technologies

MDPI

Article

When Emotions Conflict: A Reliability-Aware Framework for Arabic Multi-Label Emotion Detection

Mashary N. Alrasheedy [1,2,*], Sabrina Tiun [1,*] and Fariza Fauzi [1]

[1] Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia (UKM), Bangi 43600, Selangor, Malaysia; fariza.fauzi@ukm.edu.my
[2] Department of Computer Science, Applied College, University of Ha'il, Ha'il 55424, Saudi Arabia
* Correspondence: p105399@siswa.ukm.edu.my (M.N.A.); sabrinatiun@ukm.edu.my (S.T.)

Abstract

Arabic multi-label emotion detection (MLED) in social media remains challenging because dialectal variation, implicit affective cues, and polarity-opposed emotions may occur within the same post. Existing Arabic MLED studies have mainly emphasized thresholded predictive performance, with limited attention to whether model confidence remains reliable under emotionally conflicting conditions. In this study, we propose CONCORD-Emo (CONflict-aware Compositional Representation for Emotion Detection), a reliability-aware framework for Arabic MLED. The framework adopts established label-wise attention, mixture-of-experts routing, Monte Carlo (MC) dropout, and post hoc temperature scaling as supporting mechanisms, while its architecture-level contribution is the conflict-conditioned integration of a residual global anchor with a conflict-aware fusion gate supervised by an automatically derived polarity-conflict target. We evaluated the framework on three Arabic benchmarks: SemEval-2018-Ar, ExaAEC, and SemEval-2025-Arq using predictive and reliability-oriented criteria. CONCORD-Emo remains competitive with strong MARBERT-based baselines. On SemEval-2025-Arq, it attains point estimates of 0.471 for Jaccard, 0.606 for micro-F1, and 0.582 for macro-F1. Paired bootstrap confidence intervals show that most predictive differences include zero, whereas the lower Expected Calibration Error and Brier scores on SemEval-2018-Ar and ExaAEC are consistently supported relative to the controlled baselines. Conflict-conditioned analysis shows that polarity-conflict instances yield lower predictive performance and higher Brier scores than blended-emotion instances. Taken together, these results support a reliability-aware evaluation of Arabic MLED in which polarity conflict, calibration, uncertainty estimation, and selective prediction are examined alongside predictive performance.

Keywords: Arabic natural language processing; multi-label emotion detection; emotional conflict; uncertainty estimation; calibration; Monte Carlo dropout; selective prediction

Check for updates

Academic Editor: Hongying Meng

Received: 26 May 2026
Revised: 23 June 2026
Accepted: 29 June 2026
Published: 1 July 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Emotion detection from text supports applications in social media monitoring, affective analytics, human–computer interaction, and decision-support systems. Transformer-based language models have improved emotion classification through contextual representation learning, yet Arabic social media remains challenging because short posts may contain dialectal variation, informal spelling, implicit or figurative affective cues, and multiple emotions within the same utterance [1–5]. Multi-label emotion detection (MLED) accommodates the co-occurrence of several emotions, but not all co-occurrences have the same

affective structure. Some instances express compatible emotional blends, whereas others combine polarity-opposed cues, as may occur in contrastive discourse, sarcasm, grief narratives, or complex social commentary. A conventional multi-label classifier can assign the corresponding labels, but the predicted label set alone does not indicate whether the model is reliable under such conflict [6–8].

Reliability is particularly relevant because affective meaning is subjective and benchmark labels are commonly obtained by aggregating multiple annotations. Although prior Arabic MLED studies have improved predictive performance, aggregate metrics such as the Jaccard score, micro-F1, macro-F1, and Hamming loss do not assess whether probabilities are calibrated, whether incorrect predictions are overconfident, or whether uncertainty estimates identify high-risk instances. This motivates a joint evaluation of predictive performance and probabilistic reliability under emotionally conflicting conditions [9–14].

To address this gap, we propose CONCORD-Emo (CONflict-aware Compositional Representation for Emotion Detection), a reliability-aware framework for Arabic MLED. CONCORD-Emo models emotional conflict as a structured reliability condition rather than treating all label co-occurrences as equivalent. We distinguish the contribution of the framework from the established techniques used for its implementation. Label-wise attention and generic mixture-of-experts routing were adapted as supporting representation mechanisms, whereas Monte Carlo (MC) dropout [15] and post hoc temperature scaling [14] were used in their established uncertainty estimation and calibration roles. The architecture-level contribution is the conflict-conditioned integration of residual global anchoring with a conflict-aware fusion gate supervised by an automatically derived polarity conflict target. The empirical contribution is the operational separation of compatible emotional blending from polarity conflict and their joint evaluation using predictive, calibration, uncertainty, and selective-prediction criteria.

We further examined how emotional structure affects reliability by categorizing instances into three polarity-based groups: *single emotion*, *blended emotion*, and *polarity conflict*. The single-emotion group contains one active polarity-bearing label, the blended-emotion group contains multiple polarity-compatible labels, and the polarity-conflict group contains both positive and negative labels. These groups form an operational, dataset-specific partition of the predefined gold emotion labels rather than a universal claim about valence in Arabic language or culture. Labels with context-dependent or non-directional valence do not contribute to the positive and negative counts. The central hypothesis is that polarity-conflict instances are associated with weaker predictive reliability and elevated uncertainty than blended-emotion instances, particularly in terms of predictive entropy.

We evaluated CONCORD-Emo on SemEval-2018-Ar, ExaAEC, and SemEval-2025-Arq using complementary predictive, calibration, probabilistic, uncertainty, and selective prediction criteria [9–11].

The main contributions of this work are as follows:

- We operationalize polarity-opposed emotional conflict as a structured reliability condition and distinguish it from compatible multi-label blending through the single, blended, and conflict groups.
- We introduce a conflict-conditioned architecture that couples residual global anchoring with a conflict-aware fusion gate supervised by an automatically derived polarity-conflict target, while using established representation, uncertainty, calibration, and selective-prediction techniques in supporting roles.
- We conduct a controlled multi-benchmark reliability evaluation and quantify test-set sampling uncertainty through paired bootstrap confidence intervals for the principal model comparisons.

- The results characterize CONCORD-Emo as predictively competitive and provide interval-supported calibration and Brier-score advantages on SemEval-2018-Ar and ExaAEC. Conflict instances also exhibit higher Brier contribution and Jaccard loss than blended instances across all three datasets.

The remainder of this paper is organized as follows. Section 2 reviews related work on Arabic MLED, structured emotion modeling, uncertainty estimation, calibration and selective prediction. Section 3 presents the proposed CONCORD-Emo framework and the reliability estimation procedure. Section 4 describes the datasets, baselines, implementation settings and evaluation metrics. Section 5 presents the predictive, calibration, conflict-conditioned, ablation, and selective prediction results. Section 6 discusses the main findings and their limitations. Finally, Section 7 concludes the study and outlines the future directions.

2. Related Work

2.1. Arabic Multi-Label Emotion Detection

Arabic emotion detection has advanced through benchmark datasets, pretrained Arabic language models, and shared-task evaluations. SemEval-2018 Task 1 Affect in Tweets provided one of the most influential Arabic emotion benchmarks, including a multi-label emotion classification subtask over 11 emotions [9]. Earlier Arabic systems relied mainly on traditional machine learning, lexical features, and handcrafted representations, demonstrating the feasibility of Arabic multi-label emotion detection (MLED) on social-media text [16,17]. Later, neural and hybrid models improved predictive performance by exploiting distributed semantic representations and sequential modeling [18–20].

Dataset development has also influenced Arabic MLED research. Sarbazi-Azad et al. [10] introduced ExaAEC, a large Arabic tweet corpus containing approximately 20,000 instances annotated with ten Plutchik-inspired emotion classes. Recently, SemEval-2025 Task 11 extended multilingual emotion detection to more than 30 languages, including Algerian Arabic (Arq), with Track A formulated as a multi-label task evaluated using macro-F_1 [11]. The shared-task systems reported for this benchmark commonly relied on BERT-family fine-tuning or instruction-tuned large language models (LLMs), often combined with prompt engineering, parameter-efficient fine-tuning, ensembling, or data augmentation [11].

Recent competitive systems span several methodological families rather than a single Arabic transformer. The HTU system represents dialect-specific transformer ensembling by combining multiple DziriBERT classifiers with label-fused iterative mask-filling [21]. Multilingual and cross-lingual approaches include INFOTEC-NLP, which combines XLM-RoBERTa representations with bidirectional long short-term memory (BiLSTM) and attention; LATE-GIL-NLP, which fine-tunes mDeBERTa-v3 with optimized loss combinations; and YNU-HPCC, which translates the inputs to English and applies a DeBERTa-based prediction architecture [22–24]. The broader SemEval-2025 shared-task landscape also includes large pretrained language models, prompting, parameter-efficient fine-tuning, data augmentation, and ensembling [11]. These systems provide important predictive contexts, although most do not report probability calibration, Brier score, MC-dropout mutual information, or riskcoverage behavior under a common protocol.

The transformer era has further strengthened Arabic natural language processing (NLP). Models such as AraBERT [25], MARBERT [26], ArabicBERT [27], and CAMeL-BERT [28] provide contextual representations that are particularly useful for dialect-rich and social-media Arabic. These models are important for Arabic MLED because emotional expressions in social media often involve informal spelling, dialectal variation, implicit pragmatic cues, and short noisy contexts.

Despite these advances, most Arabic MLED studies remain primarily prediction-oriented. They usually formulate the task as multiple binary label decisions and evaluate models using aggregate metrics such as the Jaccard score, micro-F_1, macro-F_1, or Hamming loss. These metrics are necessary for measuring classification quality but do not indicate whether predicted probabilities remain reliable when a text contains ambiguous or polarity-opposed emotional cues. This limitation motivates our study, in which we treat Arabic MLED not only as a label-prediction problem but also as a reliability problem under emotionally conflicting conditions.

2.2. Label Dependency, Imbalance Handling, and Structured Emotion Modeling

MLED differs from single-label classification because the active labels within the same instance may be semantically related. In emotion detection, such relations are especially important because emotions may co-occur in meaningful patterns rather than as independent targets. Prior work has therefore explored structured modeling approaches, including attention mechanisms, graph-based representations, and label-dependency modeling, to capture inter-label relations and contextual interactions [29–31].

In Arabic MLED, imbalance and label dependency have recently received explicit attention. GTHL-Emo [32] introduced an adaptive imbalance-aware and correlation-aligned framework that combines transformer-based semantic representation, GraphSAGE-based inter-instance modeling, transformer-based label embeddings, empirical label co-occurrence alignment, and adaptive hybrid loss weighting. By jointly addressing rare-label learning and structured label correlations, GTHL-Emo demonstrated strong predictive results on the SemEval-2018-Ar, ExaAEC, and SemEval-2025-Arq datasets. This line of work shows that Arabic MLED benefits from modeling label imbalance and dependency rather than treating all emotion labels as isolated, binary targets.

However, label co-occurrence and emotional conflict are not equivalent. Some multi-label cases express compatible emotional blends, whereas others combine polarity-opposed signals, such as positive and negative emotions within the same input. A model that captures label co-occurrence may learn that two emotions often appear together; however, it does not necessarily distinguish whether this co-occurrence reflects compatible affective blending or a structurally conflicting emotional state. This distinction matters because polarity-opposed inputs may increase prediction difficulty and reduce probability reliability.

Existing structured models mainly use label relations, imbalance handling, or graph-based representation learning to improve predictive performance. They do not generally connect emotional structure to calibration, uncertainty estimation, or selective prediction. We address this gap by using CONCORD-Emo to treat polarity conflict as a reliability condition. The framework separates single-emotion, blended-emotion, and polarity-conflict instances to test whether conflict is associated with weaker predictive performance and lower probabilistic reliability.

2.3. Uncertainty Estimation and Calibration

Predictive accuracy alone is not sufficient when model confidence is used for interpretation or downstream decision-making. Neural classifiers may achieve strong classification performance while producing overconfident probability estimates, especially under ambiguity, distributional difficulty, or conflicting evidence [14]. This issue is central to emotion detection because affective meaning is subjective, context-dependent, and often expressed indirectly.

Uncertainty estimation provides one way to assess the reliability of a prediction. Existing approaches include Bayesian neural networks, deep ensembles, evidential learning, and approximate Bayesian methods [33,34]. Deep ensembles are often effective for uncertainty

estimation and calibration [35], but they are computationally expensive when applied to high-capacity transformer encoders. MC dropout offers a more practical approximation by keeping dropout active during inference and averaging multiple stochastic forward passes [15]. This enables the estimation of predictive entropy and mutual-information (MI)-based epistemic uncertainty without training multiple independent models.

Mutual information has three conceptually distinct uses in machine learning. First, as an information-theoretic quantity, it measures statistical dependence between random variables and can be used to analyze relationships among representations, labels, or other attributes [36]. Second, an estimated MI term can be incorporated into a training objective as a differentiable regularizer that encourages or suppresses a selected dependence. Incremona et al. presented this use through an uncertainty-aware MI regularizer for bias mitigation [37]. Third, in stochastic predictive models, the MI between a prediction and the stochastic model realization can be estimated from repeated predictive samples and used as an inference-time approximation of epistemic uncertainty [15]. CONCORD-Emo uses only the third role. Its MI score is computed after training from MC-dropout predictions; it is neither an estimate of pairwise dependence between emotion labels nor a differentiable term in the training objective.

Calibration complements uncertainty estimation by evaluating whether predicted probabilities are aligned with empirical correctness. Temperature scaling is a widely used post hoc calibration method that rescales logits using validation data to improve probability calibration [14]. The calibration quality can be assessed using the expected calibration error (ECE), Brier score, and reliability diagrams. These metrics are important for reliability-aware MLED because a model should not only assign plausible emotion labels but also express appropriate confidence in those labels.

2.4. Selective Prediction and Reliability-Aware Emotion Detection

Selective prediction provides an operational use for uncertainty estimates. Instead of forcing a prediction for every instance, a model may abstain from high-uncertainty cases and retain predictions with lower estimated risk [38]. This setting is particularly relevant for MLED because uncertainty may arise at both the label and instance levels. A text may contain clear evidence for one emotion while remaining ambiguous for another, or it may contain polarity-opposed cues that make the overall label set less reliable.

In Arabic MLED, uncertainty-aware evaluation remains limited compared with accuracy-oriented evaluation. Most existing studies focus on stronger encoders, class-imbalance handling, label-dependency modeling, or improved aggregate performance. Reliability measures such as predictive entropy, MI-based epistemic uncertainty, ECE, Brier score, reliability diagrams, and risk–coverage analysis are rarely treated as central evaluation dimensions. More importantly, the studies reviewed here do not explicitly formalize polarity-opposed emotional conflict as a condition for reliability analysis. This gap is the main motivation for CONCORD-Emo.

2.5. Positioning of CONCORD-Emo

The reviewed literature shows that Arabic MLED has progressed through stronger contextual encoders, transformer-based representation learning, class imbalance handling, label dependency modeling, graph-enhanced architectures, and more recently, LLM-based shared-task systems. These directions are valuable; however, they primarily address predictive performance. They do not directly answer whether model confidence remains reliable when Arabic text contains polarity-opposed emotional evidence.

The component provenance of CONCORD-Emo is explicitly distinguished from its claimed contribution. Attention-based label modeling, generic mixture-of-experts rout-

ing, MC dropout, temperature scaling, calibration metrics, and selective prediction are established techniques adopted in supporting roles rather than presented as standalone innovations [29,30,39]. The architecture-level contribution claimed in this study is the conflict-conditioned coupling of a residual global anchor with a supervised conflict-aware fusion gate. The study-level contribution is the operational separation of compatible emotional blending from polarity conflict and the joint analysis of predictive quality, calibration, probabilistic error, predictive entropy, MC-dropout mutual information, and selective prediction under these emotional structures.

The relevant comparison landscape can be organized into four broad families: Arabic-specific transformer and structured prediction systems; multilingual transformer and hybrid neural systems; ensemble or translation-based shared-task pipelines; and recent systems using larger pretrained language models, prompting, or parameter-efficient adaptation. CONCORD-Emo is closest to the first family in its use of an Arabic social media encoder, but it differs in its primary objective. The framework is designed to analyze conflict-conditioned calibration and uncertainty rather than to optimize only a shared-task classification score. Consequently, comparisons with the other families are informative for predictive positioning but do not provide controlled evidence for the reliability contribution.

We address this gap through CONCORD-Emo, a reliability-aware framework for Arabic MLED. The framework combines label-wise attention, residual global anchoring, a polarity-aware mixture-of-experts module, conflict-aware gating, and label-specific classification. We further incorporate MC-dropout uncertainty estimation, post-hoc temperature scaling, calibration metrics, and selective prediction. The model is not positioned as a universal accuracy leader over all transformer or LLM-based systems. Rather, we position it as a controlled reliability-aware architecture that remains predictively competitive while explicitly analyzing calibration and uncertainty under emotional conflict.

This positioning is distinct from that of GTHL-Emo [32]. GTHL-Emo addresses adaptive imbalance-aware optimization and correlation-aligned structural learning, whereas CONCORD-Emo focuses on conflict-aware reliability modeling and epistemic uncertainty quantification. Therefore, these two works are complementary: GTHL-Emo targets imbalance and label-dependency challenges, whereas CONCORD-Emo investigates whether emotionally conflicting inputs produce less reliable predictions and whether uncertainty and calibration can make Arabic MLED more trustworthy.

