Cherish at SemEval-2026 Task 2: Enhancing BERT-Based Models for Emotional Valence and Arousal Prediction in Ecological Essays with Personalized PLoRA and Temporal Embeddings

Cetta Reswara Parahita
School of Electrical Engineering and Informatics
Bandung Institute of Technology
Bandung, Indonesia
13521133@std.stei.itb.ac.id

Abstract

This paper describes the system developed by Team Cherish for SemEval-2026 Task 2: Predicting Variation in Emotional Valence and Arousal over Time from Ecological Essays. Our approach models emotional dynamics in user-generated text by incorporating both personalization and temporal information into a transformer-based architecture. We use RoBERTa-large as the backbone encoder and enhance it with PLoRA and a temporal embedding module. Cherish's model architecture is designed to maintain general semantic knowledge while subtly adapting to individual users and emotional shifts over varying temporal gaps. Our system achieved 13th place out of 29 teams in Subtask 1, obtaining a Pearson's r composite score of 0.596 for valence prediction and 0.505 for arousal prediction. While the team also participated in Subtask 2a, technical issues during inference led to zero variance in predictions, resulting in an undefined (NaN) official correlation score.

1 Introduction

SemEval-2026 Task 2 (Soni et al., 2026) aims to improve continuous emotion prediction from text using an English ecological longitudinal dataset. The task evaluates models on their ability to capture both individual and temporal emotional dynamics that influence emotional intensity and polarity. It consists of two main subtasks: (1) predicting the emotional valence and arousal expressed in a given text by a given user, and (2) forecasting future changes in valence and arousal for that user.

The task become important as it initiate to create model that understand how emotions evolve in real-world settings. By leveraging longitudinal dataset with self-reported affect, it enables the study of personalized and temporal emotional dynamics in natural writing. Such model will create crucial impact for different applications like mental health monitoring, well-being assessment, and affect-aware systems, where tracking emotional change over time provides more meaningful insight than single-text predictions (Zall et al., 2025).

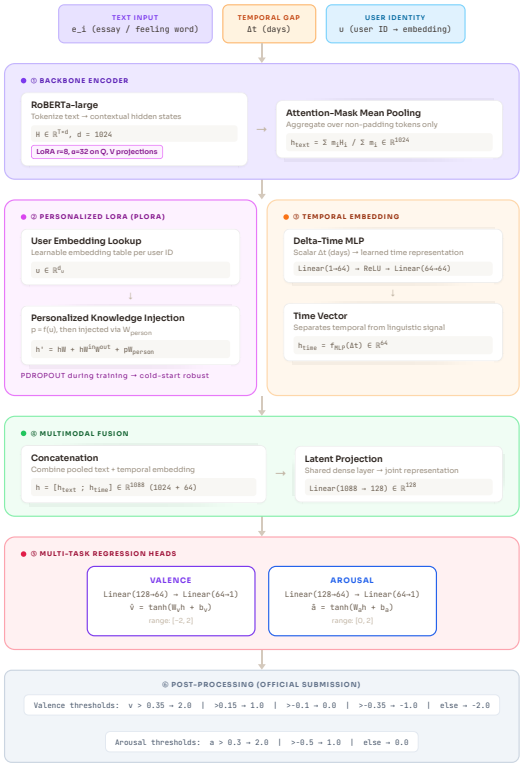

Figure 1: The Cherish Model Architecture

Acknowledging that standard language models primarily rely on lexical and semantic cues, it creates a gap between generalized text representations and the inherently personalized and temporally dynamic nature of emotional expression. In longitudinal settings, identical textual signals may reflect different affective intensities depending on the individual and the temporal context. Addressing this gap became the focus of our developed model.

Directly adapting the full backbone to user-specific patterns risks degrading its general semantic knowledge, while ignoring personalization lim-

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 2765–2770

its the model's ability to capture subjective emotional variation. To address this trade-off, Cherish's model adopt RoBERTa-large as the base encoder, providing a balanced trade-off between representational strength, stability, and computational feasibility. The base model then integrated with Personalized Low-Rank Adaptation (PLoRA), which enable user-conditioned parameter updates into selected attention neural network. We also incorporate a temporal embedding based on the delta time between consecutive texts from the same user to include the calculation of how emotional expression shifts across varying time intervals.

