NLP-FSDM at SemEval-2026 Task 2: Temporal Smoothing and CCC-MAE Optimization for Balanced Longitudinal Affect Assessment

Abdessamad Benlahbib[1], **Zouhir Essalmani**[1], **Achraf Boumhidi**[2], **Anass Fahfouh**[3], **Hamza Alami**[1]
[1] L3IA Laboratory, Faculty of Sciences Dhar EL Mehraz, USMBA, Fez, Morocco
[2]Department of Mathematics and Computer Sciences, National School of Applied Sciences Al Hoceima (ENSAH), UAE, Tetouan, Morocco
[3]Computer Science Department, Faculty of Sciences, UM5, Rabat
{abdessamad.benlahbib, zouhir.essalmani}@usmba.ac.ma,
{achraf.boumhidi, hamza.alami5}@usmba.ac.ma,
anassfahfouh@gmail.com

Abstract

This paper describes the NLP-FSDM system for SemEval-2026 Task 2, Subtask 1 on longitudinal affect assessment. The task requires predicting Valence and Arousal (V&A) scores for sequences of ecological essays and feeling words written over time. We adopt ModernBERT-large as a text encoder and formulate the task as a joint regression problem optimized using a Concordance Correlation Coefficient (CCC) loss combined with a lightly weighted Mean Absolute Error (MAE) term. To reduce variance induced by fine-tuning large transformers on relatively small user-specific datasets, we employ a three-seed ensemble. Finally, we introduce a lightweight post-inference temporal smoothing mechanism applied per user to improve within-user consistency. Our system achieves an $r_{composite}$ of 0.546 for Valence and 0.453 for Arousal, demonstrating stable cross-dimensional performance without explicitly modeling sequential dependencies.

1 Introduction

Longitudinal affect modeling differs substantially from traditional sentiment analysis (Pang et al., 2002; Benlahbib and Boumhidi, 2023; Cambria et al., 2017; Turney, 2002). Instead of predicting isolated polarity labels, systems must estimate emotional trajectories across time for individual users. In SemEval-2026 Task 2 (Soni et al., 2026), Subtask 1, participants are asked to generate real-valued Valence and Arousal scores for each text in a chronologically ordered sequence.

The evaluation combines two complementary perspectives: (1) *between-user* correlation, which measures the ability to distinguish individuals with different emotional baselines, and (2) *within-user* correlation, which evaluates how well the system captures fluctuations over time for the same individual.

The final leaderboard ranking is based on a composite correlation defined in the official evaluation script as:

$$r_{comp} = \tanh\left(\frac{\operatorname{arctanh}(r_w) + \operatorname{arctanh}(r_b)}{2}\right) \quad (1)$$

This formulation averages the two correlations in transformed space before mapping the result back to the original scale.

This evaluation setup creates an inherent trade-off. A model that focuses too much on distinguishing between users may struggle to capture changes over time, while a model that emphasizes short-term fluctuations may lose overall consistency across users. Our approach aims to balance these two aspects by combining correlation-aware training with a post-hoc temporal smoothing step.

Our system is based on ModernBERT (Warner et al., 2025) fine-tuned for multi-target regression of valence and arousal. To ensure consistency with the shared task evaluation protocol, we optimize a joint Concordance Correlation Coefficient (CCC) loss (Lin, 1989) instead of standard regression losses such as mean squared error. By directly maximizing concordance, the model learns to capture both linear association and distributional agreement between predictions and gold annotations. This evaluation-aligned training strategy improves robustness across users and temporal variations.

To ensure the reproducibility of our results, we have made the complete source code, including the training scripts and post-inference temporal

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 333–337

smoothing logic, publicly available.[1]

The remainder of this paper is organized as follows: Section 2 reviews related work in dimensional emotion modeling and longitudinal assessment. Section 3 describes the SemEval-2026 Task 2 dataset and its challenges. Section 4 details our methodology, including the ModernBERT-large backbone, our CCC-MAE joint loss function, and hardware optimization strategies. Section 5 presents our multi-seed ensemble and temporal smoothing post-processing. Section 6 reports the official results, followed by a detailed discussion of the Valence-Arousal performance gap in Section 7. Finally, Section 8 and Section 9 provide our conclusion, future work, and limitations.

2 Related Work

Emotion prediction in text has been extensively studied in both categorical and dimensional frameworks. Early work in dimensional emotion modeling often relied on lexicon-based approaches and regression over handcrafted features (Mohammad et al., 2018).

Transformer-based models such as BERT (Devlin et al., 2019) significantly improved performance in sentiment and emotion analysis by leveraging contextual representations. Subsequent research extended these models to continuous affect prediction tasks (Demszky et al., 2020).