Accordingly, CONCORD-Emo is not presented as a new uncertainty estimator, calibration method, or generic mixture of expert formulations. Its contribution is a conflict-conditioned architecture and evaluation design for Arabic MLED: the architecture preserves both label-specific and sentence-level evidence, uses polarity-derived supervision to control their fusion with expert representation, and evaluates whether polarity opposition is associated with weaker predictive and probabilistic reliability.

3. Proposed Methodology

In this section, we present CONCORD-Emo, a reliability-aware framework for Arabic multi-label emotion detection (MLED). We designed the framework to model three complementary aspects of emotionally complex Arabic text: local label-specific evidence, polarity-aware affective composition, and conflict-conditioned reliability. Rather than treating all label co-occurrences as equivalent, we use CONCORD-Emo to examine whether polarity-opposed emotional cues are associated with weaker prediction reliability, higher uncertainty, and poorer calibration.

The methodology distinguishes the proposed conflict-conditioned components from established supporting techniques. Label-wise attention and generic mixture-of-experts routing are adapted from established representation learning approaches, while

MC dropout, temperature scaling, calibration metrics, and selective prediction retain their established uncertainty, calibration, and abstention roles. The components introduced for the present conflict-conditioned design are the residual global anchor in Equations (12)– (14) and the conflict-aware fusion mechanism in Equations (19) and (20), together with the polarity-derived auxiliary supervision in Equations (26)–(28).

Figure 1 illustrates the architecture of CONCORD-Emo. We organize the framework into two connected phases. The first phase performs representation learning and prediction. A processed Arabic input is encoded using MARBERT to obtain contextual token-level representations and a global sentence representation. The token-level representations are used by the label-wise attention component to extract emotion-specific local evidence, while the global representation provides sentence-level context for residual anchoring and conflict-aware gating. The residual global anchor stabilizes the label-conditioned representation by combining local label evidence with global context. The anchored representation is then processed by a polarity-aware mixture-of-experts (MoE) module, followed by layer normalization to model positive, negative, and ambiguous/neutral affective signals. The conflict-aware gate fuses the anchored representation, normalized polarity-aware mixture, and global context before the label-specific classifier produces pre-sigmoid emotion scores, referred to as logits.

In the second phase, we perform inference-time reliability evaluation. We process the emotion logits using Monte Carlo (MC) dropout and temperature scaling to obtain calibrated multi-label probabilities. We use these probabilities, together with uncertainty estimates derived from stochastic inference, to compute reliability-oriented outputs, including predictive entropy, mutual-information-based epistemic uncertainty, Expected Calibration Error (ECE), Brier score, Area Under the Risk–Coverage Curve (AURC), and selective-prediction behavior. The following subsections provide a detailed formulation of each component.

3.1. Problem Formulation

Let $\mathcal{D} = \{(x_i, \mathbf{y}_i)\}_{i=1}^{N}$ denote an Arabic MLED dataset containing N text instances. Each input x_i is an Arabic tweet or short social media post, and $\mathbf{y}_i \in \{0,1\}^L$ is a binary multi-label vector, where L denotes the number of emotion labels. The lth element $y_{i,l} = 1$ indicates that emotion label l is present in x_i, whereas $y_{i,l} = 0$ indicates its absence.

We learn a function f_θ, parameterized by θ, that maps an input text x to a vector of emotion probabilities:

$$\mathbf{p}(x) = f_\theta(x) = [p_1(x), p_2(x), \ldots, p_L(x)], \tag{1}$$

where $p_l(x)$ denotes the predicted probability that emotion label l is active in x.

A common output-layer formulation for MLED approximates the joint label distribution as a product of label-wise Bernoulli probabilities:

$$P(\mathbf{y} \mid x) \approx \prod_{l=1}^{L} P(y_l \mid x). \tag{2}$$

This factorization enables efficient multi-label prediction through sigmoid outputs. However, it does not distinguish ordinary label co-occurrence from polarity-opposed emotional conflict. It also does not indicate whether the model's confidence is reliable when the emotional evidence is ambiguous or contradictory. CONCORD-Emo addresses this limitation by combining conflict-aware representation learning with uncertainty estimation and calibration.

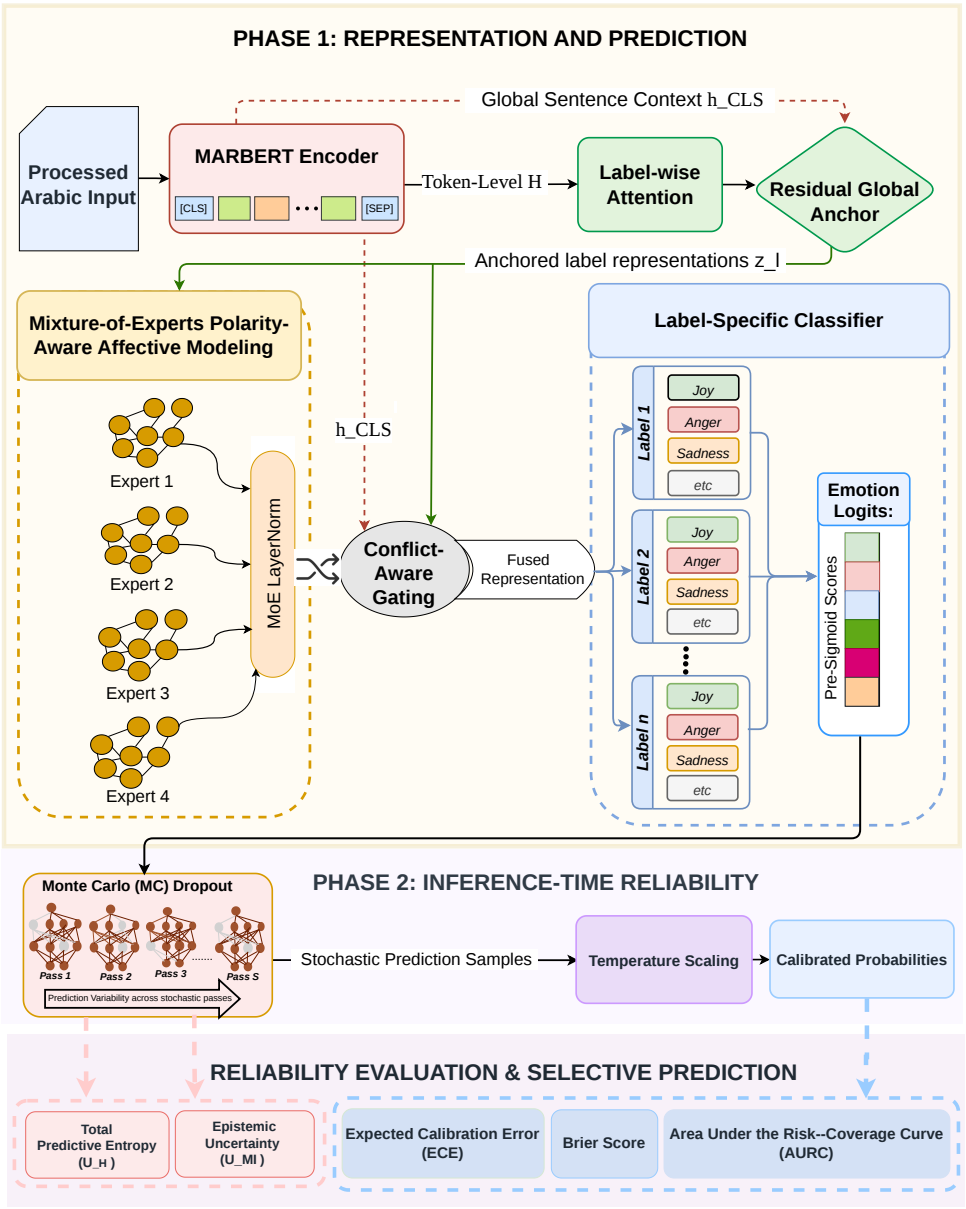


Figure 1. Overview of the proposed CONCORD-Emo framework.

3.2. Polarity Structure and Conflict Label Construction

To examine whether emotional conflict affects predictive reliability, we categorized each instance according to the polarity structure of its gold emotion labels. The taxonomy is applied to the predefined benchmark labels rather than inferred from the lexical content of the Arabic post. Therefore, it provides a deterministic label-level coding rule, but it is not presented as a culturally universal model of affective valence. The label inventories follow the definitions of SemEval-2018-Ar [9], ExaAEC [10], and SemEval-2025-Arq [11]. A binary directional axis is used only because the study hypothesis concerns the co-occurrence of clearly opposed affective signals; it is not intended to represent the full structure of emotion.

Let $\mathcal{P}$ denote the set of positive-polarity emotions and $\mathcal{N}$ the set of negative-polarity emotions. Because the evaluated datasets used different emotion inventories, we defined the polarity sets separately for each corpus. For SemEval-2018-Ar, $\mathcal{P} = \{joy, love, optimism, trust\}$ and $\mathcal{N} = \{anger, disgust, fear, pessimism, sadness\}$. For ExaAEC, $\mathcal{P} = \{joy, love, acceptance\}$ and $\mathcal{N} = \{anger, disgust, fear, sadness\}$. For SemEval-2025-Arq, $\mathcal{P} = \{joy\}$ and $\mathcal{N} = \{anger, disgust, fear, sadness\}$.

The directional criterion was deliberately conservative. Labels whose category names express favorable or affiliative affect are counted as positive, whereas labels that express adverse affect are counted as negative. Anticipation and surprise were excluded because either may accompany a positive or negative outcome, while neutral does not express directional polarity. The excluded labels do not change the polarity counts when they co-occur with the directional labels. These labels remained prediction targets in the original multi-label task and were included in the predictive, calibration, and uncertainty evaluations; they were excluded only from the positive and negative counts used to construct the polarity-structure groups. For example, instances labeled anger, joy, and surprise were assigned to the conflict group because anger and joy contributed opposing polarity signals, whereas surprise was omitted from the counts.

For a label vector $\mathbf{y}$, the number of active positive and negative labels is computed as

$$n_+(\mathbf{y}) = \sum_{l \in \mathcal{P}} y_l, \qquad n_-(\mathbf{y}) = \sum_{l \in \mathcal{N}} y_l. \tag{3}$$

We derive the conflict target automatically from the gold label vector:

$$y^{\text{conf}} = \begin{cases} 1, & \text{if } n_+(\mathbf{y}) > 0 \text{ and } n_-(\mathbf{y}) > 0, \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

For the group-wise analysis, an instance is eligible for the three-group partition when it contains at least one polarity-bearing label, that is, when $n_+ + n_- > 0$. For eligible instances, we assign the polarity structure as

$$\text{group}(x) = \begin{cases} \textit{Single}, & \text{if } n_+ + n_- = 1, \\ \textit{Blended}, & \text{if } n_+ + n_- > 1 \text{ and } (n_+ = 0 \text{ or } n_- = 0), \\ \textit{Conflict}, & \text{if } n_+ > 0 \text{ and } n_- > 0. \end{cases} \tag{5}$$

Instances containing only excluded labels have $n_+ = n_- = 0$ and are omitted from the three group-conditioned summaries. Using Equation (4), they remain non-conflict instances. The terms single and blended refer to the number and compatibility of polarity-bearing labels, not necessarily the total number of active gold labels.

The representative assignments derived from Arabic cases reported in Appendix D are summarized in Table 1.

Table 1. Representative polarity-structure assignments using qualitative cases reported in Appendix D. The grouping is determined exclusively from the gold emotion labels.

Dataset	Gold Labels	n_+	n_-	Group	Assignment Rationale
SemEval-2025-Arq	disgust	0	1	Single	One active negative-polarity label.
SemEval-2018-Ar	joy; love; optimism	3	0	Blended	Multiple active labels with compatible positive polarity.
ExaAEC	love; sadness	1	1	Conflict	Positive- and negative-polarity labels co-occur.

This formulation treats polarity conflict as an operational reliability condition. Its purpose is to define a reproducible contrast between compatible and opposed label structures, rather than to claim a universal Arabic emotion taxonomy. The empirical usefulness of

this partition is evaluated rather than assumed through group-wise predictive, calibration, probabilistic error, and uncertainty results.

3.3. Conflict-Reliability Hypothesis

The primary reliability contrast is defined as the difference between conflicting and blended instances. Both groups contain multiple polarity-bearing labels, but they differ in whether these labels have compatible or opposing polarities. Therefore, this comparison separates polarity opposition from multi-label cardinality more directly than a comparison of single instances. The single group was retained as the descriptive reference.

Let $U(x)$ denote the uncertainty score for input x. In this study, uncertainty was instantiated using predictive entropy and mutual information-based epistemic uncertainty derived from MC-dropout predictions. The central hypothesis is

$$\mathbb{E}[U(x) \mid \text{group}(x) = \textit{Conflict}] > \mathbb{E}[U(x) \mid \text{group}(x) = \textit{Blended}]. \tag{6}$$

Equation (6) is empirically evaluated as an association between polarity opposition and uncertainty, not as a causal claim. Predictive reliability may also be influenced by label imbalance, dataset size, annotation ambiguity, dialectal variation, and the model capacity.

3.4. Arabic Transformer Encoding

We tokenize each Arabic input text x and pass it through a pretrained Arabic transformer encoder. In the experiments, we use MARBERT as the encoder because of its suitability for Arabic social media and dialectal text. Let the input contain T tokens. The transformer produces contextual token representations:

$$\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_T] \in \mathbb{R}^{T \times d}, \tag{7}$$

where $\mathbf{h}_t \in \mathbb{R}^d$ denotes the contextual representation of token t, and d is the hidden dimension.

The encoder also produces a global sentence representation using the classification token:

$$\mathbf{c} = \mathbf{h}_{\text{CLS}} \in \mathbb{R}^d. \tag{8}$$

The token-level matrix $\mathbf{H}$ is used for label-wise attention, whereas the global vector $\mathbf{c}$ is used for residual anchoring and conflict-aware fusion.

3.5. Label-Wise Attention

Emotion labels may correspond to different textual cues within the same text input. To capture local label-specific evidence, CONCORD-Emo applies label-wise attention. Let $\mathbf{e}_l \in \mathbb{R}^d$ be a trainable embedding of label l. The attention score between label l and token t is computed as

$$a_{l,t} = \mathbf{e}_l^\top \mathbf{h}_t. \tag{9}$$

The attention weight is obtained using a masked softmax over valid tokens:

$$\alpha_{l,t} = \frac{\exp(a_{l,t})}{\sum_{k=1}^{T} \exp(a_{l,k})}, \tag{10}$$

where padding tokens are excluded using the attention mask during implementation.

The label-specific attentive representation is computed as

$$\mathbf{z}_l^{\text{att}} = \sum_{t=1}^{T} \alpha_{l,t} \mathbf{h}_t. \tag{11}$$

This representation allows each label to attend to the parts of the Arabic text that are most relevant to its emotional evidence.

3.6. Residual Global Anchoring

Label-wise attention focuses on token-local evidence for each emotion label. This is useful for identifying emotion-bearing cues, but it may underrepresent the sentence-level context when emotional evidence is implicit, sparse, or distributed across the input. To preserve the global context while retaining label-specific attention, we introduce a residual global anchoring mechanism. This design is motivated by residual learning, which preserves reference information through additive pathways [40], and by layer normalization, which stabilizes hidden representations during neural network training [41].

The global *[CLS]* representation is first projected into the label representation space:

$$\mathbf{g}_l = \mathbf{W}_g \mathbf{c} + \mathbf{b}_g, \tag{12}$$

where $\mathbf{W}_g$ and $\mathbf{b}_g$ are learnable projection parameters. The projected vector is shared across labels but expanded along the label dimension during implementation.

A learnable anchor gate determines how much the model should rely on local label-specific attention versus global sentence-level context:

$$\beta_l = \sigma\Big(\mathbf{w}_\beta^\top [\mathbf{z}_l^{\text{att}}; \mathbf{g}_l] + b_\beta\Big), \tag{13}$$

where $[\cdot;\cdot]$ denotes concatenation, $\sigma(\cdot)$ is the sigmoid function, and $\beta_l \in [0,1]$ is a scalar residual gate for label l.

The anchored label representation is then computed as

$$\mathbf{z}_l = \text{LayerNorm}\big(\beta_l \mathbf{z}_l^{\text{att}} + (1-\beta_l)\mathbf{g}_l\big). \tag{14}$$

Equation (14) preserves the label-wise attention while preventing the model from relying solely on sparse local cues. This mechanism stabilizes the label-conditioned representations by retaining the sentence-level context within each label representation.

3.7. Polarity-Aware Mixture-of-Experts

Mixture-of-experts (MoE) architectures combine multiple expert transformations through a learned gating function, allowing different experts to specialize in different regions of the representation space [39]. Building on this principle, we incorporate a polarity-aware MoE module to model affective composition. The anchored label representation $\mathbf{z}_l$ is passed to K expert networks:

$$\mathbf{r}_{l,k} = \phi_k(\mathbf{z}_l), \qquad k = 1, \ldots, K, \tag{15}$$

where $\phi_k(\cdot)$ denotes the transformation learned by expert k.

We compute expert mixture weights using a gating network:

$$\boldsymbol{\pi}_l = \text{softmax}(\mathbf{W}_\pi \mathbf{z}_l + \mathbf{b}_\pi), \tag{16}$$

where $\boldsymbol{\pi}_l = [\pi_{l,1}, \ldots, \pi_{l,K}]$ and $\sum_{k=1}^{K} \pi_{l,k} = 1$.

