Learning to Engage: Modeling Topic-Sensitive Reactions in Arabic Women's Online Discourse

Mabrouka Bessghaier[1] **Md. Rafiul Biswas**[2] **Shimaa Amer Ibrahim**[1]
Wajdi Zaghouani[1]

[1]Northwestern University in Qatar [2]Hamad Bin Khalifa University

{mabrouka.bessghaier, shimaa.ibrahim, wajdi.zaghouani}@northwestern.edu,
mbiswas@hbku.edu.qa

Abstract

Predicting how audiences react to Arabic social media posts requires reasoning beyond textual sentiment: reactions emerge from collective interpretation moderated by engagement dynamics and topical context. We present a multi-task learning framework that jointly learns (i) audience reaction classification (Love, Haha, Angry, Sad, Care, Wow), (ii) engagement magnitude regression (six reactions, comments, shares), and (iii) non-engagement detection. On a corpus of 158k Arabic Facebook posts spanning women's rights, gender debates, and economic empowerment, our model achieves a test macro-F1 of 72.4 and weighted-F1 of 89.1.

1 Introduction

Understanding how Arabic-speaking communities respond to discourse on women's issues requires analyzing collective audience reactions rather than only textual sentiment. On social media platforms such as Facebook, reactions (e.g., Love, Haha, Angry, Sad, Care, Wow) encode how communities collectively interpret content, revealing patterns of public opinion that vary dramatically across regional and cultural contexts. Crucially, audience reactions diverge from author sentiment: a post written with neutral, informative tone about women's workforce participation may elicit *Love* from supporters or *Angry* from opponents, while a post expressing frustration about discrimination may receive *Love* reactions from sympathetic audiences. This divergence makes audience reactions a distinct analytical target from traditional sentiment analysis, which focuses on emotions expressed by authors.

Predicting these audience reactions presents two main challenges. First, reaction distributions exhibit severe class imbalance: passive reactions like *Love* usually dominate, while emotionally charged reactions like *Angry* and *Care* remain underrepresented. Second, reactions do not occur in isolation; they correlate with distinct engagement behaviors. Posts provoking controversy attract comments; posts generating consensus accumulate *Love*; ridicule often coincides with *Haha* and sharing. Capturing these patterns requires jointly modeling reaction classification and engagement prediction, yet no prior work has explored this combination for Arabic social media.

We address this gap with two contributions. First, we present a corpus of 158k Arabic Facebook posts (2014-2024) spanning diverse domains of women's discourse, such as legal rights advocacy, feminism debates, gender identity discussions, and economic empowerment. The posts exhibit rich linguistic variation, mixing Modern Standard Arabic (MSA) with regional dialects. Second, we propose a multi-task learning (MTL) framework that jointly models (i) audience reaction classification (six classes), (ii) engagement magnitude regression (reaction counts, comments, shares), and (iii) non-engagement detection. By learning these tasks together, the model leverages engagement patterns as auxiliary signals that improve reaction prediction, particularly for underrepresented classes.

2 Related Work

Reactions as Collective Interpretation. Audience reactions on social media encode collective interpretation rather than author sentiment, making them valuable signals for understanding public opinion. Pool and Nissim (2016) introduced distant supervision using Facebook reactions for emotion detection, showing that raw reaction counts can train classifiers without manual labeling. Graziani et al. (2019) combined emotion detection and reaction prediction in a multi-task framework with logical constraints, exploiting inter-task relationships to improve accuracy. This shift from author-centric sentiment to audience-centric reactions is particularly important for analyzing discourse on contested social issues, where public response di-

Findings of the Association for Computational Linguistics: EACL 2026, pages 4846–4854

verges sharply from content polarity.

Engagement as Behavioral Signal. Beyond categorical reactions, engagement metrics (comments, shares) capture qualitatively different audience behaviors. Kim and Hwang (2025) modeled engagement as regression targets, predicting likes and comments from emotional and temporal features, demonstrating that engagement metrics reveal complementary facets of audience response. For content about sensitive topics, understanding these behavioral patterns can provide insight into how communities negotiate cultural norms and contested values. However, prior work treats engagement prediction and reaction classification as separate tasks; their joint optimization remains unexplored.