The Concordance Correlation Coefficient (CCC) has been widely used in affective computing, particularly in multimodal emotion recognition challenges such as AVEC (Valstar et al., 2016), where correlation-based objectives were shown to better align with evaluation metrics than standard MSE.

Longitudinal modeling of psychological signals has traditionally relied on recurrent neural networks, particularly Long Short-Term Memory (LSTM) architectures (Hochreiter and Schmidhuber, 1997), to capture temporal dependencies. While such models are designed to learn sequential patterns, their effectiveness can be limited in user-specific settings where only a small number of observations is available per individual. In these low-resource scenarios, complex temporal models may struggle to generalize reliably. In our experiments, we find that a simpler non-sequential encoder combined with post-hoc temporal smoothing provides a strong and stable baseline when per-user data is limited.

[1]Link to Source Code

Our work follows this pragmatic direction: rather than explicitly modeling time during training, we stabilize predictions post hoc using lightweight temporal smoothing.

3 Task and Data

The dataset consists of 2,764 training observations from 137 users. Each instance contains a text (either an ecological essay or a set of feeling words), a timestamp, and gold valence and arousal scores.

The median number of texts per user is relatively small (31), making the task particularly sensitive to overfitting at the individual level. The presence of both short feeling words and longer essays further increases variance in representation quality.

4 Methodology

4.1 Validation Strategy

To approximate the official unseen-user scenario during development, we performed user-level splitting. Users were randomly divided into training and validation sets, ensuring that no user appeared in both. This prevents artificial inflation of within-user correlation during validation.

The final submitted model was trained on the full training dataset.

4.2 Text Encoder

We adopted ModernBERT-large as our backbone. Each text is processed independently, and no temporal information is injected during encoding. This design choice reduces model complexity and mitigates overfitting risks given limited per-user data.

4.3 Loss Function

The official ranking metric is based on Pearson correlation rather than squared error. To better align training with evaluation, we implemented a CCC-based objective:

$$\mathcal{L} = 0.5(1 - \text{CCC}_V) + 0.5(1 - \text{CCC}_A) + 0.01(\text{MAE}) \quad (2)$$

The CCC term encourages agreement in mean, variance, and covariance between predictions and ground truth. The lightly weighted MAE component stabilizes optimization and anchors predictions numerically.

4.4 Optimization Details

We trained using:

- Learning rate: $2e^{-5}$

- Effective batch size: 128 (via gradient accumulation)
- Mixed precision (fp16)
- Gradient checkpointing

Although training was configured for eight epochs, validation correlation consistently peaked at epoch four. We therefore fixed early stopping at epoch four for all seeds.

4.5 Hardware Optimization and VRAM Management

To leverage the representational power of ModernBERT-large (approximately 391M parameters) within the constraints of a single **NVIDIA Tesla P100 GPU (16GB VRAM)** via the Kaggle environment, several memory-optimization techniques were employed. Given the memory-intensive nature of calculating the CCC loss across large batches, we implemented *Gradient Accumulation* with 16 steps to achieve an effective batch size of 128 while maintaining a small per-device footprint.

Furthermore, we utilized *Gradient Checkpointing* to trade computational time for memory, allowing the model to fit within the P100's memory limits during the fine-tuning process. These optimizations were essential for maintaining the stability of the correlation-based loss function, which requires sufficiently large batch sizes to accurately estimate the variance and covariance of the predicted affect scores.

5 Ensemble and Post-Processing

5.1 Multi-Seed Ensemble

Three models were trained using seeds 42, 100, and 12345. Final predictions were obtained by averaging outputs across the three runs.

5.2 Temporal Smoothing

Predictions were sorted chronologically per user and smoothed using a rolling window of size three:

$$\hat{y}_{u,t} = \frac{1}{3} \sum_{i=0}^{2} y_{u,t-i} \tag{3}$$

This reduced abrupt fluctuations caused by lexical artifacts and improved within-user consistency.

To illustrate the stabilization effect, consider a sequence of raw Arousal predictions for a single user: $[0.42, 0.55, 1.85, 0.48, 0.52]$. The sharp spike to 1.85 (perhaps triggered by a single high-intensity word like "emergency") creates a discontinuous trajectory. After applying the rolling mean ($k = 3$), the smoothed sequence becomes $[0.42, 0.48, 0.94, 0.96, 0.95]$.

By redistributing the "energy" of the spike across the surrounding timestamps, the model favors emotional inertia over instantaneous lexical artifacts. This transition from a high-variance signal to a stabilized one directly contributed to our competitive r_{within} scores, as longitudinal affect is psychologically characterized by gradual transitions rather than stochastic shifts.

