CITD@UIT at SemEval-2026 Task 2: Temporal Mixture-of-Experts for Longitudinal Valence and Arousal Prediction from Ecological Essays

Son The Phuong[1,2], My Thuy-Tra Ngo[1,2], Tri Minh Dao[1,2], Duc-Vu Nguyen[1,2]
[1]University of Information Technology, Ho Chi Minh City, Vietnam
[2]Vietnam National University, Ho Chi Minh City, Vietnam
{25210032, 25210021, 25210041}@ms.uit.edu.vn vund@uit.edu.vn

Abstract

This paper describes our participation in SemEval-2026 Task 2, which focuses on the longitudinal assessment and forecasting of emotional states through text. The challenge is divided into two primary objectives: Subtask 1, which requires estimating continuous Valence and Arousal (V&A) scores for a sequence of texts, and Subtask 2, which focuses on forecasting future emotional variations, specifically State Change (2A) and Dispositional Change (2B). To address these tasks, we propose a unified framework based on `cardiffnlp/twitter-roberta-base-sentiment-latest`[1], a transformer architecture pretrained on 124 million tweets. For all subtasks, we sort the data chronologically by `user_id` and use a sliding window approach to capture longitudinal context. We conduct extensive experiments combining this pretrained RoBERTa model with Multilayer Perceptron (MLP) and Mixture-of-Experts (MoE) architectures to optimize performance. Furthermore, we utilize both attention pooling and mean pooling on all output hidden state representations to extract richer semantic features. Our proposed system demonstrated competitive performance, officially ranking 9th in Subtask 1 and 5th in Subtask 2A among participating teams.[2]

1 Introduction

In recent years, the intersection of Natural Language Processing (NLP) and mental health has gained significant attention, particularly in detecting and monitoring emotional states through text analysis (Teferra et al., 2024). Unlike traditional discrete emotion classification, dimensional emotion analysis provides a more nuanced approach to understanding affective states through continuous scales. The Valence-Arousal (V&A) framework, where Valence represents the degree of positivity or negativity, and Arousal indicates the level of emotional activation, has proven effective in capturing the complexity of human emotions (Mendes and Martins, 2023). SemEval-2026 Task 2 addresses this by introducing "Predicting Variation in Emotional Valence and Arousal over Time from Ecological Essays" released by Soni et al. (2026), which encompasses continuous V&A estimation (Subtask 1) and forecasting future emotional variations (Subtask 2: State and Dispositional Changes). To address these complexities, we propose a unified backbone leveraging the Twitter-RoBERTa model, specifically `cardiffnlp/twitter-roberta-base-sentiment-latest`, trained on 124M tweets from January 2018 to December 2021 (Camacho-collados et al., 2022). Our architecture integrates several state-of-the-art strategies to optimize regression performance.

Separate Funnel Architecture. We employ a Separate Funnel Architecture based on Kashyap (2024); Hristov et al. (2025) to compress transformer hidden states ($768 \rightarrow 256 \rightarrow 128 \rightarrow 1$). This architecture extracts salient emotional features from ecological essays for our initial development experiments.

We extend the standard MLP head from Hristov et al. (2025) into a Mixture-of-Experts (MoE) architecture. Instead of a fixed regression path, our gating mechanism dynamically routes inputs to specialized experts for specific emotional sub-domains (Xie et al., 2025; Fedus et al., 2022). This "More Is Better" approach effectively decomposes the continuous affective space, capturing diverse emotional manifestations more robustly than monolithic models (Xie et al., 2025).

While prior works focused on standard regression losses for V&A prediction (Mendes and Martins, 2023), we propose to optimize our system using a multitask Concordance Correlation Coefficient Loss (CCCL) for separate Valence and Arousal prediction (Atmaja and Akagi, 2020). This choice directly aligns our training objective with the evaluation metric employed by SemEval-2026 Task 2, ensuring both linear correlation and scale consistency between predictions and ground truth.

We select the best model based on the performance on development sets for the final submission, and our system achieves competitive results on both subtasks

2 Related Work

Continuous Valence-Arousal (V&A) representations capture affective nuances better than traditional discrete emotion classification (Russell, 1980; Mendes

[1]Available at: https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment-latest

[2]Our code is publicly available at: https://github.com/PTSown0222/SemEval-2026-Task-2

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 167–175

and Martins, 2023). While pretrained models such as CardiffNLP's RoBERTa establish strong baselines (Camacho-collados et al., 2022), existing applications primarily focus on fixed text predictions, largely overlooking the longitudinal emotional dynamics inherent in sequential ecological essays.

To effectively process the high-dimensional embeddings generated by these transformers, recent works have transitioned from standard Multi-Layer Perceptrons to specialized regression heads. A foundational funnel architecture utilizing progressive dimensionality reduction (e.g., 768 → 256 → 128) was originally introduced by Kashyap (2024) for emotion recognition. This architecture was subsequently adapted by Hristov et al. (2025) at Stanford University to target the two-dimensional valence-arousal space for mental health prediction specifically. While providing a solid structural foundation, these single-path regression heads often struggle with highly dynamic temporal variations.