This design is motivated by the hypothesis that explicitly disentangling general semantic modeling from personalized and temporal adaptation enables more robust and context-aware emotional valence and arousal prediction, particularly in longitudinal ecological text settings.

2 Background

2.1 Task

Team Cherish participated in Subtasks 1 and 2a of SemEval-2026 Task 2.

2.1.1 Subtask 1: Longitudinal Affect Assessment

Given a chronological sequence of m texts, $e_1, \ldots, e_m$, the goal is to predict continuous Valence and Arousal (V&A) scores $(v_1, a_1), \ldots, (v_m, a_m)$ for each text, where $v \in [-2, 2]$ and $a \in [0, 2]$.

The evaluation involves two categories:
(1) Unseen users: Users not present in the training set. (b) Seen users: Users appearing in training but evaluated at future timesteps.

Example:

Input: Text ("I can't focus..."), Δ_{time}: 6.0, User ID: 17
Output: Valence: -2, Arousal: 2

2.1.2 Subtask 2: Forecasting Future Variation in Affect

Given a sequence of t texts and their gold V&A scores $(e_1, v_1, a_1), \ldots, (e_t, v_t, a_t)$, systems must forecast future emotional shifts.

Subtask 2a: State Change. This subtask requires predicting the immediate change in affect from the last observed timestep t to the next timestep $t + 1$:

$$\Delta_v^{(1)} = v_{t+1} - v_t, \quad \Delta_a^{(1)} = a_{t+1} - a_t$$

Example:

Input: e_{t-1} (v:-1.0, a:2.0), e_t (v:1.0, a:1.0)
Output: Forecasted changes $\Delta_v^{(1)} = -0.4$, $\Delta_a^{(1)} = 0.2$.

2.2 Dataset

We utilize the dataset provided by the SemEval-2026 Task 2 organizers, which consists of Ecological Momentary Assessment (EMA) texts collected in real-time from 182 users between 2021 and 2024. A key advantage of this dataset is that emotional labels are self-reported by the authors, minimizing annotator bias and capturing naturally occurring emotional microprocesses (Shiffman et al., 2008).

The full dataset contains 5,285 longitudinal observations. On average, users contributed 72.8 entries (median: 35). The training set comprises 2,764 observations from 137 users. The inclusion of both detailed essays and short "feeling words" (e.g., happy, calm) provides a diverse representation of affect across varying linguistic granularities.

Exploratory Data Analysis (EDA) revealed several data quality issues within the training set. We identified "keysmash" noise (e.g., Text ID 1636: "...content and calm. Jcjfjcjdjdncnfj...") and duplicate entries sharing identical User IDs and content. We address these issues during the preprocessing phase to ensure model robustness

3 System Overview

This experiment propose a personalized temporal emotion prediction framework that integrates PLoRA-based user adaptation and temporal embeddings into a RoBERTa-large backbone. The architecture (See Figure 1) disentangles general semantic modeling, user-specific adaptation, and temporal dynamics within a unified transformer framework. It creates robust continuous emotion prediction in longitudinal ecological settings.

3.1 Backbone Encoder

Team Cherish chooses RoBERTa-large as the textual encoder due to its strong contextual modeling capability and stable performance in regression settings. Given a tokenized input sequence $x = (x_1, \ldots, x_T)$, the encoder produces contextual hidden states $H \in \mathbb{R}^{T \times d}$, where $d = 1024$. To obtain a fixed-length representation, we apply attention-mask-aware mean pooling:

$$h_{\text{text}} = \frac{\sum_{i=1}^{T} m_i H_i}{\sum_{i=1}^{T} m_i},$$

where m_i denotes the attention mask value for token i. Mean pooling is chosen over CLS-only representations to provide a more stable aggregate signal for continuous affect regression.

3.2 Parameter-Efficient Personalization (PLoRA)

A key challenge in longitudinal emotion prediction is adapting general semantic knowledge to individual users without degrading shared representations. Fully fine-tuning the backbone for each user is computationally expensive and prone to overfitting given limited per-user data. To address this, we implement PLoRA, a plug-and-play mechanism that injects user-conditioned low-rank updates into selected attention parameters. Specifically, LoRA adapters (rank $r = 8$, $\alpha = 32$) are applied to the query and value projection matrices across all 24 attention layers of RoBERTa-large, allowing each layer to receive both task-specific and user-specific adaptation signals.

