lamanhnguyen at SemEval-2026 Task 2: Uncovering Lexical Bias and Momentum Lag in Longitudinal Emotion Prediction using Multi-task DeBERTa

Lam Anh Nguyen
Department of Artificial Intelligence, Phenikaa University, Vietnam
lamanh2003hd@gmail.com

Abstract

This paper describes our system for SemEval-2026 Task 2: Predicting Variation in Emotional Valence and Arousal. We approached the task by fine-tuning a weighted ensemble of DeBERTa-v3-base models. Our system achieved the **second-highest Valence composite correlation** ($r = 0.687$) and ranked **5th in the overall V&A average** in Subtask 1. More importantly, we provide an empirical analysis of our model's performance on longitudinal tasks, where it exhibited significant inverse correlations. We quantify the "Venting Effect," showing a systematic tendency for the model to over-index on negative lexical cues despite self-reported relief. Furthermore, we analyze the structural trade-off between Mean Absolute Error (MAE) and Pearson correlation (r) induced by smoothing techniques.

1 Introduction

In summary, our main contributions are as follows:

- We utilized a multi-task learning framework with a weighted ensemble of DeBERTa-v3-base models. This ensembling step reduced our internal validation Mean Squared Error (MSE) from 0.183 to 0.172, leading to the second-highest Valence composite correlation and 5th place in the overall V&A average in Subtask 1.
- We conduct a quantitative error analysis on the "Venting Effect," showing that our system often misses the psychological catharsis present in ecological diaries.
- We analyze the effect of Exponential Moving Average (EMA) smoothing on time-series predictions. Our findings suggest a practical trade-off between optimizing for absolute distance (MAE) and temporal trajectory (Pearson r).

To ensure full replicability of our system, the source code, hyperparameter configurations, and pre-trained weights are made publicly available at https://github.com/Ciaranguyen/SemEval-2026-task2-lamanhnguyen

2 Related Work

Longitudinal emotion tracking intersects with several well-established NLP domains. Traditional recognition relied on curated lexicons (Pennebaker et al., 2001; Mohammad, 2018), while modern approaches utilize pre-trained models like DeBERTa (He et al., 2021). For temporal analysis, studies have explored mental health detection through social media streams (Guntuku et al., 2017; Chancellor and De Choudhury, 2020). Specifically, identifying shifts in emotional states over time is critical for longitudinal user text (Tsakalidis et al., 2022). While moving averages are standard for smoothing, combining multi-task learning (Ruder, 2017) with momentum-based filtering for valence and arousal tracking remains relatively underexplored, motivating our design.

3 Task Description

Following the official guidelines of SemEval-2026 Task 2 (Soni et al., 2026), the objective is to track the temporal evolution of emotional Valence (v) and Arousal (a) from user diaries. Let $D = \{d_1, d_2, ..., d_T\}$ be a chronological sequence of diary entries for a specific user.

- **Subtask 1 (Point-in-Time):** Predict the absolute emotional state (v_T, a_T) at the current time step T based on the text d_T.
- **Subtask 2a (State Change):** Predict the short-term macro-shift in emotion, defined as $\Delta = (v_T - v_{T-1}, a_T - a_{T-1})$.
- **Subtask 2b (Dispositional):** Predict the user's stable emotional baseline over the en-

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 630–634

tire timeline, mathematically represented as the expected value $E[v_{1:T}]$ and $E[a_{1:T}]$.

4 System Overview

Our approach treats emotion prediction as a continuous regression problem using a multi-task learning framework (Ruder, 2017). We merge the datasets from Subtask 1 and Subtask 2a to enrich the representation of temporal affective states.