Beyond unimodal text approaches, multimodal studies have significantly advanced dimensional affect recognition by integrating information across speech, facial expressions, and text. These studies demonstrate that cross-attention mechanisms and multi-task learning across dimensions provide crucial cross-regularization benefits, enabling specialized fusion architectures to reach strong Concordance Correlation Coefficient (CCC) scores, such as 0.606 for valence and 0.620 for arousal on benchmark datasets (Meng et al., 2022). Inspired by the success of these cross-regularization benefits, our system explicitly moves away from standard regression losses, in favour of a multitask CCC loss to optimize textual longitudinal predictions and ensure scale consistency.

To better capture diverse affective patterns, we replace static regression heads with a Mixture-of-Experts (MoE) architecture (Fedus et al., 2022; Dai et al., 2024). This choice is motivated by the success of multi-expert frameworks in emotion recognition, notably the "More Is Better" ensemble, which achieved state-of-the-art results in the MER2025 challenge (Xie et al., 2025). By employing dynamic routing, our model is better equipped to handle the complex, long-term dispositional shifts in ecological essays, a task where traditional monolithic architectures often underperform.

3 Task Description

SemEval-2026 Task 2 focuses on modeling emotional dynamics through two primary objectives, released by Soni et al. (2026). **Subtask 1 (Longitudinal Affect Assessment)** involves predicting chronological Valence and Arousal (V&A) scores (v_i, a_i) for text sequences across both seen and unseen user groups. **Subtask 2 (Forecasting Future Affect Variation)** requires estimating changes in emotional states, categorized into: **2A (State Change)**, defined as the immediate variation $\Delta_1 = v_{t+1} - v_t$, and **2B (Dispositional Change)**, defined as the long-term shift $\Delta_{avg} = \text{avg}(v_{t+1:n}) - \text{avg}(v_{1:t})$. Full task specifications, evaluation scripts, and datasets are hosted on the CodaBench platform.

4 System Description

4.1 Data Preprocessing

Data preprocessing is essential for maintaining signal integrity in ecological essays. Beyond standard noise reduction and tokenization, we prioritize temporal alignment to preserve chronological sequences. Rather than treating posts in isolation, we restructure the dataset to capture temporal dynamics, employing distinct sequence construction strategies task-specific to the specific subtasks, as detailed below.

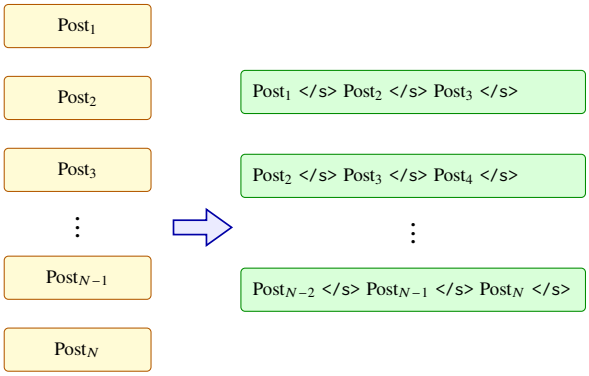

Figure 1: Generalized sliding-window concatenation for `Window_Size=3`. Consecutive posts are grouped chronologically and joined using the </s> delimiter to provide contextual history.

Sliding Window for Subtasks 1 and 2A. For Subtasks 1 and 2A, we apply a sliding-window strategy (e.g., $k = 3$) that concatenates consecutive chronological posts using the $\langle/s\rangle$ delimiter, as illustrated in Figure 1. This allows the RoBERTa encoder to capture short-term historical context for forecasting affective states.

Bipartite Grouping for Subtask 2B. To formally describe the fusion process for Subtask 2B, the input sequence is first constructed as:

$$\mathbf{x}_{input} = \texttt{<s>} \oplus \text{tokens}(G_1) \oplus \texttt{</s> </s>} \oplus \text{tokens}(G_2) \oplus \texttt{</s>} \quad (1)$$

where $\oplus$ denotes the concatenation operator. We denote the contextual embedding extracted from the RoBERTa encoder as $\mathbf{h}_{base} \in \mathbb{R}^d$. To provide the model with a quantifiable baseline of the user's past emotional state, we construct a 2-dimensional feature vector $\mathbf{v}_{past} = [\mu_{V_{G1}}, \mu_{A_{G1}}]$ representing the mean Valence and Arousal scores from the historical period (G_1). The late-fusion mechanism is then implemented by concatenating these two vectors to form a unified representation $\mathbf{z}$:

$$\mathbf{z} = [\mathbf{h}_{base}; \mathbf{v}_{past}] \quad (2)$$

where $\mathbf{z} \in \mathbb{R}^{d+2}$ serves as the input for the subsequent MLP or MoE regressor to predict the final dispositional changes $\hat{y}$:

$$\hat{y} = \text{MLP/MoE}(\mathbf{z}) \quad (3)$$

4.2 Multi-Layer Perceptron Regression Head

Following Hristov et al. (2025), we employ dual independent MLP branches integrated on top of the RoBERTa encoder to predict Valence and Arousal separately. This decoupled architecture is designed to mitigate cross-dimensional feature interference, allowing each head to specialize in the distinct linguistic signals associated with each affective dimension. Each branch adopts a bottleneck structure that progressively reduces dimensionality ($768 \rightarrow 256 \rightarrow 128 \rightarrow 1$), incorporating Layer Normalization, GELU activation, and Dropout ($p \in [0.2, 0.3]$) to enhance training stability, as illustrated in Figure 2 and detailed in Section B.

