EcoAffectTrack at SemEval-2026 Task 2: A Hierarchical DeBERTa-Transformer Framework with CCC Optimization for Longitudinal Affect Modeling

Diya Satish Kumar
Department of Computer Science and Engineering
SVNIT, Surat, India
u24cs020@coed.svnit.ac.in

Om Sujal Joshi
Department of Computer Science and Engineering
SVNIT, Surat, India
u24cs032@coed.svnit.ac.in

Abstract

Longitudinal affect modeling requires systems that can capture both the instantaneous emotional intensity of a text and the temporal evolution of a user's emotional state. In this paper, we describe *EcoAffectTrack*, our submission to SemEval-2026 Task 2: Predicting Variation in Emotional Valence and Arousal over Time from Ecological Essays. We propose a hierarchical framework that decouples text encoding from temporal forecasting. For affect assessment (Subtask 1), we fine-tune a **DeBERTa-v3** encoder optimized with a differentiable **Concordance Correlation Coefficient (CCC) Loss**, showing that direct metric optimization can lead to improved correlation compared to standard Mean Squared Error (MSE). For state change forecasting (Subtask 2A), we employ a **Transformer-based temporal forecaster with positional encoding** to mitigate inter-subject variability in emotional baselines. Our experimental results indicate that aligning the loss function with the evaluation metric and introducing task-specific temporal modeling are key factors for performance in longitudinal emotion recognition. Code is available at https://github.com/EcoAffectTrack-SemEval2026.

1 Introduction

The analysis of ecological essays presents a unique challenge in Affective Computing: unlike standard sentiment analysis—such as the specific sentiment track in the GLUE benchmark (Wang et al., 2018)—which often categorizes text into discrete buckets (positive/negative), longitudinal affect modeling requires tracking subtle, continuous fluctuations in *Valence* (positivity) and *Arousal* (activation dimension) over time. SemEval-2026 Task 2 formulates this as a regression problem across user timelines, distinguishing between a user's *disposition* (long-term trait) and their *state* (short-term mood).

A critical bottleneck in regression-based emotion detection is the choice of loss function (Atmaja and Akagi, 2021). Standard approaches typically rely on Mean Squared Error (MSE). While MSE is effective for minimizing point-wise errors, it assumes independent and identically distributed errors and fails to capture the *trend* or *variability* of emotional traces. In longitudinal tasks, preserving the "shape" of the emotional trajectory is often more important than the absolute values.

Our primary contribution is a shift from error-minimization to correlation-maximization, an approach increasingly favored in dimensional affect modeling because it prioritizes the trend of emotional traces over absolute point-wise accuracy (Lin, 1989; Atmaja and Akagi, 2021). By training our backbone encoder directly on a soft Concordance Correlation Coefficient (CCC) loss, we force the model to optimize for the competition metric itself, rather than a proxy like Mean Squared Error (MSE). Furthermore, we address the challenge of *subject variability*. Different users have different emotional baselines; a "neutral" text for one user might be "negative" for another. We propose a forecasting architecture that applies sequence-level normalization and temporal self-attention, allowing the model to learn relative emotional shifts rather than absolute embedding magnitudes. In this paper, we describe *EcoAffectTrack*, our submission to SemEval-2026 Task 2: Predicting Variation in Emotional Valence and Arousal over Time from Ecological Essays (Soni et al., 2026).

2 Related Work

Transformer-based Emotion Recognition Since the advent of BERT (Devlin et al., 2018), Transformer models have dominated emotion recognition tasks. Recent work has shown that *DeBERTa* (Decoding-enhanced BERT with Disen-

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 540–545

tangled Attention) (He et al., 2021) outperforms RoBERTa (Liu et al., 2019) and BERT on GLUE benchmarks due to its superior handling of positional semantics. However, most prior work optimizes these models using MSE or Cross-Entropy loss, which we argue is sub-optimal for continuous valence-arousal prediction.

Temporal Affect Modeling Modeling emotion over time typically involves Recurrent Neural Networks (RNNs) or Temporal Convolutional Networks (TCNs). Previous approaches in the Audio/Visual Emotion Challenge (AVEC) (Ringeval et al., 2019) have utilized LSTM-RNNs to smooth frame-level predictions. Our work differs by applying sequence-level normalization and temporal self-attention to standardize embedding scale and model emotional transitions across time.