Multi-Task Learning for Imbalance. Real-world reaction distributions exhibit severe class imbalance, where a few reaction types dominate while others remain underrepresented. MTL addresses this challenge by learning shared representations across related tasks, with auxiliary tasks providing gradient signals that improve minority class recall. Plaza-del Arco et al. (2021) showed that jointly modeling hate speech, sentiment, emotion, and target detection boosts recall for low-frequency classes. We extend this insight by using engagement regression as auxiliary supervision for reaction classification, hypothesizing that quantitative engagement patterns correlate with specific reaction types and can help discriminate minority classes.

Arabic Social Media Analysis. Arabic sentiment analysis has matured with recent work on sentiment (Ibrahim et al., 2025), hate speech (Charfi et al., 2024b), stance (Charfi et al., 2024a), and emotion detection (Biswas et al., 2025; Zaghouani et al., 2025). However, this research focuses on author-expressed sentiment rather than audience reactions. Moreover, most Arabic datasets provide only sentiment polarity labels and lack fine-grained reaction annotations or engagement metrics. Duwairi and Qarqaz (2017) assembled an Arabic Facebook comments corpus for polarity classification without leveraging reaction counts. AlShenaifi et al. (2024a,b) applied active learning and fine-tuning to improve sentiment performance but rely on coarse sentiment categories. Al-Badawi (2024) studied audience engagement patterns in Jordanian markets but did not model reaction classification or its relationship to engagement behaviors. The challenge is compounded by dialectal variation across Arab regions, where identical content may elicit divergent reactions depending on cultural and political contexts.

To our knowledge, no prior work has jointly modeled fine-grained reaction classification, engagement regression, and non-engagement detection, particularly not for Arabic social media or for discourse on culturally sensitive topics like women's rights.

3 Dataset and Task Formulation

3.1 Corpus

We analyze a corpus of approximately 158k Arabic Facebook posts (2014–2024) [1] collected via the CrowdTangle API [2], limited to publicly available pages and groups. The dataset spans a decade of discourse on women's lives, media, and sociopolitical participation across the Arab world, encompassing a mix of MSA and regional dialects (Egyptian, Levantine, Gulf, Maghrebi). Each post includes extensive metadata and engagement indicators (e.g.,page details, reactions, comments, shares, and temporal information) supporting multi-dimensional analysis of Arabic social media. In this study, we focus on the textual content (post message and related descriptions) and engagement metrics (reactions, comments, shares) to examine how content about women elicits public interaction and emotional response online.

3.2 Primary Task: Audience Response Classification

The first task models the audience's dominant emotional response to each post. Each post is assigned one of six reaction-based labels corresponding to the most frequent Facebook reaction type: Love, Haha, Wow, Sad, Angry, or Care. Crucially, this represents the collective audience sentiment toward the content rather than the emotion expressed in the text itself. Unlike traditional sentiment analysis that classifies emotions expressed by authors, we predict audience reactions, which is the collective emotional response of communities to content. For example:

- A post describing economic challenges facing women (textually negative) may receive *Love*

[1] the corpus will be made available upon request for research purposes

[2] https://transparency.meta.com/he-il/researchtools/other-datasets/crowdtangle/

reactions if audiences interpret it as advocacy they support.

- A post celebrating women's achievements (textually positive) may elicit *Angry* reactions from audiences who oppose the content's message.

- A factual, neutral post about workplace discrimination may generate polarized reactions (*Love* vs. *Angry*) depending on readers' ideological positions.

Our task is to predict this collective audience interpretation, not to classify the author's expressed emotion.

In this work, we model the dominant reaction, which is the most frequent reaction type for each post, as the primary classification target. This design choice is motivated by two considerations. First, the dominant reaction serves as a reliable, aggregate indicator of the majority audience's emotional interpretation, capturing the prevailing collective sentiment toward the content. Second, while posts may evoke mixed emotions (e.g., simultaneous *Sad* and *Angry* reactions), our auxiliary engagement regression task (Section 3.3) captures the full distribution of all six reaction counts, enabling analysis of polarization and emotional diversity. For instance, a post with 60% *Love* and 30% *Angry* reactions is labeled *Love* for classification but retains the full reaction vector for regression, allowing us to identify posts that generate contested or polarized responses. This dual approach balances interpretability (dominant emotion classification) with nuance (quantitative engagement modeling).

