ES4MLL at SemEval-2026 Task 2: Set Attention Aggregation and Recurrent Temporal Modeling for Longitudinal Affect Prediction

Andrea Lolli* **Chiara Lunazzi*** **Riccardo Coppola†** **Flavio Giobergia†**
Politecnico di Torino
*{firstname.lastname}@studenti.polito.it
†{firstname.lastname}@polito.it

Abstract

Longitudinal modelling of affect from text requires capturing both linguistic content and temporal emotional dynamics. SemEval-2026 Task 2 introduces a dataset of essays and feeling words annotated with self-reported valence and arousal scores. In this work, we propose a neural architecture that combines pretrained Transformer encoders with temporal sequence modelling to predict continuous valence and arousal over user-specific timelines. Individual texts are encoded using a Transformer-based language model and aggregated through attention-based pooling before being processed by recurrent layers to capture longitudinal dependencies. To adapt pretrained representations under limited data conditions, we explore parameter-efficient fine-tuning strategies. We make the code available at https://github.com/AndreaLolli2912/SemEval2026-EmoVA.

1 Introduction

The affective circumplex model proposes that emotions can be described in a two-dimensional space defined by *valence* and *arousal* (Russell, 1980). While most affective computing research focuses on static snapshots or external perception of emotion, real emotional experience is inherently *longitudinal* – it evolves over time, shaped by daily routines, life events, and individual patterns.

Attempts have been made to address continuous affect prediction using pre-trained Transformer architecture (Mendes and Martins, 2023) and to model long texts via hierarchical architectures with fixed-size representations and recurrent neural networks (Christ et al., 2022), effectively mapping discrete emotion labels to a continuous space remains a challenge. SemEval-2026 Task 2 (Soni et al., 2026) addresses this gap by introducing a longitudinal dataset of ecological essays and feeling words. This work focuses on **Subtask 1: Longitudinal Affect Assessment** and **Subtask 2a: Forecasting (future) Variation in Affect**. In this paper, we propose a hierarchical architecture combining BERT (Devlin et al., 2019), Set Transformer attention (Lee et al., 2019) and bidirectional LSTM for longitudinal affect modelling. We show that the proposed solution achieves remarkable results, and that it provides satisfactory performance on small datasets by employing parameter-efficient fine-tuning strategies using DoRA (Liu et al., 2024) and BitFit (Ben-Zaken et al., 2022). To address the valence-arousal difficulty gap, we design an asymmetric hybrid loss by combining Mean Squared Error and Concordance Correlation Coefficient.

2 Problem Description

The goal of the task is to correctly predict current or future valence and arousal given a collection of texts. Two different datasets are provided: training data (2,765 samples) and test data (1,738 samples and 47 samples). Each sample is provided with a variety of attributes. The correct valence/arousal values are provided for training data. Users contribute a varying number (ranging from 2 to over 200) of texts (Fig. 1), and discrete labels (Fig. 2).

Subtask 1 (ST1) requires predicting continuous valence and arousal scores for each essay, independently for each user. In Subtask 2a (ST2a), essays are provided together with preceding segments to model short-term changes in a user's emotional state. In Subtask 2b, the full history up to a split point is used to predict long-term dispositional changes in valence and arousal, capturing temporal evolution rather than absolute affective levels.

3 Proposed Methodology

We hypothesize that affective states exhibit a temporal duality. Valence reflects rapid, reactive variations triggered by immediate events and explicit emotional language. Arousal, in contrast, follows

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 720–726

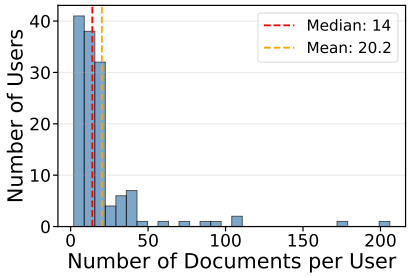

Figure 1: Number of documents per user in train data.

