SemEval-2026 Task 2: Predicting Variation in Emotional Valence and Arousal over Time from Ecological Essays

Nikita Soni[1], **H. Andrew Schwartz**[1,2], **Ryan L. Boyd**[3,4], **Phi Long Bui**[2,5], **Syeda Mahwish**[6], **August Håkan Nilsson**[7], **Adithya V Ganesan**[1,2], **Lyle Ungar**[8], **Niranjan Balasubramanian**[1], **Saif M. Mohammad**[9]

[1]Dept. of Computer Science, Stony Brook University [2]College of Connected Computing, Vanderbilt University

[3]University of Texas at Dallas, Dept. of Psychology [4]Texas Artificial Intelligence Research Institute

[5]Institute for Human-Centered AI, Stanford University [6]Cornell University [7]Oslo Metropolitan University, Oslo Business School [8]Dept. of Computer Science, University of Pennsylvania [9]National Research Council Canada

{nisoni@cs.stonybrook.edu, hansen.schwartz@vanderbilt.edu}

Abstract

We present our shared task on predicting variation in emotional valence and arousal over time from ecological essays. The shared task uses a longitudinal dataset collected over 7 data collection phases of 14-day each spanning from 2021 to 2024, consisting of real-time essays and feeling words (e.g., happy, calm, sad, etc.) written in English by U.S. service-industry workers about "how they are feeling". Each text is associated with self-reported valence (V) (0 - 4, highly negative to highly positive affect) and arousal (A) (0 - 2, low to high energy) scores. The shared task consists of three parts, Subtask (1): Longitudinal Affect Assessment, Subtask (2): Forecasting Variation in Affect as a (2a): *state change*, and (2b): *disposition change*.

The task attracted over 200 member registrations on Codabench, receiving official system submissions from 31 teams (total 104 team members), of which 28 teams (with 90 team members) submitted system description papers making it to our leaderboard. We discuss baseline results along with findings from 28 systems, highlighting the best-performing systems, a deeper analysis of performance on essays versus feeling words, and assessments for authors seen versus unseen during training. The datasets for this task are publicly available.

1 Introduction

Emotions are a fundamental aspect of human experience, shaping how people navigate their relationships, make decisions, and maintain their well-being (Cacioppo and Gardner, 1999). The widely used affective circumplex model proposes that all emotions can be described as points in a two-dimensional space defined by valence (pleasantness) and arousal (activation) (Russell and Barrett, 1999; Posner et al., 2005; see Figure 2). Computational approaches to affect assessment have benefited a range of applications, from measuring subjective well-being (Diener, 2000) to tracking

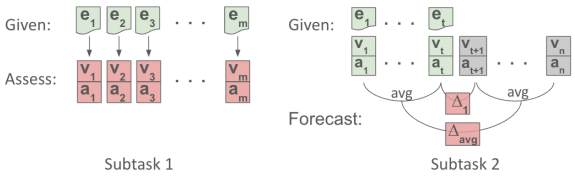

Figure 1: Tasks to assess the longitudinal affect (valence and arousal), and to forecast the change in affect, given ecological essays (and/or feeling words) written by the authors.. $\Delta_1 = v_{t+1} - v_t$, and $\Delta_{avg} = Average(v_1, .., v_{t+1}) - Average(v_t, .., v_n)$.

psychological states relevant to mental health (Eichstaedt et al., 2015; Coppersmith et al., 2017). Yet most prior NLP work on affect has relied on social media datasets in which momentary emotion is judged by third-party annotators (Mohammad et al., 2018; Preoţiuc-Pietro et al., 2016; AlZoubi et al., 2022)—an approach that conflates *emotion expression* with *emotion experience* (Vine et al., 2020; Troiano et al., 2019). Annotators can label what appears to be expressed in a social media post, but they cannot directly access the internal affective state of the person who wrote it. Moreover, emotions are inherently subjective and dynamic: they fluctuate across situations, times of day, and longer developmental timescales (Watson, 2000; Bolger et al., 2003a). To the best of our knowledge, no publicly available NLP dataset captures these characteristics—that is, no existing resource provides longitudinal, ecologically situated, *self-reported* affect labels grounded in people's own lived experience.