3.3 Auxiliary Task 1: Engagement Regression

The second task predicts a continuous 8-dimensional engagement vector capturing how content propagates and stimulates interaction beyond the categorical reactions. The vector comprises:

- **Six reaction counts:** Love, Haha, Wow, Sad, Angry, Care (log-transformed and min–max normalized)

- **Comments:** measure of discussion intensity (*active engagement*)

- **Shares:** measure of redistribution and virality

This formulation captures both the *volume* (how widely a post resonates) and the *composition* (which audience segments engage and how). For example, a post with 500 *Love* reactions but few comments suggests passive approval, whereas a post with 200 *Love* and 150 comments indicates debate or controversy. Posts with high shares and mixed reactions often correspond to contentious topics that users feel compelled to spread. All engagement variables are $\log(1+x)$-transformed and scaled to the [0,1] range. The overall engagement regression loss serves as an auxiliary supervision signal that encourages the model to learn features correlated with post popularity and emotional resonance.

3.4 Auxiliary Task 2: Non-Engagement Detection

To better capture zero-reaction posts, we introduce a binary auxiliary task that predicts whether a post received any audience reactions (*engaged* vs. *not engaged*). This component helps the model differentiate between posts that draw attention and those that remain unnoticed, addressing the natural imbalance between high- and low-engagement samples in social media data.

3.5 Challenges

The dataset presents two main challenges.

First, the reaction distribution exhibits class imbalance, with Love dominating at 70.2% (108,338 instances) and Haha at 19.9% (30,745 instances), while minority reactions Wow (1.3%), Sad (3.5%), Angry (3.3%), and Care (1.8%) collectively represent less than 10% of the dataset. Second, engagement magnitudes are heavy-tailed, reflecting the bursty nature of online virality. These properties motivate a multi-task formulation combining focal loss (for imbalance) and auxiliary engagement modeling (for robust shared representations).

4 Multi-Task Learning Framework

4.1 Problem Formulation

Given a Facebook textual post x and its engagement signals, we model three related tasks:

(i) Non-Engagement Detection $y^{(ne)} \in \{0, 1\}$ (no reactions vs. any reactions)

(ii) Audience Response Classification $y^{(cls)} \in \{1, \ldots, 6\}$ (dominant reaction: *Love, Haha, Wow, Sad, Angry, Care*),

(iii) Engagement Regression $y^{(reg)} \in \mathbb{R}^8_{\geq 0}$ (8-dimensional vector: six reaction counts + comments + shares).

The classification task predicts which emotion dominates the audience's reaction to the content rather than the sentiment expressed in the text. A post may be written in neutral language but elicit strong emotional reactions based on the topic's cultural and political resonance.

4.2 Architecture

We fine-tune MARBERTv2 (Abdul-Mageed et al., 2021) [3] due to its strong performance and coverage of dialectal Arabic. Fine-tuning allows the model to adapt its linguistic and semantic representations to the specific characteristics of social media discourse surrounding women's economic empowerment, ensuring that the learned embeddings capture domain-specific patterns of engagement and sentiment. Let $\mathbf{h} \in \mathbb{R}^{768}$ denote the [CLS] representation after dropout. Three task-specific heads are applied:

- **Non-Engagement (binary):** detects whether a post receives any audience reaction at all, distinguishing between visible engagement and audience silence. .
- **Audience Response Classification (6-way):** maps the shared 768-dimensional embedding representation to six probability outputs using a linear layer followed by a softmax activation. It captures which latent linguistic and emotional features correlate with specific response categories (e.g., love, anger, laughter).
- **Engagement Regression (8-D):** predicts eight continuous engagement values (six reaction types, comments and shares) via a linear projection followed by a ReLU activation to enforce non-negativity. It models quantitative engagement behavior, learning patterns such as posts that elicit high comment counts due to debate versus those generating reactions with limited discussion.

Using separate heads for qualitative (response type) and quantitative (engagement magnitude) modeling allows the network to specialize while sharing underlying linguistic knowledge.

[3] https://huggingface.co/UBC-NLP/MARBERTv2

4.3 Targets and Preprocessing

Text is constructed by concatenating the post *Message*, *Image Text*, *Link Text*, and *Description*. Engagement vectors are formed from the raw counts of six reactions + *Comments* + *Shares*. We apply $\log(1 + x)$ followed by per-dimension min-max scaling.

4.4 Training Objective

The overall loss combines three complementary objectives reflecting the classification, regression, and non-engagement detection subtasks.