Finally, predictions were clipped to the official competition ranges to ensure validity. Valence scores were bounded to the interval $[-2.0, 2.0]$, while Arousal scores were clipped to $[0.0, 2.0]$. These constraints prevented the model from being penalized for extreme outliers generated during the regression process.

6 Results

Metric	Valence	Arousal
$r_{composite}$	0.546	0.453
$r_{between}$	0.660	0.588
r_{within}	0.408	0.293

Table 1: Performance of NLP-FSDM on Subtask 1.

Arousal prediction proved substantially more challenging than Valence across teams. Our balanced objective prevented severe degradation in the arousal dimension, resulting in competitive cross-dimensional performance.

7 Discussion

7.1 The Valence-Arousal Paradox

A primary observation in our results is the performance discrepancy between dimensions. While our system achieved a highly competitive $r_{composite}$ of 0.453 in Arousal, outperforming the top-ranked Valence system and the official linear baseline, our Valence score (0.546) trailed slightly behind the linear(BERT) baseline (0.557).

We hypothesize that this "Baseline Paradox" arises from a trade-off between stability and sensitivity. In the ecological essay dataset, valence is often expressed through clear and explicit emotional words. While our temporal smoothing mechanism ($k = 3$) helps reduce noise, particularly in

the Arousal dimension, it may also unintentionally soften sharp and isolated shifts in Valence. As a result, sudden changes in polarity can be slightly attenuated due to the smoothing effect.

7.2 Implicit vs. Explicit Affective Cues

Our system's superior performance in Arousal suggests that the ModernBERT backbone, combined with the CCC-loss, is exceptionally capable of capturing the implicit structural and contextual cues associated with emotional intensity. Unlike Valence, which often relies on word-level sentiment, Arousal is frequently encoded in sentence structure and length. These nuances were captured more effectively by our large-scale encoder than by the simpler baseline.

7.3 Feature Limitations

The current iteration of NLP-FSDM processed text independently of the provided metadata. Analysis suggests that incorporating the `is_words` boolean and the `timestamp` delta could have significantly improved the Valence results.

In particular, short "feeling words" entries are more likely to show sudden emotional shifts than longer essays. Using the same fixed smoothing window for both types of text probably over-smoothed the short entries. In addition, we did not take into account the time gaps between consecutive timestamps. As a result, the model could not adapt its temporal behavior: emotional states should reasonably vary more across long time intervals than across short ones.

8 Conclusion

We presented the NLP-FSDM system for longitudinal affect assessment. By combining CCC-based regression, multi-seed ensembling, and lightweight temporal smoothing, we achieved balanced performance across valence and arousal without complex temporal architectures.

Future work may explore user-adaptive embeddings and direct optimization of composite correlation.

9 Limitations

Despite the robust performance of the NLP-FSDM system, several limitations warrant acknowledgment. First, our model processes each text entry independently. While we mitigate this through post-hoc temporal smoothing, the architecture lacks a native mechanism (such as recurrent layers or temporal attention) to learn long-range dependencies during the gradient descent process.

Second, the relatively small size of the dataset (2,764 training instances) limited our ability to conduct thorough ablation analyses. The "ecological essays" and "feeling words" correspond to different writing styles, yet our model handles them in the same way. This uniform treatment may miss subtle differences in how users express emotional intensity in short entries compared to longer essays.

Finally, the computational overhead of ensembling three ModernBERT-large models increases inference latency. In a real-world longitudinal monitoring application, a more distilled or computationally efficient architecture might be required to provide real-time feedback to users.

References

Abdessamad Benlahbib and Achraf Boumhidi. 2023. NLP-LISAC at SemEval-2023 task 12: Sentiment analysis for tweets expressed in African languages via transformer-based models. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 199–204, Toronto, Canada. Association for Computational Linguistics.

Erik Cambria, Soujanya Poria, Alexander Gelbukh, and Mike Thelwall. 2017. Sentiment analysis is a big suitcase. *IEEE Intelligent Systems*, 32(6):74–80.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online. 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 *NAACL*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Lawrence I-Kuei Lin. 1989. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, 45(1):255–268.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. SemEval-2018 task 1: Affect in tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 1–17, New Orleans, Louisiana. Association for Computational Linguistics.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 79–86. Association for Computational Linguistics.

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.

Peter Turney. 2002. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 417–424, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Michel Valstar, Jonathan Gratch, Björn Schuller, Fabien Ringeval, Denis Lalanne, Mercedes Torres Torres, Stefan Scherer, Giota Stratou, Roddy Cowie, and Maja Pantic. 2016. Avec 2016: Depression, mood, and emotion recognition workshop and challenge. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, AVEC '16, page 3–10, New York, NY, USA. Association for Computing Machinery.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. 2025. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2547, Vienna, Austria. Association for Computational Linguistics.