Emo-tica at SemEval-2026 Task 2: Trait–State Affect Forecaster for Longitudinal Valence and Arousal

Sadia Noor and Mehwish Fatima
School of Electrical Engineering and Computer Science (SEECS),
National University of Sciences and Technology (NUST), Islamabad, Pakistan
{snoor.msds23seecs, mehwish.fatima}@seecs.edu.pk

Abstract

Modeling longitudinal affect requires capturing both stable user tendencies and transient textual signals. For SemEval-2026 Task 2, we propose the **Trait–State Affect Forecaster (TSAF)**, which decomposes affect into persistent user traits and text-conditioned states integrated via adaptive gating. On per-text prediction (Subtask 1), TSAF achieves composite Pearson correlations of 0.645 (valence) and 0.409 (arousal), outperforming the Linear(BERT) baseline. In forecasting tasks, results reveal strong short-term affective inertia, where prior affect dominates next-step prediction, while long-term drift remains challenging under sparse supervision; TSAF shows stronger gains for arousal in this setting. Analyses across user splits and modalities highlight the benefits and limitations of explicit trait–state modeling, particularly under cold-start and short-text conditions.

1 Introduction

Dimensional affect modeling represents emotion along continuous valence and arousal axes grounded in psychological theory (Russell, 1980; Bradley and Lang, 1994). Transformer-based models perform well on text-based valence–arousal regression (Mendes and Martins, 2023), but typically treat each instance independently.

This assumption breaks in longitudinal settings, where affect evolves over time and varies across individuals. In reflective writing and mental health contexts, emotional expression reflects both stable user-level tendencies and transient contextual signals. While recent work explores behavioral sequence modeling (Ganesan et al., 2026), it rarely disentangles persistent baselines from dynamic textual effects.

Longitudinal affect data further exhibit irregular sampling, variable sequence lengths, and sparsity, limiting fully sequential approaches and motivating models that explicitly separate stable traits from context-dependent states.

SemEval-2026 Task 2 (Soni et al., 2026) formalizes this setting via (1) per-text valence–arousal prediction, (2A) short-term state change forecasting, and (2B) long-term dispositional drift prediction under seen and unseen users—requiring joint modeling of affective dynamics and user heterogeneity under sparse, irregular observations.

We propose the **Trait–State Affect Forecaster (TSAF)**, which decomposes affect into stable user traits and text-conditioned states. TSAF learns user embeddings for persistent baselines and uses a contextual encoder for momentary signals, combined via an adaptive gating mechanism for instance-specific integration. This structure aligns with metrics separating within- and between-person performance. TSAF supports per-text estimation (Subtask 1), while forecasting (Subtasks 2A/2B) is handled through feature-based temporal modeling, enabling robust generalization under heterogeneous and sparse user trajectories.

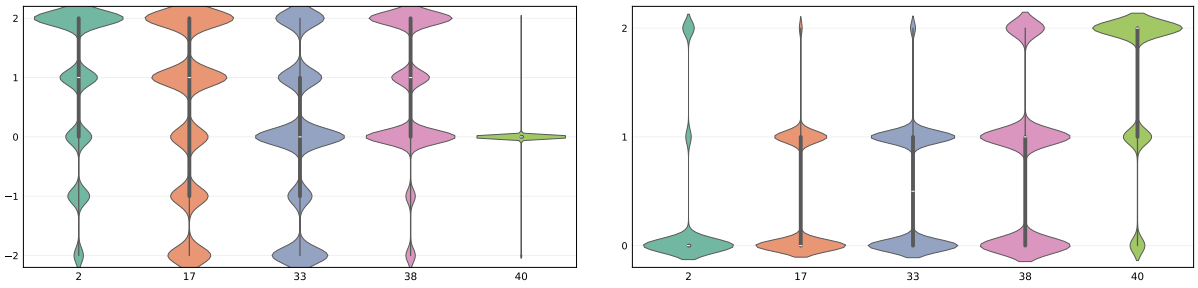
2 Related Work

Prior work spans dimensional affect modeling, longitudinal sequence modeling, and personalization via trait–state representations.

Dimensional Affect Modeling. Emotion modeling in NLP commonly uses continuous valence–arousal (V&A) representations (Russell, 1980; Bradley and Lang, 1994; Mendes and Martins, 2023; Becker et al., 2026). Early methods rely on lexicons and shallow regression (Lin et al., 2024; Mitsios et al., 2024), while transformer-based models provide strong contextualized predictions (Ahire et al., 2025). However, they largely treat texts independently, ignoring longitudinal structure.

