AFourP at SemEval-2026 Task 2: Predicting Variation in Emotional Valence and Arousal over Time from Ecological Essays

Shrika SP Thota, Lakshmi Priya Swaminatha Rao, Shivaanee SK, Thirumurugan RA, Vishal Muralidharan, Dhannya Santhakumari Madhavan
Department of Computer Science and Engineering
Sri Sivasubramaniya Nadar College of Engineering, Chennai - 603110
{shrika2310548, lakshmipriyas, shivaanee2310257}@ssn.edu.in
{thirumurugan2310277, vishal2310253, dhannyasm}@ssn.edu.in

Abstract

We describe our submission to SemEval-2026 Task 2 (Subtask 1), which asks systems to predict continuous Valence and Arousal scores from ecological diary texts. We fine-tune RoBERTa-base with a single linear regression head, treating each essay independently. Our system scores $r_{\text{composite}}$ of .679 (Valence) and .466 (Arousal) on the official test set, placing 4th on the Subtask 1 leaderboard.

1 Introduction

Deep contextualized language models have substantially changed how emotion is predicted from text. Earlier systems used polarity lexicons or shallow classifiers; pre-trained transformers encode richer contextual and pragmatic information that those approaches cannot access (Hutto and Gilbert, 2014; Mohammad and Turney, 2013; Vaswani et al., 2017; Liu et al., 2019).

SemEval-2026 Task 2 (Soni et al., 2026) introduces a unique challenge: modeling emotional variation longitudinally from ecological texts. Note that while referred to as "essays," the texts in this dataset vary widely, ranging from free-form narrative paragraphs to simple lists of feeling words. Unlike short social media posts, these entries describe lived experiences over time. The task requires predicting continuous Valence and Arousal scores for each text.

This setup introduces three modeling challenges:

1. Continuous multi-dimensional regression rather than discrete classification.
2. Longitudinal data with user-specific baselines (i.e., the natural resting state or average emotional level unique to an individual over time).
3. Narrative text containing temporally conflicting emotional cues (e.g., "I started the week feeling extremely anxious about the project, but by Friday, I was completely relaxed and overjoyed with the results").

Our approach fine-tunes RoBERTa-base with a linear regression head, treating each essay independently. We describe the architecture, training procedure, and ablation experiments, and discuss what the independent-essay assumption costs us in practice.

2 System Overview

Our system is a single-model pipeline for continuous VA regression. The key components are:

1. **Encoder:** RoBERTa-base (Liu et al., 2019), fine-tuned end-to-end on the training split.
2. **Tokenizer:** Byte-level BPE, truncated to 256 tokens.
3. **Regression head:** A single linear layer projecting the `[CLS]` representation to $\mathbb{R}^2$ (Valence, Arousal).
4. **Objective:** Mean Squared Error (MSE) minimized with AdamW.

The system treats each essay independently; no user identity or temporal ordering information is used at inference time.

Source code is available at: https://github.com/shrikathota/SemEval2026-Task2

3 Related Works and Background

Affective computing moved from discrete emotion labels to continuous dimensional scoring once annotated corpora and pre-trained models made regression-based approaches practical. Lexicon-based methods (Hutto and Gilbert, 2014; Mohammad and Turney, 2013) assign word-level sentiment scores aggregated over a text, but fail on context-sensitive phenomena such as irony and temporal sentiment reversal. Continuous VA annotation corpora—such as EmoBank (Buechel and Hahn, 2017), which provides valence, arousal, and dominance scores for 10,000 English sentences—enabled the transition to supervised regression-based affect prediction. Pre-trained transformers,

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 1695–1700

beginning with BERT (Devlin et al., 2019) and refined by RoBERTa (Liu et al., 2019) through improved pretraining objectives, encode contextual meaning via multi-head self-attention (Vaswani et al., 2017). This gives them the ability to handle irony and subtler emotional cues that lexicon-based methods miss. Continuous multi-output regression with a transformer encoder and regression head is now the common approach for predicting valence and arousal (Christ et al., 2024). SemEval-2026 Task 2 (Soni et al., 2026) applies this setup to longitudinal diary texts, where the added difficulty is that emotional trajectories vary per user and unfold over time.

3.1 Dimensional Affect and Continuous Modeling

Continuous dimensional spaces, grounded in the **circumplex model of affect** (Russell, 1980), capture emotional intensities that discrete labels flatten out. The framework places affect at a coordinate in two dimensions:

- **Valence:** Represents the horizontal axis, measuring the degree of pleasantness or "positivity" of an emotion. It ranges from negative/unpleasant (e.g., distress, sadness) to positive/pleasant (e.g., joy, serenity). In this task, valence is rated on a continuous scale of $[-2, 2]$.

