MAHED Shared Task: Multimodal Detection of Hope and Hate Emotions in Arabic Content

Wajdi Zaghouani[1], **Md. Rafiul Biswas**[2], **Mabrouka Bessghaier**[1], **Shimaa Ibrahim**[2], **Georgios Mikros**[2], **Abul Hasnat**[3], **Firoj Alam**[4],
[1]Northwestern University in Qatar, [2]Hamad Bin Khalifa University
[3]APAVI.AI, France, [4]Qatar Computing Research Institute
{mbiswas,fialm}@hbku.edu.qa, wajdi.zaghouani@northwestern.edu

Abstract

This paper presents the MAHED 2025 Shared Task on Multimodal Detection of Hope and Hate Emotions in Arabic Content, comprising three subtasks: (1) text-based classification of Arabic content into hate and hope,(2) multitask learning for joint prediction of emotions, offensive content, and hate speech and (3) multimodal detection of hateful content in Arabic memes. We provide three high-quality datasets totaling over 22,000 instances sourced from social media platforms, annotated by native Arabic speakers with Cohen's Kappa exceeding 0.85. Our evaluation attracted 46 leaderboard submissions from participants, with systems leveraging Arabic-specific pre-trained language models (AraBERT, MARBERT), large language models (GPT-4, Gemini), and multimodal fusion architectures combining CLIP vision encoders with Arabic text models. The best-performing systems achieved macro F1-scores of 0.723 (Task 1), 0.578 (Task 2), and 0.796 (Task 3), with top teams employing ensemble methods, class-weighted training, and OCR-aware multimodal fusion. Analysis reveals persistent challenges in dialectal robustness, minority class detection for hope speech, and highlights key directions for future Arabic content moderation research.

1 Introduction

Online platforms increasingly require robust systems to detect harmful and pro-social content. For Arabic, this need is compounded by dialectal diversity, code-switching, and multimodal formats (e.g., memes). Community evaluations have accelerated progress on Arabic toxicity: OSACT4 standardized offensive-language detection on Twitter, and OSACT5 extended to fine-grained hate speech, highlighting label imbalance and dialectal variation (Mubarak et al., 2020, 2022). New resources further enrich supervision, such as a multi-label Arabic corpus that jointly annotates offense, hate, emotion facets, sarcasm/humor, factuality, and perceived impact (Zaghouani et al., 2024) Surveys highlight key issues, such as implicit hate, target attribution and code-switching. They further emphasize the significance of Pretrained Language Models (PLMs), such as AraBERT and ARBERT/-MARBERT (Abdelsamie et al., 2024; Antoun et al., 2020; Abdul-Mageed et al., 2021). Beyond toxicity, detecting *hope speech* has emerged in LT-EDI shared tasks and offers complementary signals for safer moderation (Chakravarthi et al., 2022). Finally, research on multimodal harmful content shows that text-only or image-only models underperform on memes, motivating vision–language fusion; Arabic meme resources emphasize language-aware OCR and robust pipelines (Kiela et al., 2020; Alam et al., 2024b).

This paper presents the **MAHED 2025 Shared Task** on **M**ultitask **A**rabic **H**armful and **E**motional content **D**etection, comprising three subtasks: (i) **Text toxicity with hope**: classify text into *hate*, *hope*, or not_applicable; (ii) **Joint modeling**: simultaneously predict an emotion label with offensive and hate labels under an explicit hierarchy; and (iii) **Multimodal memes**: detect harmful content in image–text memes.[1]. The task is designed to investigate whether multitask and multimodal modeling improve robustness under dialectal variation, label skew, sarcasm, and noise from OCR text.

Contributions. We (1) define a three-part benchmark spanning text and memes; (2) detail datasets, label schemas, and evaluation protocols aligned with prior Arabic efforts and hope-speech literature; (3) release baseline training/evaluation code and configurations for Arabic PLMs and multimodal fusion; and (4) report results and error analyses across dialects and modalities.

[1] Exact data sources, splits, and scoring scripts are detailed in https://github.com/marsadlab/MAHED2025Dataset.git

Proceedings of The Third Arabic Natural Language Processing Conference, pages 560–574
November 8-9, 2025 ©2025 Association for Computational Linguistics

2 Related Work

Scope and definitions. We study two affective poles in Arabic: hate/offense (derogatory, dehumanizing, or abusive content) and hope (constructive, prosocial, future-oriented encouragement). We cover social media text and image memes, and acknowledge Arabic-specific challenges such as dialectal variability and code switching (Arabizi). This section positions MAHED with respect to Arabic hate/offense and hope in text, joint modeling with emotions, and multimodal detection in memes.

Arabic hate and offensive language in text. Community evaluations standardized tasks and metrics, accelerating progress. Mubarak et al. (2020) introduced Arabic offensive language detection on Twitter, and Mubarak et al. (2022) extended to finer-grained hate targets, highlighting dialectal variability and class imbalance. Beyond shared tasks, Zaghouani et al. (2024) released a 15,965 tweet multi label dataset (offense, hate, emotion facets, sarcasm/humor, factuality, perceived impact), where AraBERT style encoders outperform classical baselines; a recent survey synthesizes methods, datasets, and open challenges—including implicit hate, target attribution, and code switching—informing MAHED's taxonomy and evaluation (Abdelsamie et al., 2024). Strong Arabic PLMs such as AraBERT and ARBERT/MARBERT remain standard encoders for social media classification (Antoun et al., 2020; Abdul-Mageed et al., 2021). Overall, text-only Arabic toxicity is relatively mature, while gaps persist in dialectal robustness, implicit hate, and correlated labels under class imbalance, which MAHED targets explicitly.

Hope speech and prosocial content. Hope speech is increasingly treated as a distinct class of constructive and supportive online content in the LT and EDI communities. Shared tasks report that transformer-based models consistently outperform classical approaches for hope speech classification (Chakravarthi et al., 2022). Beyond shared tasks, work on Urdu social media shows that transformer models obtain the top macro F1 for multi-class hope and hopelessness, and that careful annotation guidelines help capture nuanced expressions of hope (Balouchzahi et al., 2025). Complementary psycholinguistic analyses indicate that hope speech displays distinctive cognitive, emotional, and communicative profiles, and that tree boosting methods such as LightGBM and CatBoost can be competitive for type-level hope classification when tuning is performed (Arif et al., 2024). Theory and experiments in social psychology connect specific emotions to prosocial behavior: emotions such as hope and gratitude can motivate helping through both intrapersonal and interpersonal pathways (van Kleef and Lelieveld, 2022), and hopeful reappraisal in distressing contexts has been shown to increase charitable giving (Brethel-Haurwitz et al., 2020). Together, these results support modeling hope as a separate target alongside hate or offense in Arabic, to avoid conflation with generic positivity and to enable evaluation of prosocial language in culturally specific settings.