The expert outputs are combined as

$$\mathbf{u}_l^{\text{MoE}} = \sum_{k=1}^{K} \pi_{l,k} \mathbf{r}_{l,k}. \tag{17}$$

We normalize the mixture output to reduce representation-scale variation:

$$\mathbf{u}_l = \mathrm{LayerNorm}\left(\mathbf{u}_l^{\mathrm{MoE}}\right). \tag{18}$$

The resulting vector $\mathbf{u}_l$ represents the normalized polarity-aware expert mixture for label l.

3.8. Conflict-Aware Gating

The polarity-aware expert mixture captures affective composition, whereas the anchored representation preserves local label-specific evidence and global sentence context. CONCORD-Emo uses a conflict-aware gate to determine how strongly the final label representation should rely on the polarity-aware mixture. For each label l, the gate is computed as

$$s_l = \sigma\left(\mathbf{w}_c^\top [\mathbf{z}_l; \mathbf{c}] + b_c\right), \tag{19}$$

where $s_l \in [0, 1]$ is the label-level conflict-aware gate, and $\mathbf{c}$ is the global sentence representation from Equation (8).

The final fused representation is computed as

$$\tilde{\mathbf{u}}_l = (1 - s_l)\mathbf{z}_l + s_l \mathbf{u}_l. \tag{20}$$

When s_l is low, the model relies more strongly on anchored label representation. When s_l is high, it assigns a greater weight to the polarity-aware expert mixture. This mechanism allows the classifier to adapt its representation according to the estimated conflict structure of the input data.

3.9. Label-Specific Classification

Instead of using a single shared linear classifier for all labels, CONCORD-Emo uses label-specific classifier parameters. For each emotion label l, the final representation $\tilde{\mathbf{u}}_l$ is mapped to a pre-sigmoid logit:

$$o_l(x) = \mathbf{w}_l^\top \tilde{\mathbf{u}}_l + b_l, \tag{21}$$

where $\mathbf{w}_l$ and b_l are the classifier weight and bias for label l.

The corresponding probability is obtained through the sigmoid function:

$$p_l(x) = \sigma(o_l(x)). \tag{22}$$

The predicted multi-label probability vector is therefore

$$\mathbf{p}(x) = [p_1(x), p_2(x), \ldots, p_L(x)]. \tag{23}$$

Binary predictions are obtained using thresholds tuned only on the validation set:

$$\hat{y}_l(x) = \begin{cases} 1, & \text{if } p_l(x) \geq \tau_l, \\ 0, & \text{otherwise,} \end{cases} \tag{24}$$

where τ_l denotes the decision threshold for label l. In the experiments, threshold tuning is performed on the validation split and is not optimized on the test set.

3.10. Training Objective

The main classification objective is binary cross-entropy (BCE), computed independently across labels:

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{L}\sum_{l=1}^{L}[y_l \log p_l(x) + (1-y_l)\log(1-p_l(x))]. \tag{25}$$

To supervise the conflict-aware gate, we use the automatically derived conflict target in Equation (4). The label-level conflict scores are pooled into an instance-level conflict score:

$$s(x) = \frac{1}{L}\sum_{l=1}^{L} s_l. \tag{26}$$

The auxiliary conflict loss is defined as

$$\mathcal{L}_{\text{conf}} = -\Big[y^{\text{conf}} \log s(x) + (1-y^{\text{conf}})\log(1-s(x))\Big]. \tag{27}$$

The final training objective is

$$\mathcal{L} = \mathcal{L}_{\text{BCE}} + \lambda_{\text{conf}}\mathcal{L}_{\text{conf}}. \tag{28}$$

We set the conflict-loss coefficient λ_{conf} to 0.05 for SemEval-2018-Ar and ExaAEC and to 0.03 for SemEval-2025-Arq. This setting reduces the strength of conflict supervision on the smaller SemEval-2025-Arq dataset while preserving conflict-aware objectives.

Algorithm 1 summarizes this training procedure. The algorithm follows the computational flow in Equations (7) and (28).

Algorithm 1: Training procedure for CONCORD-Emo.

Input: Training dataset $\mathcal{D}$, validation set $\mathcal{D}_{\text{val}}$, learning rate η, conflict-loss weight λ_{conf}

Output: Trained model parameters θ and validation-tuned thresholds τ_l

for *each training epoch* **do**
  for *each mini-batch* $(x, \mathbf{y}) \in \mathcal{D}$ **do**
    Encode input text using the Arabic transformer to obtain $\mathbf{H}$ and $\mathbf{c}$ using Equations (7) and (8);
    Compute label-wise attention scores, weights, and attentive representations using Equations (9)–(11);
    Project the global representation and compute residual global anchoring using Equations (12)–(14);
    Compute expert outputs, mixture weights, and normalized MoE representations using Equations (15)–(18);
    Compute conflict-aware gates and fused representations using Equations (19) and (20);
    Compute label logits and probabilities using Equations (21) and (22);
    Derive the conflict target using Equation (4);
    Compute $\mathcal{L}_{\text{BCE}}$, $\mathcal{L}_{\text{conf}}$, and $\mathcal{L}$ using Equations (25)–(28);
    Update θ using gradient-based optimization with learning rate η;
Tune decision thresholds τ_l on $\mathcal{D}_{\text{val}}$ using Equation (24);
return θ *and* τ_l;

3.11. Monte Carlo Dropout for Predictive Uncertainty

To estimate the predictive uncertainty, we used the MC dropout during inference. Dropout remains active, and the model performs S stochastic forward passes for the same input x. For pass s, the model produces a probability vector:

$$\mathbf{p}^{(s)}(x) = [p_1^{(s)}(x), p_2^{(s)}(x), \ldots, p_L^{(s)}(x)], \qquad s = 1, \ldots, S. \tag{29}$$

The predictive mean is computed as

$$\bar{\mathbf{p}}(x) = \frac{1}{S} \sum_{s=1}^{S} \mathbf{p}^{(s)}(x). \tag{30}$$

Predictive entropy is used to estimate total predictive uncertainty:

$$U_H(x) = \frac{1}{L} \sum_{l=1}^{L} H(\bar{p}_l(x)), \tag{31}$$

where the binary entropy function is

$$H(p) = -p \log p - (1 - p) \log(1 - p). \tag{32}$$

To approximate epistemic uncertainty, we compute mutual information (MI) across stochastic forward passes:

$$U_{MI}(x) = \frac{1}{L} \sum_{l=1}^{L} \left[H(\bar{p}_l(x)) - \frac{1}{S} \sum_{s=1}^{S} H(p_l^{(s)}(x)) \right]. \tag{33}$$

In this formulation, MI is used only as an inference-time uncertainty diagnostic. For each emotion label, it compares the entropy of the mean predictive probability with the mean entropy of the individual stochastic predictions. The label-wise values are then averaged to obtain an instance-level score. Consequently, this quantity does not estimate the dependence between pairs of emotion labels and is not optimized as a regularization term during training. Because it is computed from a finite number of MC-dropout passes, the resulting score should be interpreted as an approximate indicator of dropout-induced model disagreement rather than as an exact measure of epistemic uncertainty [15].

3.12. Temperature Scaling and Calibration Metrics

Neural classifiers may produce overconfident probability estimates even when their predictions are incorrect [14]. To improve probability calibration, we apply post hoc temperature scaling using validation data. Let $o_l(x)$ denote the pre-sigmoid logit corresponding to label l from Equation (21). The calibrated probability is computed as

$$p_l^{\text{cal}}(x) = \sigma\left(\frac{o_l(x)}{T_{\text{cal}}} \right), \tag{34}$$

where $T_{\text{cal}} > 0$ is a validation-fitted temperature parameter.

We evaluate calibration using Expected Calibration Error (ECE). Let B_m denote the set of predictions whose confidence falls into bin m, and let M be the number of bins. ECE is computed as

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{N_{\text{cal}}} |\text{acc}(B_m) - \text{conf}(B_m)|, \tag{35}$$

Here, N_{cal} denotes the number of probability–label pairs included in the calibration calculation. For ECE_{flat}, $N_{\text{cal}} = NL$. For $\text{ECE}_{\text{macro}}$, ECE is computed separately for each label using $N_{\text{cal}} = N$ and then averaged across the L labels.

In the experiments, two ECE variants are reported. ECE_{flat} was computed by pooling all label-level predictions before binning, whereas $\text{ECE}_{\text{macro}}$ was computed separately for each label and then averaged across labels. The flattened version reflects global calibration across all label decisions, whereas the macro-averaged version gives equal weight to each emotion label.

We also use the Brier score to quantify probabilistic reliability:

$$\text{Brier} = \frac{1}{NL} \sum_{i=1}^{N} \sum_{l=1}^{L} \left(p_l^{\text{cal}}(x_i) - y_{i,l} \right)^2. \tag{36}$$

ECE evaluates confidence–accuracy alignment, whereas the Brier score measures the mean squared deviation between predicted probabilities and gold labels [14].

3.13. Uncertainty-Based Selective Prediction

We use the uncertainty scores in Equations (31) and (33) for selective prediction. Selective prediction allows a model to abstain from high-uncertainty instances and retain predictions with lower estimated risk [38]. Let $U(x)$ denote either $U_H(x)$ or $U_{MI}(x)$. For a rejection threshold γ, the retained set is defined as

$$\mathcal{R}_\gamma = \{x_i : U(x_i) \leq \gamma\}. \tag{37}$$

Coverage is computed as

$$\text{Coverage}(\gamma) = \frac{|\mathcal{R}_\gamma|}{N}, \tag{38}$$

and retained-set risk is defined as

$$\text{Risk}(\gamma) = \frac{1}{|\mathcal{R}_\gamma|} \sum_{x_i \in \mathcal{R}_\gamma} \ell(\hat{\mathbf{y}}_i, \mathbf{y}_i), \tag{39}$$

where $\ell(\hat{\mathbf{y}}_i, \mathbf{y}_i)$ denotes the instance-level prediction loss. In this study, risk is computed using Jaccard loss:

$$\ell(\hat{\mathbf{y}}_i, \mathbf{y}_i) = 1 - \frac{|\hat{\mathbf{y}}_i \cap \mathbf{y}_i|}{|\hat{\mathbf{y}}_i \cup \mathbf{y}_i|}. \tag{40}$$

The Area Under the Risk–Coverage Curve (AURC) summarizes the trade-off between retained coverage and prediction risk. A lower AURC indicates that the uncertainty score is more effective at ranking high-risk predictions for rejection.

3.14. Evaluation Protocol

We evaluated CONCORD-Emo along four dimensions. First, the predictive performance was measured using the Jaccard score, micro-F_1, macro-F_1, and Hamming loss. Second, probabilistic reliability was evaluated using ECE_{flat}, $\text{ECE}_{\text{macro}}$, the Brier score, and reliability diagrams. Third, uncertainty behavior is analyzed using predictive entropy and MI-based epistemic uncertainty. Fourth, the uncertainty utility is assessed through selective prediction and AURC.

Group-wise analysis is conducted over the *single*, *blended*, and *conflict* groups defined in Equation (5). This analysis tests whether polarity-conflict instances are associated with weaker prediction quality and lower probabilistic reliability than blended-emotion

instances. We tuned thresholds, temperature scaling, and selective-prediction cutoffs using only the validation data; the test set was reserved for the final evaluation.

4. Experimental Setup

In this section, we describe the datasets, baseline models, evaluation metrics, implementation environment, and hyperparameter settings used to evaluate CONCORD-Emo. The experiments were designed to assess both the predictive performance and probabilistic reliability, with particular emphasis on how uncertainty and calibration behave under different emotional structures.

4.1. Datasets

The proposed framework was evaluated using three benchmark datasets for Arabic multi-label emotion detection (MLED).

SemEval-2018 Affect in Tweets (Arabic): (https://huggingface.co/datasets/SemEvalWorkshop/sem_eval_2018_task_1 (accessed on 20 May 2026)) The SemEval-2018 Task 1 dataset provides multilingual tweets annotated for affect analysis [9]. The Arabic subset contains tweets labeled with multiple emotion categories. Each tweet was annotated by multiple crowd workers, and the final labels were obtained through annotation aggregation, introducing inherent variability in emotion assignment.

ExaAEC Arabic Emotion Corpus: (https://github.com/exaco/exaaec (accessed on 20 May 2026) The ExaAEC dataset is a large-scale Arabic tweet corpus annotated using emotion labels inspired by the Plutchik emotion model [10]. Each tweet was annotated by multiple native Arabic speakers, and final labels were determined through agreement-based aggregation, reflecting a collective interpretation of the expressed emotions.

SemEval-2025 Multilingual Emotion Dataset (Algerian Arabic): (https://github.com/emotion-analysis-project/SemEval2025-Task11 (accessed on 20 May 2026)) The SemEval-2025 Task 11 dataset introduces multilingual resources for text-based emotion detection, including Algerian Arabic (Arq) [11]. For Track A, the task is formulated as multi-label emotion detection, where annotators selected all applicable emotion labels for each text. Final labels were derived through aggregation and agreement thresholds, making the dataset suitable for studying uncertainty under low-resource and dialectal conditions.

As shown in Table 2, the three datasets differ substantially in size, label inventory, average label cardinality, and label density. The Train, Dev., and Test columns report the nominal split sizes before deterministic preprocessing. After preprocessing, the effective Train/Dev./Test sizes used in the experiments were 2262/582/1503 for SemEval-2018-Ar, 15,961/1994/2018 for ExaAEC, and 809/88/809 for SemEval-2025-Arq. All reported model results and inferential analyses are based on these effective splits. These differences provide a useful testbed for evaluating whether CONCORD-Emo can maintain reliable predictions across diverse Arabic MLED settings.

Table 2. Statistical characteristics of the three evaluated Arabic MLED datasets.

Dataset	Labels	Train	Dev.	Test	Avg. Labels	Density	Avg. Len.	Med. Len.
SemEval-2018-Ar	11	2278	585	1518	2.267	0.206	16.510	17.000
ExaAEC	10	16,040	2005	2005	1.314	0.131	21.510	17.000
SemEval-2025-Arq	6	901	100	902	1.770	0.295	14.330	13.000

To contextualize the group-wise reliability analysis, Figure 2 summarizes the label-cardinality and emotional-structure profiles of the three datasets. The figure illustrates that the datasets differ not only in scale but also in the proportion of single-label, blended-emotion, and polarity-conflict instances. This structural variation is important because the

proposed reliability analysis is defined over instance-level emotional composition rather than only over individual label frequencies.

4.2. Baseline Models and Comparison Protocol

CONCORD-Emo was compared with controlled MARBERT-based baselines because MARBERT is specifically pretrained on large-scale Arabic social media text and is therefore a strong encoder for Arabic dialectal and informal content [26]. To separate the effect of the proposed conflict-aware architecture from generic uncertainty and calibration techniques, four controlled baselines were evaluated under the same train–development–test splits:

- **MARBERT:** A deterministic MARBERT classifier fine-tuned for multi-label emotion detection.
- **MARBERT + MC dropout:** The same MARBERT classifier evaluated with Monte Carlo (MC) dropout during inference.
- **MARBERT + temperature scaling:** The MARBERT classifier calibrated using validation-fitted post hoc temperature scaling.
- **MARBERT + MC dropout + temperature scaling:** The strongest probabilistic MARBERT baseline, combining stochastic inference and validation-based calibration.

The controlled MARBERT variants serve a specific experimental purpose: they isolate whether the proposed architecture provides predictive or reliability differences beyond Arabic transformer fine-tuning, stochastic MC-dropout inference, and post hoc temperature scaling under identical splits and metric computation. They should not be interpreted as an exhaustive comparison with all contemporary Arabic or multilingual MLED systems.

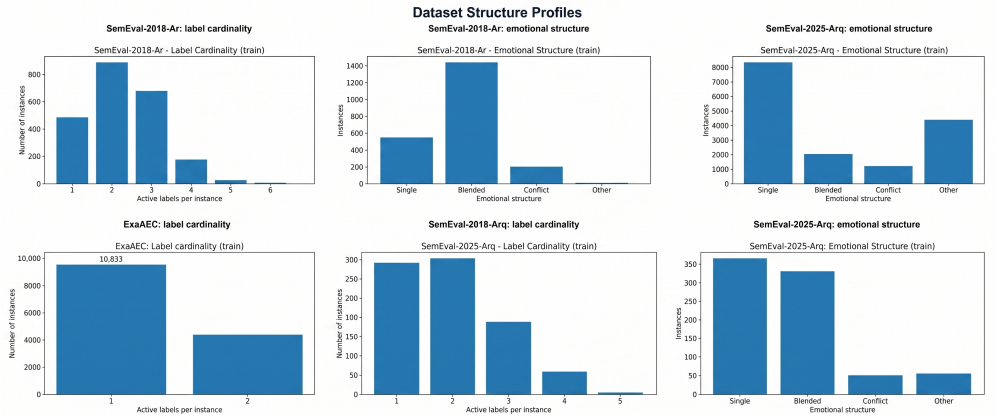


Figure 2. Dataset structure profiles for the three evaluated Arabic emotion datasets, including label-cardinality distributions and polarity-based emotional-structure categories. The emotional-structure grouping follows the formulation in Section 3.2.