Task-Specific Adaptation Given an input representation h and backbone weight matrix W, the adapted transformation is:

$$h' = hW + hW_{\text{task}}^{\text{in}} W_{\text{task}}^{\text{out}},$$

where $W_{\text{task}}^{\text{in}} \in \mathbb{R}^{d \times r}$ and $W_{\text{task}}^{\text{out}} \in \mathbb{R}^{r \times d}$ are low-rank matrices with rank $r \ll d$, enabling efficient task adaptation without modifying the original backbone parameters.

Personalized Knowledge Injection (PKI) A user embedding u is transformed through a linear projection $f(\cdot)$ to produce a personalization vector $p = f(u)$, which is injected into the representation via a learnable projection matrix W_{person}:

$$h' = hW + pW_{\text{person}}.$$

Combined Formulation The final personalized representation integrates both signals:

$$h' = hW + hW_{\text{task}}^{\text{in}} W_{\text{task}}^{\text{out}} + pW_{\text{person}}.$$

PLoRA balances shared semantic modeling, task adaptation, and user-conditioned personalization within a unified framework (Zhang et al., 2024).

Personalized Dropout and Cold-Start Handling To improve robustness in cold-start scenarios, we apply Personalized Dropout (PDropout) during training, which randomly zeros out the user embedding p with probability p_{drop} at each forward pass. This prevents the model from over-relying on user-specific signals and forces the task-specific adaptation path to remain functional independently. At inference time, unseen users are handled by setting $p = \mathbf{0}$, effectively deactivating the personalization pathway so that predictions rely solely on the shared task adaptation, preserving reasonable performance in the zero-history setting.

3.3 Temporal Modeling

Emotional states evolve over time, and the interval between consecutive texts may influence affect intensity. Rather than expecting the transformer to infer temporal structure implicitly, we explicitly model time using a learnable embedding of the delta time Δt (measured in days). A two-layer MLP maps the scalar input to a 64-dimensional representation:

$$h_{\text{time}} = f_{\text{MLP}}(\Delta t) = \mathbf{W}_2\,\text{ReLU}(\mathbf{W}_1 \Delta t + \mathbf{b}_1) + \mathbf{b}_2,$$

where $\mathbf{W}_1 \in \mathbb{R}^{1 \times 64}$ and $\mathbf{W}_2 \in \mathbb{R}^{64 \times 64}$.

The temporal embedding is concatenated with the pooled textual representation and projected to a shared latent space:

$$h = \text{Linear}_{1088 \to 128}\big([h_{\text{text}}\,;\,h_{\text{time}}]\big),$$

where $[\ ;\ ]$ denotes concatenation along the feature dimension. This design explicitly separates temporal reasoning from textual encoding, reducing interference between linguistic and time-dependent features.

3.4 Multi-Task Regression Heads

Valence and arousal capture complementary but distinct affective dimensions. Valence correlates more strongly with lexical polarity cues, whereas arousal depends more on contextual intensity and discourse dynamics. To reflect this distinction, we employ separate regression branches for each dimension. The shared latent representation $h \in \mathbb{R}^{128}$ is projected through task-specific 64-dimensional embeddings and passed through linear heads with $\tanh$ activation:

$$\hat{v} = \tanh(W_v h + b_v), \qquad \hat{a} = \tanh(W_a h + b_a).$$

This separation enables partial parameter sharing while allowing each affective dimension to specialize its prediction pathway.

3.5 Subtask 2a: Autoregressive Affect Forecasting

For Subtask 2a, we extend the Subtask 1 architecture with a sequence forecasting framework by introducing a GRU-based autoregressive decoder. Given a history of t texts, each entry is encoded using the LoRA-adapted RoBERTa-large encoder. The resulting text representations are then fused with three additional signals via concatenation, followed by a linear projection into a shared 128-dimensional space:

$$z_i = \text{Linear}\big([h_{\text{text},i}\,;\, h_{\text{time},i}\,;\, p_i\,;\, (v_i, a_i)]\big),$$

where $h_{\text{text},i}$ is the mean-pooled encoder output, $h_{\text{time},i}$ is the temporal embedding of Δt_i, $p_i = f(u)$ is the user personalization vector, and (v_i, a_i) are the observed valence–arousal scores at step i. The sequence $\{z_i\}_{i=1}^{t}$ is processed by a GRU-based temporal encoder, whose final hidden state s_t summarizes recent affect dynamics and initializes the decoder.

