi-output classification problem with a hierarchical dependency. Given an Arabic text, the system must predict: (1) an emotion from a set of 12 labels (e.g., `anger`, `joy`); (2) a binary label indicating if the text is `offensive`; and (3) if offensive, a binary label indicating if it constitutes targeted `hate`. For instance, a text translating to "You donkey, why did you forget the keys?" is labeled as {`anger`, `yes`, `not_hate`}, distinguishing general offense from targeted hate.

The task organizers provided two annotated datasets (Zaghouani et al., 2024; Biswas and Zaghouani, 2025a,b) comprising text from online sources in both Modern Standard and dialectal Arabic. Dataset statistics are detailed in Table 1 and Table 2. The primary evaluation metric for both subtasks is the macro-averaged F1-score. For a more comprehensive analysis, we also report accuracy, and macro-averaged precision and recall.

4 Methodology

Our approach involves fine-tuning both multilingual and Arabic-native Transformer models (Vaswani et al., 2017), which excel at capturing the contextual cues necessary for nuanced hate and hope speech detection. We employed distinct strategies for the Hate and Hope Speech Classification (Figure 1) and the Emotion, Offensive, and Hate Detection (Figure 2) subtasks.

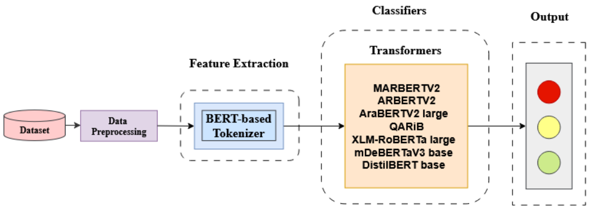


Figure 1: Schematic process for Hate and Hope Speech Classification.

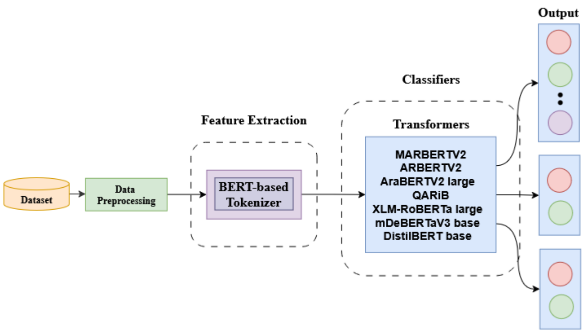


Figure 2: Schematic process for Emotion, Offensive, and Hate Detection.

4.1 Data Preprocessing

We implemented a unified text normalization pipeline for both subtasks prior to model-specific tokenization. The pipeline systematically removed URLs, user mentions, and hashtags, then normalized whitespace and filtered out non-Arabic characters. The cleaned text was subsequently processed using the AutoTokenizer corresponding to each pre-trained model. All input sequences were either padded or truncated to a fixed maximum length, generating `input_ids` and `attention_mask` tensors for model consumption.

4.2 Transformer-Based Models

Our selection of encoders was designed to evaluate a diverse range of pre-training objectives and linguistic specializations. Our model suite included Arabic-native encoders such as MARBERTV2 (UBC-NLP/MARBERTv2)[2] (Abdul-Mageed et al., 2021), ARBERTV2 (UBC-NLP/ARBERTv2)[3] (Abdul-Mageed et al., 2021), AraBERTV2 large (aubmindlab/bert-large-arabertv2)[4] (Antoun et al., 2020), and QARiB (ahmedabdelali/bert-base-qarib)[5] (Abdelali et al., 2021). These were complemented by powerful multilingual models, including XLM-RoBERTa large (FacebookAI/xlm-roberta-large)[6] (Conneau et al., 2020), mDeBERTaV3 base (microsoft/mdeberta-v3-base)[7] (He et al., 2021), and the computationally efficient DistilBERT base (distilbert/distilbert-base-multilingual-cased)[8] (Sanh et al., 2019). Each model was adapted for the downstream tasks as described below.

For Subtask 1, framed as a standard sequence classification problem, we fine-tuned each Transformer encoder by appending a sequence classification head. This head comprises a linear layer that takes the final hidden-state representation of the [CLS] token as input to produce logits for the three target classes. The entire fine-tuning process was managed using the Hugging Face Trainer API (Wolf et al., 2020), which optimized a standard Cross-Entropy Loss function. To prevent overfitting, we integrated an `EarlyStoppingCallback`, configured to monitor the macro F1-score on the official validation set and halt training after 3 epochs without improvement. The model checkpoint yielding the highest validation F1-score was preserved for the final test set evaluation.