- **Arousal:** Represents the vertical axis, measuring the level of physiological activation or "energy." It ranges from low/deactivated (e.g., boredom, calm) to high/activated (e.g., excitement, tension). In this task, arousal is rated on a continuous scale of $[0, 2]$.

Formally, we represent the affective state of a user u at time t as $y_{u,t} = (v_{u,t}, a_{u,t}) \in \mathbb{R}^2$. Shared transformer encoders with two-output regression heads have become the standard approach for predicting both dimensions jointly (Christ et al., 2024).

3.2 SemEval-2026 Task 2

SemEval-2026 Task 2 (Soni et al., 2026) asks systems to predict emotional valence and arousal variation over time from **ecological essays**. The entries are longitudinal diary-style texts written in real-world conditions—a design grounded in **Ecological Momentary Assessments** (EMA) (Shiffman et al., 2008), which capture psychological state close to the experience rather than from retrospective recall. Three difficulties follow: the labels are continuous, users have individual affective baselines (μ_u), and single entries often contain conflicting emotional cues.

4 Modeling Assumption

Each user's affect at time t can be written as $y_{u,t} = \mu_u + \delta_{u,t}$, where μ_u is a stable per-user baseline and $\delta_{u,t}$ is the essay-specific deviation. Our model does not have access to user identity—it maps each text directly to a VA score—so it cannot separate these two components. The consequence is a bias toward the training-set mean for users whose baseline sits far from the population average. We return to this in the Bias-Variance section.

5 Dataset Characterization

The SemEval-2026 Task 2 dataset (Soni et al., 2026) consists of ecological texts collected via Ecological Momentary Assessment (Shiffman et al., 2008) from participants over a multi-year period spanning 2021–2024. The training split comprises **2,764** texts from **137** users; the held-out test split comprises 1,737 texts from 91 users, with 46 users present in both splits.

5.1 User Distribution and Partitioning

The training set contains **2,764** texts from **137** users, with an average of **20.17** texts per user. Of the training entries, 1,433 (51.8%) are emotion word lists (`is_words=True`) and 1,331 (48.2%) are free-form prose. Training-set valence labels have mean 0.217 on the $[-2, 2]$ scale; arousal labels have mean 0.751 on the $[0, 2]$ scale. Across users, the standard deviation of per-user mean valence is 0.781; individual affective baselines differ substantially.

Our baseline architecture treats these texts under a conditional independence assumption—$p(y_{u,t}|X_{u,t}, X_{u,t-1}, \ldots) \approx p(y_{u,t}|X_{u,t})$—which we analyse in our discussion of representational bias.

5.2 Narrative Complexity

Texts in the dataset take two distinct forms: (1) word-list entries (51.8% of training data), such as *"tired, anxious, overwhelmed,"* which provide explicit affect signals without syntactic context; and (2) free-form prose entries (48.2%), which embed affect in narrative context. Prose entries vary considerably in emotional density. Some are direct

(e.g., *"I feel overwhelmed"*). Others are situationally implied, where a description of a professional setback carries clear affect without emotion keywords. A third pattern is temporal contrast—the author compares a past emotional state to a current one within a single entry (e.g., *"I started the week exhausted and dreading the deadline, but by Friday I felt relieved and proud"*). The hardest cases are mixed-valence texts, where positive and negative signals coexist and the model must resolve them into one VA coordinate. Self-attention handles the first three patterns reasonably well; mixed-valence prose remains the most challenging.

6 Model Architecture

6.1 RoBERTa Encoder

RoBERTa-base (Liu et al., 2019) is a 12-layer transformer encoder with 768-dimensional hidden states and approximately 125M parameters. Its multi-head self-attention handles long-range token interactions, which matters for narrative texts where emotional resolution may appear far from the triggering event.

6.2 Regression Head

The final-layer `[CLS]` representation $\mathbf{h}_{CLS} \in \mathbb{R}^{768}$ is projected to the VA space via a linear head:

$$\hat{y} = W\mathbf{h}_{CLS} + b \tag{1}$$

7 Learning Objective

We minimize Mean Squared Error (MSE) over the Valence–Arousal outputs, optimizing with AdamW and weight decay. The hyperparameters are detailed in Section 8.

8 Experimental Setup

8.1 Tokenization and Sequence Handling

We use byte-level BPE tokenization with a maximum sequence length of 256 tokens, truncating longer essays from the end. Most essays fall within 256 tokens, so truncation discards only tail content. The Tesla T4 GPU (16 GB VRAM) with batch size 8 fits either 128 or 256 tokens comfortably. The ablation (Section 11) shows 128 tokens performs comparably—slightly better on the validation set—which suggests emotional content is concentrated near the start of these entries.