The decoder then operates autoregressively:

1. At each step k, the GRU decoder receives s_{t+k-1} and the most recent VA estimate $(\hat{v}_{t+k-1}, \hat{a}_{t+k-1})$ as input.

2. A linear projection over the updated hidden state outputs a predicted increment $(\Delta\hat{v}_k, \Delta\hat{a}_k)$.

3. The next VA estimated as $(\hat{v}_{t+k}, \hat{a}_{t+k}) = (\hat{v}_{t+k-1} + \Delta\hat{v}_k,\ \hat{a}_{t+k-1} + \Delta\hat{a}_k)$.

This process repeats for n steps, enabling multi-step forecasting while preserving the personalized text-understanding backbone from Subtask 1.

4 Experimental Setup

Experiments are implemented in PyTorch using the HuggingFace Transformers and PEFT libraries.

Train Split & Validation The model is developed using the official training split with 5-fold GroupKFold cross-validation. Data are grouped by user ID to prevent user leakage between training and validation sets. The best-performing model across folds is selected based on the highest average composite Pearson's r score for valence and arousal.

Dataset Preprocessing Valence and arousal scores are normalized prior to training for regression stability. Before tokenization, we apply rule-based cleaning to reduce noise from keysmashing and irregular character patterns. Tokens longer than 15 characters are removed, and the remaining words are filtered using heuristic constraints on: (a) **vowel ratio** (> 0.2), (b) **repetition ratio** (< 0.4), (c) **character diversity** (> 0.3), and (d) **consonant ratio** (< 0.75). Cleaned texts are tokenized using the RoBERTa-large tokenizer with truncation at the maximum sequence length. Mean pooling over attention-masked hidden states is used to obtain fixed-length contextual representations.

Models We use RoBERTa-large[1] as our system backbone encoder.

Hyperparameters

- **LoRA:** rank 8, scaling factor 32, and dropout 0.1, applied to the query and value projection matrices across all attention layers.
- **Delta time:** modeled using a two-layer MLP that maps the scalar Δt to a 64-dimensional embedding, which is concatenated with the pooled textual representation and projected to a 128-dimensional latent space.
- **Prediction heads:** Separate 64-dimensional branches are used for valence and arousal, each followed by a linear regression head with $\tanh$ activation.
- **Training:** The model is trained using MSE loss and optimized with AdamW (learning rate 3×10^{-5}, weight decay 0.01, batch size 32) for 10 epochs.

Post-processing While the model produces continuous predictions, the official submission format requires discrete whole-number outputs. The threshold mapping applied during the official submission is shown in Table 1.

[1] https://huggingface.co/FacebookAI/roberta-large

Table 1: Threshold mapping for categorical submission, applied during official evaluation.

Dimension	Condition	Final Value
Valence	$v > 0.35$	2.0
Valence	$0.15 < v \leq 0.35$	1.0
Valence	$-0.1 < v \leq 0.15$	0.0
Valence	$-0.35 < v \leq -0.1$	-1.0
Valence	$v \leq -0.35$	-2.0
Arousal	$a > 0.3$	2.0
Arousal	$-0.5 < a \leq 0.3$	1.0
Arousal	$a \leq -0.5$	0.0

Table 2: Subtask 1 Results

	Valence (V)			Arousal (A)		
Metric	$r_c \uparrow$	$r_b \uparrow$	$r_w \uparrow$	$r_c \uparrow$	$r_b \uparrow$	$r_w \uparrow$
Corr	0.596	0.648*	0.538*	0.505	0.616*	0.375*
	MAE ↓			MAE ↓		
Err	0.614	1.021	0.361	0.234	0.477	–

5 Results

5.1 Subtask 1

The proposed model achieves competitive performance on both valence and arousal prediction. Detailed results are shown in Table 2. These results indicate that the model effectively captures both inter-user and intra-user affective variation, though future improvements are needed given that composite correlations remain in the 0.50-0.60 range.