Correlation Maximization in Affective Computing While standard regression often relies on MSE, recent studies in affective computing suggest that correlation-based loss functions are more robust for continuous valence-arousal prediction (Atmaja and Akagi, 2021). Techniques like Multimodal Functional Maximum Correlation (MFMC) (Zheng et al., 2025) have been proposed to maximize dependencies across signals, highlighting the benefits of optimizing for the "shape" of emotional responses rather than absolute values. Our work builds on this by integrating a differentiable CCC loss directly into a transformer-based hierarchical framework to better capture longitudinal dynamics.

3 System Architecture

Our system follows a hierarchical modeling paradigm that separates text-level affect estimation from user-level temporal and dispositional modeling. All subtasks share a common DeBERTa-based encoder backbone, while downstream architectures are specialized for their respective objectives.

3.1 Subtask 1: Affect Assessment

For text-level valence and arousal prediction, we fine-tune **microsoft/deberta-v3-base** (accessed via HuggingFace). Given an input essay, token embeddings from the final hidden layer are mean-pooled using the attention mask to produce a 768-dimensional representation.

To align training directly with the competition metric, we optimize a differentiable version of Lins Concordance Correlation Coefficient (CCC). For predictions $\hat{y}$ and ground truth y, the CCC loss is defined as:

$$\mathcal{L}_{CCC} = 1 - \frac{2\rho\sigma_{\hat{y}}\sigma_y}{\sigma_{\hat{y}}^2 + \sigma_y^2 + (\mu_{\hat{y}} - \mu_y)^2} \quad (1)$$

where μ, σ, and ρ denote mean, standard deviation, and Pearson correlation respectively. We train five independent folds across users (ensuring no user overlap between train and validation sets) and ensemble them at inference by averaging predictions. This stabilizes variance and improves robustness.

3.2 Subtask 2A: State Change Forecasting

Subtask 2A requires predicting the next-step emotional delta given a history of user essays. Instead of recurrent models, we employ a **Transformer-based temporal forecaster**.

For each user, we construct fixed-length sequences of the last $k = 8$ essay embeddings produced by the frozen Subtask 1 encoder. We set $k = 8$ empirically based on memory constraints and the average frequency of user posts within a localized emotional episode. Sequences for users with fewer than 8 historical essays are zero-padded. Each sequence has shape $[8, 768]$.

Positional Encoding Since Transformer encoders lack inherent sequential ordering, we add sinusoidal positional encoding to preserve temporal order information.

Temporal Modeling We use a 3-layer Transformer Encoder with:

- Hidden dimension: 768
- Number of heads: 8
- Feed-forward dimension: 2048
- GELU activation
- Dropout: 0.2

The final hidden representation of the last timestep is used as the summary vector of the users recent emotional trajectory.

Regression Head The summary vector is passed through a two-layer MLP mapping from $768 \rightarrow 256 \rightarrow 2$ dimensions to predict valence and arousal change. Input embeddings are normalized per sequence for stability during training. The model is trained using CCC Loss to directly optimize correlation with ground truth state changes.

3.3 Subtask 2B: Disposition Profiling

Disposition modeling differs from state forecasting in that it captures long-term user traits rather than short-term fluctuations. For each user, we aggregate up to 50 historical essay embeddings. This threshold covers the historical span of the majority of users in the dataset; timelines shorter than 50 are zero-padded.

Attention Pooling Instead of naive mean pooling, we apply a learnable attention mechanism. Given sequence embeddings $X \in R^{T \times 768}$, attention weights are computed as:

$$\alpha_t = \text{softmax}(W_2 \tanh(W_1 x_t)) \tag{2}$$

where W_1 and W_2 are learnable weight matrices and x_t is the embedding at timestep t. The final user representation is a weighted sum:

$$h = \sum_{t=1}^{T} \alpha_t x_t \tag{3}$$

This allows the model to focus on emotionally informative essays rather than treating all history equally.

Deep Regression Head The pooled representation is passed through a multi-layer perceptron mapping dimensions $768 \rightarrow 512 \rightarrow 128 \rightarrow 2$, utilizing Batch Normalization, GELU activation, and Dropout (0.3).