4.3 Mixture-of-Experts Regression Head

To robustly capture diverse emotional patterns, we replace the monolithic MLP with a Mixture-of-Experts (MoE) architecture (Dai et al., 2024; Shazeer et al., 2017). The module consists of $N = 4$ independent experts, where the final output $y(x)$ for input x is a weighted sum:

$$y(x) = \sum_{i=1}^{N} G(x)_i \cdot E_i(x) \tag{4}$$

where $E_i(x)$ represents the output of the i-th expert (typically a Feed-Forward Network), and $G(x)_i$ is the scalar routing weight assigned by the gating network to determine the contribution of expert i. We explore two distinct variants for this gating mechanism: Sparsely-Gated and Soft-Gated.

4.3.1 Sparsely-Gated Mixture-of-Experts Regressor

In the Sparse MoE configuration, we incentivizes specialization by activating only a subset of experts for each input instance (Shazeer et al., 2017; Dai et al., 2024). The gating network first computes simple routing logits via a linear projection $h(x) = xW_g$, where W_g is a trainable weight matrix. Sparsity is explicitly achieved by applying a Top-k filtering function that retains only the highest routing probabilities:

$$\text{KeepTopK}(h(x)_i, k) = \begin{cases} h(x)_i & \text{if } i \in \text{Top-}k \\ -\infty & \text{otherwise} \end{cases} \tag{5}$$

The final routing weights are subsequently obtained by applying a softmax activation over these filtered logits: $G(x) = \text{Softmax}(\text{KeepTopK}(h(x), k))$. By dynamically selecting only the top-k experts, this architecture forces individual sub-networks to specialize in processing specific affective features.

As shown in Figure 3, Section B, this architecture activates only the Top-k experts ($k < N$) per input via a linear gating network. Non-selected experts are masked with $-\infty$ to zero their weights post-Softmax. This sparse mechanism captures diverse affective patterns efficiently by computing only active experts for the final output: $y(x) = \sum_{i=1}^{N} G(x)_i \cdot E_i(x)$.

4.3.2 Soft-Gated Mixture-of-Experts Regressor

Conversely, Soft-Gating employs a dense routing strategy (Jacobs et al., 1991; Dai et al., 2024), allowing gradients to backpropagate through all experts simultaneously. Unlike the sparse variant, the gating network $G(x)$ omits Top-k filtering, applying Softmax directly to the linear transformation:

$$G(x) = \text{Softmax}(xW_g) \tag{6}$$

This ensures every expert receives a positive weight, facilitating collaborative specialization for Valence and Arousal (see Figure 4, Section B).

4.3.3 Concordance Correlation Coefficient (CCC) Loss

To align our training objective with the evaluation metrics of SemEval-2026 Task 2, we optimize the network using a multitask Concordance Correlation Coefficient (CCC) Loss. As highlighted by Atmaja and Akagi (2020) and Meng et al. (2022), the CCC Loss is superior to standard regression objectives in affect recognition as it simultaneously accounts for linear correlation and scale consistency. For a set of predictions $\hat{y}$ and ground truth labels y, the CCC is defined as:

$$\text{CCC} = \frac{2\rho\sigma_{\hat{y}}\sigma_y}{\sigma_{\hat{y}}^2 + \sigma_y^2 + (\mu_{\hat{y}} - \mu_y)^2} \tag{7}$$

where ρ is the Pearson correlation coefficient, while μ and σ^2 denote the means and variances of the respective distributions. Following the multitask learning paradigm, we minimize the joint loss for both Valence and Arousal dimensions:

$$\mathcal{L}_{total} = (1 - \text{CCC}_{Valence}) + (1 - \text{CCC}_{Arousal}) \tag{8}$$

4.4 Mean Pooling and Attention Pooling

Instead of extracting the [CLS] token in isolation, we implement two pooling mechanisms over the RoBERTa encoder's output sequence $\mathbf{H} = \{\mathbf{h}_1, \ldots, \mathbf{h}_n\}$ to create a fixed-size representation:

- **Mean Pooling:** This method computes the element-wise average of all token embeddings (Reimers and Gurevych, 2019) to capture global context:

$$\mathbf{s} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{h}_i \tag{9}$$

  where n is the sequence length and $\mathbf{h}_i$ is the i-th hidden state.

- **Attention Pooling:** This mechanism dynamically assigns importance to tokens using a learnable context vector $\mathbf{u}_w$ (Yang et al., 2016). The representation $\mathbf{s}$ is computed through the following steps:

$$\mathbf{u}_i = \text{GELU}(\mathbf{W}_w \mathbf{h}_i + \mathbf{b}_w) \tag{10}$$

$$\alpha_i = \frac{\exp(\mathbf{u}_i^\top \mathbf{u}_w)}{\sum_{j=1}^{n} \exp(\mathbf{u}_j^\top \mathbf{u}_w)} \tag{11}$$

$$\mathbf{s} = \sum_{i=1}^{n} \alpha_i \mathbf{h}_i \tag{12}$$

where α_i denotes attention weights reflecting the affective importance of each token.