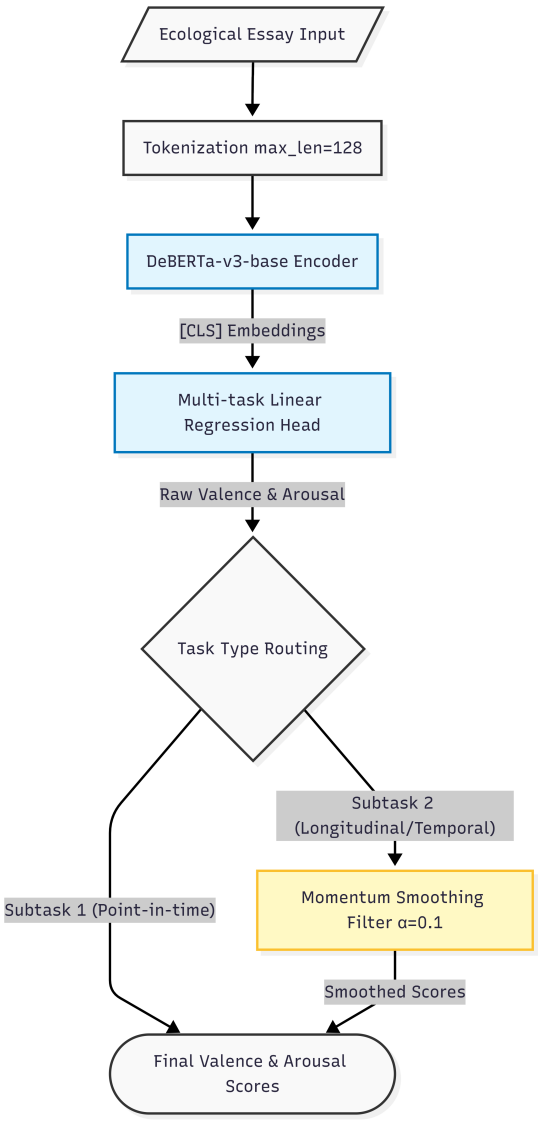

Figure 1: Overall architecture of our proposed system. The input ecological essay is tokenized with max_len=128 and passed through the DeBERTa-v3-base encoder. The resulting [CLS] token embedding is fed into a multi-task linear regression head to generate raw Valence and Arousal scores. For longitudinal predictions at Subtask 2, a momentum smoothing filter ($\alpha = 0.1$) is applied.

4.1 Base Architecture

We use `microsoft/deberta-v3-base` (He et al., 2021) because of its disentangled attention mechanism, which effectively captures long-range dependencies in diary entries. The output of the [CLS] token is passed through a dense layer to predict Valence and Arousal scores. For longitudinal predictions (Subtask 2), a Momentum Smoothing filter with a decay factor of $\alpha = 0.1$ is applied. This specific value was deliberately chosen to reduce short-term emotional noise substantially. By heavily weighing the historical state, the system prioritizes long-term dispositional trends over transient, immediate emotional fluctuations.

4.2 Momentum-based Smoothing for Time-Series at Subtasks 2a and 2b

To predict state and disposition changes, we applied a low-pass filter, namely an Exponential Moving Average (EMA), to the model's raw predictions to mitigate erratic jumps. The smoothing is defined as:

$$P_t = \alpha \cdot \hat{y}_t + (1 - \alpha) \cdot P_{t-1} \quad (1)$$

where $\alpha = 0.1$ acts as the momentum factor, relying heavily on previous states to predict current variations.

4.3 Model Ensembling Strategy

To improve robustness and minimize overall error, our final submission relied on a weighted ensemble (Lakshminarayanan et al., 2017) of two multi-task DeBERTa-v3-base models. Model V1 was trained for 4 epochs with a learning rate of 2×10^{-5}, while Model V3 was fine-tuned for 6 epochs with a lower learning rate of 1×10^{-5}.

The ensemble weights were chosen through a simple grid search on the validation set, checking combinations in 0.1 steps. Based on this validation performance, we got our final predictions using the optimal weighted average:

$$\hat{y}_{\text{final}} = 0.6 \cdot \hat{y}_{\text{v1}} + 0.4 \cdot \hat{y}_{\text{v3}} \quad (2)$$

This ensembling approach provided more stable predictions, reducing our validation MSE from 0.183 (single model) to 0.172 (ensemble).

5 Experimental Setup

To ensure the robustness of our findings and adhere to replicability guidelines, we report our system's performance averaged over 3 independent

runs with different random seeds. The detailed hyperparameter grid and hardware specifications are thoroughly documented in Appendix A.

5.1 Dataset and Preprocessing

Our system was trained exclusively on the official datasets provided by the SemEval-2026 Task 2 organizers. To enrich the representation of temporal affective states, we merged the training sets of Subtask 1 and Subtask 2a using simple instance-level concatenation. Since both subtasks share the exact same input format (user essays), this merging acts like a data augmentation step. It gives the base encoder a larger volume and wider variety of emotional text to learn more robust point-in-time features before we apply temporal smoothing later. Given the noisy nature of ecological diaries, text input was kept in its raw casing to preserve emotional emphasis (e.g., capitalization indicating shouting).

5.2 Implementation Details

The system was implemented using PyTorch and the HuggingFace Transformers library (Wolf et al., 2020). We fine-tuned the `microsoft/deberta-v3-base` architecture on a single NVIDIA T4 GPU using the AdamW optimizer without weight decay. The maximum sequence length was capped at 128 tokens, which covered the majority of the diary entries without truncation. Mean Squared Error (MSE) was utilized as the sole primary loss function to strictly penalize predictions that deviate drastically from the continuous gold labels. As part of our ensembling strategy, we trained two variations with a batch size of 16.