Longitudinal Modeling and Forecasting. Recent work models behavioral sequences (Ganesan et al., 2026) using temporal architectures such as hierarchical transformers and change-detection models

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 213–220



(a) Valence **(b)** Arousal

Figure 1: Per-user affect distributions. Valence (range $[-2, 2]$) shows higher variability, while arousal (range $[0, 2]$) is more concentrated but retains between-user differences, motivating trait-based modeling.

(Tseriotou et al., 2023; Hills et al., 2024). Forecasting remains challenging due to affective inertia and user heterogeneity (Ganesan et al., 2026; Tripodi et al., 2025). Many approaches rely on autoregressive or population-level assumptions (Tommasel et al., 2021; Matero and Schwartz, 2020), limiting personalization.

Personalization and Trait–State Modeling. Psychological theory distinguishes stable traits from transient states (Steyer et al., 1999; Wurpts, 2015). NLP personalization often uses user embeddings or identifiers (Mireshghallah et al., 2022; Soni et al., 2025; Shu, 2024), but typically entangles persistent and contextual signals (Harry et al., 2026). We instead explicitly separate trait and state components and combine them via adaptive gating, aligning with within- and between-person variance modeling (Schuurman, 2023; Hoffman and Stawski, 2009).

Overall, prior work either models text without personalization or incorporates user information without explicit trait–state decomposition. Our approach directly models and integrates both components in a structured manner aligned with longitudinal evaluation.

3 Dataset and Analysis

The SemEval-2026 Task 2 training set contains 2,764 longitudinal texts from 137 users, annotated with valence $[-2, 2]$ and arousal $[0, 2]$, along with user IDs, timestamps, and modality labels.

User Heterogeneity. User activity is highly imbalanced (2–206 texts per user; mean 20.2). Affective variability differs substantially across users (valence std: 1.05; arousal std: 0.67), with clear between-user differences in range and dispersion (Figure 1(a–b)), motivating explicit modeling of stable user traits.

Bimodal Text Structure. The dataset includes 1,433 feeling-word entries and 1,331 essays. Essays are substantially longer (up to 225 words; 75th percentile: 52), providing richer context, while feeling-word entries are sparse. Despite this, both modalities occupy similar regions in the valence–arousal space (Figure 3), indicating differences arise from expressive density rather than label distribution.

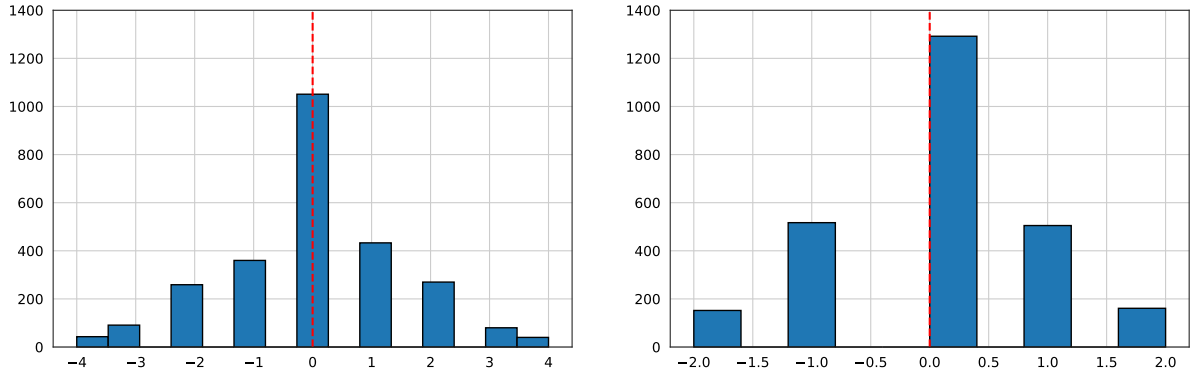
Temporal Sparsity and Inertia. Data are collected in discrete waves with uneven temporal coverage and late-phase concentration (Figure 4), introducing non-stationarity. Step-wise affect changes are strongly concentrated near zero (Figure 2(a–b)), indicating pronounced short-term inertia. Combined with irregular sampling, this limits fully sequential approaches and motivates temporally-aware but robust modeling.