5.2 Subtask 2a

Table 3 reports the official leaderboard scores for Subtask 2a. Both Pearson's r values are recorded as NaN because the model produced identical outputs across all test samples, causing the correlation to be undefined. The MAE scores reflect the magnitude of prediction error under constant output. This failure indicates that the personalization mechanism was not properly activated in the autoregressive decoding setting. Specifically, the user-embedding pathway $p = f(u)$ appears not to have propagated sufficient conditioning signal into the GRU-based decoder, causing it to collapse toward a near-constant mean prediction regardless of input context. After reflecting from latest model, we found several issue that causing the results come out as uniform value.

The "Over-Parenting" Trap at Autoregressive GRU Decoder This issue stems from excessive Teacher Forcing during training. By constantly "hand-feeding" the decoder ground-truth labels at every step, the model never developed the autonomy to correct its own errors. At inference, once this "parental" guidance was removed, small initial mistakes compounded into a total autoregressive collapse, causing the output to drift into a static, meaningless plateau. To resolve this, future iterations should implement scheduled sampling, gradually forcing the model to rely on its own predictions during training to build resilience.

Table 3: Subtask 2a Results

	Valence (V)		Arousal (A)	
Metric	$r \uparrow$	MAE ↓	$r \uparrow$	MAE ↓
Score	NaN	1.565	NaN	2.130

The Identity Blindness at Personalized Embedding Table (Input Layer) The model was unable to recognize User IDs absent from the training set. Without a robust cold-start guard, the embedding layer attempted to look up raw IDs that exceeded the table's dimensions, resulting in null or "garbage" personalization signals. Consequently, the model became "blind" to individual traits and defaulted to a generic population mean. To mitigate this, future work should implement Personalized Dropout (PDropout) during training to simulate unseen users and utilize a dedicated "Unknown" embedding slot to provide a stable, learned baseline for new users at inference time.

6 Conclusion

Modeling a continuous emotion prediction from text has several significant challenges. Emotional perception varies across individuals, while current language models inherently encode generalized patterns learned from large-scale data. This creates a gap between population-level representations and user-specific affective dynamics.

In this experiment, we addressed this challenge by extending a personalized RoBERTa-based architecture with a lightweight autoregressive decoder to model future valence-arousal trajectories. The system demonstrates competitive performance in Subtask 1, effectively capturing both between-user and within-user variability through the integration of textual, temporal, and user-level signals.

However, the results of Subtask 2a reveal a critical limitation. When personalization mechanisms are not properly activated, predictions collapse to near-constant outputs. This highlights the sensi-

tivity of personalized architectures to conditioning signals and the importance of robust identity modeling.

In future work, we aim to strengthen the personalization of our model through improved user conditioning strategies, more stable adaptation mechanisms, and better handling of cold-start scenarios. Additionally, exploring more expressive temporal decoders or personalization techniques may further enhance the modeling of complex affect dynamics.

Acknowledgments

At the heart of this journey lies my gratitude to Allah SWT, the Most Gracious and Most Merciful. His infinite blessings provided me with the clarity and resilience needed to navigate the simultaneous challenges I faced during this period. This competition, marking my maiden entry into the international AI research community, was made possible only through the ease and strength He granted me.

Besides, I want to express my sincere gratitude for Dr. Fariska Zakhralativa Ruskanda. This milestone would not have been possible without her mentorship that served as a compass throughout this experiment. Her technical insights and constant encouragement to step onto the global stage were invaluable to my growth.

To my parents and family, thank you for being my sanctuary and for your endless prayers that sustained me through every hurdle. Finally, to my friends, thank you for your companionship and support. This achievement is a testament to the love and support of everyone who stood by me.

References

Saul Shiffman, Arthur A. Stone, and Michael R. Hufford. 2008. Ecological momentary assessment. *Annual Review of Clinical Psychology*, 4:1–32.

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.

Raziyeh Zall, Alireza Kheyrkhah, Erik Cambria, Zahra Naseri, and M. Reza Kangavari. 2025. Intelligent agents with emotional intelligence: Current trends, challenges, and future prospects. *arXiv preprint*, arXiv:2511.20657.

Y. Zhang, J. Wang, L.-C. Yu, D. Xu, and X. Zhang. 2024. Personalized lora for human-centered text understanding. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence (AAAI-24)*, volume 38, pages 19588–19596. AAAI Press.