To provide broader predictive context, we also report published results from traditional and neural Arabic MLED systems, Arabic-specific transformer and structured models, multilingual transformer hybrids, ensemble and translation-based pipelines, and official SemEval-2025 leaderboard systems in Appendix B. These external results are not treated as controlled baselines because the corresponding studies may differ in split construction, preprocessing, threshold selection, training resources, external data, translation, model scale, augmentation, prompting, parameter-efficient fine-tuning, and ensembling. They also generally report only thresholded classification metrics and do not provide the calibrated probabilities or per-instance stochastic outputs required to reproduce ECE, Brier score, mutual-information uncertainty, or risk–coverage analysis.

4.3. Evaluation Metrics

The models were evaluated using classification- and reliability-oriented criteria.

Classification metrics: Jaccard similarity, micro-F_1, and macro-F_1 were reported. Jaccard similarity measures the overlap between the predicted and gold label sets. Micro-F_1 captures aggregate label-level performance across all decisions, whereas macro-F_1 assigns equal weight to each emotion label and is therefore informative under class imbalance. For SemEval-2025-Arq, macro-F_1 is also important because the official Track A metric is average macro-F_1 [11].

Calibration metrics: Expected Calibration Error (ECE), defined in Equation (35), was used to measure confidence–accuracy alignment. The Brier score, defined in Equation (36), was also reported as a probabilistic reliability metric that measures the mean squared deviation between predicted probabilities and gold labels [42].

Uncertainty metrics: Predictive entropy and mutual information (MI), defined in Equations (31) and (33), were used to quantify total predictive uncertainty and approximate dropout-induced epistemic disagreement, respectively. The MI score is computed independently for each binary emotion output and then averaged across labels; it is not a measure of dependence among emotion labels.

Selective-prediction metrics: Uncertainty utility was evaluated using risk–coverage analysis [38]. The Area Under the Risk–Coverage Curve (AURC) summarizes the trade-off between retained-set coverage and prediction risk.

Conflict-conditioned analysis: To test the hypothesis formalized in Equation (6), predictive performance, calibration, Brier score, predictive entropy, and MI were analyzed across the single-emotion, blended-emotion, and polarity-conflict groups defined in Equation (5).

4.4. Statistical Inference

For each controlled model comparison, we quantified test-set sampling uncertainty using 10,000 paired non-parametric bootstrap resamples with replacement and a fixed bootstrap seed of 42. The saved per-instance outputs of CONCORD-Emo and each MARBERT-based baseline were aligned using unique test-instance identifiers, and the same sampled indices were applied to both systems in every bootstrap replicate. Jaccard score, micro-averaged F1, macro-averaged F1, flattened Expected Calibration Error, macro-averaged Expected Calibration Error, Brier score, and predictive-entropy-based area under the risk–coverage curve were recomputed from the complete resampled predictions and probabilities.

For each metric, the reported difference is CONCORD-Emo minus the controlled baseline. Positive differences favor CONCORD-Emo for Jaccard and F1 metrics, whereas negative differences favor CONCORD-Emo for Expected Calibration Error, Brier score, and area under the risk–coverage curve. We report two-sided 95-percent percentile bootstrap confidence intervals. A difference is described as interval-supported only when its confidence interval excludes zero.

Mutual-information-based area under the risk–coverage curve comparisons were evaluated only when both systems produced non-degenerate stochastic mutual-information values. They were treated as not applicable for deterministic baselines whose mutual-information scores were identically zero.

For the conflict-versus-blended analysis, the two disjoint groups were resampled independently using 10,000 bootstrap replicates. Two-sided Mann–Whitney U tests and rank-biserial effect sizes were also computed as complementary distributional summaries. The bootstrap intervals are used as the primary inferential evidence.

These analyses quantify uncertainty due to resampling the evaluated test instances while holding the trained checkpoint, thresholds, temperature, and model configuration fixed. They do not quantify variability across independent training seeds.

4.5. Implementation Details

All experiments were implemented in Python using PyTorch 2.5.1 [43] and HuggingFace Transformers library [44]. MARBERT was used as the main Arabic transformer encoder in CONCORD-Emo and in the controlled MARBERT-based baselines because it was pretrained on large-scale Arabic social media text and is therefore suitable for informal and dialectal Arabic inputs [26].

All datasets were processed using the Arabic text normalization pipeline. The preprocessing included removing uniform resource locators (URLs), user mentions, repeated noisy symbols, and non-textual artifacts; preserving hashtag content where applicable; removing diacritics and Tatweel; normalizing common Arabic orthographic variants; and standardizing the whitespace. The cleaned texts were then tokenized using the corresponding transformer tokenizer and were padded or truncated to a maximum length of 128 tokens. Instances without any active emotion labels were excluded from the model training and evaluation to preserve the multi-label emotion detection setting.

The polarity conflict target was not manually annotated. Instead, it was automatically derived from the gold emotion labels using the polarity-count formulation in Equations (3) and (4). This ensured that the conflict-aware supervision and group-wise reliability analysis used the same deterministic polarity definition across all datasets.

All models were trained using the AdamW optimizer with linear warmup and learning-rate scheduling. Early stopping was applied using the development set to reduce overfitting. The best checkpoint was selected according to the development set performance and then evaluated once on the held-out test set. Decision thresholds were tuned only on development predictions using a grid search from 0.05 to 0.95, with a step size of 0.05. No threshold, calibration parameter, or model hyperparameter was tuned on the test set.

During inference, Monte Carlo (MC) dropout was applied using 10 stochastic forward passes for the uncertainty-aware configurations. Temperature scaling was fitted using development logits only and then applied to the test predictions. Selective prediction was evaluated after final inference by ranking instances according to predictive entropy and mutual information. All experiments were conducted with a fixed random seed of 42 on a Google Colab (https://colab.research.google.com/) paid graphics processing unit (GPU) environment using an NVIDIA A100-class accelerator.

4.6. Hyperparameter Configuration

The main hyperparameters used in the experiments are listed in Table 3. We applied the same methodological protocol across all datasets, with limited dataset-specific settings used only to account for differences in the dataset size and training stability. In particular, ExaAEC used a larger batch size because of its larger training set, whereas SemEval-2025-Arq used a smaller conflict loss weight because of its smaller development set and higher sensitivity to conflict supervision.

Table 3. Main hyperparameter configuration used for CONCORD-Emo.

Hyperparameter	Selected Value
Transformer encoder	MARBERT
Maximum sequence length	128 tokens
Early stopping patience	4 epochs
Optimizer	AdamW
Warmup ratio	0.10
Weight decay	0.02
Gradient clipping	1.0

Table 3. *Cont.*

Hyperparameter	Selected Value
Dropout rate	0.10
Number of MoE experts	4
Expert hidden dimension	256
Batch size	8 for SemEval-2018-Ar and SemEval-2025-Arq; 16 for ExaAEC
Learning rate	1×10^{-5} for SemEval-2018-Ar and SemEval-2025-Arq; 2×10^{-5} for ExaAEC
Conflict-loss weight λ_{conf}	0.05 for SemEval-2018-Ar and ExaAEC; 0.03 for SemEval-2025-Arq
MC dropout passes S	10

5. Results

We evaluate CONCORD-Emo from six complementary perspectives. First, we compare predictive performance against controlled MARBERT-based baselines under the same data splits and evaluation protocol. Second, we examine calibration and probabilistic reliability using Expected Calibration Error (ECE), Brier score, reliability diagrams, and risk–coverage behavior. Third, we analyze whether polarity-conflict instances exhibit lower predictive reliability than blended or single-emotion instances. Fourth, we evaluate the contribution of the main CONCORD-Emo components through ablation analysis. Fifth, we test whether uncertainty estimates can support selective prediction by identifying lower-risk retained predictions. Finally, we position the results relative to accuracy-oriented Arabic MLED studies, while reporting detailed external comparisons in Appendix B.

5.1. Predictive Performance Under Controlled Baselines

Table 4 reports the predictive performance of CONCORD-Emo against controlled MARBERT-based baselines across the three evaluated datasets. We use this comparison as the primary predictive-performance evaluation because all systems were trained and evaluated under the same data splits, preprocessing pipeline, validation-only threshold tuning, calibration protocol, and metric computation. Dataset-specific optimization settings are applied consistently within each dataset, as summarized in Table 3.

The results show that CONCORD-Emo remains competitive with strong MARBERT-based baselines, although its predictive advantage is dependent on the dataset. On SemEval-2025-Arq, CONCORD-Emo achieved the strongest overall predictive profile, with a Jaccard score of 0.471, micro-F_1 of 0.606, and macro-F_1 of 0.582. These values constitute the strongest combined point estimate profile among the controlled systems on SemEval-2025-Arq; however, the paired confidence intervals relative to the strongest per-metric comparator include zero.

On SemEval-2018-Ar, CONCORD-Emo performed close to the strongest MARBERT-based baseline. The differences between CONCORD-Emo and MARBERT were small for Jaccard and micro-F_1, and CONCORD-Emo remained nearly tied with the strongest macro-F_1 result. This indicates that adding reliability-oriented components does not substantially compromise the predictive performance of this benchmark.

Table 4. Predictive performance of CONCORD-Emo against controlled MARBERT-based baselines. Higher values indicate better performance. Bold indicates the best result within each dataset and metric.

Dataset	Model	Jaccard	Micro-F_1	Macro-F_1
ExaAEC	MARBERT	**0.619**	0.666	0.638
	MARBERT+MC	0.615	0.665	0.641
	MARBERT+Temp	0.619	0.667	0.638
	MARBERT+MC+Temp	0.617	**0.667**	**0.644**
	CONCORD-Emo	0.616	0.663	0.627
SemEval-2018-Ar	MARBERT	**0.582**	**0.696**	0.605
	MARBERT+MC	0.573	0.689	0.604
	MARBERT+Temp	0.577	0.691	**0.607**
	MARBERT+MC+Temp	0.571	0.688	0.604
	CONCORD-Emo	0.575	0.695	0.606
SemEval-2025-Arq	MARBERT	0.471	0.597	0.574
	MARBERT+MC	0.464	0.597	0.563
	MARBERT+Temp	0.471	0.600	0.576
	MARBERT+MC+Temp	0.456	0.574	0.549
	CONCORD-Emo	**0.471**	**0.606**	**0.582**

Note: MC = Monte Carlo dropout; Temp = temperature scaling.

On ExaAEC, CONCORD-Emo does not exceed the strongest MARBERT-based baselines in raw predictive metrics. MARBERT obtains the highest Jaccard score, whereas MARBERT+MC+Temp obtains the strongest micro-F_1 and macro-F_1. We therefore interpret the ExaAEC predictive result as competitive rather than superior. The main contribution of CONCORD-Emo on this dataset is better reflected in calibration and probabilistic reliability, which we examine in the next subsection.

5.2. Calibration and Probabilistic Reliability

Table 5 presents the calibration and probabilistic reliability results for CONCORD-Emo and the controlled MARBERT-based baselines. Unlike the predictive metrics in Table 4, these metrics evaluate the quality of predicted probabilities rather than the correctness of thresholded label decisions. Lower values of ECE, Brier score, and Area Under the Risk–Coverage Curve (AURC) indicate stronger probabilistic reliability.

The point estimates in Table 5 show a clearer calibration and probabilistic quality advantage for CONCORD-Emo than for the thresholded predictive metrics. On ExaAEC, CONCORD-Emo obtained the lowest flattened and macro-averaged ECE values, both 0.074, and the lowest Brier score of 0.074. It also obtained the lowest predictive-entropy-based AURC of 0.220. Among the systems using MC dropout, CONCORD-Emo obtained the lowest MI-based AURC point estimate of 0.230.

A similar calibration pattern appears in SemEval-2018-Ar. CONCORD-Emo obtains the lowest flattened ECE of 0.039, the lowest macro-averaged ECE of 0.054, and the lowest Brier score of 0.091. The AURC point estimates are more mixed: deterministic MARBERT obtains the lowest predictive-entropy-based AURC, whereas CONCORD-Emo obtains the lowest MI-based AURC among the systems using MC dropout. The paired confidence intervals for the AURC differences include zero; therefore, no interval-supported ranking advantage is claimed on this dataset.

Table 5. Calibration and probabilistic reliability of CONCORD-Emo against controlled MARBERT-based baselines. ECE_{flat} denotes flattened Expected Calibration Error, ECE_{macro} denotes macro-averaged Expected Calibration Error, Brier denotes the Brier score, $AURC_{U_H}$ denotes the Area Under the Risk–Coverage Curve when instances are ranked by predictive entropy, and $AURC_{U_{MI}}$ denotes the corresponding quantity under mutual-information-based epistemic-uncertainty ranking. Lower values are better. A dash indicates that $AURC_{U_{MI}}$ is not applicable because the corresponding deterministic model does not produce stochastic mutual-information estimates.

Dataset	Model	ECE flat	ECE macro	Brier	AURC U_H	AURC U_{MI}
ExaAEC	MARBERT	0.127	0.127	0.094	0.239	–
	MARBERT+MC	0.136	0.136	0.093	0.238	0.241
	MARBERT+Temp	0.126	0.126	0.094	0.239	–
	MARBERT+MC+Temp	0.134	0.134	0.093	0.237	0.241
	CONCORD-Emo	**0.074**	**0.074**	**0.074**	**0.220**	**0.230**
SemEval-2018-Ar	MARBERT	0.053	0.065	0.098	**0.293**	–
	MARBERT+MC	0.060	0.066	0.095	0.298	0.305
	MARBERT+Temp	0.062	0.068	0.097	0.295	–
	MARBERT+MC+Temp	0.070	0.074	0.095	0.301	0.308
	CONCORD-Emo	**0.039**	**0.054**	**0.091**	0.303	**0.304**
SemEval-2025-Arq	MARBERT	0.102	0.112	0.189	**0.459**	–
	MARBERT+MC	0.070	0.085	0.183	0.471	0.489
	MARBERT+Temp	**0.063**	0.083	0.180	0.464	–
	MARBERT+MC+Temp	0.066	**0.076**	0.179	0.472	0.489
	CONCORD-Emo	0.074	0.087	**0.177**	0.471	**0.484**

Note: MC = Monte Carlo dropout; Temp = temperature scaling. Bold values indicate the best result in each column.

On SemEval-2025-Arq, MARBERT+Temp obtains the lowest flattened ECE, and MARBERT+MC+Temp obtains the lowest macro-averaged ECE. CONCORD-Emo obtains the lowest Brier-score point estimate and the lowest MI-based AURC point estimate among the systems using MC dropout. However, the paired bootstrap analysis shows that the strongest Brier-score difference includes zero and that the temperature-scaled baselines retain interval-supported ECE advantages. The reliability profile on this dataset is therefore mixed rather than uniformly favorable to CONCORD-Emo.

Figure 3 presents reliability diagrams for the three datasets. These diagrams complement Table 5 by visualizing the relationship between the predicted confidence and empirical accuracy after calibration.

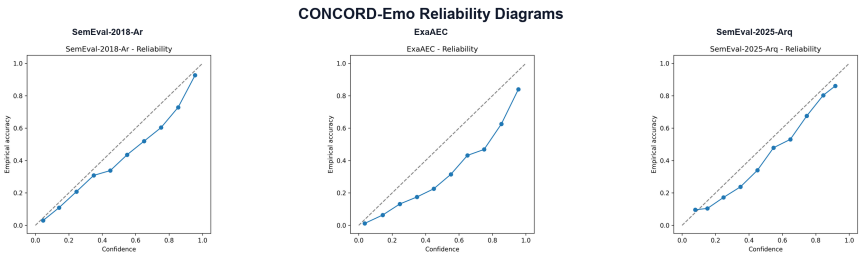


Figure 3. Reliability diagrams for CONCORD-Emo across three evaluated datasets. The diagonal reference line represents perfect calibration, where predicted confidence matches empirical accuracy.

5.3. Paired Bootstrap Analysis of Controlled Comparisons

Tables 4 and 5 report point estimates, whereas Table 6 summarizes paired bootstrap differences against the strongest controlled point-estimate baseline for each dataset and metric. Full comparisons against all four controlled baselines are reported in Appendix C.

Table 6. Paired bootstrap comparison of CONCORD-Emo with the strongest controlled point-estimate baseline for each dataset and metric. The reported difference is CONCORD-Emo minus the baseline, followed by its 95% percentile bootstrap confidence interval based on 10,000 paired test-instance resamples.