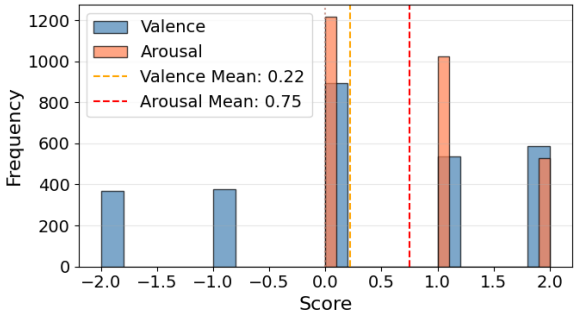

Figure 2: Distribution of Valence and Arousal scores in the training data.

a slower, inertial dynamic influenced by routines, persistent stress, and general well-being. Taking this into account, the proposed solution consists in one encoder-only pretrained Transformer, combined with attention-based aggregation and recurrent neural network (LSTM).

3.1 Data Preprocessing

Efficiently processing longitudinal data requires addressing the variability in both the number of documents per user and the length of individual texts. To enable parallelized batch processing, we implement a hierarchical padding strategy that standardizes input dimensions into a (B, S, L) tensor, where B is the batch size, S is the maximum number of documents per user, and L is the maximum token length. We generate a binary sequence mask $M_{seq} \in \{0, 1\}^{B \times S}$ to distinguish between genuine user texts and padding. This mask is propagated throughout the network to ensure that padded positions do not contribute to attention computations or the loss function. Due to the quadratic complexity of attention and memory constraints, we limit the maximum sequence length to $L = 128$.

3.2 Model Architecture

Our architecture is a hierarchical neural pipeline that transforms longitudinal text into continuous valence and arousal predictions. It comprises four modules: (1) a Transformer encoder, (2) attention-based aggregation, (3) temporal modeling, and (4) a prediction head.

3.2.1 Encoder Backbone

We employ `bert-base-uncased` as semantic encoder, which produces contextualized token embeddings for each input sequence. To reduce computational cost and prevent overfitting, we adopt Parameter-Efficient Fine-Tuning (PEFT) strategies described in Section 3.4.

3.2.2 Attention-based Aggregation (ISAB & PMA)

To aggregate variable-length token sequences into fixed-size document representations, we adopt the Set Transformer framework (Lee et al., 2019).

Specifically, we use Induced Set Attention Blocks (ISAB) to efficiently model interactions among tokens through a set of learnable inducing points, reducing the computational complexity of standard self-attention. The resulting representations are then pooled using Pooling by Multihead Attention (PMA), which employs a set of learnable seed vectors to produce a fixed number of aggregated features.

In our configuration, we use $m = 32$ inducing points and $k = 8$ seed vectors. The final pooled representation is flattened and used as input to the temporal modeling component.

3.2.3 Temporal Modeling and Inference

To capture the longitudinal dynamics of affect, the sequence of document embeddings is processed by an LSTM with a hidden dimension of 256. We fix the depth to 1 layer to mitigate overfitting. We treat directionality as a hyperparameter, evaluating both Unidirectional and Bidirectional configurations. The resulting temporal states are projected via two separate linear heads to predict valence and arousal respectively.

3.2.4 Multimodal Fusion

To address the forecasting objective of ST2, we adapted the temporal modeling architecture to integrate historical user data. We introduced two key modifications.

Multimodal Input Fusion: We employ a feature-level fusion strategy to combine textual and affective signals. At each timestep t, the aggregated document embedding is concatenated with the corresponding historical valence and arousal values.

This enables the LSTM to jointly model the linguistic content and the user's affective trajectory.

State-Aware Prediction Head: The prediction mechanism is conditioned on the most recent observable state. We concatenate the final LSTM hidden state with the last known affective scores before the output projection. This design explicitly informs the model of the prior state, effectively simplifying the task to predicting the *variation* (or delta) required to reach the future state, rather than regressing the absolute values from scratch.

3.3 Optimization Objective

Predicting continuous affect presents two objectives of interest: minimizing absolute error (magnitude accuracy) and preserving structural trends (correlation). A model trained solely on Mean Squared Error (MSE) tends to regress toward the mean, while pure correlation objectives can be numerically unstable.

The Zero-Variance Problem The gradient of correlation-based losses is inversely proportional to the prediction standard deviation $\sigma_{\hat{y}}$. In early training, if the model predicts near-constant values gradients can explode, causing training instability.