Focal Loss (Lin et al., 2017) is applied to the audience response classification head to address class imbalance by (1) assigning higher penalties to minority classes (e.g., misclassifying *Angry* incurs ten times the cost of *Love*) and (2) down-weighting easy examples. We set $\gamma = 2$ and use inverse-frequency class weighting.

Mean Squared Error (MSE) is used for engagement prediction, measuring the average squared difference between predicted and true engagement values across the 8 dimensions.

Cross-Entropy Loss (CE) is applied to the non-engagement detection head, a binary classification task distinguishing posts with reactions (engaged) from those without (not engaged).

The total loss is defined as:

$$L = L_{focal} + \lambda L_{MSE} + \mu L_{CE}$$

where $\lambda = 0.5$ controls the engagement regression weight, $\mu = 0.25$ controls the non-engagement detection weight, and L_{CE} is the cross-entropy loss for binary non-engagement classification.

4.4.1 Optimization and Early Stopping

We optimize with AdamW (learning rate 2×10^{-5}), linear warmup (10% of total steps), and gradient clipping ($\|\nabla\| \leq 1.0$). Mini-batch size is 16 and max sequence length is 128. We split the data into 70%/15%/15% (train/validation/test) with stratification by $y^{(cls)}$. Early stopping is based on validation Macro-F1 for the 6-way classification head.

4.4.2 Evaluation Metrics

We report Macro and Weighted-F1 for the 6-way classification, accuracy for non-engagement detection, and analyze the regression head via Pearson correlations between predicted response probabilities and engagement dimensions.

4.4.3 Implementation Details

We use HuggingFace Transformers for MARBERTv2 and tokenization, and PyTorch for training. Class weights α_c are computed via inverse class frequency on the training split. All experiments are run with fixed random seeds (42) for data splits. Hyperparameter grids for λ and μ are evaluated on the validation set, with optimal values $\lambda = 0.5$ (engagement regression weight) and $\mu = 0.25$ (non-engagement detection weight). Best models (by validation Macro-F1) are checkpointed and used for final test reporting.

5 Evaluation

The optimized multi-task model was assessed on the held-out test set, incorporating all three tasks: audience reaction classification, engagement regression, and non-engagement detection.

Reaction Classification The model achieves a test macro-F1 of 72.4 and weighted-F1 of 89.1 in the six-way reaction classification task, substantially improving overall balance across classes. Notably, minority classes exhibit marked gains: Care (61.8), Wow (56.4), and Angry (56.3).

Engagement Regression The regression head achieves an average mean squared error (MSE) of 0.04 across all eight engagement dimensions (six reaction counts, comments, and shares), indicating precise modeling of quantitative engagement.

Non-Engagement Detection The binary non-engagement classifier attains 91.7% accuracy in identifying posts that received no reactions.

Ablation Study Ablating auxiliary components degrades classification:

- Without engagement regression: macro-F1 drops to 67.3 (-5.1).
- Without comments in regression: macro-F1 drops to 69.6 (-2.8).
- Without shares in regression: macro-F1 drops to 71.0 (-1.4).
- Classification-only training: macro-F1 of 64.1 (-8.3).

These results confirm that engagement supervision (especially comments) provides critical signals for improving minority reaction recall.

Reaction-Engagement Correlations Predicted reaction probabilities exhibit strong Pearson correlations with engagement metrics (all $p < .01$): Love-Reactions (0.78) (i.e., when the model assigns a high probability to "Love" posts actually receive more total reactions, indicating that the model correctly captures overall audience approval), Love-Comments (0.34), Love–Shares (0.41), Angry-Comments (0.52), Sad-Comments (0.47), and Haha-Shares (0.31). These patterns validate the interpretability benefits of joint training.

Analysis The results demonstrate that integrating engagement regression and non-engagement detection into reaction classification significantly enhances model performance, particularly on underrepresented reaction types. By incorporating quantitative engagement signals, the model learns behaviorally grounded features that are inaccessible through text alone. The regression outputs also enable identification of polarized posts: for instance, posts with high variance in predicted reaction probabilities (e.g., 45% *Love*, 40% *Angry*) correspond to contested topics that fragment audiences, while low-variance predictions indicate consensus. Frequent errors include misinterpreted dialectal sarcasm, confusion between Sad and Angry reactions, and failures when visual or temporal cues, which are absent from text, drive audience engagement.