Emotion analysis in Arabic. Arabic emotion analysis has progressed in both text and speech, enabling fine-grained affect modeling. For social content, resources such as ArPanEmo support recognition of multiple emotions, plus neutral, and allow multi-class setups (Althobaiti, 2023). In speech, the King Saud University Emotions corpus and related datasets demonstrate that speaker gender, emotion type, and their interaction affect perception and recognition, and they provide a basis for statistical and perceptual analyses (Meftah et al., 2018, 2021). Studies on Arabic dialects report strong performance with standard classifiers, as well as with prosodic and spectral features. For example, support vector machines provide about 77 percent accuracy on Saudi dialect data (Aljuhani et al., 2021), long-term average spectrum and wavelet features yield improvements for Egyptian Arabic (Abdel-Hamid, 2020), and multistage classification schemes offer reasonable gains (Poorna and Nair, 2019). Earlier studies based on TV show speech, along with subsequent surveys, highlight the consistent roles of pitch, intensity, speaking rate, and mel-frequency cepstral coefficients (MFCCs), while also underscoring the open challenges of achieving cross-speaker and cross-dialect generalization (Klaylat et al., 2018; Meddeb et al., 2017; Nasr et al., 2024). Evidence from perceptual research indicates that prosody and lexical semantics contribute through separate yet intertwined channels, with prosodic dominance often observed (Ben-David et al., 2016). In parallel, corpus-based studies of Arabic vocabulary in religious texts highlight a wide lexical space for emotional expression, underscoring the need for culturally informed annotation and modeling choices (Salsabila et al., 2024). These findings motivate the integration of emotion signals with toxicity and prosociality labels. Additionally, in order to address label imbalance and better capture minority classes such as hope, multi-label or ordinal objectives can be adopted.

Multitask and multi-label modeling. Given correlated labels (for example, hate ⇒ offense; emotion ↔ toxicity), joint learning can improve minority classes via shared representations. In Arabic, multitask architectures that combine offense/hate with sentiment or related signals improved robustness on OSACT style data (Abu Farha and Magdy, 2020; Djandji et al., 2020). MAHED follows this paradigm in Task 2 by jointly predicting emotions, offensive content, and hate under an explicit label hierarchy.

Multimodal harmful content and Arabic memes. The Hateful Memes benchmark demonstrated the insufficiency of unimodal baselines and popularized vision–language fusion (Kiela et al., 2020). Subsequent efforts, such as MultiOFF and the SemEval 2022 MAMI task, further highlighted the benefits of fusing text and image and incorporating subtype labels (Suryawanshi et al., 2020; Fersini et al., 2022). For Arabic, Alam et al. (2024b) introduced ArMeme, a manually annotated meme dataset targeting propagandistic techniques, and established text-image fusion as essential baselines for Arabic script and domains. Building on this trend, MAHED extends the scope to Arabic memes by evaluating OCR-aware text–image fusion for both hate and hope, while leaving speech and video analysis out of scope for this edition.

Summary and link to design. From 2020 to 2025, Arabic hate/offense matured via shared tasks and PLMs, affect resources expanded, and hope remained comparatively under-resourced in Arabic. Multitask and multimodal fusion approaches have been consistently beneficial. In response, MAHED unifies hate, offense, and hope annotations for Arabic text, investigates joint learning with emotions to improve the representation of minority classes, and extends its scope to OCR-aware text–image fusion, with particular attention to dialect variation and code-switching.

3 Tasks and Datasets

The MAHED shared task consists of three subtasks: **(1)** Text-based Hope and Hate Speech Classification, **(2)** Multitask Learning for Emotion, Offensive Content, and Hate Detection, and **(3)** Multimodal Hateful Meme Detection. All content in the related datasets was sourced from public social media platforms, anonymized to protect user privacy, and annotated by native Arabic speakers. The annotation process achieved a high inter-annotator agreement, with a Cohen's Kappa score exceeding 0.85, indicating strong consistency among annotators.

Data Partition	Label	Count	Dist.
Train (6,890)	Hate	1,301	18.9%
	Hope	1,892	27.5%
	NA	3,697	53.7%
Dev (1,476)	Hate	261	17.7%
	Hope	409	27.7%
	NA	806	54.6%
Test (1,477)	Hate	287	19.4%
	Hope	422	28.6%
	NA	768	52%

Table 1: Distribution of class labels in the Task 1 dataset. NA: not_applicable

3.1 Task 1 : Text-based Hope and Hate Speech Classification

Task: The objective of the first task is to develop a model that classifies Arabic text into one of three categories: *"hate"*, *"hope"*, and *"not_applicable"*. In this context, *hate* refers to expressions that contain offensive, discriminatory, or harmful language directed toward individuals or groups based on features such as religion, nationality, ethnicity, or other protected characteristics. *Hope* refers to expressions of positive emotional content, including aspirational, motivational, or future-oriented messages, as well as statements that convey optimism, gratitude, or encouragement. The *not_applicable* category includes all remaining cases that do not contain explicit hate or hope content.

Dataset: The dataset used for this task consists of 9,843 high-quality Arabic text instances that have been carefully prepared for classification into the *"hate"*, *"hope"*, and *"not_applicable"* categories. The data is divided into three subsets: 6,890 samples for training, 1,476 for validation, and 1,477 for testing. The dataset have been obtained from the combination of three high quality datasets (Zaghouani et al., 2024; Zaghouani and Biswas, 2025b,a). Table 1 presents the label distribution across the training, validation, and test sets, reporting both the number of instances in each category and their relative proportions.

3.2 Task 2: Multitask Learning for Emotion, Offensive Content, and Hate Detection

Task. The second task addresses multitask learning for joint emotion, offensive language, and hate speech detection in Arabic text. The objectives of this task are (i) predicting a single emotion label from a predefined list of 12 emotions (*neutral, anger, anticipation, disgust, fear, joy, love, optimism, pessimism, sadness, surprise, trust*), (ii) determining whether the text is offensive (*yes/no*), and (iii) if offensive, deciding if the text is hate speech (*hate* vs. *not_hate*). This order reflects the hierarchical relationship between offensiveness and hate since all hate speech is offensive, but not all offensive content is hate speech. Specifically, texts labeled as *hate* contain offensive content directed at an identity group (e.g., religion, nationality, ethnicity, or gender). In contrast, texts labeled as *not_hate* may also be offensive but do not target specific identities, such as instances of casual or profane language without identity-based targeting.

Dataset. The dataset for this task comprises 8,515 high-quality annotated Arabic text instances, prepared for joint classification of emotions, offensive language, and hate speech. Three high quality data sources were used for curation of this shared task datasets (Zaghouani et al., 2024; Zaghouani and Biswas, 2025b,a). It is divided into three subsets: 5,960 samples for training, 1,277 for validation, and 1,278 for testing. Each instance is labeled with three layers of information aligned with the task objectives: (i) one emotion from the 12 categories, (ii) an offensiveness label (*yes/no*), and (iii) for offensive texts, a hate label distinguishing between *hate* and *not_hate*. Table 2 summarizes the distribution of these label categories across the training, validation, and test sets.