Dataset	Metric	Comparator	Difference [95% CI]	Interval Interpretation
Predictive metrics				
SemEval-2018-Ar	Jaccard	MARBERT	−0.0070 [−0.0181, 0.0041]	CI includes zero
SemEval-2018-Ar	Micro-F_1	MARBERT	−0.0012 [−0.0102, 0.0078]	CI includes zero
SemEval-2018-Ar	Macro-F_1	MARBERT+Temp	−0.0012 [−0.0155, 0.0129]	CI includes zero
ExaAEC	Jaccard	MARBERT	−0.0033 [−0.0164, 0.0100]	CI includes zero
ExaAEC	Micro-F_1	MARBERT+MC+Temp	−0.0040 [−0.0162, 0.0082]	CI includes zero
ExaAEC	Macro-F_1	MARBERT+MC+Temp	−0.0170 [−0.0332, −0.0012]	Baseline favored
SemEval-2025-Arq	Jaccard	MARBERT+Temp	0.0002 [−0.0185, 0.0195]	CI includes zero
SemEval-2025-Arq	Micro-F_1	MARBERT+Temp	0.0058 [−0.0098, 0.0215]	CI includes zero
SemEval-2025-Arq	Macro-F_1	MARBERT+Temp	0.0054 [−0.0115, 0.0223]	CI includes zero
Calibration and probabilistic metrics				
SemEval-2018-Ar	ECE_{flat}	MARBERT	−0.0140 [−0.0180, −0.0119]	CONCORD-Emo favored
SemEval-2018-Ar	ECE_{macro}	MARBERT	−0.0108 [−0.0138, −0.0056]	CONCORD-Emo favored
SemEval-2018-Ar	Brier score	MARBERT+MC	−0.0040 [−0.0057, −0.0022]	CONCORD-Emo favored
ExaAEC	ECE_{flat}	MARBERT+Temp	−0.0515 [−0.0529, −0.0501]	CONCORD-Emo favored
ExaAEC	ECE_{macro}	MARBERT+Temp	−0.0514 [−0.0525, −0.0494]	CONCORD-Emo favored
ExaAEC	Brier score	MARBERT+MC+Temp	−0.0189 [−0.0207, −0.0172]	CONCORD-Emo favored
SemEval-2025-Arq	ECE_{flat}	MARBERT+Temp	0.0113 [0.0032, 0.0161]	Baseline favored
SemEval-2025-Arq	ECE_{macro}	MARBERT+MC+Temp	0.0108 [0.0011, 0.0187]	Baseline favored
SemEval-2025-Arq	Brier score	MARBERT+MC+Temp	−0.0019 [−0.0053, 0.0014]	CI includes zero
Selective reliability				
SemEval-2018-Ar	$AURC_{U_H}$	MARBERT	0.0095 [−0.0020, 0.0211]	CI includes zero
ExaAEC	$AURC_{U_H}$	MARBERT+MC+Temp	−0.0171 [−0.0300, −0.0048]	CONCORD-Emo favored
SemEval-2025-Arq	$AURC_{U_H}$	MARBERT	0.0115 [−0.0199, 0.0431]	CI includes zero

Note: CI = confidence interval; ECE = Expected Calibration Error; AURC = Area Under the Risk–Coverage Curve. Positive differences favor CONCORD-Emo for Jaccard and F1 metrics, whereas negative differences favor CONCORD-Emo for ECE, Brier score, and AURC. The strongest comparator is selected descriptively from the controlled point estimates. Full pairwise intervals against all controlled baselines are presented in Appendix C.

On SemEval-2018-Ar, all paired predictive-performance intervals include zero. In contrast, the ECE and Brier-score intervals exclude zero in favor of CONCORD-Emo against every controlled baseline. The predictive-entropy-based AURC intervals include zero.

On ExaAEC, the Jaccard and micro-averaged F1 intervals included zero across controlled comparisons. The macro-averaged F1 comparison with MARBERT+MC+Temp favored the baseline. Conversely, every ECE, Brier score, and predictive-entropy-based AURC interval excluded zero in favor of CONCORD-Emo.

On SemEval-2025-Arq, CONCORD-Emo has interval-supported gains in macro-averaged F1 relative to MARBERT+MC and in micro- and macro-averaged F1 relative to MARBERT+MC+Temp. Nevertheless, its predictive intervals relative to the strongest point-estimate comparator include zero. Calibration remains mixed: CONCORD-Emo improves ECE and Brier score relative to uncalibrated MARBERT, while the temperature-scaled baselines retain lower ECE values. The Brier-score difference from the strongest controlled baseline includes zero.

Valid AURC comparisons based on mutual-information ranking are restricted to systems using MC dropout, and all corresponding confidence intervals include zero.

5.4. Conflict-Conditioned Reliability Analysis

Table 7 reports the predictive and reliability behavior of CONCORD-Emo across the polarity-based emotional structures defined in Equation (5). We separated *single*, *blended*, and *conflict* instances to test whether polarity-opposed emotional composition was associated with weaker predictive reliability. Only polarity-eligible instances were included in the group-conditioned summaries. Instances containing exclusively excluded labels

were omitted; therefore, the group counts did not necessarily sum to the complete test set size. The primary inferential contrast was between conflict and blended.

Table 7. Conflict-conditioned predictive and reliability analysis of CONCORD-Emo. Here, n denotes the number of eligible test instances assigned to each emotional-structure group; instances with no active polarity-bearing gold label are excluded from the group-conditioned summaries. $\mathrm{ECE}_{\mathrm{macro}}$ denotes macro-averaged Expected Calibration Error, Brier denotes the Brier score, and U_H denotes predictive entropy. Lower values are better for $\mathrm{ECE}_{\mathrm{macro}}$ and Brier score, whereas higher U_H values indicate greater total predictive uncertainty.

Dataset	Group	n	Jaccard	Micro-F_1	Macro-F_1	$\mathrm{ECE}_{\mathrm{macro}}$	Brier	U_H
SemEval-2025-Arq	Single	362	0.446	0.541	0.482	0.171	0.166	0.797
	Blended	334	0.529	0.674	0.561	0.153	0.180	0.799
	Conflict	48	0.352	0.506	0.466	0.266	0.265	0.807
SemEval-2018-Ar	Single	357	0.461	0.545	0.447	0.135	0.083	0.473
	Blended	978	0.644	0.754	0.639	0.056	0.084	0.441
	Conflict	160	0.427	0.566	0.484	0.107	0.150	0.532
ExaAEC	Single	1064	0.593	0.634	0.504	0.089	0.078	0.408
	Blended	248	0.694	0.787	0.511	0.095	0.070	0.425
	Conflict	153	0.516	0.634	0.423	0.107	0.102	0.457

The results show a consistent degradation in predictive performance for polarity-conflict instances compared to blended-emotion instances. On SemEval-2025-Arq, the conflict group obtained a Jaccard score of 0.352 and micro-F_1 of 0.506, compared with 0.529 and 0.674 for the blended group. On SemEval-2018-Ar, the same pattern appears: Jaccard decreases from 0.644 for blended instances to 0.427 for conflict instances, and micro-F_1 decreases from 0.754 to 0.566. On ExaAEC, conflict instances again performed worse than blended instances, with Jaccard decreasing from 0.694 to 0.516 and micro-F_1 decreasing from 0.787 to 0.634.

The reliability metrics follow the same general pattern. Conflict instances obtain higher Brier scores than blended instances across all three datasets: 0.265 versus 0.180 on SemEval-2025-Arq, 0.150 versus 0.084 on SemEval-2018-Ar, and 0.102 versus 0.070 on ExaAEC. This shows that polarity-conflict cases are not only harder to classify but also less reliable probabilistically.

Predictive entropy is numerically higher for conflict instances across all three datasets, but the confidence interval analysis qualifies this pattern. The conflict-minus-blended difference is 0.0912 with a 95 percent confidence interval of [0.0759, 0.1064] on SemEval-2018-Ar and 0.0316 [0.0111, 0.0520] on ExaAEC. On SemEval-2025-Arq, the difference is only 0.0072 [−0.0136, 0.0280], so the interval includes zero. Predictive entropy therefore separates conflict from blended instances on SemEval-2018-Ar and ExaAEC but not on the smaller Arq test set. Figure 4 visualizes the predictive-entropy distribution across the Single, Blended, and Conflict groups.

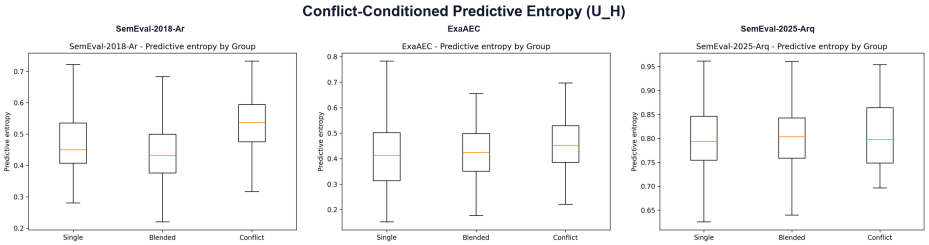


Figure 4. Predictive entropy (U_H) across single, blended, and conflict groups for the three evaluated datasets. Here, U_H denotes total predictive uncertainty estimated from Monte Carlo dropout predictions. Higher values indicate greater total predictive uncertainty.

The bootstrap analysis provides consistent support for the prediction-error and probabilistic-quality findings. Conflict instances have higher Brier contribution than blended instances on SemEval-2018-Ar, ExaAEC, and SemEval-2025-Arq, with differences of 0.0652 [0.0529, 0.0775], 0.0319 [0.0180, 0.0457], and 0.0853 [0.0557, 0.1162], respectively. Jaccard loss is also higher under conflict on all three datasets, with differences of 0.2166 [0.1752, 0.2569], 0.1775 [0.1113, 0.2429], and 0.1772 [0.1060, 0.2463], respectively.

The uncertainty findings are less uniform. MI-based uncertainty is higher for conflict instances on SemEval-2018-Ar, is not separated from blended instances on ExaAEC, and is lower for conflict instances on SemEval-2025-Arq. These results support polarity conflict as a robust indicator of higher prediction error and poorer probabilistic quality, while showing that neither predictive entropy nor MI provides a dataset-invariant conflict signal. Full confidence intervals, Mann–Whitney tests, and rank-biserial effect sizes are reported in Appendix C.

5.5. Ablation Study

Table 8 reports the ablation results for the main components of CONCORD-Emo. Each ablation removes or disables one targeted component while preserving the remaining architecture and evaluation protocol. Unlike the controlled baseline comparison, this table reports the actual metric values for the full proposed model and each ablated variant. We interpret the results in terms of predictive–reliability trade-offs because several components are designed to support calibration, uncertainty estimation, and conflict-conditioned reliability rather than to maximize thresholded classification scores alone.

Table 8. Ablation results for CONCORD-Emo. The table reports actual metric values for the full proposed model and each ablated variant. Jaccard, micro-F_1, and macro-F_1 are higher-is-better metrics, whereas ECE_{flat} and Brier are lower-is-better metrics. ECE_{flat} denotes flattened Expected Calibration Error, and Brier denotes the Brier score. Bold indicates the best result within each dataset and metric.

Dataset	Variant	Jaccard	Micro-F_1	Macro-F_1	ECE_{flat}	Brier
SemEval-2025-Arq	w/o Global Anchor	0.469	0.603	**0.593**	0.080	0.181
	w/o MoE LayerNorm	0.455	0.591	0.569	0.099	0.187
	w/o Conflict-Aware Gate	0.456	0.599	0.583	0.082	0.183
	w/o Conflict Loss	0.467	0.596	0.577	**0.071**	**0.176**
	w/o Temperature Scaling	0.471	0.601	0.581	0.135	0.195
	w/o MC Dropout	0.461	0.599	0.581	0.073	0.179
	CONCORD-Emo (Proposed)	**0.471**	**0.606**	0.582	0.074	0.177
SemEval-2018-Ar	w/o Global Anchor	0.558	0.682	0.592	0.048	0.092
	w/o MoE LayerNorm	0.573	0.689	0.600	0.052	0.093
	w/o Conflict-Aware Gate	0.565	0.686	**0.609**	0.063	0.093
	w/o Conflict Loss	0.561	0.683	0.600	0.052	0.093
	w/o Temperature Scaling	0.564	0.684	0.599	0.071	0.098
	w/o MC Dropout	0.567	0.688	0.600	**0.036**	0.093
	CONCORD-Emo (Proposed)	**0.575**	**0.695**	0.606	0.039	**0.091**
ExaAEC	w/o Global Anchor	0.620	**0.669**	**0.641**	0.132	0.094
	w/o MoE LayerNorm	0.612	0.660	0.627	0.072	0.074
	w/o Conflict-Aware Gate	0.618	0.664	0.627	**0.061**	**0.071**
	w/o Conflict Loss	0.611	0.663	0.639	0.130	0.094
	w/o Temperature Scaling	0.615	0.665	0.629	0.062	0.078
	w/o MC Dropout	**0.625**	0.667	0.637	0.070	0.076
	CONCORD-Emo (Proposed)	0.616	0.663	0.627	0.074	0.074

Note: w/o = without; MoE = mixture of experts; MC = Monte Carlo.

The ablation results show that CONCORD-Emo does not dominate every ablated variant on every individual metric. Instead, the component effects vary across datasets and reveal a predictive–reliability trade-off. This pattern is expected because reliability-aware components may improve probability quality, uncertainty estimation, or conflict-sensitive behavior without necessarily improving every thresholded classification metric.

On SemEval-2025-Arq, removing the global anchor increased the macro-F_1 from 0.582 to 0.593; however, it did not improve the overall predictive profile: Jaccard decreased from 0.471 to 0.469, and micro-F_1 decreased from 0.606 to 0.603. Therefore, we do not interpret the higher macro-F_1 as evidence against the global anchor. Rather, it suggests that local label-wise attention can sometimes favor minority-label recall, whereas the residual anchor preserves stronger label-set consistency and micro-level performance.

The effect of MC dropout is also metric-dependent. On ExaAEC, disabling MC dropout increased Jaccard from 0.616 to 0.625 and macro-F_1 from 0.627 to 0.637. This suggests that deterministic inference can produce sharper thresholded decisions on the larger ExaAEC dataset. However, the same variant produced a weaker Brier score than CONCORD-Emo and removed the stochastic forward-pass mechanism required for mutual-information-based epistemic uncertainty estimation. We therefore retained MC dropout because it supports the uncertainty and selective-prediction analyses that are central to the reliability-aware formulation, not because it consistently improves raw accuracy.

On SemEval-2018-Ar, CONCORD-Emo obtained the best Jaccard, micro-F_1, and Brier scores among the ablation variants. Although the variant without the conflict-aware gate yields a slightly higher macro-F_1, and the variant without MC dropout yields a slightly lower ECE_{flat}, the full model provides the strongest balance between label-set accuracy and probabilistic quality.

The ExaAEC ablations show different effects for the two conflict-related components. Removing the auxiliary conflict loss substantially worsened the ECE and Brier scores, whereas removing the conflict-aware gate produced slightly lower ECE and Brier values. Therefore, the calibration contribution cannot be uniformly attributed to the gate.

Overall, the ablation study supports CONCORD-Emo as a reliable and consistent architecture. Its components should not be interpreted as independent modules that improve all metrics. Instead, they jointly support conflict-aware representation, calibrated probability estimation, uncertainty quantification, and selective prediction.

5.6. Selective Prediction

Table 9 reports the retained set performance of CONCORD-Emo at different coverage levels after uncertainty-based filtering.

Table 9. Selective-prediction results for CONCORD-Emo using predictive entropy (U_H) as the uncertainty ranking criterion. Coverage denotes the retained proportion of test instances after rejecting high-uncertainty samples. Jaccard, micro-F_1, and macro-F_1 are computed only on the retained subset, and higher values indicate better retained-set performance.

Dataset	Coverage	Retained n	Jaccard	Micro-F_1	Macro-F_1
SemEval-2025-Arq	1.00	809	0.471	0.606	0.582
	0.80	648	0.497	0.629	0.605
	0.60	486	0.510	0.642	0.615
	0.50	405	0.519	0.650	0.620
SemEval-2018-Ar	1.00	1503	0.575	0.695	0.606
	0.80	1203	0.623	0.740	0.609
	0.60	902	0.660	0.772	0.608
	0.50	752	0.677	0.786	0.605
ExaAEC	1.00	2018	0.616	0.663	0.627
	0.80	1615	0.677	0.718	0.681
	0.60	1211	0.734	0.768	0.714
	0.50	1009	0.766	0.795	0.710

The results show that selective prediction improves the retained-set performance across the evaluated datasets. When high-uncertainty instances are rejected, the retained predictions generally achieve higher Jaccard and micro-F_1 scores than the full test set. This

indicates that the uncertainty estimates are not merely descriptive; they provide a useful ranking signal for separating lower-risk from higher-risk predictions.

This improvement is particularly important for the reliability-centered aim of CONCORD-Emo. In practical Arabic emotion analysis, a model that can identify uncertain or high-risk cases is more useful than one that always outputs predictions without indicating reliability. Therefore, selective prediction complements the calibration and Brier score results by showing how uncertainty estimates can support abstention from less reliable predictions.

At the same time, selective prediction should not be interpreted as eliminating model error. It improves the retained-set quality by reducing coverage, which creates a trade-off between prediction availability and reliability. For this reason, we interpret selective prediction as an operational reliability tool rather than as a replacement for predictive-performance evaluation.

Figure 5 presents the corresponding risk–coverage curves. The curves summarize the trade-off between retained-set coverage and prediction risk, complementing the discrete coverage levels reported in Table 9.

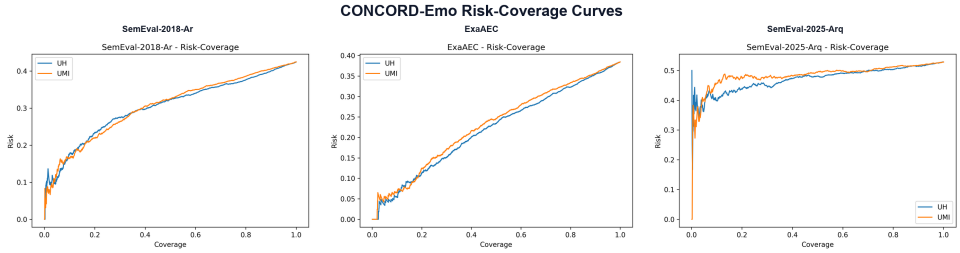


Figure 5. Risk–coverage curves for CONCORD-Emo across the three evaluated datasets. Coverage denotes the proportion of retained test instances, while risk denotes the prediction error on the retained subset. Lower curves indicate stronger selective-prediction behavior.