Loss Function Justification: While CCC was highly effective for Subtasks 1 and 2A due to the importance of capturing emotional trajectories, Subtask 2B predicts a single static disposition value per user. For static point-estimation where the concept of "trajectory shape" does not apply, minimizing absolute error via MSE remains the standard and most empirically stable approach.

4 Experimental Setup

4.1 Data Processing

All essays are tokenized using the DeBERTa-v3 SentencePiece tokenizer with a maximum sequence length of 256 tokens. This covers over 98% of essays without truncation.

Modality Awareness via Text Augmentation The dataset features heterogeneous inputs, comprising both free-form ecological essays and shorter "feeling word" lists. To ensure our representations remain robust across both short and long inputs, we dynamically apply text augmentation during data loading. For instances flagged as word lists, we prepend the contextual phrase "*I am feeling...*" to the input sequence. This transforms disjointed affective descriptors into syntactically valid sentences, allowing the DeBERTa encoder to process both modalities with consistent attention patterns.

User-level temporal sequences are constructed from cached encoder embeddings to reduce recomputation and ensure consistency across subtasks.

4.2 Training Configuration

Hyperparameters were selected via grid search on the validation folds. All experiments were conducted with fixed random seeds for reproducibility.

Subtask 1

- Optimizer: AdamW
- Learning rate: $2e^{-5}$
- Weight decay: 0.01
- 5-fold grouped cross-validation (by user ID)

Subtask 2A

- Optimizer: AdamW
- Learning rate: $1e^{-4}$
- Weight decay: $1e^{-2}$
- Early stopping with patience 7
- ReduceLROnPlateau scheduler

Subtask 2B

- Optimizer: AdamW
- Learning rate: $5e^{-5}$
- Weight decay: $1e^{-2}$
- Early stopping with patience 10

Objective	Valence (CCC)	Arousal (CCC)
MSE Loss	0.52	0.48
Huber Loss	0.55	0.51
CCC Loss	**0.71**	**0.65**

Table 1: Cross-validation results showing the superiority of metric-aligned training.

Subtask	Valence (r)	Arousal (r)
Subtask 1	0.663	0.373
Subtask 2A	-0.243	-0.011
Subtask 2B	-0.243	0.226

Table 2: Official leaderboard performance of EcoAffectTrack on SemEval-2026 Task 2.

4.3 Official Submission

The final system used a 5-fold ensemble for Subtask 1, a Transformer-based forecaster for Subtask 2A (window size 8), and a deep attention profiling network for Subtask 2B. The submitted system achieved **9th place overall** on the official SemEval-2026 Task 2 leaderboard.

5 Evaluation and Discussion

5.1 Ablation Study: Loss Functions

To validate our core contribution, we conducted an ablation study comparing standard MSE Loss against our proposed CCC Loss on Subtask 1.

As shown in Table 1, training with CCC Loss yielded a substantial improvement ($> 35\%$ relative gain) in correlation metrics. Qualitative analysis revealed that MSE-trained models tended to predict near the mean (low variance), a statistical phenomenon known as "regression to the mean" (Bland and Altman, 1994). In contrast, CCC-trained models successfully captured the peaks and valleys of the emotional signal.

5.2 Baseline Comparison

We first implemented a vanilla BERT (2018) baseline trained with Mean Squared Error (MSE) for all subtasks. This baseline directly fine-tuned the encoder for regression without explicit temporal modeling or user-level aggregation. While the baseline achieved reasonable correlation for Subtask 1 (Valence CCC of 0.41 compared to our proposed 0.71), it struggled significantly in Subtasks 2A and 2B.

Replacing BERT with DeBERTa-v3 (He et al., 2021) improved contextual representation quality and yielded consistent gains in valence and arousal correlation. Furthermore, switching from MSE to CCC Loss (Lin, 1989) substantially improved metric alignment. Finally, introducing a Transformer-based temporal forecaster (Vaswani et al., 2017) and an attention-based disposition profiling network enabled task-specific modeling, resulting in improved robustness compared to the naive baseline.

5.3 Official Leaderboard Results

As detailed in Table 2, our system ranked 9th overall on the official SemEval-2026 Task 2 leaderboard. While Subtask 1 performance was competitive, Subtasks 2A and 2B proved substantially more challenging due to limited historical context and high inter-user variability.