6 Conclusion

We presented a multi-task learning framework that jointly models audience reaction classification, engagement regression, and non-engagement detection for Arabic social media. Applied to 158k posts on women's discourse, our approach achieves macro-F1 of 72.4 and weighted-F1 of 89.1. Ablation studies confirm that engagement supervision is critical, improving macro-F1 by 8.3 points, with disproportionate benefits for minority classes. The framework reveals systematic correlations: *Angry* predictions correlate with comments (r=0.52), indicating debate, while *Love* correlates with reaction volume (r=0.78), indicating consensus. By modeling reactions as collective interpretation rather than textual sentiment, our work provides insights into how discourse on women's issues resonates across Arabic-speaking communities. Future work should explore temporal dynamics and extend to other sensitive topics.

7 Limitations

Data Scope and Representativeness This study relies exclusively on publicly available Facebook posts collected via the CrowdTangle API. While Facebook remains a dominant platform in the Arab world, our findings may not generalize to other social media ecosystems (e.g., Twitter/X, Instagram, TikTok) where content formats, audience demographics, and engagement mechanisms differ substantially. The dataset spans 2014–2024, but temporal coverage is uneven due to API access restrictions and platform policy changes. We acknowledge that our corpus may overrepresent certain geographic regions, socioeconomic groups, and organizational voices (e.g., NGOs, media outlets) while underrepresenting grassroots and individual users. The decision to focus on pages rather than personal profiles introduces selection bias toward institutional discourse, potentially missing critical dimensions of everyday conversation about women's empowerment.

Engagement Metrics as Proxies Although engagement metrics (reactions, comments, shares) serve as useful behavioral proxies for audience response, they imperfectly reflect underlying user stance, sentiment, and genuine opinion. These signals can be heavily mediated by multiple confounding factors: (1) *algorithmic curation*, where Facebook's ranking algorithms preferentially surface certain content types, inflating their visibility and engagement; (2) *social desirability bias*, where users may react publicly in ways that conform to perceived group norms rather than expressing private views; (3) *platform affordances*, where the discrete set of six reactions constrains emotional expression and may not capture the full spectrum of audience response; (4) *coordinated inauthentic behavior*, including bot activity, brigading, and organized campaigns that artificially inflate or suppress engagement; and (5) *temporal dynamics*, where early reactions can cascade and influence subsequent engagement patterns through social proof mechanisms. Our model does not explicitly account for these sources of noise and cannot distinguish genuine sentiment from strategically manipulated engagement.

Linguistic and Cultural Limitations Despite MARBERTv2's coverage of dialectal Arabic, our model's performance varies across regional varieties. The training data's dialect distribution may not reflect the true linguistic diversity of Arabic-speaking women's discourse, potentially disadvantaging speakers of underrepresented varieties (e.g., Hassaniya, Sudanese Arabic). Moreover, the model processes only textual content, ignoring visual information (images, videos), which often carries critical cultural and emotional cues in social media posts. Sarcasm, irony, and humor-prevalent in Arabic digital discourse-remain difficult to detect, particularly when they rely on visual-textual interplay or shared cultural knowledge not encoded in text. The model also lacks awareness of temporal context (e.g., posts during political crises, religious holidays, or viral movements) that can dramatically shift interpretation and engagement patterns.

Methodological Constraints Our multi-task framework jointly optimizes reaction classification, engagement regression, and non-engagement detection, but this design imposes architectural constraints. The shared encoder may learn representations that prioritize majority tasks at the expense of minority ones, and the fixed task weighting (controlled by hyperparameter λ) may not adapt optimally across domains or time periods. We evaluate on a single 70/15/15 train-dev-test split; cross-validation or temporal holdout evaluation would provide more robust performance estimates. The model does not incorporate user-level features (e.g., author demographics, network position, historical engagement patterns) or page-level attributes (e.g., follower count, posting frequency), which could improve prediction but raise additional privacy and fairness concerns.

Generalizability and External Validity The focus on women's empowerment discourse means our model is optimized for this specific topical domain. Performance may degrade substantially when applied to other content areas (e.g., sports, entertainment, political news) with different linguistic registers, engagement norms, and audience compositions. While our test set spans a decade, we do not evaluate robustness to distribution shift over time (e.g., training on 2014–2019 and testing on 2020–2024). Societal attitudes toward women's issues, platform features, and user behavior have evolved considerably during this period, and our static model does not adapt to these changes. Deploying such models in real-world content moderation, recommendation, or analytical systems without careful domain adaptation and continuous

monitoring risks producing misleading or harmful predictions.