Overall, the selective-prediction results demonstrate that predictive entropy provides a useful ranking signal for identifying lower-risk predictions. This supports the reliability-oriented formulation of CONCORD-Emo by showing that uncertainty estimates are not only descriptive but also useful for controlling the retained-set prediction quality.

5.7. Positioning Relative to Published Arabic MLED Systems

We used two complementary comparison levels. Tables 4–9 provide the primary controlled evidence because all compared systems use the same data splits, preprocessing, validation-only threshold tuning, and metric computation. Appendix B provides broader contextual comparisons with published systems, but those results are interpreted only using the metrics reported by their original studies.

For SemEval-2018-Ar, the external comparison spanned random and support vector machine baselines, shared-task systems, recurrent and hybrid neural architectures, ensemble-oriented methods, and the structurally informed GTHL-Emo framework. CONCORD-Emo remains within the range of strong published systems but does not obtain the highest reported values for every predictive metric.

For ExaAEC, published comparisons are more limited because fewer directly related systems report results on this corpus. CONCORD-Emo exceeded the original BiLSTM with the Embeddings from Language Models (ELMo) benchmark in the micro-averaged F1 score, whereas GTHL-Emo reported stronger Jaccard, micro-averaged F1, and macro-averaged F1 results. We interpret this comparison as contextual predictive evidence rather than a controlled architectural comparison.

SemEval-2025-Arq provides the broadest and most recent comparison. The systems summarized in Appendix B include the dialect-specific DziriBERT ensemble submitted by HTU, the XLM-RoBERTa and BiLSTM-attention hybrid submitted by INFOTEC-NLP, the mDeBERTa-v3 system submitted by LATE-GIL-NLP, the translation-based DeBERTa approach submitted by YNU-HPCC and GTHL-Emo, and the official JNLP and Pai leaderboard results [11,21–24,32]. CONCORD-Emo exceeded several of these reported systems in the macro-averaged F1 score but remained below the strongest official leaderboard entries.

These external comparisons cannot establish superiority under a common experimental setting. Differences in training data, preprocessing, augmentation, translation, prompting, model scale, ensembling, thresholding, and test protocol may contribute to the reported scores. Moreover, the external studies do not provide directly comparable Expected Calibration Error, Brier score, predictive entropy, mutual-information uncertainty, or selective-prediction results. Therefore, we used them to establish predictive context while restricting claims about calibration and uncertainty effectiveness to the controlled MARBERT-based experiments.

The qualitative examples in Appendix D further illustrate representative reliability behaviors, including correct low-uncertainty predictions, high-uncertainty errors, polarity conflict errors, and overconfident misclassifications. We used these examples as illustrative evidence only; the main reliability conclusions were drawn from the aggregate controlled results.

Additional class-wise predictive behavior, contextual external comparisons, paired bootstrap confidence-interval plots, and qualitative reliability examples are reported in Appendices A–D.

6. Discussion

6.1. Predictive Competitiveness and Reliability-Centered Contribution

We position CONCORD-Emo as a reliability-aware framework for Arabic multi-label emotion detection rather than as a uniformly superior predictive classifier. This distinction is central to the interpretation of the results. The Jaccard score, micro-F_1, and macro-F_1 establish whether the model remains predictively adequate, whereas the expected calibration error (ECE), Brier score, predictive entropy, mutual-information-based epistemic uncertainty, and risk–coverage behavior evaluate whether the model's probability estimates remain reliable under difficult emotional conditions.

The paired bootstrap analysis characterizes CONCORD-Emo as predictively competitive rather than uniformly superior. On SemEval-2018-Ar, all controlled predictive-performance intervals include zero. On ExaAEC, most predictive intervals also include zero, although MARBERT+MC+Temp retains an interval-supported macro-F1 advantage. On SemEval-2025-Arq, CONCORD-Emo has interval-supported micro-F1 and macro-F1 gains over selected MC-based baselines, but its differences from the strongest per-metric controlled comparator include zero. Predictive adequacy is therefore established without claiming general predictive dominance.

This interpretation is important for Arabic social-media emotion detection. Social-media posts often contain implicit, mixed, dialectal, or contrastive emotional cues. In such cases, high classification scores alone are insufficient; a model may assign plausible labels while still producing poorly calibrated probabilities or failing to identify unreliable instances. Our results support a broader evaluation perspective in which predictive performance is considered jointly with probability quality, uncertainty behavior, and selective-prediction utility.

6.2. Polarity Conflict as a Reliability Condition

A central finding of this study is that polarity-conflict instances are consistently less reliable than blended-emotion instances. Across the evaluated datasets, the conflict group shows weaker predictive performance and higher Brier scores than the blended group. This pattern indicates that the prediction difficulty in Arabic MLED is not determined only by the number of active labels but also by the affective relationship among those labels. Same-polarity emotional blending and polarity-opposed emotional composition represent different reliability conditions.

This finding supports the main hypothesis of the study. Multi-label emotion detection is often treated as a problem of assigning all applicable labels; however, this view does not distinguish compatible co-occurrence from polarity opposition. CONCORD-Emo makes this distinction explicit by separating *single*, *blended*, and *conflict* instances. The results show that polarity conflict is not merely another form of label co-occurrence; it is associated with lower predictive reliability and weaker probabilistic quality.

The consistency of this direction across the three datasets with different label inventories provides empirical support for the usefulness of the operational partition as a reliability analysis device. It does not establish a universal or culture-independent polarity taxonomy for Arabic emotions.

Interval analysis shows that uncertainty behavior is dataset-dependent. Predictive entropy separates conflict from blended instances on SemEval-2018-Ar and ExaAEC, but its interval includes zero on SemEval-2025-Arq. MI-based uncertainty is higher under conflict on SemEval-2018-Ar, is not separated on ExaAEC, and is lower under conflict on SemEval-2025-Arq. In contrast, higher Brier contribution and Jaccard loss under conflict are supported across all three datasets. Therefore, polarity conflict is a robust indicator of poorer prediction and probability quality, whereas its relation to a particular uncertainty statistic is not universal.

6.3. Calibration, Brier Score, and Selective Prediction

The paired confidence intervals strengthen the reliability interpretation of SemEval-2018-Ar and ExaAEC. For both datasets, every controlled ECE and Brier-score comparison excludes zero in favor of CONCORD-Emo. ExaAEC also shows lower predictive-entropy-based AURC than every controlled baseline. SemEval-2025-Arq is more heterogeneous: CONCORD-Emo improves over uncalibrated MARBERT and obtains a lower Brier score than MARBERT+MC, whereas temperature-scaled baselines retain lower ECE values. The reliability advantage is therefore strong on two datasets and metric-dependent on the smaller Arq benchmark.

The Brier score is especially informative in this study because it evaluates the quality of predicted probabilities against multi-label targets. Conflict instances consistently yield higher Brier scores than blended instances, confirming that polarity-opposed examples are not only harder to classify but also harder to estimate probabilistically. This supports the argument that reliability-aware Arabic MLED requires more than aggregate classificationmetrics.

Selective prediction provides a practical use for uncertainty estimation. When high-uncertainty instances are rejected, the retained set performance improves across the evaluated datasets. The strongest retained-set gain appears on ExaAEC, where the Jaccard and micro-F_1 increase substantially after the uncertainty-based rejection. This shows that predictive entropy is not merely descriptive; it provides a ranking signal for identifying lower-risk predictions. In applications such as social media monitoring, content moderation, and human-in-the-loop affective analysis, this behavior is important because uncertain predictions can be deferred rather than treated as equally reliable.

At the same time, selective prediction introduces a coverage–reliability trade-off. Higher retained-set performance is achieved by reducing coverage, not by eliminating the model error. Therefore, we interpret selective prediction as an operational reliability mechanism rather than a replacement for predictive-performance evaluation.

6.4. Relation to Prior Arabic MLED Systems

The expanded comparison places CONCORD-Emo within the heterogeneous Arabic MLED landscape. Earlier SemEval-2018 systems included feature-based classifiers, recurrent architectures, hybrid neural models, and ensemble-oriented approaches. More recent systems include Arabic-specific structured transformers such as GTHL-Emo, dialect-specific transformer ensembles such as HTU, multilingual transformer hybrids such as INFOTEC-NLP and LATE-GIL-NLP, cross-lingual translation pipelines such as YNU-HPCC, and shared-task systems that use larger pre-trained models, prompting, parameter-efficient adaptation, augmentation, or ensembling.

These approaches answer a different experimental question from the controlled comparison in this study. They test how strongly a particular system performs under its own training resources and shared-task configuration, whereas the MARBERT-based experiments isolate architectural and reliability differences under a common protocol. Therefore, the two comparison levels are complementary rather than interchangeable.

The external results showed that CONCORD-Emo was predictively competitive but not uniformly superior. GTHL-Emo remains stronger on some SemEval-2018-Ar and ExaAEC metrics, while the top official SemEval-2025-Arq systems remain higher in macro-F_1. Conversely, CONCORD-Emo exceeded the original ExaAEC benchmark and several SemEval-2025-Arq systems. These mixed results support a cautious interpretation: predictive adequacy is established, but the principal contribution is the controlled analysis of calibration, probabilistic error, conflict-conditioned uncertainty, and selective predictions.

A direct reliability comparison with external systems is currently unavailable because their publications do not report the probability outputs and stochastic predictions needed to compute the same ECE, Brier, mutual-information, and risk–coverage measures. Accordingly, the manuscript does not claim universal superiority over Arabic-specific transformers, multilingual systems, ensembles, or LLM-based approaches.

6.5. Limitations and Future Directions

Several limitations should be acknowledged. First, the conflict definition is based on polarity opposition and therefore does not capture all forms of affective complexity. The positive and negative mappings are corpus-specific operational codings of benchmark labels, not culture-independent judgments about the valence of Arabic expressions. Excluding anticipation, surprise, and neutrality reduces obvious borderline assignments, but discourse context and culturally situated interpretation may still affect how labels such as trust or acceptance are understood. Sarcasm, pragmatic reversal, emotional intensity, and other culturally specific affective expressions may also influence reliability, but they are not explicitly modeled in the current formulation. Future work should validate polarity mappings with Arabic-speaking annotators or continuous valence judgments when such annotations are available.

Second, the datasets provide aggregated labels instead of annotator-level distributions. Consequently, we cannot directly distinguish the uncertainty caused by model limitations from the ambiguity arising from human disagreement. Access to annotator-level labels would allow a more direct analysis of aleatoric uncertainty in Arabic emotion annotations.

Third, the reliability comparison was conducted primarily against controlled MARBERT-based baselines. This is appropriate for isolating the contribution of CONCORD-Emo under

a unified protocol, but it also means that direct comparison with published systems is limited to predictive metrics. Most related Arabic MLED studies do not report ECE, Brier score, predictive entropy, mutual-information-based epistemic uncertainty, or risk–coverage analysis on the same datasets, which prevents direct external reliability comparison.

External comparison is also constrained by protocol heterogeneity. Published systems may use different data partitions, preprocessing rules, external corpora, translation, augmentation, prompting, parameter-efficient adaptation, model scales, thresholding procedures, and ensembles. Therefore, their reported classification scores cannot be combined with the controlled results, as though they arose from the same experiment. In addition, the absence of released calibrated probabilities and stochastic test outputs prevents the retrospective computation of directly comparable ECE, Brier score, mutual-information uncertainty, and risk–coverage curves. Future work should benchmark Arabic-specific transformers, multilingual encoders, ensemble systems, and LLM-based approaches under a common reliability-evaluation protocol.

Fourth, the MI score is derived from a finite set of MC-dropout predictions and therefore inherits the limitations of the dropout-based posterior approximation. In the multi-label setting, CONCORD-Emo computes binary label-wise MI terms and averages them to obtain an instance-level score. The resulting quantity does not represent the joint MI of the complete emotion-label vector and does not measure the dependence among emotion labels. It is also a post-training diagnostic rather than a differentiable regularizer. Accordingly, it should be interpreted as a complementary proxy for dropout-induced epistemic disagreement rather than an exact Bayesian uncertainty estimate. The dataset-dependent behavior of this score further supports cautious interpretation.

The reported bootstrap confidence intervals quantify test-instance sampling variability conditional on the saved trained checkpoints. Because the experiments used a fixed training seed, these intervals do not capture variability arising from parameter initialization, mini-batch order, or repeated model training. Multi-seed retraining would provide a broader estimate of training-procedure variability but would require substantially greater computation across the three datasets, CONCORD-Emo, the four controlled baselines, and the complete ablation suite.

Finally, some overconfident errors remain, as illustrated in the qualitative examples reported in Appendix D. This confirms that calibration and uncertainty estimation reduce, but do not eliminate, prediction unreliability. Future studies should investigate stronger uncertainty estimation methods, annotator-disagreement modeling, and reliability-aware learning objectives that directly penalize overconfident errors under ambiguous or polarity-conflicting emotional conditions.

Overall, the results support the main claim of this study: polarity conflict is a meaningful indicator of reduced reliability in Arabic MLED, and reliability-aware evaluation provides information that is not captured by the classification accuracy alone. CONCORD-Emo advances this direction by integrating conflict-aware representation, probability calibration, MC-dropout uncertainty estimation, and selective prediction within a unified Arabic MLEDframework.

7. Conclusions

We proposed CONCORD-Emo, a reliability-aware framework for Arabic multi-label emotion detection (MLED) that integrates label-wise attention, residual global anchoring, polarity-aware mixture-of-experts modeling, conflict-aware gating, MC dropout, and temperature-based calibration. The contribution of this study lies in the conflict-conditioned representation and evaluation design; label-wise attention, generic mixture-of-experts routing, MC dropout, temperature scaling, and selective prediction were used as established

supporting techniques rather than presented as new methods. The framework was designed to examine not only whether emotion labels are predicted correctly, but also whether model confidence remains reliable under emotionally complex and polarity-conflicting inputs.

Across SemEval-2018-Ar, ExaAEC, and SemEval-2025-Arq, CONCORD-Emo remained competitive with controlled MARBERT-based systems. The paired bootstrap intervals showed that most predictive differences included the zero. The principal interval-supported advantages appear in the calibration and probabilistic quality on SemEval-2018-Ar and ExaAEC, whereas SemEval-2025-Arq exhibits a mixed pattern in which predictive and Brier-score point estimates are competitive, but the temperature-scaled baselines retain a lower expected calibration error (ECE).

Conflict-conditioned analysis provides more consistent evidence for prediction error and probabilistic quality than for uncertainty alone. Conflict instances show higher Brier contribution and Jaccard loss than blended instances on all three datasets. Predictive entropy separates the groups on SemEval-2018-Ar and ExaAEC but not on SemEval-2025-Arq, while MI-based disagreement varies in both magnitude and direction across datasets. Therefore, polarity conflict is a meaningful reliability condition, but no single uncertainty statistic captures it uniformly.

The selective-prediction analysis further indicates that predictive entropy provides a useful ranking signal for retained-set evaluation. Rejecting high-uncertainty instances improves retained-set performance, although this gain is obtained by reducing coverage. MI-based epistemic disagreement provides a complementary but more dataset-dependent signal. The ablation results further show that CONCORD-Emo should be interpreted through a predictive–reliability trade-off rather than through uniform gains on every individual metric.

Overall, CONCORD-Emo contributes a reliability-aware evaluation perspective for Arabic MLED. The findings show that polarity structure, calibration, probabilistic error, uncertainty estimation, and selective prediction provide complementary evidence for understanding model behavior in emotionally complex Arabic text. Future work should investigate annotator-level disagreement, stronger uncertainty estimation methods, reliability-aware learning objectives, and broader validation across additional Arabic dialects, domains, and large-language-model-based emotion detection systems.

Author Contributions: Conceptualization, M.N.A. and S.T.; methodology, M.N.A., S.T. and F.F.; software, M.N.A.; validation, M.N.A. and F.F.; formal analysis, M.N.A.; investigation, M.N.A.; data curation, M.N.A.; visualization, M.N.A.; writing—original draft preparation, M.N.A.; writing—review and editing, S.T. and F.F.; supervision, S.T. and F.F. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The datasets are publicly available at the following links: SemEval-2018 Task 1: https://huggingface.co/datasets/SemEvalWorkshop/sem_eval_2018_task_1 (accessed on 20 May 2026); SemEval-2025 Task 11: https://github.com/emotion-analysis-project/SemEval2025-Task11 (accessed on 20 May 2026); ExaAEC Dataset: https://github.com/exaco/exaaec (accessed on 20 May 2026).

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AURC	Area under the risk–coverage curve
BCE	Binary cross-entropy
BERT	Bidirectional Encoder Representations from Transformers
BiLSTM	Bidirectional long short-term memory
CI	Confidence interval
ECE	Expected Calibration Error
ELMo	Embeddings from Language Models
GPU	Graphics processing unit
LLM	Large language model
MC	Monte Carlo
MI	Mutual information
MLED	Multi-label emotion detection
MoE	Mixture of experts
NLP	Natural language processing
URL	Uniform resource locator
SVM	Support vector machine

Appendix A. Class-Wise Predictive Behavior

Figure A1 shows the class-wise F_1 profiles of CONCORD-Emo across the three evaluated datasets. This analysis complements the aggregate Jaccard, micro-F_1, and macro-F_1 results reported in the main text by showing how the predictive behavior varies across individual emotion labels. The figure is included as diagnostic evidence only; the main contribution of this study remains the reliability-oriented analysis based on calibration, Brier score, uncertainty estimation, conflict-conditioned reliability, and selective prediction.