These properties motivate trait–state decomposition, modality-aware representations, and non-sequential temporal modeling under sparse, heterogeneous user trajectories.

4 Trait–State Affect Forecaster (TSAF)

Figure 5 illustrates the proposed **Trait–State Affect Forecaster (TSAF)**, which decomposes affect prediction into a *trait* component capturing stable user baselines and a *state* component modeling text-dependent fluctuations. These correspond to between-user and within-user variation and are combined via an adaptive gating mechanism. While TSAF refers to the full system submitted across all subtasks, the neural trait–state model with adaptive gating applies specifically to Subtask 1. For Subtasks 2A and 2B, temporal forecasting is handled by separate feature-based models (Ridge regression and LightGBM), described in Sections 4.2 and 4.3. Code is publicly available[1].

[1] Code repository



(a) Valence change

(b) Arousal change

Figure 2: Distribution of step-wise affect changes (Subtask 2A). Both valence and arousal are strongly concentrated near $\Delta = 0$, indicating pronounced short-term affective inertia.

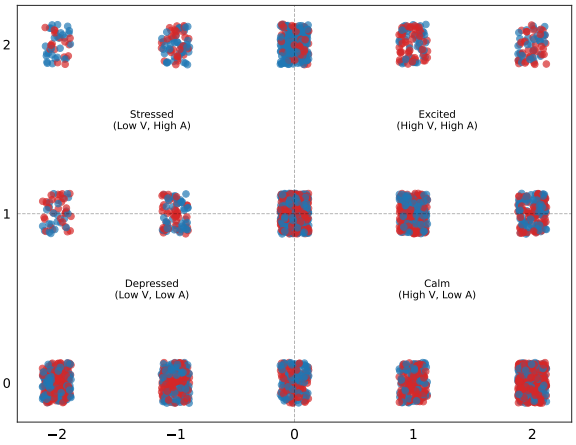


Figure 3: Valence–arousal coverage by modality. Both modalities span similar affective regions despite large length differences, indicating variation arises from expressive density rather than label distribution.

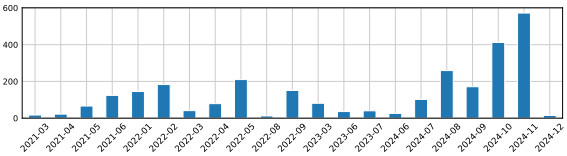


Figure 4: Monthly distribution of training texts showing uneven activity with late-phase concentration, indicating wave-based collection and temporal imbalance.

4.1 Subtask 1: Trait–State Affect Modeling

Given user u and text $x_{u,t}$, TSAF predicts valence and arousal as a gated mixture of trait and state signals.

State Representation. Each text is encoded using DistilBERT-base-uncased with masked mean pooling, yielding $\mathbf{h}_{u,t} = \text{Encoder}(x_{u,t}) \in \mathbb{R}^d$.

Trait Representation. Each user u is assigned a learned embedding $\mathbf{z}_u \in \mathbb{R}^k$, trained jointly to capture stable affective baselines. Unseen users share a common UNK embedding.

Regression Heads. Separate linear heads produce affect estimates $\hat{\mathbf{y}}^{\text{state}}_{u,t} = f_{\text{state}}(\mathbf{h}_{u,t})$ and $\hat{\mathbf{y}}^{\text{trait}}_{u} = f_{\text{trait}}(\mathbf{z}_u)$, where $\hat{\mathbf{y}} \in \mathbb{R}^2$ denotes valence and arousal.

Adaptive Gating. TSAF learns a gate $g_{u,t} = \sigma(W_g[\mathbf{h}_{u,t}; \mathbf{z}_u] + b_g)$, where $g_{u,t} \in (0, 1)$ is a scalar gating coefficient. The final prediction is computed as $\hat{\mathbf{y}}_{u,t} = g_{u,t}\hat{\mathbf{y}}^{\text{state}}_{u,t} + (1 - g_{u,t})\hat{\mathbf{y}}^{\text{trait}}_{u}$, dynamically balancing user baselines and contextual cues.

Training. We minimize mean squared error $\mathcal{L} = \frac{1}{N}\sum_{u,t} \|\hat{\mathbf{y}}_{u,t} - \mathbf{y}_{u,t}\|_2^2$ and select models using the official composite Pearson correlation, which captures both within-user and between-user performance.