This shared task (hosted at SemEval-2026 (Ghosh et al., 2026)) aims to address this gap by introducing a **longitudinal** dataset ('ecological essays' and 'feeling words') collected over multiple years (2021 to 2024) and consisting of real-time essays and feeling words written by U.S. service industry workers in response to the prompt "how are you feeling." These

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 3731–3752

chronological texts represent *ecologically embedded affect*—language produced in the natural flow of daily life rather than under laboratory conditions (Trampe et al., 2015)—and each text is paired with the author's own self-reported valence and arousal scores over an affective circumplex (Figure 2). Because the data span repeated assessments within individuals over time, they allow us to analyze changes in affect over shorter timescales (e.g., different times of day, different days) as well as longer periods (e.g., across months and years), capturing both within-person dynamics and between-person differences in affective trajectories.

This task represents a shift toward modeling emotion as a lived, dynamic experience rather than an annotated perception of expressed emotion or a single static snapshot. Unlike social media datasets, which often must rely on performative or the "perceived" affect as annotated by third parties, the longitudinal 'ecological essays'and 'feeling words' offer introspective, self-reported data grounded in naturalistic, real-world settings. This allows for the development of models that not only generalize across individuals but also adapt to the emotional rhythms of a single person over time, crucial for building tools that are emotionally intelligent and personalized. The data's temporal depth and free-text format provide a rare opportunity to uncover how subtle verbal behavioral cues track with self-identified internal affective states, enabling research into affective trajectories, tipping points, and resilience. By modeling how emotions unfold and are conveyed in self-described experience — rather than merely how they are perceived by outside parties — we open the door to next-generation affective systems that can proactively support mental health, augment therapeutic processes and interventions, and bridge into smarter, ambient, emotion-aware technologies.

The shared task consists of three parts, Subtask 1: Longitudinal Affect Assessment, Subtask 2a: Forecasting Variation in Affect as a *state change*, and Subtask 2b: Forecasting Variation in Affect as a *disposition change*. Subtask 1 had held out ecological essays and feeling words from authors seen during training as well as unseen authors. Each team could participate in one or more of these subtasks to assess affect. Our official evaluation metrics were a composite pearson correlation for subtask 1 combining the pearson correlation for between-author and within-author, and pearson correlation for subtask 2a and 2b (refer Section 4.3). We rank the participating teams for each subtask by averaging the correlations for valence and arousal respectively. Our task attracted over 200 participants on Cod- abench, with 104 participants submitting an official submissions of which 90 participants (28 teams) submitted system description papers. Across systems, fine-tuning transformer-based models were prevalent for affect assessment performance, where arousal prediction proved to be a relatively harder task than valence prediction. Forecasting tasks were inherently difficult, where systems highlighted the importance of explicit temporal modeling and the challenging problem of predicting longer-term dispositional change. All task details and resources are available on the task website and GitHub page[1].

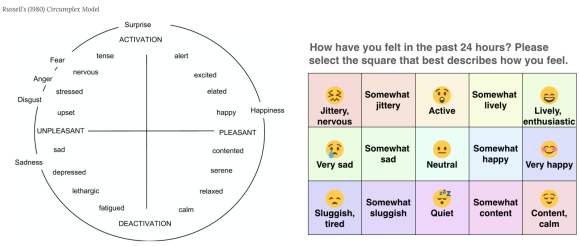

Figure 2: Left: Affective circumplex model reflecting the two-dimensional space of valence (pleasantness) and arousal (activation) used to describe emotions. Right: The affective circumplex grid used to self-report affect during the collection of 'ecological essays' and 'feeling words' dataset.

2 Related Work

The NLP community has huge collections of datasets labeled with "sentiment" (e.g., how people like a given movie (McAuley and Leskovec, 2013) or product (Asghar, 2016), based on the number of stars they give it), and recent work shows such sentiment is better assessed within the context of the author (Soni et al., 2026). However, there is a dearth of datasets that are labeled with actual first-person affective experience. While third-party annotations are useful for understanding perceived emotional meaning helping applications such as conversational affect modeling (Mohammad et al., 2018), self-reports along with longitudinal and ecological language can give us insights on first-order experiences and emotions. Studying such data will

[1] https://semeval2026task2.github.io/SemEval-2026-Task2/, https://github.com/semeval2026task2/EmotionValArouTimeVariation2026

allow the community to understand the similarity and difference between first-party experiences and third-party annotations.

Furthermore, prior work has placed increased importance on processing language within the context of the author (Soni et al., 2022, 2024a), time (Matero et al., 2021), and situation (Soni et al., 2025b) to better assess affect (Soni et al., 2025a; Singh et al., 2025) and mental health (Ganesan et al., 2022; Varadarajan et al., 2024; Mangalik et al., 2024). Temporal patterns and anomalies in human behavior exhibited through their language can be essential for well-being and mental health assessments (Soni et al., 2024b) as well as assist in measuring the dynamic human states and more stable traits (Singh et al., 2025). The dataset in our shared task as well as the formulation of our Subtask 2 to forecast variations in affect provides an opportunity to deepen research in the directions of longitudinal affect and mental health. Being able to assess affect longitudinally, can also offer a step towards assessment tools for personalized mental health care, and will eventually allow better connection to objective behaviors.