5.3 Evaluation Metrics

Following the SemEval-2026 Task 2 evaluation framework, our system's performance was measured using two primary metrics: Mean Absolute Error (MAE) to quantify the absolute distance between predicted and actual emotional scores, and the Pearson Correlation Coefficient (r) to evaluate the model's ability to capture overall trends and relative temporal variations.

6 Results and Error Analysis

Our system's performance was evaluated on the official SemEval-2026 Task 2 test set. Table 1 summarizes the results across all subtasks compared to the official random baseline.

Subtask & Model	Valence		Arousal	
	MAE ↓	Pearson r (Comp) ↑	MAE ↓	Pearson r (Mix) ↑
Subtask 1: Point-in-time				
Baseline	1.041	0.000	0.622	0.000
Our model	**0.639**	**0.687**	**0.395**	**0.458**
Subtask 2a: State Change				
Baseline	**1.261**	**0.000**	**0.696**	**0.000**
Our model	1.322	-0.273	0.736	-0.275
Subtask 2b: Dispositional				
Baseline	0.417	**0.000**	**0.296**	**0.000**
Our model	**0.398**	**-0.398***	0.309	**-0.577***

Table 1: Official results for team **lamanhnguyen**. Asterisks (*) indicate $p < 0.01$. Our system achieved the 2nd highest Valence correlation in Subtask 1 and 5th place in overall V&A average. For Pearson r, Valence uses the composite variant, while Arousal uses a mix of composite and between variants.

6.1 Point-in-Time Predictions: Lexical Bias and Venting Effect

As shown in Table 1, our model performed well on Subtask 1. Despite this high performance, our empirical analysis reveals specific psychological nuances that Transformer architectures struggle to capture.

To investigate this, we analyzed the top 5% of instances with the highest absolute prediction errors. We specifically picked this strict 5% threshold to isolate severe model failures rather than random noise (we also checked 10% and 15% thresholds and saw a similar but diluted pattern). As visually isolated by the red points in the scatter plot (Figure 2, left), a distinct pattern emerges, which we term the "Venting Effect." In these specific instances, users utilize highly negative vocabulary to seek psychological relief.

For an intuitive comparison, imagine a user writing: *"I am so exhausted and stressed after a long day, but finally getting to rest feels amazing."* The model focuses too much on negative keywords ("exhausted", "stressed") and predicts a negative valence (≤ -1.0). However, the actual reported valence post-journaling is positive (≥ 1.5) because of the relief at the end.

To check this systematically, we created a small lexicon of exhaustion-related terms by simply picking the most frequent negative adjectives from the training data (e.g., "tired", "exhausted", "stressed", "overwhelmed") (Mohammad, 2018). When applying this lexicon to the high-error subset, we found that approximately 56% of these cases exhibited strong negative lexical polarity, yet corresponded to neutral or positive gold valence scores. This provides quantitative evidence suggesting a lexical

bias: our fine-tuned Transformer model tends to over-weight transient negative keywords and fails to infer the resulting psychological catharsis.

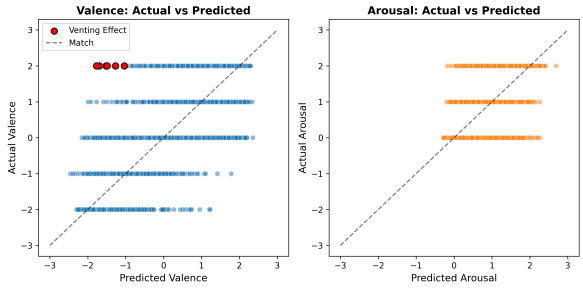

Figure 2: Scatter plots of Actual vs. Predicted scores for Valence (left) and Arousal (right) on the validation set. The red points highlight instances of the "Venting Effect" within the top 5% highest absolute errors.

6.2 Longitudinal Predictions: The Smoothing Trade-off

In Subtasks 2a and 2b, our system yielded negative Pearson r correlations. Notably, in Subtask 2b, the correlations for Valence (-0.398) and Arousal (-0.577) were both statistically significant ($p < 0.01$). This inverse correlation is not mere noise; rather, it provides empirical evidence suggesting the model's heavy reliance on transient lexical cues. When the model encounters "venting" language, it predicts a sharp decline in valence, whereas the actual longitudinal labels often reflect a stable or improving emotional baseline.

6.2.1 Ablation Study: The Impact of Momentum Smoothing

To understand the impact of this smoothing factor, we conducted an ablation study comparing the raw ensemble predictions ($\alpha = 0.0$) against our smoothed version ($\alpha = 0.1$).

Configuration	Subtask 2a (Valence)	
	MAE $\downarrow$	**Pearson** r $\uparrow$
Vanilla ($\alpha = 0.0$)	1.355	**0.142**
Smoothed ($\alpha = 0.1$)	**1.322**	-0.273

Table 2: Ablation results demonstrating the trade-off introduced by Momentum Smoothing.