In contrast, for Subtask 2, we addressed the task's explicit hierarchical dependency by designing a cascaded pipeline of three independently optimized classifiers. This modular design avoids the potential negative interference of joint multi-task optimization and allows each model to specialize. The pipeline consists of: an Emotion Classifier (12-class), an Offensive Classifier (binary), and a Hate Classifier (binary). The Hate classifier was trained exclusively on the subset of training data labeled as Offensive. During inference, test instances are processed in parallel by the Emotion and Offensive models; instances classified as Offensive are then routed to the Hate classifier for the final prediction. Each model in this pipeline was fine-tuned

[2]https://huggingface.co/UBC-NLP/MARBERTv2
[3]https://huggingface.co/UBC-NLP/ARBERTv2
[4]https://huggingface.co/aubmindlab/bert-large-arabertv2
[5]https://huggingface.co/ahmedabdelali/bert-base-qarib
[6]https://huggingface.co/FacebookAI/xlm-roberta-large
[7]https://huggingface.co/microsoft/mdeberta-v3-base
[8]https://huggingface.co/distilbert/distilbert-base-multilingual-cased

using a custom PyTorch loop, employing a class-weighted Cross-Entropy Loss to counteract severe label imbalance. Model selection for each of the three components was based on the highest macro F1-score achieved on the validation dataset.

All experiments were conducted with the AdamW optimizer (Loshchilov and Hutter, 2017) and utilized mixed-precision (FP16) training for computational efficiency. The specific hyperparameters for all models are detailed in Table 3.

Model	LR	WD	BS	EP
Subtask 1: Hate and Hope Classification				
MARBERTV2	2e-5	0.01	32	10
ARBERTV2	2e-5	0.01	32	10
AraBERTV2 large	1e-5	0.01	32	7
QARiB	2e-5	0.01	32	10
XLM-RoBERTa large	2e-5	0.01	16	10
mDeBERTaV3 base	2e-5	0.01	16	10
DistilBERT base	2e-5	0.01	16	10
Subtask 2: Emotion, Offensive, Hate				
MARBERTV2	2e-5	-	16	8
ARBERTV2	2e-5	-	16	8
AraBERTV2	2e-5	-	16	8
QARiB	2e-5	-	16	8
XLM-RoBERTa large	2e-5	-	16	8
mDeBERTaV3 base	2e-5	-	16	8
DistilBERT base	2e-5	-	16	8

Table 3: Hyperparameters used for fine-tuning. LR: Learning Rate, WD: Weight Decay, BS: Per-device Batch Size, EP: Max Epochs.

5 Result Analysis

This section presents the performance of our Transformer-based models on the MAHED 2025 shared task. All models were evaluated using the official metrics: accuracy, and macro-averaged precision, recall, and F1-score, with the macro F1-score serving as the primary metric for comparison. The comprehensive results for both subtasks are detailed in Table 4.

In Subtask 1, the Arabic-native models demonstrated a clear advantage over their multilingual counterparts. ARBERTv2 emerged as the top-performing system, achieving the highest macro F1-score of 0.6824 and the best accuracy of 0.6879. This strong performance is likely attributable to its pre-training on a large corpus of Arabic social media and web data, which aligns closely with the task's domain. Notably, MARBERTv2 secured the highest precision at 0.6824, indicating its proficiency in correctly identifying positive instances, albeit with a slightly lower overall F1-score. Other Arabic-specific models like QARiB and the multilingual mDeBERTaV3 base also delivered competitive results, underscoring the effectiveness of modern Transformer architectures. Conversely, AraBERTv2 large and DistilBERT base lagged behind, suggesting that either model scale or pre-training objective was less suited to this specific classification challenge.