4.2 Subtask 2A: State Change Forecasting

For each user at time t, the feature-based forecasting model predicts next-step affect change $\Delta^{\text{state}}_{u,t}$ using strictly past information. Features include current affect $(v_{u,t}, a_{u,t})$, first-order differences, rolling statistics (window=3), time gaps, phase indicators, and modality. We capture the temporal dependencies via lagged affect features and short-window statistics rather than learned sequential representations. We train a Ridge regression model for valence and a LightGBM model for arousal. Users without a history default to the global mean change. We use feature-based models because short-term affect is largely driven by previous states. Also, given short and irregular user histories, sequence models tend to overfit and provide little additional benefit.

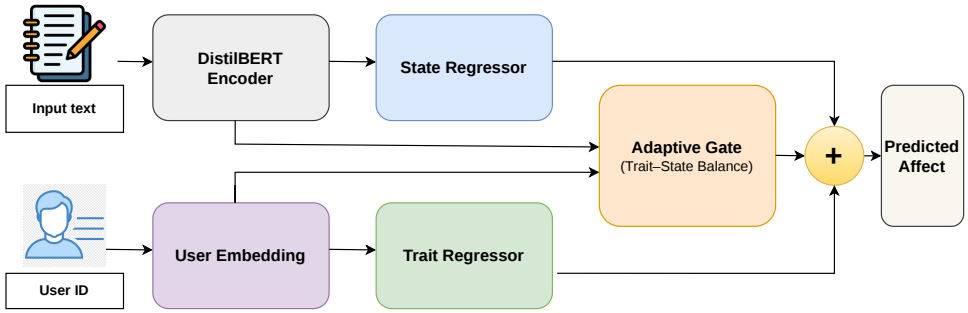


Figure 5: Trait–State Affect Forecaster (TSAF). Texts are encoded with DistilBERT to produce state representations, while learned user embeddings encode stable trait baselines. An adaptive gate computes a weighted combination of trait and state predictions to predict valence and arousal.

4.3 Subtask 2B: Dispositional Change Modeling

To capture long-term affect drift, the Ridge-based model aggregates first-half (early) user entries into summary statistics: mean, standard deviation, extrema, linear trend, entry count, and temporal span. Separate Ridge regressors predict valence ($\alpha = 100$) and arousal ($\alpha = 1$), with stronger regularization for valence to mitigate overfitting under limited user-level data. This aggregation-based approach reflects the limited number of observations per user and focuses on capturing coarse-grained trends rather than fine-grained temporal dynamics.

5 Experiments

5.1 Data Splits

For Subtask 1, we exclude users with fewer than six entries to ensure reliable within-user estimates. We hold out 15% of users as unseen. For seen users, we train on the earliest 80% of texts and validate on the most recent 20%, preserving temporal order. For Subtask 2A, we reserve the final labeled transition per user for validation. For Subtask 2B, we split users 80/20 and retrain on all labeled users before test prediction.

5.2 Implementation Details

We tokenize text using DistilBERT-base-uncased (max length 256) without additional normalization and order samples chronologically per user. We compute all temporal features strictly from past observations to prevent leakage. We use pretrained Hugging Face models with PyTorch. We set the user embedding dimension to 64 and include a shared UNK embedding for unseen users. We implement structured models using scikit-learn and LightGBM.

5.3 Training and Evaluation

For Subtask 1, we minimize MSE using AdamW (learning rate 2×10^{-5}, batch size 16) for up to 20 epochs with early stopping (patience 3). We select models based on validation composite Pearson r. For Subtask 2A, we model valence change using Ridge regression (cross-validated α) and arousal change using LightGBM (300 rounds, learning rate 0.05). For Subtask 2B, we select Ridge regularization via grid search over $\alpha \in \{0.01, 0.1, 1, 10, 100\}$, yielding $\alpha = 100$ (valence) and $\alpha = 1$ (arousal). We follow the official SemEval-2026 Task 2 evaluation protocol and report Pearson r and MAE. For Subtask 1, we additionally report within- and between-person Pearson r, combined via Fisher-z averaging.