Label	**Train**	**Val**	**Test**
Emotion			
Neutral	661	137	128
Anger	1,551	331	327
Anticipation	491	121	120
Disgust	777	153	167
Fear	53	9	13
Joy	533	120	135
Love	593	135	117
Optimism	419	88	79
Pessimism	194	54	39
Sadness	335	54	68
Surprise	143	28	33
Confidence (Trust)	210	47	52
Offensive			
Yes	1,744	363	370
No	4,216	914	908
Hate (if offensive)			
Hate	303	68	69
Not hate	1,441	294	301
Total	**5,960**	**1,277**	**1,278**

Table 2: Label distribution in the Task 2 dataset across training, validation, and test splits.

3.3 Task 3 : Multimodal Hateful Meme Detection

Task. The objective of this subtask is to *determine whether a meme—comprising both textual and visual content—is hateful or not*, formulated as a binary classification problem. Participants were allowed to adopt any experimental setup, leveraging text-only, image-only, or multimodal approaches.

Dataset. For this subtask, the dataset is derived from prior work (Hasanain et al., 2024; Alam et al., 2024c,a) and comprises 3,562 memes, including the final evaluation test set. These memes were collected from diverse social media platforms such as Facebook, Twitter, Instagram, and Pinterest. The textual content within the memes was extracted using an off-the-shelf OCR tool[2], followed by manual post-editing to ensure accuracy.

Hateful meme annotations for the training and development sets were obtained through a hybrid approach, combining multiple large language models (LLMs) replicating human annotation approaches. The test set (referred to as dev-test) was fully human-annotated. For the shared task, we additionally constructed a new test split, adhering to the data collection methodology and annotation guidelines described in (Alam et al., 2024c).

4 Results

This section reports the leaderboard results for each of the three subtasks, including the team rankings and their corresponding Macro F1-scores.

4.1 Task 1

Task 1 received a total of 28 submissions. The baseline system, a BERT-based model, achieved a Macro F1-score of 0.53, providing a reference point for evaluating participant systems. As shown in Table 3, *HTU* (Saleh and Biltawi, 2025) achieved

[2] https://github.com/JaidedAI/EasyOCR

the highest performance with a Macro F1-score of 0.723. Their system combined multiple Arabic models (ArabicDeBERTa-DA, BERT-MSA, MARBERTv2) in an ensemble, which allowed them to capture variation across dialects and improved robustness. *NYUAD* (AlDahoul and Zaki, 2025), the second-ranked team with 0.721 F1-score, leveraged large language models by fine-tuning GPT-4o-mini and Gemini Flash 2.5 alongside Google text embeddings with an SVM classifier, and fused predictions through majority voting, which helped them handle subjective and dialectal confusions. *AAA* (Elzainy et al., 2025) and *NguyenTriet* (Nguyen and Dang, 2025a) shared third place with an F1-score of 0.707. AAA systematically evaluated multiple transformer encoders and found that MARBERT was the most effective. NguyenTriet, by contrast, used a carefully preprocessed dataset and built an ensemble of Arabic-specific BERT encoders with soft-voting fusion.

LoveHeaven (Nguyen and Dang, 2025b) achieved strong results (0.703) by ensembling AraBERT-Twitter variants and incorporating attention-based features. *IRIT_HOPE* (Moudjari et al., 2025) also ranked among the top systems (with 0.701), combining token-level augmentation with pragmatic features derived from multiple sources (MAHED, MLMA, and synthetic data). *phucclone** likewise delivered a competitive performance, securing a place within the top seven.

Beyond the top-performing group, several other teams achieved competitive results. For instance, *novatriee**, *CUET_Zahra_Duo* (Alam et al., 2025) (which fine-tuned AraBERTv2-large with optimized early stopping), *ahmedabdou** and *TranTranUIT* (Tran and Dang, 2025), all scored near 0.69. *TranTranUIT* focused on dialect sensitivity and cross-lingual generalization, applying extensive data augmentation strategies including backtranslation , EDA-based transformations, and noise reduction. They fine-tuned AraBERTv2, AraBERT-Twitter, and XLM-RoBERTa, combining them in a soft-voting ensemble.

Teams clustered in the 0.64–0.69 range included *SmolLab_SEU* (Rahman et al., 2025), which experimented with several Arabic-native and multilingual transformers, and *Quasar* (Chowdhury and Chowdhury, 2025), which combined text normalization with data augmentation and large models. Other teams in this group were *CIC-NLP* (Obiadoh et al., 2025), *ANLPers* (Yasser et al., 2025), *sudo_apt**, *Muhammad Annas Shaikh**, *michaelibrahim**, *minhtriet**, *nguyenminhtriet**, *Baoflowin502* (Bao and Thin, 2025), *KALAM* (Hameed and Al-Fuqaha, 2025), *AraNLP* (Khalil and El-Kassas, 2025), and *turabusmani**. The lowest-ranked group — including *ANLP-UniSo* (El Abed et al., 2025), *REGLAT* (Ashraf et al., 2025), *shadmansaleh**, and *AyahVerse* (Rashid and Khalil, 2025) — scored below 0.60.

Rank	Team	F1-score
1	**HTU**	**0.723**
2	**NYUAD**	**0.721**
3	**AAA**	**0.707**
3	**NguyenTriet**	**0.707**
4	LoveHeaven	0.703
5	IRIT_HOPE	0.701
6	phucclone*	0.700
7	novatriee*	0.698
8	CUET_Zahra_Duo	0.695
9	ahmedabdou*	0.695
10	trantranuit	0.694
11	SmolLab_SEU	0.682
12	Quasar	0.674
13	CIC-NLP	0.673
14	ANLPers	0.672
15	sudo_apt*	0.671
16	Muhammad Annas Shaikh*	0.669
17	michaelibrahim*	0.665
18	minhtriet*	0.659
18	nguyenminhtriet*	0.659
19	Baoflowin502	0.651
20	KALAM	0.650
20	AraNLP	0.650
21	turabusmani*	0.647
22	ANLP-UniSo	0.595
23	REGLAT	0.579
baseline	Baseline model	0.53
25	shadmansaleh*	0.483
25	AyahVerse	0.481

*The corresponding papers were not submitted.

Table 3: Task 1 results with team rankings

4.2 Task 2

Task 2 received a total of 11 submissions. The baseline system, built with an AraBERT model, achieved a Macro F1-score of 0.50. As shown in Table 4, *NYUAD* ranked first with a Macro F1-score of 0.578. Their system trained three fine-tuned GPT-4o-mini models, each specialized for emotion, offensive, and hate detection sub-tasks. They further addressed class imbalance by oversampling the “hate” class fivefold. *NguyenTriet*, in second place with 0.553, developed a hierarchical cascade architecture where predictions from emotion classification were fed into offensiveness detection, which in turn informed hate detection. They relied on ensembling MARBERTv2 and AraBERT-Twitter with soft voting at each stage. Rigorous text normalization (emoji demojization, diacritic removal, URL/stopword filtering) and class-weighted training with cosine learning-rate scheduling improved their ability to handle imbalance and dialectal variation. *HTU* placed third with 0.535, proposing a Retrospective Reader with an ALBERT approach. Their system first produced an initial prediction and then used retrospective verification to refine the classification, which helped reduce false positives. *CUET_823* (Dhar and Mallik, 2025), ranking fourth with 0.518, applied Meta-Llama-3.1-8B with instruction tuning and quantization (LoRA + 4-bit) for efficiency. They used a two-stage prompt-based approach that enabled zero- and few-shot adaptability. Finally, *SmolLab_SEU* finished in the top five with 0.514, building three separate classifiers for emotion, offensive, and hate detection using a wide range of pretrained models (MARBERTv2, ARBERTv2, AraBERTv2-large, QARiB, XLM-RoBERTa-large, mDeBERTaV3-base, DistilBERT-base). The remaining teams, including Quasar, deleted_user_25186*, KALAM, turabusmani*, MultiMinds (Debnath et al., 2025), and ashfaq*, scored between 0.33 and 0.48. These systems struggled with borderline distinctions between offensive and hate, as well as imbalanced data, highlighting the difficulty of this subtask compared to Task 1.