5 Experimental Setup

To systematically evaluate the effectiveness of different architectural components, we define five distinct model configurations for all subtasks:

- Model 1: RoBERTa Cardiff MLP
- Model 2: RoBERTa Cardiff Sparse MoEs and Mean Pooling
- Model 3: RoBERTa Cardiff Sparse MoEs and Attention Pooling
- Model 4: RoBERTa Cardiff Soft Gating MoEs and Mean Pooling
- Model 5: RoBERTa Cardiff Soft Gating MoEs and Attention Pooling

We split the official dataset by `user_id` using `GroupShuffleSplit`, creating training (85%) and local development (15%) sets to prevent user-level leakage. Final blind-test predictions were generated from models trained on this split. Main hyperparameters are listed in Table 6 (Section A). All experiments were conducted with PyTorch and HuggingFace Transformers on Kaggle using a single NVIDIA P100 GPU.

6 Results

6.1 Development Results

6.1.1 Subtask 1 – Longitudinal Affect Assessment

Table 1: Performance comparison on Subtask 1

Model/ Metrics	r_{comp}	mae_{comp}
Model 1	0.539	0.607
Model 2	0.587	0.538
Model 3	0.623	0.534
Model 4	**0.649**	**0.527**
Model 5	0.619	0.531

As shown in Table 1, we evaluate models using Average Pearson correlation (r_{comp}) and Average Mean Absolute Error (mae_{comp}). Model 4 (Soft-Gated MoEs + Mean Pooling) achieved the best performance, with the highest r_{comp} (0.649) and lowest mae_{comp} (0.527). This indicates that soft gating with mean pooling captures emotional states more effectively than the RoBERTa MLP baseline. Therefore, Model 4 was selected for Subtask 1.

6.1.2 Subtask 2A – State Change

Table 2: Performance comparison on Subtask 2A (State Change)

Model/ Metrics	r_{avg}	mae_{avg}
Model 1	**0.624**	0.971
Model 2	0.606	0.819
Model 3	**0.623**	**0.801**
Model 4	0.615	0.805
Model 5	0.608	0.874

As shown in Table 2, we evaluate models using Average Pearson correlation (r_{avg}) and Average Mean Absolute Error (mae_{avg}). Although Model 1 achieved a slightly higher r_{avg} (0.624), it had much higher error (mae_{avg} = 0.971). Model 3 (Sparse Experts + Attention Pooling) offered the best trade-off, with the lowest mae_{avg} (0.801) and nearly identical correlation (r_{avg} = 0.623). Therefore, Model 3 was selected for Subtask 2A.

6.1.3 Subtask 2B – Long-term State Change

Table 3: Performance comparison on Subtask 2B (Disposition Change)

Model/ Metrics	r_{avg}	mae_{avg}
Model 1	0.243	0.649
Model 2	**0.327**	**0.668**
Model 3	0.263	0.737
Model 4	0.287	0.759
Model 5	**0.327**	0.835

As shown in Table 3, we evaluate models using Average Pearson correlation (r_{avg}) and Average Mean Absolute Error (mae_{avg}). Model 2 (Sparse MoEs + Mean Pooling) was selected as the best configuration, achieving the highest correlation (r_{avg} = 0.327) with better stability (mae_{avg} = 0.668) than Model 5, which had the same correlation but higher error. Although Model 1 achieved slightly lower MAE, its weaker correlation (0.243) made Model 2 the best balance for forecasting long-term dispositional changes.

6.2 Official Test Results and Discussion

Table 4: Main results of the CITD@UIT system on the official SemEval-2026 Task 2 test set, reported using Pearson correlation (r) for Valence, Arousal, and their average (r_{avg}).

Subtask	$r_{valence}$	$r_{arousal}$	r_{avg}
Subtask 1	0.637	0.489	0.563
Subtask 2A	0.629	0.633	0.631
Subtask 2B	-0.169	-0.060	-0.114

Table 5: Detailed performance breakdown for Subtask 1 on the official blind test set. r_{bet} and r_{within} denote between-person and within-person correlations, respectively, while r_{comp} denotes $r_{composite}$. Bold values indicate the best performance in each subgroup; * indicates $p < 0.01$, and indicates $p < 0.05$ (significantly different from 0.0).

Category	Valence (V)				Arousal (A)			
	$r_{comp} \uparrow$	$r_{bet} \uparrow$	$r_{within} \uparrow$	$MAE \downarrow$	$r_{comp} \uparrow$	$r_{bet} \uparrow$	$r_{within} \uparrow$	$MAE \downarrow$
Seen User	**0.639**	0.740*	0.510*	0.563	0.405	0.508*	0.291*	0.421
Unseen User	0.620	0.666*	0.570*	0.436	**0.555**	0.690*	0.381*	0.400
Words Only	**0.693**	0.783*	0.576*	0.513	**0.526**	0.574*	0.476*	0.414
Essay Only	0.586	0.643*	0.522*	0.585	0.397	0.565*	0.199*	0.465

We evaluated the best validation configurations on the official blind test set: Model 4 for Subtask 1, Model 3 for Subtask 2A, and Model 2 for Subtask 2B. As summarized in Table 4, our systems achieved competitive results on Subtask 1 and Subtask 2A, while Subtask 2B remained substantially more challenging due to the difficulty of forecasting long-term dispositional changes. For a more detailed analysis of Subtask 1, particularly its longitudinal dynamics across user groups and text conditions, Table 5 presents an extended performance breakdown.