Concordance Correlation Coefficient (CCC) To address this, we employ Lin's Concordance Correlation Coefficient, which measures both correlation and agreement in mean/variance:

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

By including the ground truth variance σ_y^2 in the denominator, CCC remains stable even when $\sigma_{\hat{y}} \approx 0$.

Combined Loss for ST1 Our loss function combines MSE for magnitude accuracy and CCC for structural alignment:

$$\mathcal{L}_{dim}(y, \hat{y}) = \lambda \cdot \mathcal{L}_{MSE}(y, \hat{y}) + (1 - \lambda) \cdot (1 - \text{CCC}(y, \hat{y})) \quad (2)$$

The total loss is computed separately for valence and arousal:

$$\mathcal{L}_{total} = \omega_v \mathcal{L}_{dim}(v, \hat{v}) + (1 - \omega_v)\mathcal{L}_{dim}(a, \hat{a}) \quad (3)$$

We set $\lambda = 0.15$, favoring CCC over MSE: a model trained primarily on MSE tends to regress toward the target mean, whereas correlation-based losses better align with the official evaluation metrics. The dimension weight $\omega_v = 0.2$ allocates 80% of the loss to arousal, which we consistently observed to be harder to predict across all configurations. Both hyperparameters were selected empirically; a systematic search was not feasible due to computational constraints.

3.4 Parameter-Efficient Fine-Tuning

Full fine-tuning of large language models on small longitudinal datasets often leads to overfitting. Although few-shot approaches have been shown to generalize well from limited amounts of data (Giobergia et al., 2024), they are typically designed for instance-level predictions and lack an explicit mechanism to model temporal dependencies across sequences of texts. Instead, we adopt a fine-tuning approach, by freezing the majority of the encoder parameters and making use of two efficient adaptation strategies.

Bias-term Fine-tuning (BitFit): Following Ben-Zaken et al. (2022), we freeze all attention and feed-forward weight matrices, training *only* the bias vectors. This approach drastically reduces the number of trainable parameters while allowing the model to shift activation distributions to align with the affect prediction task.

Weight-Decomposed Low-Rank Adaptation (DoRA): To capture more complex dependencies than bias updates allow, we employ DoRA (Liu et al., 2024), which decomposes the pre-trained weights into magnitude and direction components. We apply DoRA to all linear layers with a rank of $r = 16$ and a scaling factor $\alpha = 32$.

3.5 Implementation Details

The system is trained using mixed-precision computation (FP16) to improve efficiency. Due to the memory overhead of the Set Transformer and LSTM unrolling, we use a micro-batch size of 1 with gradient accumulation to simulate an effective batch size of 32. Gradient clipping (1.0) is applied to stabilize optimization. Using PEFT strategies, we apply a differential optimization schedule. We use a lower learning rate (2×10^{-6}) for the pre-trained encoder parameters and a higher rate (1×10^{-4}) for the randomly initialized task-specific modules (ISAB, PMA, LSTM, prediction head). This prevents catastrophic forgetting in the backbone while allowing the new attention mechanisms to converge efficiently. All hyperparameters, including loss weighting coefficients (λ, ω_v) and

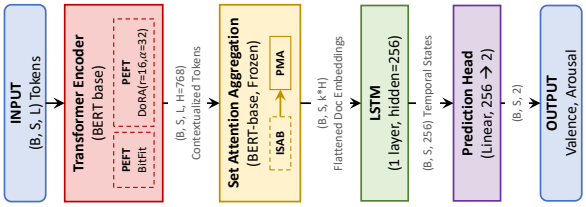

Figure 3: Overview of the proposed architecture.

pooling configurations, are managed via experiment configuration files to ensure reproducibility.

We present the main results obtained using the proposed pipeline, detailing the impact of various hyperparameter configurations on overall performance. Model evaluation was performed using the official SemEval evaluation metrics on the validation set, which differ based on the subtask. The official metrics are summarized in the subsequent section. Results are summarized in Table 1.