Rank	Team	F1-score
1	**NYUAD**	**0.578**
2	**NguyenTriet**	**0.553**
3	**HTU**	**0.535**
4	**CUET_823**	**0.518**
5	**SmolLab_SEU**	**0.514**
baseline	Baseline model	0.50
6	Quasar	0.480
7	deleted_user_25186*	0.459
8	Kalam	0.434
9	turabusmani*	0.398
10	MultiMinds	0.349
11	ashfaq*	0.336

*The corresponding papers were not submitted.

Table 4: Task 2 results with team rankings and Macro F1-scores

4.3 Task 3

Task 3 received a total of 7 submissions. The baseline multimodal hateful-meme detection system obtained a Macro F1-score of 0.70. As shown in Table 5, *NYUAD* achieved the best performance with a Macro F1-score of 0.796, the highest across all subtasks. The next two teams, *yassirEA* (0.750) (El Attar, 2025) and *Araminds* (0.744) (Zaytoon et al., 2025), also performed strongly, both surpassing 0.74. *thinkingNodes* (Safwan, 2025) followed in fourth place with 0.718, while *Muhammad Annas Shaikh** and *joy_2004114* (Das et al., 2025) obtained mid-range scores of 0.684 and 0.629, respectively. *MultiMinds* ranked last with 0.497.

Rank	Team	F1-score
1	**NYUAD**	**0.796**
2	**yassirEA**	**0.750**
3	**Araminds**	**0.744**
4	**thinkingNodes**	**0.718**
baseline	Baseline Model	0.70
5	Muhammad-Annas Shaikh*	0.684
6	joy_2004114	0.629
7	MultiMinds	0.497

*The corresponding papers were not submitted.

Table 5: Task 3 results with team rankings and Macro F1-scores

5 System Description

5.1 Data Preprocessing Techniques

The most common preprocessing steps applied by teams are summarized below:

- **Tokenization** (8 teams: SmolLab_SEU, AAA, KALAM, AraNLP, HTU, REGLAT, MultiMinds, NYUAD): Segmenting text into tokens for compatibility with deep learning models.
- **Remove URLs** (6 teams: NguyenTriet, SmolLab_SEU, KALAM, AraNLP, REGLAT, MultiMinds): Eliminating hyperlinks to reduce noise in social media text.
- **Remove Mentions/Hashtags** (5 teams: NguyenTriet, SmolLab_SEU, REGLAT, LoveHeaven, Araminds): Stripping social media markers that encode metadata rather than content.
- **Lowercasing/Normalization** (4 teams: NguyenTriet, SmolLab_SEU, KALAM, MultiMinds): Standardizing case and script

forms to reduce vocabulary redundancy.

- **AraBERT Preprocessing** (4 teams: AraNLP, CIC-NLP, LoveHeaven, AyahVerse): Using an Arabic-specific pipeline for diacritic removal, normalization, and script unification.

5.2 Feature Engineering

Text-based Tasks (Task 1 and Task 2) For the text-only tasks, teams employed both traditional vectorization methods and deep contextual embeddings:

- Google text embeddings with SVM (NYUAD): Used pretrained Google text embeddings as input to an SVM classifier, providing a strong baseline with fixed semantic representations.
- Ensemble of Arabic-specific BERT encoders (NguyenTriet): Combined outputs from MARBERTv2 and related encoders to improve robustness across dialectal variation.
- TF–IDF and Embedding-based Features (KALAM, REGLAT): Leveraged classical TF–IDF along with embeddings from AraBERT, CAMeL-BERT, and MARBERT; in some cases, attention-based features were added to capture contextual cues.
- Bag-of-Words and Morphological Features (trantranuit, CIC-NLP): Applied n-gram BoW features, enriched with morphological features such as POS tags, verb patterns, and affixes.
- Attention-based Features (KALAM, Muhammad Annas Shaikh, LoveHeaven): Extracted attention weights from transformer models as features, highlighting salient contextual dependencies.
- Augmentation and Scaling (IRIT_HOPE): Introduced token-level augmentation and normalized log feature scaling to improve robustness and feature balance.
- Linguistic Features and Normalization (CIC-NLP, Quasar): Integrated handcrafted linguistic signals and normalization of diacritics to reduce noise in Arabic text.

Multimodal Task (Task 3) For the multimodal setting (image + text), teams explored fusion strategies combining visual and textual embeddings:

- Google Multimodal Embeddings: Used 512-dimensional embeddings for both image and text, fused via element-wise averaging or concatenation.
- Pretrained Encoders with Fusion (CLIP, MARBERT): Extracted features from CLIP-ViT (vision) and MARBERT (text), projecting them into a shared space and applying cross-attention or gated fusion strategies.
- Dual-encoder Architectures: Combined text and image encoders with late fusion, optimizing with binary cross-entropy and contrastive losses to align modalities.
- Hybrid Fusion Models: Used CLIP ViT-B/32 features with text embeddings (e.g., DistilBERT) and fused them using cross-attention modules.
- Advanced Fusion (MARBERTv2 + CLIP ViT-L/14): Explored multiple fusion mechanisms, including transformers, early concatenation, bilinear pooling, and cross-attention for joint representation learning.

5.3 Data Augmentation

Text-based Tasks (Task 1 and Task 2) For the text-only tasks, teams experimented with different augmentation strategies, although many reported limited or no improvement.

- Synonym replacement and back-translation were applied to increase lexical diversity, though in some cases they did not yield performance gains.
- Synthetic Minority Over-sampling Technique (SMOTE) and oversampling were used to generate synthetic minority samples, balancing class distributions and reducing bias in training data.
- Easy Data Augmentation (EDA) techniques such as random insertion, swapping, deletion, and synonym replacement were employed to expand the dataset with simple transformations.
- Bigram augmentation and contextual embeddings were explored to introduce variation at both the lexical and semantic levels.
- Some teams leveraged external synthetic and multilingual datasets (e.g., MAHED, MLMA) to supplement training and cover dialectal variation.

Multimodal Task (Task 3) In the multimodal meme classification task, augmentation targeted both text and image modalities.