6.2.1 Subtask 1 – Longitudinal Affect Assessment

Performance. Our system achieved a highly competitive rank of 9th out of 26 teams, with an official r_{avg} of 0.563 in Table 4, significantly surpassing the 0.428 baseline. We observed a high degree of stability during the transition from the development phase ($r_{avg} = 0.649$) to the official blind test set. This relatively minor performance degradation validates the effectiveness of the Soft-Gated MoE architecture (Model 4) in learning robust, continuous affective representations.

Generalization to Unseen Users. A standout finding in Table 5 is the model's performance on *Unseen Users*. Notably, in the Arousal dimension, the correlation for Unseen Users (0.555) surpassed that of Seen Users (0.405). This provides strong empirical evidence that our MoE framework successfully extracts universal affective features rather than over-specializing in specific user histories or memorizing historical biases, directly addressing concerns regarding potential overfitting.

Signal Dilution and Contextual Window. The granular analysis reveals a significant performance gap between *Words Only* ($r_{comp} = 0.693$ for Valence) and *Essay Only* ($r_{comp} = 0.586$). We attribute this to a "signal dilution" effect, where the neutral narrative elements inherent in long-form ecological essays obscure dense emotional cues. This finding justifies our selection of a smaller window size ($k = 4$) for Subtask 1; by prioritizing immediate temporal context, the model effectively captures momentary affect while mitigating the noise introduced by irrelevant distant history.

Trait versus State Tracking. Our results show that the between-person correlation (r_{bet} up to 0.783) consistently outperforms the within-person correlation ($r_{within} \approx 0.510 - 0.576$). This indicates that while the system is exceptionally proficient at identifying a user's valence tracking subtle, high-frequency arousal fluctuations within a single individual remains a systemic challenge. This also explains our architectural decision to utilize separate regression heads; the independent MLP branches prevent the optimization of the more stable Valence dimension from being hindered by the higher noise levels inherent in within-person Arousal tracking.

Error Analysis. Despite the overall success, Arousal prediction remains more challenging than Valence, as reflected in the lower r_{within} scores. In the absence of explicit historical priors in the input, the model occasionally over-relies on lexical cues. It struggles when emotional intensity is conveyed through implicit situational context, such as descriptive stagnation, subtle shifts in narrative pace, or even non-lexical markers such as "!!!, grrrr!!!, @@QWERT", rather than standard vocabulary. Furthermore, preliminary internal benchmarks indicated that while the multi-layer bottleneck MLP ($768 \rightarrow 256 \rightarrow 128 \rightarrow 1$) adds complexity, it is essential for providing the non-linear expressivity required to resolve these intricate affective signals.

6.2.2 Subtask 2A – State Change

Performance. Our system achieved a distinguished Top 5 finish among 15 participating teams, with an official r_{avg} of 0.631. A remarkable observation is that the model's performance on the blind test set achieved its validation score ($r_{avg} = 0.623$). This positive delta (+0.008) underscores the exceptional generalization capability of the Sparse MoE + Attention Pooling architecture (Model 3) in tracking relative emotional shifts.

Strategic Temporal Context. A key differentiator for Subtask 2A was our decision to utilize a larger window size ($k = 8$), doubling the context used in Subtask 1. While momentary affect (Subtask 1) requires immediate focus to avoid narrative noise, detecting a "state change" is inherently longitudinal. Establishing a reliable emotional baseline requires a broader historical

horizon; thus, the $k = 8$ configuration provides the necessary temporal depth to distinguish between fleeting emotional fluctuations and meaningful shifts in a user's affective trajectory.

The Role of Attention Pooling. The effectiveness of Attention Pooling in this subtask highlights its ability to capture non-linear temporal cues. Unlike Mean Pooling, the Attention mechanism dynamically weights specific historical segments, allowing the model to focus on pivotal "turning points" in the ecological essays. This synergy between Sparse Experts (which specialize in different emotional patterns) and Attention Pooling (which identifies critical moments) proves highly effective for the state change regression task.

Error Analysis. Despite its high performance, the system occasionally struggles with subtle emotional transitions and the presence of sarcasm. In these complex linguistic scenarios, the Attention mechanism may over-focus on localized "trigger words" or non-lexical marker terms such as repetitive punctuation that may not represent the broader context. This often leads to an over-prediction of emotional intensity or a miscalculation of state change magnitudes. Furthermore, while the current architecture is robust, it lacks an explicit memory component for tracking emotional evolution over extended sequences, a limitation we aim to address in future work.

6.2.3 Subtask 2B – Long-term State Change

The Performance. Subtask 2B proved to be the most formidable challenge in the competition. While our Model 2 (Sparse MoEs + Mean Pooling) achieved a promising $r_{avg} = 0.327$ during the validation phase, it experienced a severe decline to -0.114 on the official blind test set. This delta of 0.441 suggests that the model failed to generalize the long-term dispositional trends.