Model	Accuracy	Precision	Recall	F1 Score
Subtask 1: Hate and Hope Speech Classification				
MARBERTv2	0.6804	**0.6824**	0.6562	0.6665
ARBERTv2	**0.6879**	0.6794	**0.6939**	**0.6824**
AraBERTv2 large	0.6269	0.6547	0.5714	0.5802
QARiB	0.6770	0.6664	0.6831	0.6738
XLM-RoBERTa large	0.6567	0.6514	0.6652	0.6554
mDeBERTaV3 base	0.6798	0.6716	0.6794	0.6729
DistilBERT base	0.6330	0.6258	0.6124	0.6110
Subtask 2: Emotion, Offensive, and Hate Detection				
MARBERTv2	0.7272	0.5040	0.5163	0.5078
ARBERTv2	0.7089	**0.5316**	**0.5257**	**0.5142**
AraBERTv2 large	0.6922	0.4765	0.4575	0.4593
QARiB	**0.7415**	0.5259	0.4943	0.4915
XLM-RoBERTa large	0.6896	0.4609	0.4564	0.4506
mDeBERTaV3 base	0.6907	0.4498	0.4619	0.4504
DistilBERT base	0.6468	0.3761	0.3801	0.3749

Table 4: Performance comparison of all evaluated models for Subtask 1 and Subtask 2. The best score in each column is highlighted in **bold**.

For the more complex, multi-output Subtask 2, ARBERTv2 once again demonstrated superior performance, leading across all macro-F1 (0.5142), precision (0.5316), and recall (0.5257) metrics. Its consistent success across both subtasks highlights the model's robustness and its ability to generalize well to related but distinct classification problems. MARBERTv2 followed closely with an F1-score of 0.5078. An interesting observation is the performance of QARiB, which achieved the highest accuracy (0.7415) but a lower F1-score of 0.4915. This discrepancy suggests the model may have excelled at predicting the majority classes (e.g., neutral emotion, no offensive) but struggled with the less frequent, yet critical, minority classes, reinforcing the importance of the macro F1-score as the primary evaluation metric in imbalanced scenarios.

Overall, our results indicate a distinct performance advantage for Arabic-native models pre-trained on diverse, user-generated content for both hate/hope speech detection and nuanced emotion classification. The performance gap between the two subtasks, with F1-scores being considerably lower in Subtask 2, underscores the inherent difficulty of the multi-output, hierarchically-dependent classification challenge. A detailed error analysis is provided in Appendix A.

6 Conclusion

In this paper, we presented our systems for the MAHED 2025 shared task, systematically evaluating Arabic-native and multilingual Transformer models on hope, hate, and emotion detection. Our findings consistently demonstrate the superiority of Arabic-native encoders, with our **ARBERTv2**-based system emerging as the top-performing model across both subtasks, achieving a macro F1-score of **0.682** (11th place) in Subtask 1 and **0.514** (5th place) in the more complex Subtask 2. The success of our cascaded classification pipeline in Subtask 2 underscores the value of modular models for hierarchical problems, though error analysis revealed persistent challenges in distinguishing nuanced emotional states and overcoming severe class imbalance, particularly for identifying targeted hate speech. Ultimately, this work contributes a robust benchmark comparing Arabic-native and multilingual models, affirming that domain- and language-specific pretraining remains crucial for tackling the subtleties of affective computing in Arabic social media.

Limitations

Our study is constrained by several limitations. Severe class imbalance, particularly in Subtask 2, significantly impacted our model's ability to detect the minority *hate* class, resulting in a conservative bias and a high number of false negatives. Our models also struggled with semantic nuance, often misclassifying subtle expressions of *hope* as neutral and confusing strong negative sentiment with targeted *hate* speech. The dataset, while valuable, may not fully capture the evolving nature of coded language across diverse Arabic dialects. Finally, our work was confined to the text modality, leaving the rich contextual information from the full multimodal task unexplored.

Acknowledgments

This work was supported by Southeast University, Bangladesh.

References

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

Sufian Kedir Abdella and Worku Jifara Sori. 2024. Detection of emotions in afan oromo social media texts using deep learning method. *Ethiopian Journal of Science and Sustainable Development*, 11(1):70–84.

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.

Badriyya B Al-Onazi, Hassan Alshamrani, Fatimah Okleh Aldaajeh, Amira Sayed A Aziz, and Mohammed Rizwanullah. 2023. Modified seagull optimization with deep learning for affect classification in arabic tweets. *IEEE Access*, 11:98958–98968.

Hussein Farooq Tayeb Al-Saadawi and Resul Das. 2024. Ter-ca-wgnn: trimodel emotion recognition using cumulative attribute-weighted graph neural network. *Applied Sciences*, 14(6):2252.

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.