6 Results

Tables 1 and 2 report official test results. We focus on Pearson correlation (r), the primary ranking metric, and report MAE for completeness. We compare against official baselines: Linear(BERT) (L-BERT), Linear(BERT; previous), linear(prev) (L-Prev), Rand-M, and Rand-Z.

6.1 Subtask 1: Longitudinal Affect Modeling

TSAF achieves $r = 0.645$ (valence) and $r = 0.409$ (arousal), improving over L-BERT by +0.088 and +0.110, respectively. Gains are consistent across between- and within-user metrics, with stronger improvements in within-user alignment, indicating improved modeling of user-specific dynamics.

Figure 6 visualizes performance across user splits and modalities. Detailed results are provided in Appendix A.

Model	Comp	Between	Within
Valence (r)			
L-BERT	.557	.659	.435
Rand-M	.000	.028	.000
TSAF	**.645**	**.705**	**.577**
Valence (MAE)			
L-BERT	.743	.472	.886
Rand-M	.000	.627	1.041
TSAF	**.685**	**.472**	**.822**
Arousal (r)			
L-BERT	.299	.343	.253
Rand-M	.000	.096	.000
TSAF	**.409**	**.430**	**.388**
Arousal (MAE)			
L-BERT	.459	.311	.585
Rand-M	.488	.326	.622
TSAF	**.407**	**.275**	**.524**

Table 1: Subtask 1 test performance. TSAF improves correlation and reduces error across both affect dimensions.

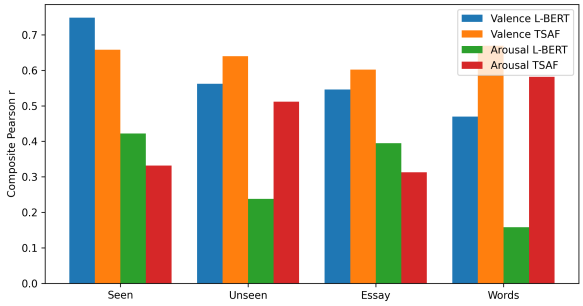


Figure 6: Composite Pearson correlation (r) across user splits and modalities. TSAF improves under unseen users and short inputs, while L-BERT remains competitive on essays.

6.1.1 Analysis

Seen vs. Unseen Users. For seen users, L-BERT achieves higher valence correlation, especially for between-user effects, indicating strong implicit modeling of stable variance. In contrast, TSAF improves composite correlation under unseen users for both valence and arousal, demonstrating stronger robustness under cold-start conditions.

Within vs. Between-User Effects. TSAF improves within-user correlation for unseen users while maintaining competitive between-user performance, confirming that explicit trait–state decomposition enhances modeling of user-specific variation without degrading global alignment.

Modality Effects. TSAF consistently outperforms L-BERT on feeling-word entries, where sparse inputs benefit from adaptive trait–state balancing. For essays, L-BERT remains competitive, particularly for between-user correlation, suggesting that longer texts already encode stable affective signals.

Model	r_V	r_A	MAE_V	MAE_A
Subtask 2A: State Change				
L-Prev	**.615**	**.670**	**1.168**	**.638**
L-BERT+Prev	.430	.405	1.251	.708
Rand-Z	.000	.000	1.261	.696
TSAF	.424	.355	1.297	.842
Subtask 2B: Dispositional Change				
L-Prev	**.434**	**.584**	**.406**	**.286**
L-BERT+Prev	-.029	.019	.436	.305
Rand-Z	.000	.000	.417	.296
TSAF	.257	.418	.461	.298

Table 2: Subtask 2 forecasting performance.

TSAF shows larger gains for arousal than valence, suggesting that activation dynamics are more amenable to explicit trait–state decomposition under sparse, longitudinal supervision.

6.2 Subtask 2: Forecasting

Subtask 2A: Short-Term Forecasting. TSAF achieves $r = 0.424$ (valence) and $r = 0.355$ (arousal), outperforming random baselines but trailing L-Prev. This confirms strong short-term affective inertia: recent affect dominates next-step changes, limiting gains from more expressive models.

Subtask 2B: Long-Term Drift. TSAF achieves $r = 0.257$ (valence) and $r = 0.418$ (arousal), ranking 4th overall and 3rd for arousal. Unlike Subtask 2A, long-term prediction benefits from aggregated user statistics rather than short-term autoregressive signals. Stronger arousal performance suggests that activation trends are more structurally recoverable than valence drift under sparse supervision.