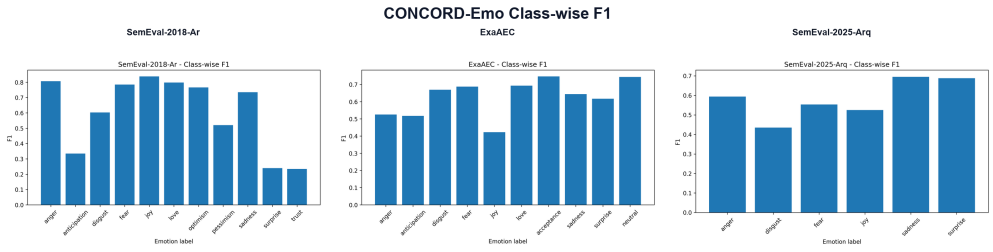


Figure A1. Class-wise F_1 profiles of CONCORD-Emo across SemEval-2018-Ar, ExaAEC, and SemEval-2025-Arq. Class-wise F_1 denotes the label-specific harmonic mean of precision and recall computed independently for each emotion label.

The class-wise profiles show that CONCORD-Emo did not perform uniformly across all emotion labels. This behavior is expected in Arabic multi-label emotion detection because the evaluated datasets contain heterogeneous label frequencies, different degrees of label co-occurrence, and varying levels of polarity conflicts. Labels with clearer lexical or contextual evidence tend to produce stronger class-wise F_1, whereas sparse or semantically overlapping labels exhibit weaker performance.

The variation across labels also explains why macro-F_1 behaves differently from Jaccard and micro-F_1 in the main results of this study. Jaccard and micro-F_1 are more influenced by frequent labels and overall label-set overlap, whereas macro-F_1 gives equal weight to each label and is therefore more sensitive to low-support or difficult emotion categories. The class-wise figure confirms that CONCORD-Emo should be interpreted as a

reliability-aware framework with competitive predictive behavior rather than a model that uniformly maximizes every label-specific classification score.

Overall, the class-wise analysis supports the interpretation adopted in the main manuscript: CONCORD-Emo maintains reasonable predictive adequacy across labels, but its primary contribution lies in modeling and evaluating reliability under uncertainty and polarity conflict. The observed label-level variation further motivates the use of calibration and selective-prediction analyses, because aggregate classification metrics alone do not fully characterize model reliability across heterogeneous emotion categories.

Appendix B. Contextual Comparison with Published Arabic MLED Systems

This appendix provides a broader contextual comparison with published systems on the same Arabic emotion benchmarks. The represented approaches include traditional and shared-task baselines, recurrent and hybrid neural models, Arabic-specific transformers, structured and ensemble-based models, multilingual transformer hybrids, translation-based systems, and recent official shared-task entries. The comparison is limited to the classification metrics reported in the original sources, including Jaccard accuracy, micro-F_1, and macro-F_1.

The external scores are not treated as controlled experimental baselines. The corresponding systems may differ in data partitioning, preprocessing, model architecture, training resources, external data, thresholding, augmentation, translation, prompting, parameter-efficient fine-tuning, ensembling, and evaluation protocol. They also do not report a common set of probability-calibration or uncertainty measures. Consequently, the tables below establish only the predictive context of CONCORD-Emo. The primary evidence for its reliability-oriented contribution remains the controlled MARBERT-based comparison, conflict-conditioned analysis, calibration evaluation, and selective-prediction results in the main text.

Table A1 reports the contextual comparison on SemEval-2018-Ar. CONCORD-Emo achieves competitive classification performance relative to prior Arabic MLED systems, particularly in micro-F_1 and macro-F_1. GTHL-Emo reports the strongest micro-F_1, while Elfaik et al. [45] and Mansy et al. [46] report the highest Jaccard accuracy and macro-F_1, respectively. These results indicate that CONCORD-Emo remains within the range of strong published Arabic MLED systems, while its main distinction lies in the reliability-oriented analyses reported in the main text.

Table A1. Contextual comparison on SemEval-2018-Ar. Jaccard Acc. denotes Jaccard accuracy; micro-F_1 and macro-F_1 denote micro-averaged and macro-averaged F1-scores, respectively. Best results in each column are shown in bold. NR indicates that the metric was not reported in the corresponding source.

Model	Jaccard Acc. (%)	Micro-F_1 (%)	Macro-F_1 (%)
Baseline: Random [9]	17.70	NR	NR
MEDIAN Team [9]	25.40	NR	NR
Baseline: SVM-Unigrams [9]	38.00	NR	NR
UNCC Team [47]	44.60	NR	NR
Tw-StAR Team [17]	46.50	NR	NR
PARTNA Team [9]	48.40	NR	NR
EMA Team [16]	48.90	61.80	46.10
Khalil et al. [18]	49.80	61.50	44.00
Alswaidan and Menai [19]	51.20	63.10	50.20
Samy et al. [20]	53.20	49.50	64.80
Elfaik et al. [48]	53.82	NR	NR

Table A1. *Cont.*

Model	Jaccard Acc. (%)	Micro-F_1 (%)	Macro-F_1 (%)
Mansy et al. [46]	54.00	52.70	**70.10**
Elfaik et al. [45]	**60.00**	52.00	35.00
GTHL-Emo [32]	58.70	**71.02**	60.83
CONCORD-Emo (Proposed)	57.53	69.50	60.56

Note: SVM = support vector machine; NR = not reported.

Table A2 reports the contextual comparison of ExaAEC. CONCORD-Emo exceeds the original ExaAEC BiLSTM+ELMo benchmark in micro-F_1, but GTHL-Emo reports a stronger raw predictive performance across the available classification metrics. This difference reflects the distinct aims of the two systems: GTHL-Emo was developed for imbalance-aware and correlation-aligned predictive learning, whereas CONCORD-Emo evaluates conflict-aware reliability, calibration, and uncertainty behavior.

Table A2. Contextual comparison on ExaAEC. Jaccard Acc. denotes Jaccard accuracy; micro-F_1 and macro-F_1 denote micro-averaged and macro-averaged F1-scores, respectively. Best results in each column are shown in bold. NR indicates that the metric was not reported in the corresponding source.

Model	Jaccard Acc. (%)	Micro-F_1 (%)	Macro-F_1 (%)
Sarbazi-Azad et al. [10] (BiLSTM+ELMo)	NR	65.00	NR
GTHL-Emo [32]	**65.99**	**70.72**	**68.71**
CONCORD-Emo (Proposed)	61.55	66.34	62.67

Note: BiLSTM = bidirectional long short-term memory; ELMo = Embeddings from Language Models; NR = not reported.

Table A3 reports the contextual comparison of SemEval-2025 Track A for Algerian Arabic (Arq). Macro-F_1 is used because it is the official Track A metric of the challenge. The table covers several methodological families: the HTU dialect-specific DziriBERT ensemble, the INFOTEC-NLP XLM-RoBERTa hybrid, the LATE-GIL-NLP mDeBERTa-v3 system, the YNU-HPCC translation-based DeBERTa system, the structured GTHL-Emo model, and the official JNLP and Pai leaderboard results [11,21–24,32]. CONCORD-Emo exceeded the official baselines, GTHL-Emo, and several selected shared-task systems but remained below the strongest official entries. Because the systems differ in training strategy and resource usage, this ranking is reported as contextual predictive positioning rather than controlled evidence of architectural superiority.

Table A3. Contextual comparison on SemEval-2025 Track A for Algerian Arabic (Arq). Macro-F_1 is the official Track A metric. Best result is shown in bold.

Model	Macro-F_1 (%)
Official RoBERTa baseline [11]	41.40
Official majority baseline [11]	44.50
YNU-HPCC Team [24]	44.40
LATE-GIL-NLP Team [23]	48.60
HTU Team [21]	51.20
INFOTEC-NLP Team [22]	51.70
GTHL-Emo [32]	56.69
JNLP Team [11]	64.10
Pai Team [11]	**66.90**
CONCORD-Emo (Proposed)	58.17

Taken together, the external comparisons show that CONCORD-Emo is competitive with accuracy-oriented Arabic MLED systems but is not uniformly the highest-scoring model in the thresholded classification metrics. Therefore, this appendix only supports the contextual predictive positioning of the model. The central contribution of CONCORD-Emo remains the reliability-oriented analysis in the main text, including the Expected Calibration Error (ECE), Brier score, predictive entropy, mutual-information-based epistemic uncertainty, conflict-conditioned reliability, and uncertainty-based selective prediction.

Appendix C. Paired Bootstrap Confidence Interval Analysis

This appendix reports the complete interval-based analysis of the controlled model comparisons and the conflict-versus-blended reliability contrast. The point estimate in each forest plot is the difference between CONCORD-Emo and the corresponding controlled baseline, and the horizontal bar is the 95-percent paired bootstrap confidence interval. Green intervals exclude zero in favor of CONCORD-Emo, red intervals exclude zero in favor of the baseline, and gray intervals include zero.

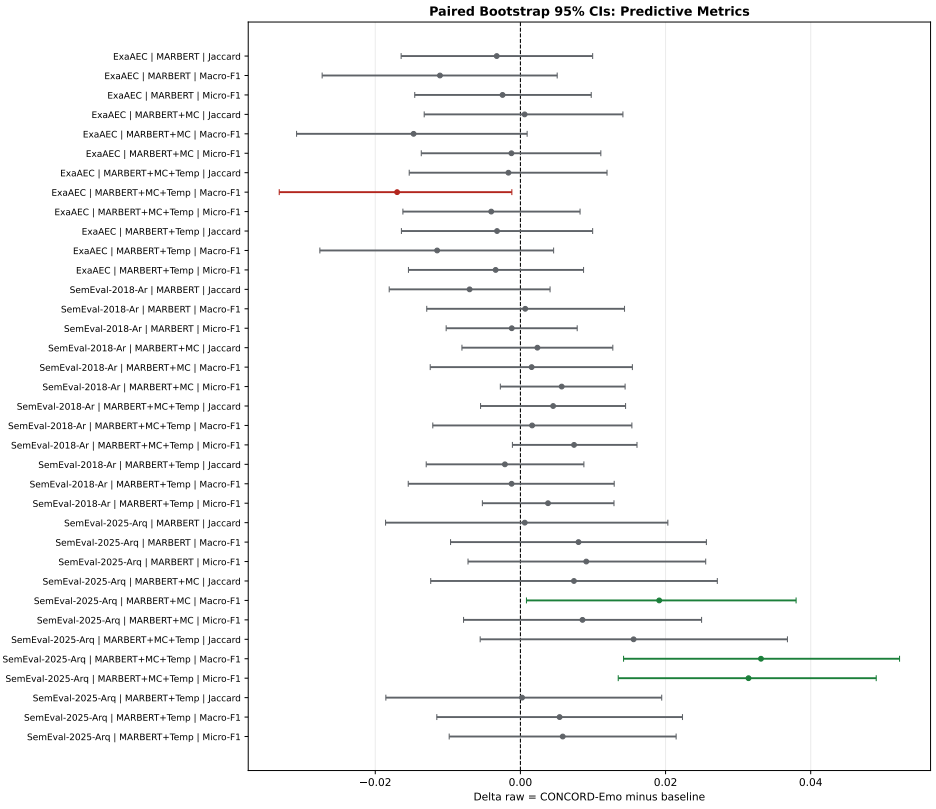


Figure A2. Paired bootstrap 95% confidence intervals for predictive-metric differences between CONCORD-Emo and the controlled MARBERT-based baselines. The difference is CONCORD-Emo minus baseline; positive values favor CONCORD-Emo. Green intervals exclude zero in favor of CONCORD-Emo, red intervals exclude zero in favor of the baseline, and gray intervals include zero.

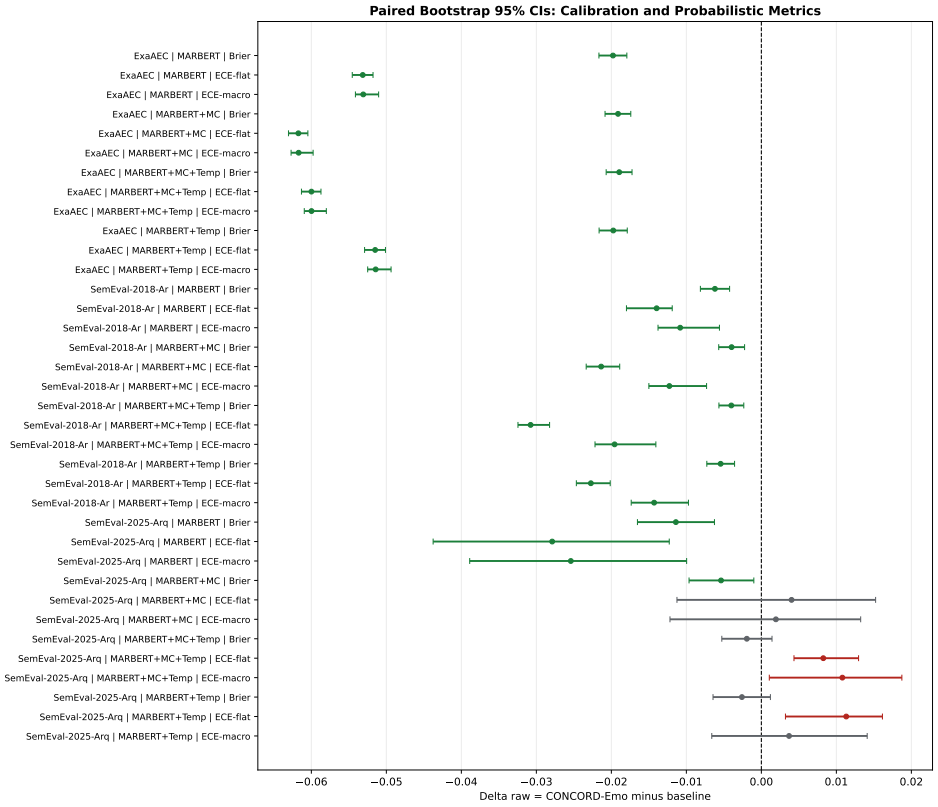


Figure A3. Paired bootstrap 95% confidence intervals for calibration and probabilistic-metric differences. The difference is CONCORD-Emo minus baseline; negative values favor CONCORD-Emo because lower Expected Calibration Error and Brier score are better. Green intervals exclude zero in favor of CONCORD-Emo, red intervals exclude zero in favor of the baseline, and gray intervals include zero.

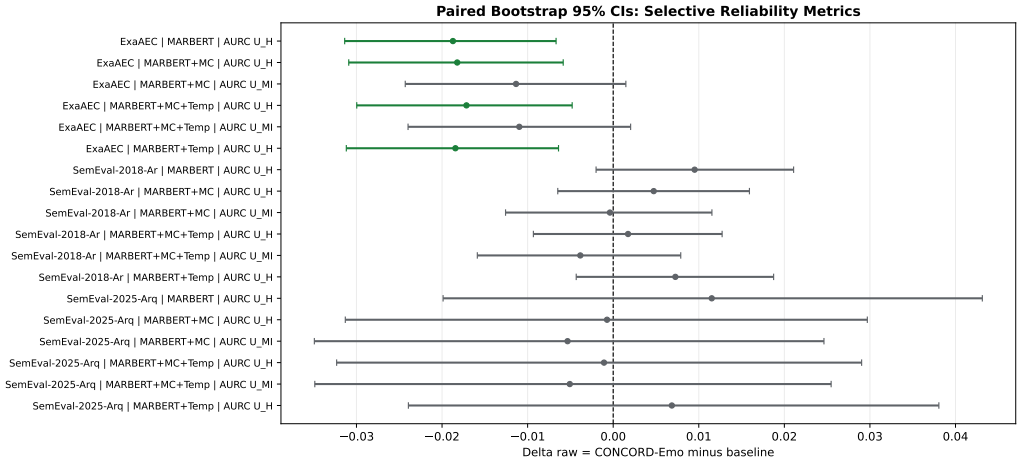


Figure A4. Paired bootstrap 95% confidence intervals for selective-reliability differences. The difference is CONCORD-Emo minus baseline; negative values favor CONCORD-Emo because lower Area Under the Risk–Coverage Curve is better. Mutual-information-based comparisons are omitted for deterministic baselines with degenerate mutual-information values. Green intervals exclude zero in favor of CONCORD-Emo and gray intervals include zero.

Table A4. Conflict-versus-blended statistical analysis for CONCORD-Emo. The difference is the conflict mean minus the blended mean. Confidence intervals are 95% percentile intervals based on 10,000 independent-group bootstrap resamples. The reported p values are two-sided Mann–Whitney U test results, and r_{rb} denotes the rank-biserial effect size.