- Oversampling of hate memes was performed up to nine times to alleviate class imbalance and strengthen minority-class learning.
- Image-based augmentation included rotation, scaling, perspective shifts, color jitter, gamma

Type	Model	T1	T2	T3	Key advantage
Transformer	AraBERT/v2	✓	✓		Arabic morphology
	MARBERT/v2	✓	✓	✓	Noisy social text
	CAMeL-BERT	✓			Robust baseline
	QARiB	✓			News/social adapted
	XLM-RoBERTa	✓	✓		Multilingual
	DistilBERT	✓		✓	Lightweight
	DeBERTa variants	✓			Better attention
Vision	CLIP (ViT)			✓	Vision-text align
	ResNet/ResNeXt			✓	Visual backbone
LLM/VLM	GPT-4	✓	✓	✓	Few-shot learning
	Gemini	✓	✓	✓	Multimodal reason
	LLaMA			✓	Finetuned branch
	Gemma			✓	Compact VLM
	Qwen			✓	Multilingual VLM

Table 6: Model families used across tasks

correction, noise, blurring, distortions, shadows, fog effects, and crop–resize operations.

- Text within memes was augmented using OCR-based extraction followed by synonym replacement, character-level dropout, and back-translation between Arabic and English.
- Some teams focused augmentation specifically on hate-class examples, ensuring that rare cases were better represented in multimodal training.

5.4 Model Usages Across Tasks

Task 1: Text-based Hope and Hate Speech Classification. Teams primarily used Arabic-centric transformers (AraBERT, MARBERT, CAMeL-BERT, QARiB, XLM-R) to obtain context-aware sentence embeddings robust to morphology, code-mixing, and informal orthography. These encoders work well for short, noisy social posts where pragmatic cues and dialectal markers are crucial. LLMs (e.g., GPT-4, Gemini) appeared as auxiliary backbones or zero/few-shot components, valued for broad world knowledge and flexible prompting when labeled data are limited.

Task 2: Multitask Emotion/Offense/Hate. A shared transformer encoder with lightweight task heads provides a compact way to model related label spaces, enabling representation sharing across emotion, offensive content, and hate signals. This setup simplifies training pipelines and reduces overfitting via shared inductive biases; LLMs help unify task instructions and can serve as promptable controllers for multi-objective finetuning.

Task 3: Multimodal Hateful Meme Detection. Vision–language stacks (CLIP/ViT + Arabic text encoders) align image and text into a shared semantic space so that cross-modal cues—caption sarcasm, visual symbols, and text overlays—can be interpreted jointly. LLM/VLM components (Gemini, LLaMA, Gemma, Qwen) are useful where reasoning over both modalities or following structured prompts improves recognition of subtle or template-driven hateful content.

5.5 Training Configurations and Rationale

Drop-in Recipes (Space-Efficient, Reproducible)

Recipe: Text Hope/Hate

Encoder: AraBERTv2 or MARBERTv2; max length 256; batch 16; LR 2×10^{-5} (AdamW, WD 0.01), 10% warmup, cosine decay, FP16, grad clip 1.0. Class-weighted CE; early stopping on macro-F1 (patience 3); 5-fold stratified CV; select best checkpoint by macro-F1.

Recipe: Multitask (Emotion/Offense/Hate)

Shared encoder (AraBERT/MARBERT) with multi-head classifiers; batch 16 (grad accum 2); LR 1×10^{-5}; warmup 10%, cosine schedule; FP16. Class-weighted CE; early stopping on macro-F1. Tune per-head dropout/epochs via Optuna; optional LR multiplier ($\approx$1.8) for heads.

Recipe: Multimodal Memes

Text: MARBERTv2 [CLS] or DistilBERT tokens; Image: CLIP ViT-B/32 (or ViT-L/14). Project to 512-d; fuse by concatenation or cross-attention. Batch 16–32 (per-device 2–4 for large VLMs); LR text/vision 2×10^{-5}, fusion head 1×10^{-3}; AdamW (WD 10^{-4}), linear or cosine schedule; FP16, grad clip 1.0. Loss: weighted BCE/CE, focal-loss trial. Early stopping with patience 5–15; oversample minority class.

Use Cases

- *Macro-F1 selection, class-weighted losses, and oversampling* address severe label imbalance (hate/hope and multimodal memes), prioritizing minority-class recall without inflating accuracy.

Task	Typical Backbones	Epochs	Batch Size	Seq. Length	Learning Rate	Optimizer & Strategy
Hope/Hate (Text)	AraBERT, MARBERT, CAMeLBERT, XLM-RoBERTa, QARiB, ArabicDeBERTa	2–10 (ES:3–5)	16–32	128–256	310^{-6}–110^{-5}	AdamW, cosine scheduler, FP16
Multitask (Text)	AraBERT, MARBERT variants	3–10	8–16	128	110^{-5}–210^{-5}	AdamW, warmup/cosine, FP16
Multimodal Memes	CLIP ViT + MARBERT; VLMs (Gemma, Qwen, Paligemma)	5–40 (ES) VLM:10	16–32 2–4 (VLM)	Variable	Text 110^{-5} Vision 210^{-5} VLM 510^{-6}	AdamW, gradient clip 1.0, cross-attention fusion

Table 7: Typical training settings distilled from submitted systems across tasks. ES = early stopping, VLM = vision-language model.

- *Warmup + cosine/linear schedules with AdamW* stabilize finetuning of large encoders and prevent early-step divergence; *weight decay and dropout* regularize under limited data.
- *FP16 and gradient clipping* improve memory efficiency and prevent exploding gradients, which is critical in multimodal or multitask finetuning.
- *Shared encoders with task heads (multitask)* reuse domain signals (emotion, offense, hate) and conserve parameters; LR multipliers let heads adapt faster without overfitting the encoder.
- *CLIP+Arabic encoders with projection/fusion* capture cross-modal interactions in memes; aligning to a 512-d shared space simplifies fusion while retaining modality-specific strengths.
- *CV and Optuna* provide robust, reproducible hyperparameters without exhaustive grids; reporting the validation macro-F1 criterion ensures consistent model selection.

6 Conclusions and Future Work

The MAHED 2025 shared task establishes comprehensive benchmarks for Arabic content moderation across textual and multimodal formats. With 46 participating teams, the evaluation demonstrates consistent improvements over baselines, achieving macro F1-scores of 0.723 (Task 1), 0.578 (Task 2), and 0.796 (Task 3). Top systems leveraged Arabic-specific PLMs (AraBERT, MARBERT), ensemble methods, and OCR-aware multimodal fusion.

Key Challenges: Our analysis reveals persistent limitations: (i) dialectal robustness gaps of up to 34% in error cases, with Gulf and Levantine expressions frequently misclassified; (ii) minority class detection difficulties, particularly for hope speech (average recall: 0.52); (iii) OCR noise contributing to 28% of multimodal errors; and (iv) Task 2's hierarchical multitask complexity, where conflicting optimization pressures across emotion, offense, and hate detection yielded the lowest performance (0.578 F1).

Future Directions: Critical research priorities include: dialect-invariant representations through cross-dialectal augmentation and adversarial training; culturally-grounded hope speech annotation with contrastive learning objectives; Arabic-specific scene text recognition for stylized fonts; and uncertainty-aware multitask architectures. Evaluation methodology should incorporate dialectal breakdowns, calibration analysis, and fairness auditing.