Data Processing and Feature Fusion Imbalance. Our late-fusion mechanism concatenates the high-dimensional RoBERTa embedding $\mathbf{h}_{base} \in \mathbb{R}^{768}$ with a low-dimensional historical vector $\mathbf{v}_{past} \in \mathbb{R}^2$. We suspect this creates a dimensionality mismatch where the model either over-relies on the historical prior. This bias essentially forces the model to mirror past states, thereby overshadowing the subtle textual cues that indicate a long-term dispositional shift.

Potential Expert Collapse in Long Contexts. In the Sparse MoE architecture, the gating network $G(\mathbf{z})$ is responsible for routing tokens to specialized experts. However, across long-term ecological essays, we observed signs of expert collapse, where the gating distribution becomes overly sparse or biased toward a single generalist expert. This prevents the model from utilizing specialized sub-networks to capture the non-linear evolution of a user's disposition over multi-turn interactions.

Overfitting to Validation Trajectories. The longitudinal nature of Task 2B involves a limited number of unique users. There is a high probability that the model overfitted to specific emotional trajectories present in the training/validation splits. When faced with the blind test set which likely contains a significant distribution shift or different annotation variances the patterns learned by the MoE became counter-productive, leading to the observed negative correlation.

7 Conclusion

In conclusion, we presented a RoBERTa-based MoE framework for SemEval-2026 Task 2, achieving 9th place on Subtask 1 and 5th place on Subtask 2A. The results show that adapting MoE routing strategies (Soft/Sparse) and pooling mechanisms (Mean/Attention) to task-specific temporal dynamics can substantially improve emotion forecasting performance. However, although the framework is effective for short-term prediction tasks, it remains less robust for long-term dispositional forecasting (Subtask 2B), mainly due to limited long-range dependency modeling and sensitivity to distribution shifts. Future work will investigate memory-augmented transformers and temporal graph neural networks to better model the subtle longitudinal evolution of human disposition.

Acknowledgement

This research was supported by the VNUHCM-University of Information Technology's Scientific Research Support Fund. We thank the anonymous reviewers for their time and helpful suggestions that improved the quality of the paper.

References

Bagus Tris Atmaja and Masato Akagi. 2020. Combined learning of speech intensity and emotional valence-arousal. *IEEE Access*, 8:192801–192813.

Jose Camacho-collados, Kiamehr Rezaee, Talayeh Riahi, Asahi Ushio, Daniel Loureiro, Dimosthenis Antypas, Joanne Boisson, Luis Espinosa Anke, Fangyu Liu, Eugenio Martínez Cámara, and 1 others. 2022. TweetNLP: Cutting-edge natural language processing for social media. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–49, Abu Dhabi, UAE. Association for Computational Linguistics.

Damai Dai, Chengqi Deng, Chenggang Zhao, R.X. Xu, Huazuo Gao, Deli Chen, Jiashi Li, Wangding Zeng, Xingkai Yu, Y. Wu, Zhenda Xie, Y.K. Li, Panpan Huang, Fuli Luo, Chong Ruan, Zhifang Sui, and Wenfeng Liang. 2024. DeepSeekMoE: Towards ultimate expert specialization in mixture-of-experts language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23:1–40.

Christo Hristov, Ion Martinis, Yalcin Tur, Kevina Wang, and Miko Rimer. 2025. Text and valence-arousal: A two-dimensional foundational approach for mental health prediction. Technical report, Stanford University. Technical Report, CS 277 / BIODS 271.

Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. 1991. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87.

Arun Kashyap. 2024. Mental health chatbot using roberta and gemini. https://github.com/kashyaparun25/Mental-Health-Chatbot-using-RoBERTa-and-Gemini. GitHub repository.

Gonçalo Azevedo Mendes and Bruno Martins. 2023. Quantifying valence and arousal in text with multilingual pre-trained transformers. *arXiv preprint arXiv:2302.14021*.

Liyu Meng, Yuchen Liu, Xiaolong Liu, Zhaopei Huang, Wenqiang Jiang, Tenggan Zhang, Chuanhe Liu, and Qin Jin. 2022. Valence and arousal estimation based on multimodal temporal-aware features for videos in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 2360–2367.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

James A. Russell. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations (ICLR)*.

Nikita Soni, H. Andrew Schwartz, Ryan L. Boyd, Phi Long Bui, Syeda Mahwish, August Håkan Nilsson, Adithya V Ganesan, Lyle Ungar, Niranjan Balasubramanian, and Saif M. Mohammad. 2026. SemEval-2026 task 2: Predicting variation in emotional valence and arousal over time from ecological essays. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*. Association for Computational Linguistics.

Bazen Gashaw Teferra, Alice Rueda, Hilary Pang, Richard Valenzano, Reza Samavi, Sridhar Krishnan, and Venkat Bhat. 2024. Screening for depression using natural language processing: Literature review. *Interactive Journal of Medical Research*, 13:e55067.

Jun Xie, Yingjian Zhu, Feng Chen, Zhenghao Zhang, Xiaohui Fan, Hongzhu Yi, Xinming Wang, Chen Yu, Yue Bi, Zhaoran Zhao, Xiongjun Guan, and Zhepeng Wang. 2025. More is better: A MoE-based emotion recognition framework with human preference alignment. In *Proceedings of the 3rd International Workshop on Multimodal and Responsible Affective Computing (MRAC '25)*, pages 2–7. ACM.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489.