3.6 Official Metrics

To evaluate the model's performance, we adopt the official metrics defined for the task, which rely on three correlation measures:

1. **Between-user correlation** estimates the model's ability to capture the differences between users. For each user u we compute the mean predicted score $\hat{y}_m = \{\text{mean}_{t \in u}(\hat{y}_{u,t})\}_{u=1}^{N}$ and the mean gold score $y_m = \{\text{mean}_{t \in u}(y_{u,t})\}_{u=1}^{N}$ across all associated texts. Pearson r is then calculated as

$$r_{\text{b}}(\{\hat{y}_{u,t}\}, \{y_{u,t}\}) = r(\hat{y}_m, y_m).$$

2. **Within-user correlation** assesses the model's ability to capture variations within a single user's data. We compute the correlation for each user u across their texts t and then report the mean value across the entire user collection

$$r_{\text{w}} = \text{mean}_{\forall u}(r_u(\hat{y}_u, y_u)).$$

where $\hat{y}_u = \{\hat{y}_{u,t}\}_{t \in u}$, $y_u = \{y_{u,t}\}_{t \in u}$ and $r_u = r_u(\hat{y}_u, y_u)$.

3. **Composite correlation** combines Between-user and Within-user correlations using Fisher's z-transformation:

$$r_{\text{comp}} = \tanh\left(\frac{\tanh^{-1}(r_{\text{b}}) + \tanh^{-1}(r_{\text{w}})}{2}\right)$$

The leaderboard ranking is determined by r_{comp}.

Additionally, for for Subtask 2 (ST2) evaluation relies on the Pearson correlation computed between the predicted and gold-state change values per user. In both tasks, the overall score is the average of the Valence and Arousal correlations.

3.7 ST1: Affect Assessment

The best performing model achieves a composite score of $r = 0.6945$ (Sim.10) on the evaluation set. This configuration combines a `BERT-base-uncased` encoder adapted via DoRA, ISAB with 32 inducing points, PMA with $k = 8$ seeds and a unidirectional LSTM for temporal modeling. Compared to the baseline relying on simple mean pooling, our analysis reveals three key differences:

Impact of Attention Mechanism: Attention based aggregation provided significant gains. PMA alone (Sim.3 and Sim.4) substantially outperformed the mean-pooling baseline (Sim.1, Sim.2). Adding ISAB (32 points) further improved the score. Notably, PMA primarily improved Valence correlation, while ISAB filters noise in the user's textual history, thereby enhancing the more challenging Arousal dimension.

PEFT:While DoRA alone achieved the highest validation score (r=0.6945), the combination of DoRA and BitFit yielded the best test performance (r=0.5333), suggesting BitFit's bias-only updates may act as an implicit regularizer that improves generalization.

Bidirectional temporal modeling LSTM improved a simpler baseline model, but unidirectional LSTM performed better when paired with ISAB. We attribute this to functional redundancy: since ISAB already handles global attention, the LSTM only needs to capture local sequential progression. This unidirectional setup also reduces the number of parameters, mitigating overfitting.

3.8 ST2a: Affect Forecasting

For the forecast objective, the models were evaluated using Pearson's correlation (r). The experiments reveal a significant divergence in architectural requirements compared to ST1. On the validation set, the strongest performance was achieved by Sim. 2 ($r = 0.6017$) which relies on a simple Bidirectional LSTM over mean-pooled embedding without any PEFT or Set Attention Mechanism. Notably our ablation studies demonstrate that introducing Set Attention (ISAB and PMA) leads to severe performance degradation in forecasting.