Impact: The released datasets (22,000+ instances, Cohen's Kappa >0.85), baseline implementations, and comprehensive analysis provide a reproducible foundation for Arabic content safety research. While significant progress was demonstrated, the identified challenges underscore the need for culturally-informed approaches that address Arabic's unique linguistic and cultural characteristics.

7 Limitations

The MAHED shared task has several inherent constraints: (i) focus on social media data excludes formal Arabic domains; (ii) binary hope/hate categories oversimplify the prosocial-harmful spectrum; (iii) hierarchical multitask design in Task 2 introduces conflicting optimization pressures; (iv) OCR-dependent multimodal processing creates systematic extraction errors; and (v) annotation guidelines may not fully capture dialectal and cultural diversity across Arabic-speaking regions.

Acknowledgments

We thank all participating teams for their contributions, as well as the organizers of the Arabic NLP 2025 Conference, and the broader community for their support. The authors thank our annotation team for ensuring the authenticity and appropriateness of our resources. The contribution of this work is supported by the NPRP grant 14C-0916-210015 from the Qatar National Research Fund, part of the Qatar Research Development and Innovation Council (QRDI). The findings achieved herein are solely the responsibility of the authors.

References

Lamiaa Abdel-Hamid. 2020. Egyptian arabic speech emotion recognition using prosodic, spectral and wavelet features. *Speech Communication*, 122:19–30.

Mahmoud Mohamed Abdelsamie, Shahira Shaaban Azab, and Hesham A. Hefny. 2024. A comprehensive review on arabic offensive language and hate speech detection on social media: methods, challenges and solutions. *Social Network Analysis and Mining*, 14(1):111.

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.

Ibrahim Abu Farha and Walid Magdy. 2020. Multitask learning for Arabic offensive language and hate-speech detection. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 86–90, Marseille, France. European Language Resource Association.

Firoj Alam, Md Rafiul Biswas, Uzair Shah, Wajdi Zaghouani, and Georgios Mikros. 2024a. Propaganda to hate: A multimodal analysis of arabic memes with multi-agent llms. In *International Conference on Web Information Systems Engineering*, pages 380–390. Springer.

Firoj Alam, Abul Hasnat, Fatema Ahmad, Md. Arid Hasan, and Maram Hasanain. 2024b. ArMeme: Propagandistic content in Arabic memes. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21071–21090, Miami, Florida, USA. Association for Computational Linguistics.

Firoj Alam, Abul Hasnat, Fatema Ahmed, Md Arid Hasan, and Maram Hasanain. 2024c. ArMeme: Propagandistic content in arabic memes. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*.

Walisa Alam, Mehreen Rahman, Shawly Ahsan, and Mohammed Moshiul Hoque. 2025. Cuet_zahra_duo@mahed 2025: Hate and hope speech detection in arabic social media content using transformer. In *Proceedings of the Third Arabic Natural Language Processing Conference (ArabicNLP 2025)*, Suzhou, China. Association for Computational Linguistics.

Nouar AlDahoul and Yasir Zaki. 2025. Nyuad at mahed shared task: Detecting hope, hate, and emotion in arabic textual speech and multi-modal memes using large language models. In *Proceedings of the Third Arabic Natural Language Processing Conference (ArabicNLP 2025)*, Suzhou, China. Association for Computational Linguistics.

R. H. Aljuhani, A. Alshutayri, and H. Alahdal. 2021. Arabic speech emotion recognition from saudi dialect corpus. *IEEE Access*, 9:127081–127085.

Maha Jarallah Althobaiti. 2023. An open-source dataset for arabic fine-grained emotion recognition of online content amid covid-19 pandemic. *Data in Brief*, 51:109745.

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.

Muhammad Arif, Moein Shahiki Tash, Ainaz Jamshidi, Fida Ullah, Iqra Ameer, Jugal Kalita, Alexander Gelbukh, and Fazlourrahman Balouchzahi. 2024. Analyzing hope speech from psycholinguistic and emotional perspectives. *Scientific Reports*, 14(1):23548.

Nsrin Ashraf, Mariam Labib, Tarek Elshishtawy, and Hamada Nayel. 2025. Reglat at mahed shared task: A hybrid ensemble-based system for arabic hate speech detection. In *Proceedings of the Third Arabic Natural Language Processing Conference (ArabicNLP 2025)*, Suzhou, China. Association for Computational Linguistics.

F. Balouchzahi and 1 others. 2025. Urduhope: Analysis of hope and hopelessness in urdu texts. *Knowledge-Based Systems*, 308:112746.

Nguyen Minh Bao and Dang Van Thin. 2025. Baoflowin502 at mahed shared task: Text-based hate and hope speech classification. In *Proceedings of the*

Third Arabic Natural Language Processing Conference (ArabicNLP 2025), Suzhou, China. Association for Computational Linguistics.

Boaz M. Ben-David, Nandini Multani, Vered Shakuf, Frank Rudzicz, and Pascal H. H. van Lieshout. 2016. Prosody and semantics are separate but not separable channels in the perception of emotional speech: Test for rating of emotions in speech. *Journal of Speech, Language, and Hearing Research*, 59(1):72–89.

Kristin M. Brethel-Haurwitz, Maria Stoianova, and Abigail A. Marsh. 2020. Empathic emotion regulation in prosocial behaviour and altruism. *Cognition and Emotion*, 34(8):1532–1548.

Bharathi Raja Chakravarthi, Vigneshwaran Muralidaran, Ruba Priyadharshini, Subalalitha Chinnaudayar Navaneethakrishnan, John P. McCrae, Miguel Ángel García, Salud María Jiménez-Zafra, Rafael Valencia-García, Prasanna Kumar Kumaresan, Rahul Ponnusamy, Daniel García-Baena, and José Antonio García-Díaz. 2022. Overview of the shared task on hope speech detection for equality, diversity and inclusion. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 378–388, Dublin, Ireland. Association for Computational Linguistics.

Md Sagor Chowdhury and Adiba Fairooz Chowdhury. 2025. Quasar at mahed shared task : Decoding emotions and offense in arabic text using llm and transformer-based approaches. In *Proceedings of the Third Arabic Natural Language Processing Conference (ArabicNLP 2025)*, Suzhou, China. Association for Computational Linguistics.

Joy Das, Alamgir Hossain, and Mohammed Moshiul Hoque. 2025. joy_2004114 at mahed shared task : Filtering hate speech from memes using a multimodal fusion-based approach. In *Proceedings of the Third Arabic Natural Language Processing Conference (ArabicNLP 2025)*, Suzhou, China. Association for Computational Linguistics.

Riddhiman Swanan Debnath, Abdul Wadud Shakib, and Md Saiful Islam. 2025. Multiminds at mahed 2025: Multimodal and multitask approaches for detecting emotional, hate, and offensive speech in arabic content. In *Proceedings of the Third Arabic Natural Language Processing Conference (ArabicNLP 2025)*, Suzhou, China. Association for Computational Linguistics.