6.3 Ablation on Validation Set

On the validation split (Section 5.1), we compare four variants: state-only, trait-only, concatenation of user and text embeddings, and gated fusion (TSAF). The state-only model performs strongly, with mean $r \approx 0.64$ and only a small seen–unseen gap, showing that contextual text representations provide the dominant predictive signal. In contrast, the trait-only model fails to converge, confirming that user-level signals alone are insufficient for accurate affect prediction. Simple concatenation achieves the highest validation performance (mean $r \approx 0.65$), but it also increases the seen–unseen gap, indicating greater reliance on user-specific parameters and reduced generalization. TSAF attains slightly lower peak performance (mean $r \approx 0.62$), reflecting the trade-off introduced by explicitly modeling both trait and state components, while

providing a structured and interpretable integration of user-level and contextual signals. Overall, these results show that affect prediction in this dataset is primarily state-driven, while trait information remains complementary rather than sufficient on its own. The trait–state decomposition offers a principled personalization mechanism, but its benefit depends on balancing expressivity and generalization.

To quantify generalization to unseen users, we compare seen and unseen splits. TSAF achieves a composite correlation of 0.6247 for seen users and 0.5506 for unseen users, a modest drop of 0.0741, which indicates limited degradation under the shared UNK embedding and strong dependence on contextual text signals. We also evaluate few-shot adaptation with $k = 3$ samples; performance drops further to $r = 0.4335$, below the non-adapted unseen score, suggesting that small-sample adaptation overfits rather than improves personalization.

6.4 Error Analysis

We analyze validation errors to characterize TSAF's inductive biases and failure modes.

Overall Difficulty. Valence remains harder to predict than arousal (MAE: 0.77 vs. 0.48), reflecting its wider range and higher variability. This increases regression difficulty and amplifies sensitivity to user-level baselines.

Extreme Valence Underestimation. For strongly negative instances ($v < -1.5$, $n = 141$), TSAF shows a mean positive bias of +1.01, indicating systematic underestimation of extreme negativity. This reflects regression toward user-level trait components, which regularize predictions but attenuate high-magnitude affect.

Affective Inertia and Spikes. For large arousal changes ($|\Delta a| > 1.0$), MAE increases to 0.69 (vs. 0.48 overall), indicating that abrupt transitions are difficult to capture. The model better captures gradual trends than sharp state changes, consistent with strong affective inertia.

Polarity Reversals. High-error cases frequently involve polarity reversals, where strongly positive or negative labels are predicted with opposite sign. These errors often occur in longer essays with mixed or contrastive emotional cues, where stable trait signals can dominate localized textual indicators.

Overall, errors reveal a structural trade-off: trait modeling improves global user alignment and robustness but reduces sensitivity to extreme or rapidly changing affective states.

7 Conclusions

TSAF introduces a structured trait–state decomposition for longitudinal valence and arousal modeling in SemEval-2026 Task 2, combining stable user embeddings with text-conditioned signals via adaptive gating. The model improves over L-BERT on per-text prediction, with analysis showing that affect is primarily state-driven while trait information provides complementary gains, particularly under cold-start conditions. Improvements are more pronounced for arousal, suggesting stronger alignment with stable user tendencies. Forecasting remains challenging due to strong short-term inertia and limited supervision for long-term drift. Overall, TSAF offers a principled and interpretable approach to personalization, while future work should explore sequence-aware methods to better capture temporal dynamics under sparse and irregular observations.

Limitations

Despite improvements over the baseline in Subtask 1, several limitations remain. Personalization relies on learned user embeddings, with unseen users sharing a single UNK representation, which constrains user-specific adaptation, although the modest performance drop suggests predictions remain largely text-driven. The model also treats each text independently and does not leverage short-term affect history. For Subtask 2A, reliance on engineered temporal features instead of learned sequence models limits the capture of fine-grained transitions, while Subtask 2B summarizes user trajectories using simple aggregates under limited data, restricting long-term modeling. Additionally, the dataset's bimodal structure—long essays versus sparse feeling-word entries—introduces variation in expressive detail, potentially leading to uneven performance across modalities. Overall, while trait–state decomposition improves longitudinal affect modeling, more expressive sequence-based approaches are needed to better capture temporal dynamics under sparse and irregular observations.