Computational Resources and Reproducibility Fine-tuning large transformer models requires substantial computational resources (GPU access, memory, training time) that may not be available to all researchers, particularly those in under-resourced regions. While we report fixed random seeds, neural network training can exhibit variance across hardware configurations and software versions. We have not released our dataset due to ethical and legal considerations (see Ethics Statement), which limits full reproducibility. Future work should explore smaller, more efficient architectures and consider the environmental cost of large-scale model training.

8 Ethics Statement

Data Collection and Privacy All data were collected from publicly available Facebook pages and groups via the CrowdTangle API before its discontinuation in August 2024. We restricted collection to content marked as "public" by page administrators, adhering to Meta's Terms of Service and the API's usage policies. No personal user profiles, private messages, or restricted content were accessed. We have permanently removed all personally identifiable information (names, usernames, profile links) from the dataset. However, we acknowledge that "public" data does not imply informed consent for research use, and individuals may not anticipate that their reactions to posts will be analyzed in aggregate. The notion of contextual integrity suggests that users' expectations of privacy vary across social contexts, and repurposing platform data for academic research may violate these expectations even when technically permissible.

Informed Consent and Ethical Considerations Users who reacted to posts in our dataset did not provide explicit consent for their engagement to be included in a research corpus. While legal frameworks in many jurisdictions classify publicly observable social media activity as non-human-subjects research, ethical best practices increasingly emphasize the importance of meaningful consent and respect for user autonomy. We argue that aggregating reactions at the post level—without tracking individual users across posts or inferring personal attributes—reduces privacy risks compared to user-level analysis. Nonetheless, researchers working with such data must weigh the societal benefits of studying public discourse against potential harms, including erosion of trust in digital platforms and contribution to surveillance culture.

Representation, Bias, and Fairness Our dataset reflects existing biases in who participates in online discussions about women's issues, which content becomes visible, and which voices are amplified or marginalized. Women from marginalized communities, including those in conflict zones, rural areas, or under authoritarian regimes, may be underrepresented due to digital access barriers, censorship, or safety concerns. Our model learns from these biased data and will reproduce and potentially amplify such biases in its predictions. For instance, if posts advocating for gender equality receive disproportionately negative reactions in certain regions, the model may learn to associate feminist content with anger or rejection, reinforcing stereotypes. Deploying such models without careful auditing could harm advocacy efforts, misrepresent public opinion, or justify discriminatory practices. We strongly caution against using our model to make high-stakes decisions (e.g., content moderation, funding allocation, policy recommendations) without human oversight and domain expertise.

Cultural Sensitivity and Context Women's empowerment encompasses deeply sensitive topics, including reproductive rights, domestic violence, legal discrimination, and economic autonomy, that vary in social acceptability across Arab societies. Our analysis aggregates engagement across diverse cultural and political contexts, potentially obscuring important regional differences and flattening complex debates into simplified emotional categories. We recognize that Western feminist frameworks do not universally apply, and our interpretation of engagement patterns must be attentive to local norms, values, and power dynamics. Misinterpreting audience reactions without cultural competence risks perpetuating Orientalist stereotypes or undermining grassroots movements.

Dual-Use and Potential for Misuse While our research aims to advance understanding of online discourse about women's issues, the techniques developed here could be misappropriated for harmful purposes. Authoritarian regimes or anti-feminist actors could use reaction prediction models to optimize disinformation campaigns, suppress advocacy

content, or identify and target activists. Corporations might exploit engagement prediction to manipulate users' emotions or amplify divisive content for profit. We have deliberately withheld certain implementation details and will not publicly release trained model weights without establishing appropriate access controls. Researchers and practitioners interested in applying our methods should carefully consider the potential for harm in their specific deployment context and implement safeguards against misuse.

Stakeholder Impact and Accountability Our work directly concerns women's rights advocates, feminist organizations, social media platform operators, and policymakers. We have not engaged in participatory research design or consulted with these stakeholders during model development. Future work should incorporate community input to ensure that research priorities align with the needs and values of those most affected. If our findings inform content moderation policies or algorithmic curation, there must be transparent governance mechanisms, appeals processes, and accountability structures to address harms. We advocate for third-party audits, public reporting of model performance across demographic groups, and ongoing monitoring for unintended consequences.