Sim. #	PEFT	ISAB	PMA	Bi-LSTM	Validation Set (r)			Test Set (r)		
					Val.	Aro.	Avg.	Val.	Aro.	Avg.
ST1: Longitudinal Affect Assessment — Composite correlation (r_{comp})										
1	–	–	–	No	0.6770	0.4218	0.5494	–	–	–
2	–	–	–	Yes	0.6830	0.4561	0.5695	–	–	–
3	–	–	8	No	0.7987	0.4797	0.6393	–	–	–
4	–	–	8	Yes	0.7706	0.4547	0.6126	–	–	–
5	–	16	8	No	0.7867	0.5397	0.6632	–	–	–
6	–	16	8	Yes	0.7583	0.5228	0.6405	–	–	–
7	–	32	8	No	0.7995	**0.5662**	0.6803	**0.6798**	0.3419	0.5109
8	–	32	8	Yes	0.7579	0.5551	0.6565	–	–	–
9	BitFit	32	8	No	0.7740	0.5630	0.6690	0.6742	0.3596	0.5169
10	DoRA	32	8	No	**0.8241**	0.5649	**0.6945**	0.6564	0.3762	0.5163
11	DoRA and BitFit	32	8	No	0.7369	0.5457	0.6710	0.6552	**0.4114**	**0.5333**
ST2a: Forecasting Variation in Affect — Per-user Pearson (r_{user})										
1	–	–	–	No	0.5401	0.6188	0.5795	0.6469	0.5157	0.5813
2	–	–	–	Yes	0.5623	0.6411	**0.6017**	0.6374	0.5107	0.5740
3	–	–	8	No	0.4197	0.5039	0.4618	–	–	–
4	–	–	8	Yes	0.4453	0.5474	0.4963	–	–	–
5	–	32	8	No	0.4753	0.4853	0.4803	–	–	–
6	BitFit	–	–	No	0.5415	0.6155	0.5785	0.6448	0.5201	0.5825
7	BitFit	–	–	Yes	**0.5630**	0.6380	0.6005	0.6370	0.5096	0.5733
8	DoRA	–	–	No	0.5318	0.6322	0.5820	0.6428	0.4739	0.5583
9	DoRA	–	–	Yes	0.5336	**0.6538**	0.5947	**0.6627**	**0.5107**	**0.5867**

Table 1: Ablation results on the validation set for ST1 and ST2a, with official Test Set performance reported for selected configurations. Metrics shown are r_{comp} (ST1) and per-user Pearson r (ST2a). ISAB and PMA columns indicate inducing points and seeds. **Bold** marks the best per metric within each task.

Rank	Team	Val.	Aro.	Avg.
ST1 — r_{comp}				
1	UKP_Psycontrol	0.667	**0.554**	**0.611**
3	cclin	0.647	0.527	0.587
5	lamanhnguyen	**0.687**	0.458	0.573
–	ES4MLL† (Sim. 11)	0.655	0.411	0.533
ST2a — Per-user Pearson r				
1	UKP_Psycontrol	0.675	**0.683**	**0.679**
2	YNU	**0.692**	0.647	0.669
3	UAlberta	0.615	0.674	0.645
5	Ajman University	0.615	0.670	0.642
–	ES4MLL† (Sim. 9)	**0.663**	0.511	0.587

Table 2: Comparison with official leaderboard results. **Bold** marks the best value per column within each subtask. †Post-hoc evaluation on gold labels released after the competition deadline; not an official submission.

The evaluation on official test set highlighted the necessity of PEFT (DoRA) also in the forecast setting. The architecture combining a bidirectional LSTM with DoRA (Sim. 9) achieved the highest overall test score ($r = 0.5867$) and delivered the strongest Valence prediction on the test set. These findings suggest that while continuous affect assessment (ST1) benefits from complex attention-based noise filtering, forecasting future emotional trajectories (ST2a) relies predominantly on strong temporal modeling (Bi-LSTM) coupled with parameter-efficient fine-tuning (DoRA).

3.9 Performance on Test Data

Following the ablation study, we selected the best-performing configurations for each subtask and retrained the model on the full training set. Specifically, for each subtask, we used the number of training epochs corresponding to the best validation score observed during ablation, i.e. the epoch at which early stopping would have triggered. The retrained models were then evaluated against the official gold labels released by the task organizers.

3.10 Comparison with Official Leaderboard

To contextualize our results, we compare our best test set configurations against the official challenge leaderboard in Table 2.[1] On ST1, our best configuration (Sim. 11, DoRA and BitFit) achieves $r_{\text{comp}} = 0.533$, below the winning system (0.611). The gap is primarily driven by arousal (0.411 vs. 0.554), consistent with the difficulty asymmetry observed throughout our ablation study. Valence

[1]Our scores are obtained by evaluating on the gold labels released after the competition deadline and were not submitted through the official evaluation portal.

performance (0.655) remains competitive with mid-table systems. On ST2a, our architecture (Sim. 9, DoRA with Bi-LSTM) reaches an average score of 0.587, which would place it in the upper tier of the leaderboard. Notably, our valence correlation (0.663) ranks second overall, surpassed only by YNU (0.692). The arousal gap (0.511 vs. 0.683) again constitutes the main limitation, reinforcing the finding that arousal modeling remains the primary challenge across both subtasks.