Dataset	Metric	Conflict	Blended	Difference [95% CI]	p	r_{rb}
SemEval-2018-Ar	U_H	0.5317	0.4406	0.0912 [0.0759, 0.1064]	8.05×10^{-26}	0.518
SemEval-2018-Ar	U_{MI}	0.0203	0.0155	0.0048 [0.0032, 0.0063]	9.96×10^{-13}	0.351
SemEval-2018-Ar	Brier contribution	0.1496	0.0844	0.0652 [0.0529, 0.0775]	5.85×10^{-26}	0.519
SemEval-2018-Ar	Jaccard loss	0.5728	0.3561	0.2166 [0.1752, 0.2569]	4.07×10^{-21}	0.458
SemEval-2018-Ar	Hamming error	0.2250	0.1247	0.1003 [0.0792, 0.1209]	3.59×10^{-22}	0.461
SemEval-2018-Ar	Confidence	0.8423	0.8791	−0.0367 [−0.0437, −0.0299]	1.67×10^{-22}	−0.481
ExaAEC	U_H	0.4571	0.4255	0.0316 [0.0111, 0.0520]	0.004	0.171
ExaAEC	U_{MI}	0.0140	0.0128	0.0012 [−0.0009, 0.0034]	0.204	0.076
ExaAEC	Brier contribution	0.1022	0.0703	0.0319 [0.0180, 0.0457]	3.53×10^{-6}	0.276
ExaAEC	Jaccard loss	0.4837	0.3061	0.1775 [0.1113, 0.2429]	1.84×10^{-7}	0.295
ExaAEC	Hamming error	0.1320	0.0794	0.0526 [0.0325, 0.0731]	2.47×10^{-7}	0.290
ExaAEC	Confidence	0.8683	0.8840	−0.0156 [−0.0248, −0.0064]	0.001	−0.189
SemEval-2025-Arq	U_H	0.8066	0.7994	0.0072 [−0.0136, 0.0280]	0.643	0.041
SemEval-2025-Arq	U_{MI}	0.0162	0.0198	−0.0035 [−0.0059, −0.0009]	0.002	−0.274
SemEval-2025-Arq	Brier contribution	0.2650	0.1797	0.0853 [0.0557, 0.1162]	1.20×10^{-7}	0.472
SemEval-2025-Arq	Jaccard loss	0.6479	0.4708	0.1772 [0.1060, 0.2463]	7.62×10^{-6}	0.396
SemEval-2025-Arq	Hamming error	0.4132	0.2839	0.1293 [0.0721, 0.1869]	3.69×10^{-5}	0.356
SemEval-2025-Arq	Confidence	0.7155	0.7185	−0.0031 [−0.0178, 0.0119]	0.854	−0.016

Note: Mann–Whitney U test p values are unadjusted and are reported as complementary distributional evidence; the bootstrap confidence intervals constitute the primary inferential analysis.

Appendix D. Qualitative Reliability Examples

The qualitative examples in Table A5 provide illustrative case-level evidence for the reliability behavior of CONCORD-Emo. The examples were selected to cover correct low-uncertainty predictions, high-uncertainty errors, polarity-conflict errors, and overconfident misclassifications. These cases are not used as primary evidence for the model's effectiveness; rather, they complement the aggregate quantitative results by showing how confidence and uncertainty behave for representative Arabic inputs.

These cases also illustrate the group-assignment rule. The disgust-only case is single. The joy–love–optimism case is blended, whereas the love–sadness case is conflict. In the anger–joy–surprise case, surprise is omitted from the polarity counts, so the opposing anger and joy labels determine the conflict assignment.

Table A5. Qualitative reliability examples from CONCORD-Emo. The Arabic input texts are shown as image snippets. U_H denotes predictive entropy, and U_{MI} denotes mutual-information-based epistemic uncertainty estimated from Monte Carlo (MC) dropout.

Dataset	Case Type	Arabic Text	Gold Labels	Predicted Labels	Conf.	U_H	U_{MI}
SemEval-2025-Arq	High-uncertainty error	الراوي عمر ما كانش راحله اللي الناس دايرينه ولد صغيرولا قبيح . ما كانش علاباله بيهم.	disgust	anger; fear	0.602	0.962	0.032
SemEval-2025-Arq	Conflict error	منصوريه هدرت بزاف اليوم، شويه بزاف. اسمحولي قالتها بلعقل . بصح مارانيش حابه تشدوني. راني شكرتكم و سلمت عليكم. لازم نروح.	joy; sadness	disgust; sadness; surprise	0.606	0.954	0.030
SemEval-2018-Ar	Correct low uncertainty	ل اهلي ل صديقاتي ل كل شخص لطيف مر بحياتي ل متابعيني كل عام وانتو بخير كل عام والفرحه تغمركم وابتسامتكم الجميله م يعكرها حزن	joy; love; optimism	joy; love; optimism	0.964	0.220	0.002
SemEval-2018-Ar	Conflict error	اكيد من باب الفكاهه لان مفيش انسان عاقل يقول هكي	anger; joy; surprise	anger; anticipation; disgust; pessimism	0.732	0.733	0.042

Table A5. *Cont.*

Dataset	Case Type	Arabic Text	Gold Labels	Predicted Labels	Conf.	U_H	U_{MI}
ExaAEC	Conflict correct	يا قلبي المشتاق مافاد قصيدي الله كتبها فراق و الله مو ب يدي ؟	love; sadness	love; sadness	0.936	0.328	0.005
ExaAEC	Overconfident error	تمويل شخصي من الي الف بدون كفيل بدون تحويل راتب تقدم ب طلب قرضك الشخصي الان الرياض جده سكاكا القريات الجوف الطايف الصندوق العقاري صباح الثلاثاء	acceptance	neutral	0.973	0.176	0.002

Note: Conf. = confidence.

The qualitative examples illustrate complementary reliability behaviors that are consistent with the aggregate results. The two SemEval-2025-Arq error cases show high predictive entropy, with $U_H = 0.962$ for the high-uncertainty error and $U_H = 0.954$ for the conflict error. These values indicate that the model assigns an unstable probability mass when the final label set is incorrect or when polarity-opposed labels are involved. The SemEval-2018-Ar conflict error shows a similar pattern, with elevated entropy ($U_H = 0.733$) and the highest mutual-information value among the selected examples ($U_{MI} = 0.042$), suggesting stronger stochastic disagreement under this conflict-structured case.

In contrast, the SemEval-2018-Ar correct low-uncertainty example shows high confidence (0.964), low predictive entropy ($U_H = 0.220$), and low mutual information ($U_{MI} = 0.002$), indicating a stable prediction when the emotional evidence is correctly captured. The ExaAEC conflict-correct example further shows that polarity conflict does not necessarily imply prediction failure; the model correctly predicts both *love* and *sadness* with high confidence and low uncertainty, respectively. However, the ExaAEC overconfident error demonstrates a limitation of uncertainty estimation: the model predicts an incorrect label with high confidence (0.973) and low entropy ($U_H = 0.176$). This case reinforces the need to evaluate the calibration and Brier scores alongside entropy-based uncertainty, since not all incorrect predictions are detected by high uncertainty.

References

1. Plaza-del Arco, F.M.; Cercas Curry, A.A.; Cercas Curry, A.; Hovy, D. Emotion Analysis in NLP: Trends, Gaps and Roadmap for Future Directions. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), Turin, Italy, 22–24 May 2024; pp. 5696–5710.
2. Poria, S.; Cambria, E.; Bajpai, R.; Hussain, A. A Review of Affective Computing: From Unimodal Analysis to Multimodal Fusion. *Inf. Fusion* **2017**, *37*, 98–125. [CrossRef]
3. Al-Ayyoub, M.; Khamaiseh, A.A.; Jararweh, Y.; Al-Kabi, M.N. A Comprehensive Survey of Arabic Sentiment Analysis. *Inf. Process. Manag.* **2019**, *56*, 320–342. [CrossRef]
4. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186.
5. Alhuzali, H.; Ananiadou, S. SpanEmo: Casting Multi-label Emotion Classification as Span-Prediction. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, Online, 19–23 April 2021; pp. 1573–1584.
6. Demszky, D.; Movshovitz-Attias, D.; Ko, J.; Cowen, A.; Nemade, G.; Ravi, S. GoEmotions: A Dataset of Fine-Grained Emotions. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 4040–4054.
7. Chauhan, D.S.; Dhanush, S.R.; Ekbal, A.; Bhattacharyya, P. Sentiment and Emotion Help Sarcasm? A Multi-task Learning Framework for Multi-Modal Sarcasm, Sentiment and Emotion Analysis. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 6351–6357.
8. Farha, I.A.; Magdy, W. From Arabic Sentiment Analysis to Sarcasm Detection: The ArSarcasm Dataset. In Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection, Marseilles, France, 12 May 2020; pp. 32–39.
9. Mohammad, S.M.; Bravo-Marquez, F.; Salameh, M.; Kiritchenko, S. SemEval-2018 Task 1: Affect in Tweets. In Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval 2018), New Orleans, LA, USA, 5–6 June 2018; pp. 1–17.

10. Sarbazi-Azad, S.; Akbari, A.; Khazeni, M. ExaAEC: A New Multi-label Emotion Classification Corpus in Arabic Tweets. In Proceedings of the 11th International Conference on Computer Engineering and Knowledge (ICCKE 2021), Mashhad, Iran, 28–29 October 2021; pp. 465–470.
11. Muhammad, S.H.; Ousidhoum, N.; Abdulmumin, I.; Wahle, J.P.; Ruas, T.; Beloucif, M.; de Kock, C.; Surange, N.; Teodorescu, D.; Ahmad, I.S.; et al. BRIGHTER: Bridging the Gap in Human-Annotated Textual Emotion Recognition Datasets for 28 Languages. *arXiv* **2025**, arXiv:2502.11926.
12. Wu, W.; Zhang, C.; Woodland, P. Estimating the Uncertainty in Emotion Attributes Using Deep Evidential Regression. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Toronto, ON, Canada, 9–14 July 2023; pp. 15681–15695.
13. Seoni, S.; Jahmunah, V.; Salvi, M.; Barua, P.D.; Molinari, F.; Acharya, U.R. Application of Uncertainty Quantification to Artificial Intelligence in Healthcare: A Review of Last Decade (2013–2023). *Comput. Biol. Med.* **2023**, *165*, 107441. [CrossRef] [PubMed]
14. Guo, C.; Pleiss, G.; Sun, Y.; Weinberger, K.Q. On Calibration of Modern Neural Networks. In Proceedings of the 34th International Conference on Machine Learning (ICML 2017), Sydney, NSW, Australia, 6–11 August 2017; pp. 1321–1330.
15. Gal, Y.; Ghahramani, Z. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In Proceedings of the 33rd International Conference on Machine Learning (ICML 2016), New York, NY, USA, 19–24 June 2016; pp. 1050–1059.
16. Badaro, G.; El Jundi, O.; Khaddaj, A.; Maarouf, A.; Kain, R.; Hajj, H.; El-Hajj, W. EMA at SemEval-2018 Task 1: Emotion Mining for Arabic. In Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval 2018), New Orleans, LA, USA, 5–6 June 2018; pp. 236–244.
17. Mulki, H.; Ali, C.B.; Haddad, H.; Babaoğlu, I. TW-Star at SemEval-2018 Task 1: Preprocessing Impact on Multi-label Emotion Classification. In Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval 2018), New Orleans, LA, USA, 5–6 June 2018; pp. 167–171.
18. Khalil, E.A.H.; Houby, E.M.F.E.; Mohamed, H.K. Deep Learning for Emotion Analysis in Arabic Tweets. *J. Big Data* **2021**, *8*, 136. [CrossRef]
19. Alswaidan, N.; Menai, M.E.B. Hybrid Feature Model for Emotion Recognition in Arabic Text. *IEEE Access* **2020**, *8*, 37843–37854. [CrossRef]
20. Samy, A.E.; El-Beltagy, S.R.; Hassanien, E. A Context Integrated Model for Multi-label Emotion Detection. *Procedia Comput. Sci.* **2018**, *142*, 61–71. [CrossRef]
21. Saleh, A.; Biltawi, M. HTU at SemEval-2025 Task 11: Divide and Conquer – Multi-Label Emotion Classification Using 6 DziriBERTs Submodels with Label-fused Iterative Mask Filling Technique for Low-Resource Data Augmentation. In Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025), Turin, Italy, 31 July–1 August 2025; pp. 675–683.
22. Santos-Rodriguez, E.; Graff, M. INFOTEC-NLP at SemEval-2025 Task 11: A Case Study on Transformer-Based Models and Bag of Words. In Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025), Turin, Italy, 31 July–1 August 2025; pp. 350–356.
23. Gómez-Adorno, H.; Sierra, G.; Sierra-Casiano, V.; Canchola-Hernández, D.; Tovar-Cortés, J.; Solís-Vilchis, R.; Salazar, G. LATE-GIL-NLP at SemEval-2025 Task 11: Multi-Language Emotion Detection and Intensity Classification Using Transformer Models with Optimized Loss Functions for Imbalanced Data. In Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025), Turin, Italy, 31 July–1 August 2025; pp. 666–674.
24. Yang, H.; Wang, J.; Zhang, X. YNU-HPCC at SemEval-2025 Task 11: Bridging the Gap in Text-Based Emotion Using Multiple Prediction Headers. In Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025), Turin, Italy, 31 July–1 August 2025; pp. 83–89.
25. Antoun, W.; Baly, F.; Hajj, H. 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, Marseilles, France, 12 May 2020; pp. 9–15.
26. Abdul-Mageed, M.; Elmadany, A. 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), Online, 1–6 August 2021; pp. 7088–7105.
27. Safaya, A.; Abdullatif, M.; Yuret, D. KUISAIL at SemEval-2020 Task 12: BERT-CNN for Offensive Speech Identification in Social Media. In Proceedings of the Fourteenth Workshop on Semantic Evaluation (SemEval 2020), Barcelona, Spain (Online), 12–13 December 2020; pp. 2054–2059.
28. Inoue, G.; Alhafni, B.; Baimukan, N.; Bouamor, H.; Habash, N. The Interplay of Variant, Size, and Task Type in Arabic Pre-trained Language Models. In Proceedings of the Sixth Arabic Natural Language Processing Workshop, Kyiv, Ukraine (Online), 19 April 2021; pp. 92–104.
29. Fei, H.; Zhang, Y.; Ren, Y.; Ji, D. Latent Emotion Memory for Multi-label Emotion Classification. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 7692–7699.

30. Zhu, Y.; Wu, O. Elementary Discourse Units with Sparse Attention for Multi-label Emotion Classification. *Knowl.-Based Syst.* **2022**, *240*, 108114. [CrossRef]
31. Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Liò, P.; Bengio, Y. Graph Attention Networks. In Proceedings of the 6th International Conference on Learning Representations (ICLR 2018), Vancouver, BC, Canada, 30 April–3 May 2018; pp. 1–12.
32. Alrasheedy, M.N.; Tiun, S.; Fauzi, F. GTHL-Emo: Adaptive Imbalance-Aware and Correlation-Aligned Training for Arabic Multi-Label Emotion Detection. *Electronics* **2026**, *15*, 1169.
33. Abdar, M.; Pourpanah, F.; Hussain, S.; Rezazadegan, D.; Liu, L.; Ghavamzadeh, M.; Fieguth, P.; Cao, X.; Khosravi, A.; Acharya, U.R.; et al. A Review of Uncertainty Quantification in Deep Learning: Techniques, Applications and Challenges. *Inf. Fusion* **2021**, *76*, 243–297. [CrossRef]
34. Gawlikowski, J.; Tassi, C.R.N.; Ali, M.; Lee, J.; Humt, M.; Feng, J.; Kruspe, A.; Triebel, R.; Jung, P.; Roscher, R.; et al. A Survey of Uncertainty in Deep Neural Networks. *Artif. Intell. Rev.* **2023**, *56*, 1513–1589. [CrossRef]
35. Lakshminarayanan, B.; Pritzel, A.; Blundell, C. Simple and Scalable Predictive Uncertainty Estimation Using Deep Ensembles. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 6402–6413.
36. Shannon, C.E. A Mathematical Theory of Communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423. [CrossRef]
37. Incremona, A.; Pozzi, A.; Guiscardi, A.; Tessera, D. A differentiable and uncertainty-aware mutual information regularizer for bias mitigation. *Neurocomputing* **2026**, *669*, 132498.
38. Geifman, Y.; El-Yaniv, R. Selective Classification for Deep Neural Networks. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 4885–4894.
39. Shazeer, N.; Mirhoseini, A.; Maziarz, K.; Davis, A.; Le, Q.; Hinton, G.; Dean, J. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv* **2017**, arXiv:1701.06538.
40. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [CrossRef]
41. Ba, J.L.; Kiros, J.R.; Hinton, G.E. Layer Normalization. *arXiv* **2016**, arXiv:1607.06450.
42. Brier, G.W. Verification of Forecasts Expressed in Terms of Probability. *Mon. Weather. Rev.* **1950**, *78*, 1–3. [CrossRef]
43. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 8024–8035.
44. Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. Transformers: State-of-the-art Natural Language Processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Online, 16–20 November 2020; pp. 38–45.
45. Elfaik, H. Leveraging Feature-level Fusion Representations and Attentional Bidirectional RNN–CNN Deep Models for Arabic Affect Analysis on Twitter. *J. King Saud Univ.–Comput. Inf. Sci.* **2023**, *35*, 462–482. [CrossRef]
46. Mansy, A.; Rady, S.; Gharib, T. An Ensemble Deep Learning Approach for Emotion Detection in Arabic Tweets. *Int. J. Adv. Comput. Sci. Appl.* **2022**, *13*, 980–989. [CrossRef]
47. Abdullah, M.; Shaikh, S. TeamUNCC at SemEval-2018 Task 1: Emotion Detection in English and Arabic Tweets Using Deep Learning. In Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval 2018), New Orleans, LA, USA, 5–6 June 2018; pp. 350–357.
48. Elfaik, H. Combining Context-Aware Embeddings and an Attentional Deep Learning Model for Arabic Affect Analysis on Twitter. *IEEE Access* **2021**, *9*, 111214–111230. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.