Ratnajit Dhar and Arpita Mallik. 2025. Cuet-823 at mahed 2025 shared task: Large language model-based framework for emotion, offensive, and hate detection in arabic. In *Proceedings of the Third Arabic Natural Language Processing Conference (ArabicNLP 2025)*, Suzhou, China. Association for Computational Linguistics.

Marc Djandji, Fady Baly, Wissam Antoun, and Hazem Hajj. 2020. Multi-task learning using AraBert for offensive language detection. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 97–101, Marseille, France. European Language Resource Association.

Yasmine El Abed, Mariem Ben Arbia, Saoussen Ben Chaabene, and Omar Trigui. 2025. Anlp-uniso at mahed shared task: Detection of hate and hope speech in arabic social media based on xlm-roberta and logistic regre. In *Proceedings of the Third Arabic Natural Language Processing Conference (ArabicNLP 2025)*, Suzhou, China. Association for Computational Linguistics.

Yassir El Attar. 2025. Yassirea at mahed 2025: Fusion-based multimodal models for arabic hate meme detection. In *Proceedings of the Third Arabic Natural Language Processing Conference (ArabicNLP 2025)*, Suzhou, China. Association for Computational Linguistics.

Ahmed Elzainy, Hazem Abdelsalam, Ahmed Samir, and Mohamed Amin. 2025. Aaa at mahed shared task: A systematic encoder evaluation for arabic hope and hate speech classification. In *Proceedings of the Third Arabic Natural Language Processing Conference (ArabicNLP 2025)*, Suzhou, China. Association for Computational Linguistics.

Elisabetta Fersini, Francesca Gasparini, Giulia Rizzi, Aurora Saibene, Berta Chulvi, Paolo Rosso, Alyssa Lees, and Jeffrey Sorensen. 2022. SemEval-2022 task 5: Multimedia automatic misogyny identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 533–549, Seattle, United States. Association for Computational Linguistics.

Saad Hameed and Ala Al-Fuqaha. 2025. Kalam at mahed shared task 2025: Transformer-based approaches for arabic sentiment classification and stance detection. In *Proceedings of the Third Arabic Natural Language Processing Conference (ArabicNLP 2025)*, Suzhou, China. Association for Computational Linguistics.

Maram Hasanain, Md. Arid Hasan, Fatema Ahmed, Reem Suwaileh, Md. Rafiul Biswas, Wajdi Zaghouani, and Firoj Alam. 2024. ArAIEval Shared Task: Propagandistic techniques detection in unimodal and multimodal arabic content. In *Proceedings of the Second Arabic Natural Language Processing Conference*, Bangkok. ACL.

Enas A. Hakim Khalil and Wafaa S. El-Kassas. 2025. Aranlp at mahed 2025 shared task: Using arabert for text-based hate and hope speech classification. In *Proceedings of the Third Arabic Natural Language Processing Conference (ArabicNLP 2025)*, Suzhou, China. Association for Computational Linguistics.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: detecting hate speech in multimodal memes.

In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Samira Klaylat, Zeina Osman, and Rached Zantout. 2018. Emotion recognition in arabic speech. *Analog Integrated Circuits and Signal Processing*, 96(2):185–198.

Mohamed Meddeb, Rima Malka, and Mohamed Ali Hammami. 2017. Building and analysing emotion corpus of the arabic speech. In *2017 1st International Workshop on Arabic Script Analysis and Recognition (ASAR)*, pages 18–22.

Ali Hamid Meftah, Yousef Ajami Alotaibi, and Sid-Ahmed Selouani. 2018. Evaluation of an arabic speech corpus of emotions: A perceptual and statistical analysis. *IEEE Access*, 6:72845–72861.

Ali Hamid Meftah, Mohammad A. Qamhan, Yasser M. Seddiq, Yousef Ajami Alotaibi, and Sid-Ahmed Selouani. 2021. King saud university emotions corpus: Construction, analysis, evaluation, and comparison. *IEEE Access*, 9:54201–54219.

Leila Moudjari, Mélissa Hacene Cherkaski, and Farah Benamara. 2025. Descartes_hope at mahed shared task 2025: Integrating pragmatic features for arabic hope and hate speech detection. In *Proceedings of the Third Arabic Natural Language Processing Conference (ArabicNLP 2025)*, Suzhou, China. Association for Computational Linguistics.

Hamdy Mubarak, Hend Al-Khalifa, and Abdulmohsen Al-Thubaity. 2022. Overview of OSACT5 shared task on Arabic offensive language and hate speech detection. In *Proceedinsg of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools with Shared Tasks on Qur'an QA and Fine-Grained Hate Speech Detection*, pages 162–166, Marseille, France. European Language Resources Association.

Hamdy Mubarak, Kareem Darwish, Walid Magdy, Tamer Elsayed, and Hend Al-Khalifa. 2020. Overview of OSACT4 Arabic offensive language detection shared task. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 48–52, Marseille, France. European Language Resource Association.

L. Nasr, M. Hani, A. Harkous, Y. Al Khalil, A. Abou Daya, H. Hajj, and R. El-Khoury. 2024. Survey on arabic speech emotion recognition. *International Journal of Speech Technology*. Online first.

Minh Triet Nguyen and Van Thin Dang. 2025a. Nguyentriet at mahed shared task: Ensemble of arabic bert models with hierarchical prediction and soft voting for text-based hope and hate detection. In *Proceedings of the Third Arabic Natural Language Processing Conference (ArabicNLP 2025)*, Suzhou, China. Association for Computational Linguistics.

Thien Bao Nguyen and Van Thin Dang. 2025b. Loveheaven at mahed 2025: Text-based hate and hope speech classification using arabert-twitter ensemble. In *Proceedings of the Third Arabic Natural Language Processing Conference (ArabicNLP 2025)*, Suzhou, China. Association for Computational Linguistics.

A. E. Obiadoh, O.J Abiola, T.D. Ogunleye, B.A. Tewodros, and T.O Abiola. 2025. Cic-nlp at mahed 2025 task 1:assessing the role of bigram augmentation in multiclass arabic hate and hope speech classification. In *Proceedings of the Third Arabic Natural Language Processing Conference (ArabicNLP 2025)*, Suzhou, China. Association for Computational Linguistics.

S. S. Poorna and G. J. Nair. 2019. Multistage classification scheme to enhance speech emotion recognition. *International Journal of Speech Technology*, 22(2):327–340.

Md. Abdur Rahman, Md. Sabbir Dewan, Md. Tofael Ahmed Bhuiyan, and Md. Ashiqur Rahman. 2025. Smollab_seu at mahed shared task: Do arabic-native encoders surpass multilingual models in detecting the nuances of hope, hate, and emotion? In *Proceedings of the Third Arabic Natural Language Processing Conference (ArabicNLP 2025)*, Suzhou, China. Association for Computational Linguistics.

Ibad-ur-Rehman Rashid and Muhammad Hashir Khalil. 2025. Ayahverse at mahed shared task: Fine-tuning arabicbert with preprocessing for hope and hate detection. In *Proceedings of the Third Arabic Natural Language Processing Conference (ArabicNLP 2025)*, Suzhou, China. Association for Computational Linguistics.