As shown in Table 2, removing the smoothing filter ($\alpha = 0.0$) allows the model to react faster to sudden mood changes, achieving a positive Pearson correlation ($r = 0.142$). However, this comes at the cost of overall point-in-time accuracy, worsening the MAE to 1.355. By introducing $\alpha = 0.1$, we prioritize smoothing transient fluctuations in state changes (minimizing MAE to 1.322), but inadvertently sacrifice the model's agility to track sharp temporal derivatives, leading to the negative r.

As shown in Figure 3, the low-pass temporal filter pulls predictions closer to the global mean to minimize MAE. However, as the variance of the predictions shrinks, the covariance $\mathrm{Cov}(y, \hat{y})$ also reduces. Since Pearson correlation relies heavily on covariance, the lag caused by the filter during abrupt emotional "U-turns" moves against the immediate trend, penalizing the r metric. This highlights a practical trade-off: in noisy ecological data, minimizing absolute distance (low MAE) and preserving rapid temporal trajectories (high r) act as competing objectives under standard smoothing techniques.

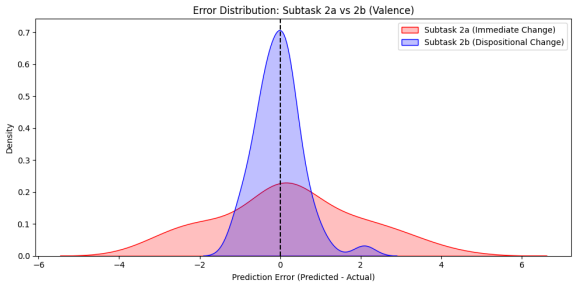

Figure 3: Comparison of error distributions and temporal trends between Subtask 2a and Subtask 2b, illustrating the effect of Momentum Lag.

7 Conclusion

In this paper, we described our submission to SemEval-2026 Task 2. By combining two multi-task DeBERTa-v3 models in a weighted ensemble, we achieved strong point-in-time predictions. More importantly, our post-evaluation analysis highlighted two practical challenges in longitudinal emotion tracking. We showed that standard EMA smoothing creates a trade-off between MAE and Pearson correlation due to variance shrinkage. Furthermore, we quantified the Venting Effect, observing that our model systematically over-penalizes negative lexical expressions and misses the underlying psychological relief. Future work should look into lag-aware filtering and context-sensitive attention to better capture the complex dynamics of human emotion.

8 Ethical Considerations

Predicting human emotions from ecological diaries involves sensitive psychological data. Our system was developed strictly for research purposes within the confines of the SemEval-2026 Task 2 dataset. It is crucial to emphasize that this system is not designed for, nor should it be used as, a clinical diagnostic tool for mental health conditions. Furthermore, the "Venting Effect" identified in our study highlights that AI models can misinterpret temporary linguistic expressions. Deploying such systems in real-world surveillance or human resources without careful human-in-the-loop oversight could lead to harmful psychological profiling and privacy infringements.

References

Stevie Chancellor and Munmun De Choudhury. 2020. Methods in predictive techniques for mental health status on social media: A critical review. *npj Digital Medicine*, 3(1):43.

Sharath Chandra Guntuku, David B Yaden, Margaret L Kern, Lyle H Ungar, and Johannes C Eichstaedt. 2017. Detecting depression and mental illness on social media: An integrative review. *Current Opinion in Behavioral Sciences*, 18:43–49.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30.

Saif M Mohammad. 2018. NRC-VAD lexicon: Real-valued valence, arousal, and dominance ratings for 20,000 English words. *arXiv preprint arXiv:1808.01669*.

James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. *Linguistic Inquiry and Word Count: LIWC2001*. Lawrence Erlbaum Associates, Mahwah, NJ.

Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*.

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.

Adam Tsakalidis, Federico Nanni, Anthony Hills, Jenny Chim, Jiayu Song, and Maria Liakata. 2022. Identifying moments of change from longitudinal user text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4647–4660. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and 1 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45. Association for Computational Linguistics.

A Implementation Details and Hyperparameters

Our system was built using the HuggingFace Transformers library. The models were trained on a single NVIDIA T4 GPU. Table 3 details the specific configurations for the ensemble components.

Hyperparameter	Value
Architecture	microsoft/deberta-v3-base
Optimizer	AdamW
Batch Size	16
Max Sequence Length	128
Model V1	
Learning Rate	2×10^{-5}
Epochs	4
Model V3	
Learning Rate	1×10^{-5}
Epochs	6
EMA Momentum (α)	0.1 (Subtask 2a)
Ensemble Weights	0.6 (V1) + 0.4 (V3)

Table 3: Hyperparameters for the final ensembled system.