A Implementation Details

Table 6 summarizes the core training hyperparameters used across all three subtasks. We adjusted sequence length, window size, batch size, and learning rate according to the specific characteristics and computational requirements of each prediction setting.

Table 6: Hyperparameters

Subtask	Seq. Len.	Win. Size	Batch	LR
Subtask 1	384	4	16	2e-5
Subtask 2A	512	8	16	2e-5
Subtask 2B	512	–	4	3e-5

Subtask	Warmup	Opt.	N_Experts	Epochs
Subtask 1	0.1	AdamW	4	8
Subtask 2A	0.1	AdamW	4	8
Subtask 2B	0.1	AdamW	4	8

B Detailed Model Architectures

Figures 2, 4, and 4 illustrate the three prediction heads explored in this study: a RoBERTa-based multi-task regression architecture, a sparsely gated Mixture-of-Experts (MoE) head with Top-K routing, and a soft-gated MoE head with dense expert aggregation, respectively. Together, these architectures represent progressively more flexible strategies for modeling valence and arousal prediction from shared contextual encoder representations.

C Official Competition Results

This section summarizes the official results of SemEval-2026 Task 2 across all three subtasks, where Table 7 reports the rankings of top teams together with our official submission, **CITD@UIT**, which placed 9th of 26 teams in Subtask 1, 6th of 15 teams in Subtask 2A, and 11th of 12 teams in Subtask 2B, demonstrating competitive performance on dimensional affect modeling and short-term emotional state forecasting despite the substantially greater difficulty of long-term dispositional prediction tasks under realistic evaluation settings overall in practice across diverse users and temporal conditions.

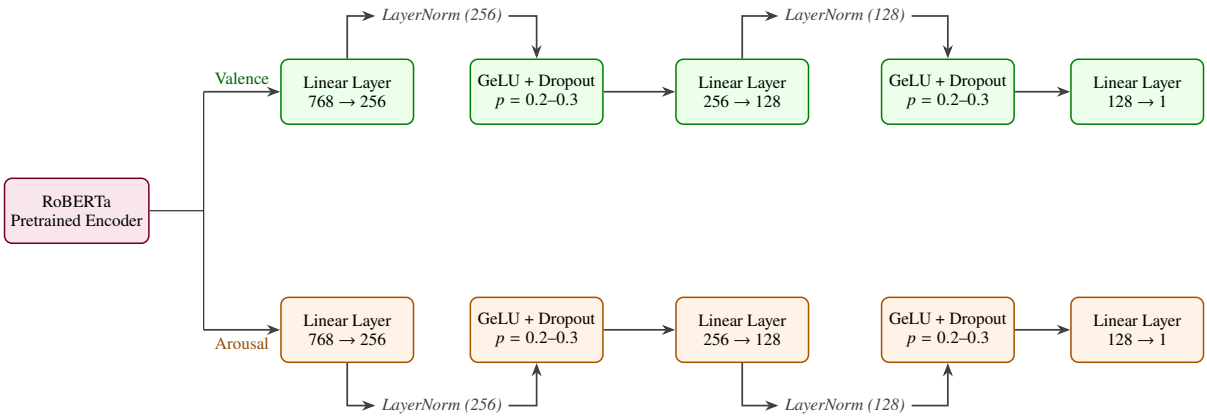

Figure 2: Architecture of the RoBERTa-based multi-task regression framework. A shared pretrained encoder is followed by two task-specific prediction heads for valence and arousal estimation, each composed of linear layers, LayerNorm, GeLU activation, and dropout regularization.

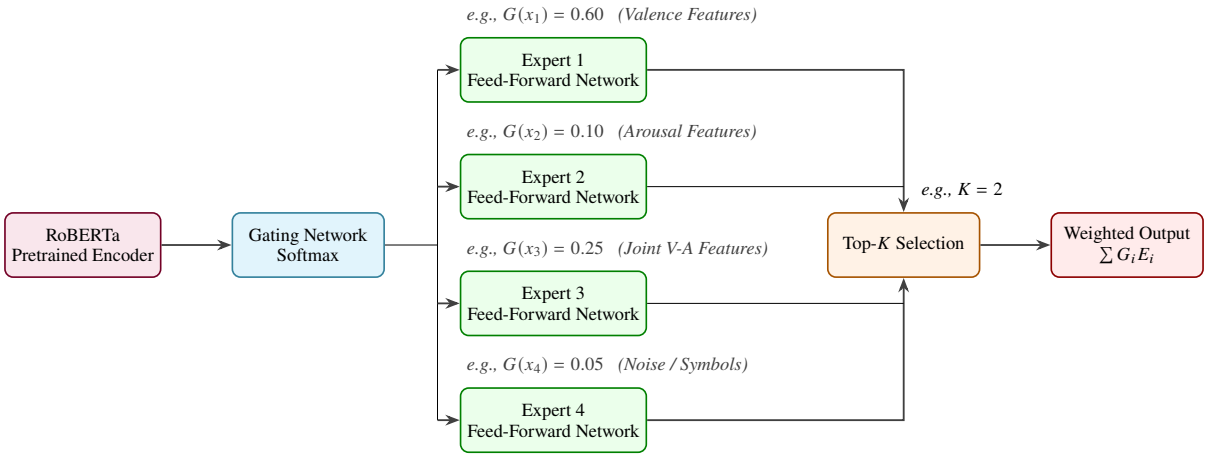

Figure 3: Architecture of the sparsely gated Mixture-of-Experts prediction head. A gating network assigns routing scores to multiple experts, after which Top-K selection activates only the most relevant experts to produce a weighted aggregated output. Numerical scores and feature descriptions shown above experts are illustrative examples only.