Acknowledgments

We thank the organizers of SemEval-2026 Task 2 for designing the shared task and providing the dataset. We also thank the reviewers for their constructive feedback and valuable suggestions, which improved this work.

References

Vrushank Ahire, Kunal Shah, Mudasir Nazir Khan, Nikhil Pakhale, Lownish Rai Sookha, M. A. Ganaie, and Abhinav Dhall. 2025. Maven: Multi-modal attention for valence-arousal emotion network. *Preprint*, arXiv:2503.12623.

Jonas Becker, Liang-Chih Yu, Shamsuddeen Hassan Muhammad, Jan Philip Wahle, Terry Ruas, Idris Abdulmumin, Lung-Hao Lee, Nelson Odhiambo, Lilian Wanzare, Wen-Ni Liu, Tzu-Mi Lin, Zhe-Yu Xu, Ying-Lung Lin, Jin Wang, Maryam Ibrahim Mukhtar, Bela Gipp, and Saif M. Mohammad. 2026. Dimstance: Multilingual datasets for dimensional stance analysis. *Preprint*, arXiv:2601.21483.

Margaret M. Bradley and Peter J. Lang. 1994. Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, 25(1):49–59.

Adithya V Ganesan, Vasudha Varadarajan, Oscar NE Kjell, Whitney R Ringwald, Scott Feltman, Benjamin J Luft, Roman Kotov, Ryan L Boyd, and H Andrew Schwartz. 2026. From word sequences to behavioral sequences: Adapting modeling and evaluation paradigms for longitudinal nlp. *Preprint*, arXiv:2601.07988.

Tamunotonye Harry, Ivoline Ngong, Chima Nweke, Yuanyuan Feng, and Joseph Near. 2026. Beyond fixed psychological personas: State beats trait, but language models are state-blind. *Preprint*, arXiv:2601.15395.

Anthony Hills, Talia Tseriotou, Xenia Miscouridou, Adam Tsakalidis, and Maria Liakata. 2024. Exciting mood changes: A time-aware hierarchical transformer for change detection modelling. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12526–12537, Bangkok, Thailand. Association for Computational Linguistics.

Lesa Hoffman and Robert S. Stawski. 2009. Persons as contexts: Evaluating between-person and within-person effects in longitudinal analysis. *Research in Human Development*, 6(2-3):97–120.

Diefan Lin, Yi Wen, Weishi Wang, and Yan Su. 2024. Enhanced sentiment intensity regression through lora fine-tuning on llama 3. *IEEE Access*, PP:1–1.

Matthew Matero and H. Andrew Schwartz. 2020. Autoregressive affective language forecasting: A self-supervised task. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2913–2923, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Gonçalo Azevedo Mendes and Bruno Martins. 2023. Quantifying valence and arousal in text with multilingual pre-trained transformers. In *Advances in Information Retrieval: 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2–6, 2023, Proceedings, Part I*, page 84–100, Berlin, Heidelberg. Springer-Verlag.

Fatemehsadat Mireshghallah, Vaishnavi Shrivastava, Milad Shokouhi, Taylor Berg-Kirkpatrick, Robert Sim, and Dimitrios Dimitriadis. 2022. UserIdentifier: Implicit user representations for simple and effective personalized sentiment analysis. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3449–3456, Seattle, United States. Association for Computational Linguistics.

Michail Mitsios, Georgios Vamvoukakis, Georgia Maniati, Nikolaos Ellinas, Georgios Dimitriou, Konstantinos Markopoulos, Panos Kakoulidis, Alexandra Vioni, Myrsini Christidou, Junkwang Oh, Gunu Jho, Inchul Hwang, Georgios Vardaxoglou, Aimilios Chalamandaris, Pirros Tsiakoulis, and Spyros Raptis. 2024. Improved text emotion prediction using combined valence and arousal ordinal classification. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 808–813, Mexico City, Mexico. Association for Computational Linguistics.

James A. Russell. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178.

Noémi Schuurman. 2023. A "within/between problem" primer: About (not) separating within-person variance and between-person variance in psychology. OSF Preprint.

Chang Shu. 2024. Enhanced personalized text generation using user embeddings and attention mechanisms. *Applied and Computational Engineering*, 107(1):61–72.

Nikita Soni, Pranav Chitale, Khushboo Singh, Niranjan Balasubramanian, and H. Andrew Schwartz. 2025. Evaluation of LLMs-based hidden states as author representations for psychological human-centered NLP tasks. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7673–7682, Albuquerque, New Mexico. 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.