8.2 Hyperparameters

- Learning rate: 2×10^{-5}
- Batch size: 8
- Epochs: 3
- Optimizer: AdamW

Training performed on Tesla T4 GPU.

9 Training Dynamics

Epoch	Training MSE
1	0.7725
2	0.5939
3	0.5013

Table 1: Training set MSE (combined Valence and Arousal) per epoch for RoBERTa-base fine-tuned over 3 epochs.

Training was stopped after 3 epochs as a fixed choice; no separate validation set was monitored during training, and additional epochs were not tested. The monotonic decrease indicates stable convergence without oscillation or divergence.

10 Multi-Seed Robustness

Seed	Valence MSE	Arousal MSE
42	0.4981	0.5127
123	0.4927	0.5098
2026	0.5019	0.5154

Table 2: Validation-set MSE for Valence and Arousal across three random seeds (RoBERTa-base, 3 epochs). Lower is better. Official leaderboard results are reported in Section 12.

The variance across seeds is low, consistent with a stable fine-tuning procedure.

11 Ablation Study

Configuration	r_V	r_A	r_{comp}	MAE
Linear Head	0.7309	0.5943	0.6626	0.6009
2-layer MLP	0.7191	0.5789	0.6490	0.5998
Frozen Encoder	0.1803	0.0996	0.1400	0.8609
Max Length 128	0.7487	0.6083	0.6785	0.5766

Table 3: Validation-set ablation results (RoBERTa-base, 3 epochs, 80/20 split). r_V, r_A: Pearson r for Valence and Arousal; r_{comp}: average of r_V and r_A; MAE: average of per-dimension Mean Absolute Error. Higher r and lower MAE are better.

Each configuration tests one design decision. The linear head (r_{comp} = 0.6626) is the simplest option. Adding a 2-layer MLP drops performance to r_{comp} = 0.6490 with nearly identical MAE (0.5998). Extra capacity appears to hurt on a training set this small (2,764 examples).

Freezing the encoder is more damaging: r_{comp} falls to 0.1400. Off-the-shelf RoBERTa representations are not sufficient; the encoder must be fine-tuned.

The 128-token run (r_{comp} = 0.6785, MAE = 0.5766) outperforms all others including the 256-token baseline. Emotionally relevant content in these entries is apparently front-loaded.

12 Results

12.1 Evaluation Metrics

The official evaluation uses two primary metrics (Soni et al., 2026): **Pearson correlation** (r), measuring the linear association between predicted and gold-standard scores; and **Mean Absolute Error** (MAE), measuring the average absolute deviation. The primary leaderboard ranking is determined by $r_{\text{composite}}$, the Pearson r reported separately for Valence and Arousal. These metrics are appropriate for continuous affect regression because they reward both correct relative ordering and accurate magnitude of emotional scores, unlike classification accuracy or MSE alone.

12.2 Official Leaderboard Performance

Table 4 reports our official submission results on the **test set** as scored by the SemEval-2026 Task 2 leaderboard. The official evaluation metrics are Pearson correlation (r), Mean Absolute Error (MAE), and a composite score ($r_{\text{composite}}$) combining both dimensions.

Team	r_{comp} (V)	r_{comp} (A)	Avg
UKP_Psycontrol	.667	.554	**.611**
YNU	.677	.528	.603
cclin	.647	.527	.587
AFourP (ours)	**.679**	**.466**	**.573**

Table 4: Official Subtask 1 leaderboard results ($r_{\text{composite}}$ for Valence and Arousal, and their average). AFourP ranks 4th overall.

12.3 Comparison with Organizer Baselines

Table 5 compares our system against the baselines provided by the task organizers (Soni et al., 2026) on the test set.

System	$r_{\text{composite}}$ (V)	$r_{\text{composite}}$ (A)
linear(BERT)	.557	.299
AFourP (ours)	.679	.466

Table 5: Comparison with organizer-provided Subtask 1 baselines on the official test set (Soni et al., 2026).

Note: Tables 1–3 use an 80/20 internal validation split and measure training behavior and design choices. Tables 4 and 5 report the official metrics on the held-out test set.

13 Bias-Variance Considerations

We analyze generalization from two angles: variance across random initializations, and bias introduced by the independence assumption.

13.1 Estimator Variance Across Seeds

Our multi-seed experiments (Table 2) show low variability: Valence MSE ranges from 0.4927 to 0.5019 ($\sigma \approx 0.0038$) and Arousal MSE from 0.5098 to 0.5154 ($\sigma \approx 0.0023$). The fine-tuning procedure is stable—not surprising given that RoBERTa's pretraining already provides a strong initialization before any task-specific gradient steps.