4 Discussion and Conclusion

We presented a flexible neural architecture for longitudinal affect modeling that combines BERT encoding, Set Transformer attention (ISAB/PMA), and recurrent temporal modeling. Our best configurations achieved correlations of $r = 0.6802$ (validation) and $r = 0.5420$ (test) for Subtask 1, and $r = 0.6017$ (validation) and $r = 0.5867$ (test) for Subtask 2a. Three factors proved critical for performance: (1) task-specific aggregation, where attention-based pooling (PMA) over mean pooling for document aggregation, with ISAB providing additional gains by filtering noisy historical context (Subtask 1), but severely degrades forecasting (Subtask 2a), which instead requires simple mean pooling to preserve sequential continuity; (2) DoRA adaptation of the encoder, which outperformed both frozen backbones and BitFit; (3) unidirectional LSTMs when combined with ISAB, avoiding functional redundancy with the attention mechanism (Subtask1) while pure Bidirectional LSTMs are essential to track historical trajectories without attention (Subtask 2a). The architecture presented for Subtask 2a scales well to Subtask 2b (Dispositional change), where the core component of the architecture is a prediction head that processes deltas. The PMA module could also be employed to summarize the observed sequence of events into a fixed-length vector. Consequently, the model would track dispositional changes in affect by simply changing the reference point in the input from the last observed point to the mean of the sequence. One of the main limitation of the proposed method is related to the assumption regarding the uniformity of observation points in longitudinal data, whereas EMA data is generally uneven, with varying gaps between user entries. Standard LSTM may struggle to capture the changes in emotional states over time. Our architecture uses BERT-base-uncased as the encoder backbone. We also experimented with DeBERTa-v3-base, DistilRoBERTa, and ELECTRA-base, but these exceeded the GPU memory available on free-tier Google Colab even with mixed-precision training and gradient checkpointing enabled. Future works should focus on time-aware modeling, such as temporal embedding or Time-LSTM (Zhu et al., 2017)), to better address these irregularity and enhance the robustness of the longitudinal affect trajectories modeling the temporal decay of emotion over time (Puccetti et al., 2022).

References

Elad Ben-Zaken, Shauli Ravfogel, and Yoav Goldberg. 2022. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. *Preprint*, arXiv:2106.10199.

Lukas Christ, Shahin Amiriparian, Manuel Milling, Ilhan Aslan, and Björn W. Schuller. 2022. Automatic emotion modelling in written stories. *Preprint*, arXiv:2212.11382.

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 (long and short papers)*, pages 4171–4186.

Flavio Giobergia, Alkis Koudounas, and Elena Baralis. 2024. Large language models-aided literature reviews: A study on few-shot relevance classification. In *2024 IEEE 18th International Conference on Application of Information and Communication Technologies (AICT)*, pages 1–5. IEEE.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. 2019. Set transformer: A framework for attention-based permutation-invariant neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, pages 3744–3753.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. 2024. Dora: Weight-decomposed low-rank adaptation. *Preprint*, arXiv:2402.09353.

Gonçalo Azevedo Mendes and Bruno Martins. 2023. Quantifying valence and arousal in text with multilingual pre-trained transformers. *Preprint*, arXiv:2302.14021.

Nikki A. Puccetti, William J. Villano, Jonathan P. Fadok, and Aaron S. Heller. 2022. Temporal dynamics of affect in the brain: Evidence from human imaging and animal models. *Neuroscience & Biobehavioral Reviews*, 133.

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

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.

Yu Zhu, Hao Li, Yikang Liao, Beidou Wang, Ziyu Guan, Haifeng Liu, and Deng Cai. 2017. What to do next: Modeling user behaviors by time-lstm. In *IJCAI*, volume 17, pages 3602–3608. Melbourne, VIC.