Rolf Steyer, Manfred Schmitt, and Michael Eid. 1999. Latent state-trait theory and research in personality and individual differences. *European Journal of Personality*, 13(5):389–408.

Antonela Tommasel, Andrés Diaz-Pace, Juan Manuel Rodriguez, and Daniela Godoy. 2021. Forecasting

mental health and emotions based on social media expressions during the covid-19 pandemic. *Information Discovery and Delivery*, 49(3):259–268.

Ignacio J. Tripodi, Greg Buda, Margaret Meagher, and Elizabeth A. Olson. 2025. Assessing effective de-escalation of crisis conversations using transformer-based models and trend statistics. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 29763–29777, Suzhou, China. Association for Computational Linguistics.

Talia Tseriotou, Adam Tsakalidis, Peter Foster, Terence Lyons, and Maria Liakata. 2023. Sequential path signature networks for personalised longitudinal language modeling. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5016–5031, Toronto, Canada. Association for Computational Linguistics.

Ingrid C. Wurpts. 2015. Performance of contextual multilevel models for comparing between-person and within-person effects. *Multivariate Behavioral Research*, 50(6):721.

A Detailed Subtask 1 Results

This section provides a full breakdown of Subtask 1 performance across user splits (seen vs. unseen) and text modalities (essay vs. feeling-word).

We report performance using Pearson correlation (r) and mean absolute error (MAE). In addition to overall (composite) scores, we compute separate metrics to distinguish between inter-user and intra-user performance.

Metric Definitions. Composite correlation (r_{comp}) is computed over all instances. Between-user correlation ($r_{between}$) is computed over per-user averages, capturing inter-user variation, while within-user correlation (r_{within}) is computed after centering values per user, capturing intra-user temporal dynamics. MAE is computed analogously for each setting, with lower values indicating better performance.

These metrics provide complementary perspectives: correlation evaluates ranking consistency, while MAE measures absolute prediction accuracy.

As shown in Tables 3 and 4, L-BERT performs strongly on seen users, particularly for between-user correlation, indicating reliance on learned user-specific patterns, whereas TSAF shows substantial improvements for unseen users and sparse inputs, demonstrating stronger generalization under limited context. TSAF shows larger gains for arousal than valence and often achieves higher correlation without consistently reducing MAE, indicating better ranking but less accurate prediction values.

Split	Valence		Arousal	
	L-BERT	TSAF	L-BERT	TSAF
Composite (r_{comp})				
Seen Users	.748	.658	.422	.332
Unseen Users	.562	**.640**	.238	**.512**
Essay Only	.546	**.602**	.395	.313
Words Only	.470	**.669**	.158	**.582**
Between-user ($r_{between}$)				
Seen Users	.851	.732	.592	.383
Unseen Users	.624	**.691**	.258	**.540**
Essay Only	.629	**.652**	.395	.296
Words Only	.570	**.738**	.124	**.638**
Within-user (r_{within})				
Seen Users	.588	.569	.215	**.278**
Unseen Users	.494	**.583**	.238	**.483**
Essay Only	.451	**.547**	.395	.330
Words Only	.357	**.586**	.191	**.520**

Table 3: Subtask 1 correlation performance across user splits and modalities for valence and arousal. Higher values indicate better performance.

Split	Valence		Arousal	
	L-BERT	TSAF	L-BERT	TSAF
Composite (MAE_{comp})				
Seen Users	**.598**	.710	**.408**	.415
Unseen Users	.772	**.660**	.488	**.398**
Essay Only	**.710**	.745	.468	**.437**
Words Only	.823	**.603**	.487	**.389**
Between-user ($MAE_{between}$)				
Seen Users	**.354**	.543	**.268**	.288
Unseen Users	.411	**.399**	.309	**.261**
Essay Only	**.548**	.563	.358	**.298**
Words Only	.641	**.419**	.384	**.292**
Within-user (MAE_{within})				
Seen Users	**.766**	.823	.531	**.529**
Unseen Users	.924	**.822**	.633	**.519**
Essay Only	**.820**	.858	.566	**.558**
Words Only	.918	**.739**	.578	**.479**

Table 4: Subtask 1 MAE performance across user splits and modalities for valence and arousal. Lower values indicate better performance.