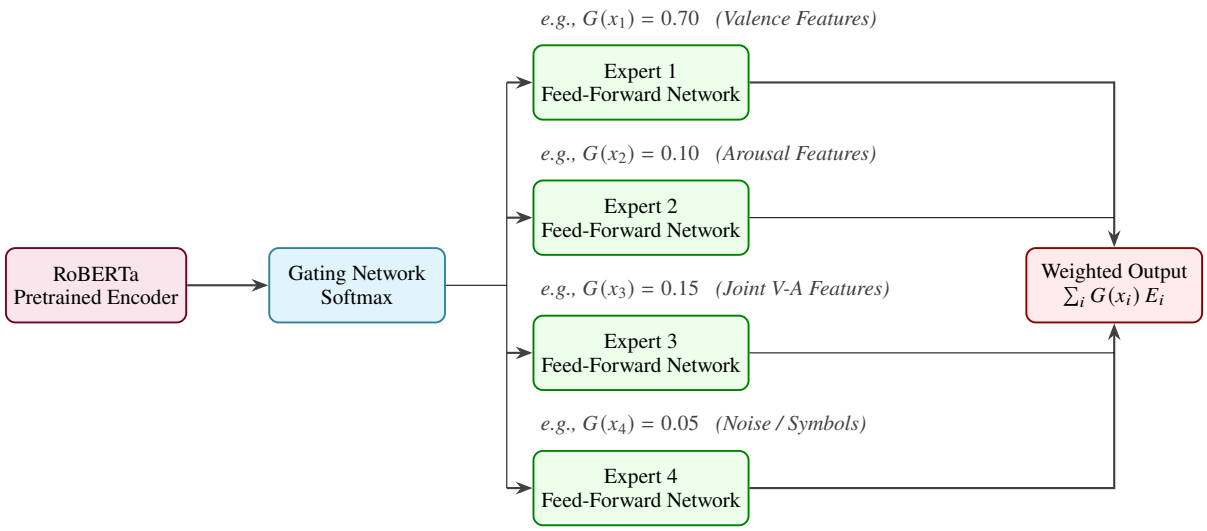

Figure 4: Architecture of the soft-gated Mixture-of-Experts prediction head. A gating network assigns dense routing weights to all experts, allowing every expert to contribute to the final prediction through a weighted summation. Numerical scores and feature descriptions shown above experts are illustrative examples only.

Table 7: Official leaderboard results for all three shared subtasks.

Top 10 Official Leaderboard for Subtask 1 Among 26 Participating Teams

Team	Valence (V)	Arousal (A)	V&A_average
	$r_{composite}$	$r_{composite}$	
UKP_Psycontrol	0.667	0.554	**0.611**
YNU	0.677	0.528	0.603
cclin	0.647	0.527	0.587
AFourP	0.679	0.466	0.573
lamanhnguyen	0.687	0.458	0.573
CSIRO-LT	0.656	0.488	0.572
CuriosAI	0.683	0.451	0.567
Bison AI4PC	0.665	0.468	0.567
CITD@UIT*	**0.637**	**0.489**	**0.563**
mcmaster4z03	0.665	0.460	0.562

Top 10 Official Leaderboard for Subtask 2A Among 15 Participating Teams

Team	Valence (V)	Arousal (A)	V&A_average
	r	r	
UKP_Psycontrol	0.675	0.683	**0.679**
YNU	0.692	0.647	0.669
UAlberta	0.615	0.674	0.645
linear(prev) (Baseline)	0.615	0.670	0.643
Ajman University	0.615	0.670	0.642
CITD@UIT*	**0.629**	**0.633**	**0.631**
CSIRO-LT	0.621	0.477	0.549
AI4PC - Howard U.	0.597	0.413	0.505
linear(BERT; prev) (Baseline)	0.430	0.405	0.418
Emo-tica	0.424	0.355	0.390

Top 10 Official Leaderboard for Subtask 2B Among 12 Participating Teams

Team	Valence (V)	Arousal (A)	V&A_average
	r	r	
linear(prev) (Baseline)	0.434	0.584	**0.509**
UAlberta	0.405	0.602	0.503
NLPGroup8	0.354	0.388	0.371
Emo-tica	0.257	0.418	0.337
AI4PC - Howard U.	0.046	0.348	0.197
Ajman University	-0.124	0.456	0.166
AGI	0.086	-0.081	0.003
rand (Baseline)	0.000	0.000	0.000
linear(BERT; prev) (Baseline)	-0.029	0.019	-0.005
EcoAffectTrack	-0.243	0.226	-0.009
	...		
CITD@UIT (11th)*	**-0.169**	**-0.060**	**-0.114**