The results indicate that while metric-aligned optimization significantly improved text-level affect prediction (Subtask 1), forecasting and disposition modeling remain challenging problems. In particular, the negative correlations observed in Subtask 2A suggest that modeling short-term emotional change is highly sensitive to limited sequence history and data sparsity. Subtask 2B performance, while modest, demonstrates the benefit of attention-based aggregation over naive averaging approaches.

5.4 Error Analysis and Modality Impact

To contextualize the challenges encountered in temporal forecasting (Subtask 2A), we analyzed the impact of linguistic structure and input modality on prediction quality. The dataset consists of both narrative essays and affect-word lists, and our observations align with the hypothesis that narrative text introduces significant semantic variability that is difficult to model.

As shown in Table 3, arousal prediction is highly sensitive to syntactic complexity. Dense contextual embeddings may struggle when emotional states are conveyed implicitly through events or relational contexts rather than explicit affective phrasing. We hypothesize that texts rich in descriptive structure stabilize arousal predictions, whereas verb- and pronoun-heavy passages obfuscate the underlying emotional activation level.

Furthermore, the negative correlation in Subtask 2A is largely attributable to our choice of sequence window ($k = 8$). For users with sparse data, zero-padding diluted the temporal sig-

Input Text Example	Pred. Arousal	True Arousal
"I am feeling overwhelmed, stressed, exhausted" (Augmented Word List)	0.81	0.85
"It rained heavily today. The bus was late, which made the morning quite a blur, but the coffee helped." (Narrative)	0.22	0.65

Table 3: Qualitative comparison of arousal predictions showing model sensitivity to implicit narrative contexts.

nal. Even with sequence-level normalization, local windows of $k = 8$ often captured localized noise rather than the macro-emotional trend, causing the forecaster to predict inverse trajectories compared to the gold labels.

Limitations

A primary limitation of this work is the strict parameterization of historical context ($k = 8$ for Subtask 2A, up to 50 for Subtask 2B). This hard constraint restricted the model's ability to contextualize state changes for highly active users and resulted in signal dilution for sparse users. Furthermore, relying on pre-trained models like DeBERTa intrinsically assumes a standard grammatical structure; as demonstrated in our error analysis, highly idiosyncratic narrative styles found in ecological essays often obfuscate accurate arousal predictions.

6 Ethical Considerations

The development of longitudinal affect models raises important privacy and ethical concerns. While the dataset is anonymized, models that track emotional states over time could potentially be used for surveillance or manipulative targeting. We emphasize that such systems should only be deployed with explicit user consent and transparency. Furthermore, our model relies on pre-trained language models (DeBERTa) which may inherit demographic biases present in their training corpora; care must be taken to ensure the model does not exhibit demographic bias in its emotion predictions.

7 Conclusion

We presented a comprehensive framework for longitudinal affect assessment. Our experiments demonstrate that *how* a model is trained (loss function) is often more critical than *which* model is used. By optimizing directly for concordance and introducing task-specific temporal and aggregation modules, *EcoAffectTrack* demonstrates competitive performance for longitudinal affect modeling. In future iterations, we aim to integrate explicit psycholinguistic features (e.g., LIWC categories), part-of-speech frequencies, and trainable user-trait embeddings into the regression framework to better stabilize predictions against narrative complexity and inter-subject baseline shifts.

References

Bagus Tris Atmaja and Masato Akagi. 2021. Evaluation of error- and correlation-based loss functions for multitask learning dimensional speech emotion recognition. In *Journal of Physics: Conference Series*, volume 1896, page 012004. IOP Publishing.

J Martin Bland and Douglas G Altman. 1994. Regression towards the mean. *BMJ*, 308(6942):1499.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Lawrence I-Kuei Lin. 1989. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, pages 255–268.

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. *arXiv preprint arXiv:1907.11692*.

Fabien Ringeval, Björn Schuller, Michel Valstar, Nicholas Cummins, Roddy Cowie, Leili Tavabi, Maximilian Schmitt, Sina Alisamir, Shahin Amiriparian, Eva-Maria Messner, and 1 others. 2019. Avec 2019 workshop and challenge: State-of-mind, depressing with ai, and cross-cultural affect recognition. In *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*.

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, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*.

Deyang Zheng and 1 others. 2025. Multimodal functional maximum correlation for emotion recognition. In *arXiv preprint arXiv:2512.23076*.