Md. Rafiul Biswas and Wajdi Zaghouani. 2025a. An annotated corpus of arabic tweets for hate speech analysis. *CoRR*, abs/2505.11969.

Md. Rafiul Biswas and Wajdi Zaghouani. 2025b. Emohopespeech: An annotated dataset of emotions and hope speech in english and arabic. *CoRR*, abs/2505.11959.

El Mehdi Cherrat, Hassan Ouahi, Abdellatif BEKKAR, and 1 others. 2024. Sentiment analysis from texts written in standard arabic and moroccan dialect based on deep learning approaches. *International Journal of Computing and Digital Systems*, 16(1):447–458.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Hanane Elfaik and 1 others. 2023. Leveraging feature-level fusion representations and attentional bidirectional rnn-cnn deep models for arabic affect analysis on twitter. *Journal of King Saud University-Computer and Information Sciences*, 35(1):462–482.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *Preprint*, arXiv:2111.09543.

D Sami Khafaga, Maheshwari Auvdaiappan, K Deepa, Mohamed Abouhawwash, and F Khalid Karim. 2023. Deep learning for depression detection using twitter data. *Intelligent Automation & Soft Computing*, 36(2):1301–1313.

Muhammad Shahid Khan, Muhammad Shahid Iqbal Malik, and Aamer Nadeem. 2024. Detection of violence incitation expressions in urdu tweets using convolutional neural network. *Expert Systems with Applications*, 245:123174.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Hamdy Mubarak, Kareem Darwish, and Walid Magdy. 2017. Abusive language detection on arabic social media. In *First Workshop on Abusive Language Online 2017*, pages 52–56. Association for Computational Linguistics (ACL).

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

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 (ArabicNLP 2025)*, Suzhou, China. Association for Computational Linguistics.

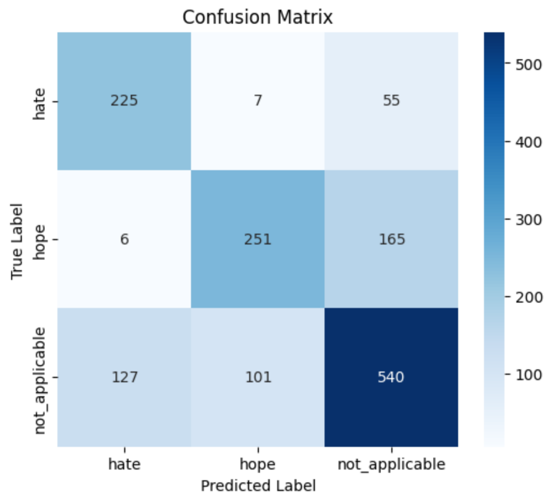
Wajdi Zaghouani, Hamdy Mubarak, and Md. Rafiul Biswas. 2024. So hateful! building a multi-label hate speech annotated Arabic dataset. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15044–15055, Torino, Italia. ELRA and ICCL.

A Error Analysis

We conducted a quantitative and qualitative error analysis of our best model, ARBERTv2, on the test set to understand its performance and limitations.