Itbaan Safwan. 2025. Thinking nodes at mahed: A comparative study of multimodal architectures for arabic hateful meme detection. In *Proceedings of the Third Arabic Natural Language Processing Conference (ArabicNLP 2025)*, Suzhou, China. Association for Computational Linguistics.

Abdallah Saleh and Mariam Biltawi. 2025. Htu at mahed shared task: Ensemble-based classification of arabic hate and hope speech using pre-trained dialectal arabic models. In *Proceedings of the Third Arabic Natural Language Processing Conference (ArabicNLP 2025)*, Suzhou, China. Association for Computational Linguistics.

Octa Syakila Salsabila, Sri Melati Rahmayani, and Isti Yuniastuti. 2024. Content analysis of arabic vocabulary in al quran for the improvement of emotional intelligence. *LiNGUA: Jurnal Ilmu Bahasa dan Sastra*, 19(1):97–108.

Shardul Suryawanshi, Bharathi Raja Chakravarthi, Mihael Arcan, and Paul Buitelaar. 2020. Multimodal meme dataset (MultiOFF) for identifying offensive content in image and text. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 32–41, Marseille, France. European Language Resources Association (ELRA).

Tran Tran, Trinh and Thin Dang, Van. 2025. Trantranuit at mahed shared task: Multilingual transformer ensemble with advanced data augmentation and optuna-based hyperparameter optimization. In *Proceedings of the Third Arabic Natural Language Processing Conference (ArabicNLP 2025)*, Suzhou, China. Association for Computational Linguistics.

Gerben A. van Kleef and Gert-Jan Lelieveld. 2022. Moving the self and others to do good: The emotional underpinnings of prosocial behavior. *Current Opinion in Psychology*, 44:80–88. Epub 2021-08-31.

Al-Habashi Yasser, Sibaee1 Serry, Nacar Omer, Ammar Adel, and Wadii Boulila1. 2025. Anlpers at mahed shared task: From hate to hope: Boosting arabic text classification. In *Proceedings of the Third Arabic Natural Language Processing Conference (ArabicNLP 2025)*, Suzhou, China. Association for Computational Linguistics.

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, 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.

Mohamed Zaytoon, Ahmed Salem, Ahmed Sakr, and Hossam Elkordi. 2025. Araminds at mahed 2025: Leveraging vision-language models and contrastive multi-task learning for multimodal hate speech detection. In *Proceedings of the Third Arabic Natural Language Processing Conference (ArabicNLP 2025)*, Suzhou, China. Association for Computational Linguistics.

8 Appendix

Table 8: Task 1: Text-based Hate/Hope/Emotion Detection

Team	Models	Notable Methods	Compute
NYUAD	GPT-4o-mini, Gemini Flash 2.5 + SVM	Google text embeddings + SVM	OpenAI platform
NguyenTriet	MARBERTv2, AraBERTv0.2-Twitter	Arabic cleanup, ensemble	Tesla P100 (Kaggle)
SmolLab_SEU	MARBERTv2, ARBERTv2, AraBERTv2-large, XLM-R, mDeBERTaV3	Multi-model ensemble	Kaggle P100
AAA	MARBERT, AraBERT-Twitter, XLM-RoBERTa	Arabic tokenization	Tesla V100
KALAM	TF-IDF+LR, AraBERT, CAMeL-BERT, MARBERT	TF-IDF + embeddings + attention	24 GB GPU
AraNLP	AraBERT v0.2-Twitter	AraBERTPreprocessor + 5-fold CV	Google Colab L4
HTU	ArabicDeBERTa-DA, BERT-MSA, MARBERTv2	—	—
REGLAT	AraBERTv2, CAMeL-BERT + SVM/LR	TF-IDF + embeddings, majority voting	Colab GPU
ANLP-UniSo	XLM-RoBERTa, LSTM	SMOTE augmentation	—
trantranuit	AraBERT, XLM-RoBERTa	BoW + TF-IDF + morphological features	Kaggle P100
CIC-NLP	MARBERT	Linguistic + BoW features	RTX 3800, 32 GB RAM
CUET_Zahra_Duo	AraBERTv2-large	Contextual embedding + early stopping	Tesla T4 (32 GB total)
IRIT_HOPE	bert-base-arabertv02-twitter	Token-level augmentation, multi-embedding	—
LoveHeaven	bert-base-arabertv02(-twitter)	Attention-based features	Kaggle P100
AyahVerse	AraBERT	Embeddings + EDA (synonym/back-translation)	—
baoflowin502	AraBERTv2, CAMeL-BERT, BERT Arabic	—	Kaggle P100
Quasar	xlm-roberta-large, gemma-7b, qwen2.5-14b-instruct	Diacritics normalization + synonym balancing	—
TranTranUIT	AraBERTv2, AraBERT-Twitter, XLM-RoBERTa	Dialect sensitivity, cross-lingual + back-translation	—

Table 9: Task 2: Multitask Text Classification

Team	Models	Multitask Setup	Compute
NYUAD	GPT-4o-mini (3 models)	Parallel: separate per sub-task	OpenAI platform
MultiMinds	SVM, XGBoost, AraBERT, GPT-5	Parallel multi-head shared encoder	Colab (6 GB)
NguyenTriet	MARBERTv2, AraBERTv0.2-Twitter	Sequential cascade: Emotion→Offensive→Hate	Kaggle P100
SmolLab_SEU	MARBERTv2, ARBERTv2, XLM-RoBERTa-large	Sequential cascade (3 classifiers)	Kaggle P100
KALAM	CAMeL-BERT, MARBERT, AraBERT	Single-task fine-tuning	24 GB GPU
HTU	Retrospective Reader, ALBERT	—	—
CUET_823	Meta-Llama-3.1-8B	—	Kaggle GPU (16 GB)
Quasar	qwen2.5-14B, gemma-7b, AraBERTv2	—	—

Table 10: Task 3: Multimodal Meme Classification

Team	Models	Fusion / Approach	Compute
NYUAD	GPT-4o-mini, Gemini Flash 2.5, Llama 3.2-11B, Paligemma2	Multimodal embeddings + oversampling	OpenAI + Vertex AI

Team	**Models**	**Fusion / Approach**	**Compute**
thinkingNodes	CLIP-ViT-B/32 + MARBERT	Cross-attention, CNN fusion, contrastive CLIP-Arabic	Kaggle T4 (15 GB)
Araminds	Qwen2.5-1.5B+ResNet / MARBERTv2+ResNet, Gemma3-4B	Dual-encoder + contrastive + VLM ensemble	RTX 3090
MultiMinds	CLIP ViT-B/32 + DistilBERT	ELU-Net cross-attention fusion	Google Colab (6.2 GB)
yassirea	MARBERTv2 + CLIP-Large (ViT-L/14)	4-way fusion + heavy augmentation	RTX 6000 Ada (48 GB)
Muhammad Annas Shaikh	EfficientNet-B0 + AraBERT	—	—
CUET_NLP	mBERT + InceptionResNetV2	—	—
joy_2004114	mBERT, AraBERT, InceptionNet	—	—