13.2 Model Bias from Independent Essay Modeling

The bigger issue is that our architecture treats each essay independently, ignoring temporal order and user identity. Formally, the conditional independence assumption is:

$$p(y_{u,t} \mid X_{u,t}, X_{u,t-1}, \dots) \approx p(y_{u,t} \mid X_{u,t}) \quad (2)$$

Since emotional baselines vary substantially across users—the standard deviation of per-user mean valence is 0.781—failure to model μ_u explicitly will systematically mis-estimate users at the emotional extremes.

13.3 Overparameterization and Implicit Regularization

RoBERTa-base has roughly 125M parameters for 2,764 training examples, yet the generalization gap is small: training MSE after 3 epochs is 0.5013 (Table 1) versus 0.497–0.515 on the validation set (Table 2). RoBERTa's large pretraining corpus

provides a useful prior for affective tasks; weight decay handles the rest.

14 Limitations

Our approach has four concrete limitations, most of which trace back to the same root cause: the model processes each essay without any knowledge of who wrote it or what came before.

14.1 Absence of Explicit Temporal Modeling

The dataset has longitudinal structure—each user contributes on average 20 entries—but we do not use it. Our system scores each text in isolation, with no access to what the same user wrote previously. A model that carries a hidden state $h_{u,t-1}$ forward across entries could capture gradual emotional drift; ours cannot.

14.2 Lack of User-Specific Calibration

Users have different affective baselines. In the training set, the standard deviation of per-user mean valence is 0.781 (on a $[-2, 2]$ scale); individual baselines clearly differ. A model that processes texts without knowing which user wrote them will systematically over- or under-predict for individuals at the emotional extremes.

14.3 Continuous Scaling for Varied Texts

The training set is evenly divided between emotion word lists (51.8%) and free-form prose (48.2%). The linear projection from `[CLS]` to VA scores may not equally handle these structurally different input types—a word list such as *"tired, anxious"* and a multi-sentence narrative may produce very different `[CLS]` representations that a single linear layer cannot disambiguate into the same continuous VA space.

14.4 Attention vs. Emotional Causality

Attention weights indicate which tokens the model focuses on, but not why those tokens drove a particular VA score. For the 48.2% of prose entries that may contain temporal contrasts or mixed-valence signals, we cannot determine from attention alone whether the model correctly resolved the final emotional state or averaged conflicting cues. This ambiguity limits interpretability for failure cases.

15 Ethical Considerations

The SemEval-2026 Task 2 dataset consists of longitudinal emotional self-reports, which are sensitive by nature.

The dataset is anonymized and participants consented to their texts being used for evaluation. Deploying automated emotion inference outside a research setting raises privacy concerns this submission does not address. These predictions are not clinical assessments—a continuous VA score derived from a diary entry reflects relative affective content in text, not a person's psychological state.

16 Future Work

A hierarchical encoder would be the simplest way to add temporal context: run each essay through the current model, then pass the per-essay representations through a second model ordered by time. This would track emotional drift without retraining the text encoder.

A simpler fix for the per-user baseline problem is a learned user embedding $g(u)$:

$$\hat{y}_{u,t} = f(X_{u,t}) + g(u) \tag{3}$$

This requires knowing user identity at test time, which is available for the 46 users shared between training and test splits.

Two other directions seem worth trying. Training on label deltas rather than absolute scores might better capture variation, which is what the task actually measures. Contrastive objectives could make the embedding space more meaningful—texts with similar affect should cluster together, not just share close label values. Bayesian or ensemble methods could add uncertainty estimates, which would matter when annotator agreement is low.

17 Conclusion

We fine-tuned RoBERTa-base with a linear regression head for continuous VA prediction from ecological diary texts, placing 4th on the SemEval-2026 Task 2 Subtask 1 leaderboard ($r_{\text{composite}}$ of .679 Valence, .466 Arousal).

Ablation results show one thing clearly: the encoder must be fine-tuned—freezing it collapses performance. The 128-token run actually outperforms 256 on the validation set, which was an unexpected finding after submission. Training is stable across random seeds. The main limitation runs through all the analysis: the model has no knowledge of who wrote the essay or what they wrote before.

Modeling user sequences, rather than individual essays, is the obvious path forward.

Acknowledgments

We thank the SemEval-2026 organizers (Soni et al., 2026).

References

Sven Buechel and Udo Hahn. 2017. EmoBank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 578–585, Valencia, Spain. Association for Computational Linguistics.

Lukas Christ, Shahin Amiriparian, Manuel Milling, Ilhan Aslan, and Björn Schuller. 2024. Modeling emotional trajectories in written stories utilizing transformers and weakly-supervised learning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7144–7159, Bangkok, Thailand. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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*, pages 4171–4186. Association for Computational Linguistics.

C. Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1):216–225.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

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

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.