A.1 Quantitative Analysis

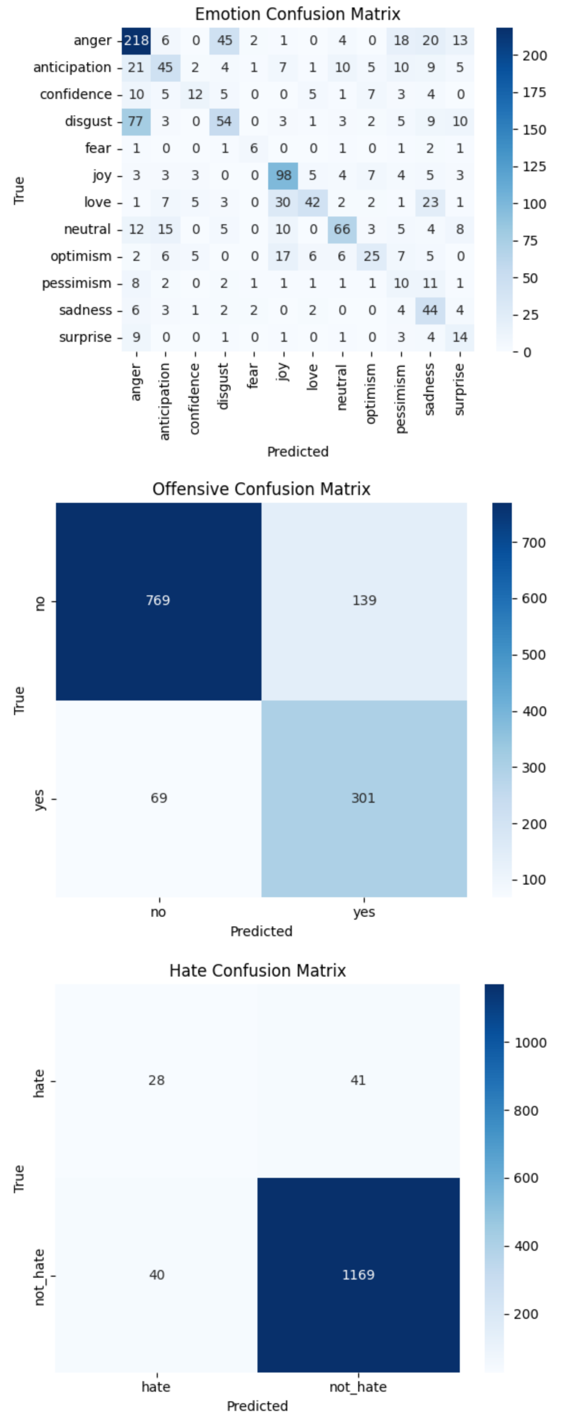
For Subtask 1, Figure 3 reveals key performance patterns. The model performs well on the `not_applicable` (540 true positives), hope (251), and `hate` (225) classes. However, it struggles with nuance, misclassifying 165 hope instances as `not_applicable`. Additionally, it misclassifies 127 `not_applicable` cases as `hate`, suggesting an oversensitivity to strong negative language.



Subtask 1

Figure 3: Confusion matrix of the proposed model (ARBERTv2) for Hate and Hope Speech Classification.

For Subtask 2, Figure 4 shows the challenges at each stage of our cascaded pipeline. In **Emotion Detection**, the model excels at high-frequency classes like `anger` (218) and `joy` (98) but struggles with fine-grained distinctions, often confusing `optimism` with `neutral` (25) or `joy` (17). In **Offensive Detection**, the model shows a conservative bias, missing 139 offensive instances (false negatives) while correctly identifying 301. Finally, severe data imbalance in the **Hate Detection** stage heavily impacts performance; the model misclassifies 41 hate cases as `not_hate` while correctly identifying only 28, showing its difficulty in distinguishing targeted hate from general offensiveness.



Subtask 2

Figure 4: Confusion matrices of the proposed model (ARBERTv2) for Emotion, Offensive, and Hate Detection.

A.2 Qualitative Analysis

Qualitative analysis of misclassifications reveals further limitations of ARBERTv2. For Subtask 1 (Figure 5), a politically charged text implying hostility was misclassified as `not_applicable` instead of `hate`, highlighting the model's difficulty with implicit threats that lack explicit slurs. For Subtask 2 (Figure 6), a text containing an expletive was mislabeled as `neutral` instead of `anger`. The formal phrasing seemingly overrode the informal expletive, highlighting challenges with mixed-tone sentences.

These observations confirm the system's primary weaknesses: handling nuanced language, distinguishing related emotions, and overcoming data imbalance, especially for targeted hate speech detection.

Subtask 1

Text Sample	Actual	Predicted
ترا شعور مرعب (What a terrifying feeling)	not_applicable	not_applicable
.ينشبك حلمك بحلمي (May your dream intertwine with my dream)	hope	hope
هنحاكم السيسي لانه كمم الافواه : (We will judge/prosecute Sisi because he silenced the mouths/voices.)	hate	not_applicable

Figure 5: Few examples of predictions produced by the proposed ARBERTv2 model on Subtask 1.

Subtask 2

Text Sample	Actual	Predicted
انا لحبيبي و حبيبي ألي (I belong to my beloved, and my beloved belongs to me.)	love,no,	love,no,
عدى خمس شهور انت متخيل؟ (Five months have passed, can you imagine?)	surprise,no,	surprise,no,
الجدير بالذكر أنه كسم #فاران (It is noteworthy that it's bullshit #Varane.)	anger,yes,not_hate	neutral,yes,not_hate

Figure 6: Few examples of predictions produced by the proposed ARBERTv2 model on Subtask 2.