Transparency and Open Science We are committed to transparency in our methods and findings. However, we have chosen not to release our full dataset publicly due to concerns about re-identification risks, potential misuse, and respect for users' contextual privacy expectations. Researchers may request access for non-commercial, academic purposes through a formal data sharing agreement that specifies permitted uses, requires IRB approval (or equivalent ethical review), and prohibits attempts to de-anonymize individuals. We provide detailed model architecture descriptions, hyperparameters, and evaluation protocols to facilitate replication with alternative data sources. We encourage the NLP community to develop consensus guidelines for responsible sharing of social media corpora, balancing scientific progress with ethical obligations.

Institutional Review and Compliance This research was conducted in accordance with our institution's policies on human subjects research. We comply with the EU General Data Protection Regulation (GDPR), which applies to data from European users, and we have implemented data minimization principles to retain only information necessary for our research objectives.

9 Acknowledgments

This study was supported by the grant NPRP14C0916-210015, awarded by the Qatar Research, Development and Innovation Council (QRDI).

References

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021. Arbert & marbert: Deep bidirectional transformers for arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7088–7105, Online. Association for Computational Linguistics.

Mohammed Al-Badawi. 2024. English and arabic language trends in social media marketing: Analyzing linguistic patterns and their impact on audience engagement in zarqa, jordan market. In *Frontiers of Human Centricity in the Artificial Intelligence-Driven Society 5.0*, pages 969–974. Springer.

Noura AlShenaifi, Nora Al-Twairesh, and Hend Al-Khalifa. 2024a. Fine-tuning marbert for arabic stance detection. In *Proceedings of ArabicNLP 2024*, pages 856–861.

Noura AlShenaifi, Nora Al-Twairesh, and Hend Al-Khalifa. 2024b. Llm-in-the-loop active learning for arabic sentiment analysis. *arXiv preprint arXiv:2509.23515*.

Md Rafiul Biswas, Shimaa Ibrahim, Mabrouka Bessghaier, and Wajdi Zaghouani. 2025. Evaluation of pretrained and instruction-based pretrained models for emotion detection in arabic social media text. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing-Natural Language Processing in the Generative AI Era*, pages 158–165.

Anis Charfi, Mabrouka Ben-Sghaier, Andria Samy Raouf Atalla, Raghda Akasheh, Sara Al-Emadi, and Wajdi Zaghouani. 2024a. Marasta: A multi-dialectal arabic cross-domain stance corpus. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11060–11069.

Anis Charfi, Mabrouka Besghaier, Raghda Akasheh, Andria Atalla, and Wajdi Zaghouani. 2024b. Hate speech detection with adhar: a multi-dialectal hate speech corpus in arabic. *Frontiers in Artificial Intelligence*, 7:1391472.

Rehab Duwairi and Iman Qarqaz. 2017. Collecting and processing arabic facebook comments for sentiment analysis. *International Journal of Computer Applications*, 170(8):1–6.

Lisa Graziani, Stefano Melacci, and Marco Gori. 2019. Jointly learning to detect emotions and predict facebook reactions. In *International Conference on Artificial Neural Networks*, pages 185–195. Springer.

Shimaa Amer Ibrahim, Mabrouka Bessghaier, and Wajdi Zaghouani. 2025. Ahasis shared task: Hybrid lexicon-augmented arabert model for sentiment detection in arabic dialects. In *Proceedings of the Shared Task on Sentiment Analysis for Arabic Dialects*, pages 29–34.

Yunwoo Kim and Junhyuk Hwang. 2025. Predicting social media engagement from emotional and temporal features. *arXiv preprint arXiv:2508.21650*.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

Flor Miriam Plaza-del Arco, Sercan Halat, Sebastian Padó, and Roman Klinger. 2021. Multi-task learning with sentiment, emotion, and target detection to recognize hate speech and offensive language. *arXiv preprint arXiv:2109.10255*.

Chris Pool and Malvina Nissim. 2016. Distant supervision for emotion detection using facebook reactions. In *Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media (PEOPLES)*, pages 30–39, Osaka, Japan. The COLING 2016 Organizing Committee.

Wajdi Zaghouani, Md Rafiul Biswas, Mabrouka Bessghaier, Shimaa Ibrahim, George 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: Shared Tasks*, pages 560–574.