3 Data

The dataset consists of *chronological texts* ('ecological essays' and 'feeling words' (using their own words e.g., happy, calm, sad, etc.)) written in *English* by U.S. service industry workers collected over 7 data collection phases from 2021 to 2024. Each collection phase consisted of 14-day periods where each participant was prompted to write an essay on how they are feeling up to three times a day. In addition, the participants were also asked to fill in the affective circumplex grid selecting exactly one cell (on the right of Figure 2), providing a direct measure of experienced valence (V) (highly negative (0) to highly positive (4) along the X-axis) and arousal (A) (low (0) to high (2) along the Y-axis). Therefore, V & A scores were derived from participants' self-reported current mood state, instead of being assigned by third-party annotators. Each of the texts is associated with self-reported valence (0–4, highly negative to highly positive affect) and arousal (0–2, low to high energy) scores.

The way we describe and assign emotion words/labels to our feelings is a rich psychological process that is diagnostic of how our past experiences influence our current perception of how we relate to the world. Therefore, we include the feeling words data as well in our training and evaluation sets. Additionally, this increases the number of examples. We suggested our shared task participants to consider modeling feeling-words separately or jointly with the essays.

The dataset underwent a systematic cleaning process. Raw text data collected across multiple collection phases were processed separately in a standardized fashion, ensuring consistent formatting for timestamps/timezone information (retaining the user's local timezone), user identifiers, and survey responses. Entries with duplicate identifiers or timestamps were identified and removed to maintain uniqueness and accuracy, retaining the first and most complete entry for any given essay/feeling words prompt. The data was then filtered to exclude incomplete records, particularly those lacking necessary affective ratings (valence and arousal). Sensitive or personally identifiable information within text responses was removed using regular expressions to detect and eliminate patterns such as email addresses and URLs. The cleaned dataset was finally structured into a unified format suitable for longitudinal affect analysis.

The data collection study participants included in the dataset had a minimum of three *chronological texts* (essays and feeling words). Each text contains either an essay of at least five words or 'feeling words' with at least three words. Each text is associated with 'valence' and 'arousal' labels. This yielded 5,285 longitudinal texts written by 182 authors. In total, the average number of texts per user are 72.8 (average essays per users: 53.1; average feeling-words per user: 48.3), and the median number of texts per user are 35 (median essays per users: 18.00; median feeling-words per user: 18.00). The dataset structure with selected rows from the data as examples are shown in Appendix Table 6, and data statistics are presented in Appendix Tables 7, 8, and 9.

4 Task Description

The task consists of multiple subtasks and participants could participate in one or more of these subtasks to assess affect.

4.1 Subtasks

Subtask 1—Longitudinal Affect Assessment: Given a sequence of m essays and/or feeling words as texts, e_1 ... e_m in chronological order, this subtask includes producing V & A scores, (v_1, a_1)

... (v_m, a_m), one pair for each text (refer Figure 1). The training set includes sequences of texts and their associated V & A scores from multiple people. For this subtask, there are two types of test set data: (1) *unseen users* – those for whom no data was seen during training and (2) *seen users*– those that were also in the training data (test set including texts over a future period of time). The test data is marked to identify seen and unseen users.

Subtask 2a and 2b—Forecasting (future) Variation in Affect: Given a sequence of the first t essays and/or feeling words along with their associated V & A scores, this subtask includes forecasting two changes in V & A scores (refer Figure 1): (a) *state change*: from the last timestep observed to the next: $\Delta_1 = v_{t+1} - v_t$, (b) *dispositional change*: from the average observed to the average of an equally-sized timespan in the future: $\Delta_{avg} = \text{avg}(v_{t+1:n}) - \text{avg}(v_{1:t})$.

4.2 Task Organization

We used Codabench[2] as the data release and system submission platform, and a website and Google Groups to disseminate information throughout the competition. Our task attracted 211 member registrations on Codabench, with 351 submissions. We received official system submissions from 32 teams (total of 104 team members), of which 28 teams (a total of 90 team members) submitted system description papers making it to our leaderboard, having affiliations from different parts of the world, as shown in Figure 3. We provided participants with a pilot dataset in the beginning to help them understand the task and dataset structure. We also held online sessions during the training phase and later for paper writing guidance. During the training phase, participants were allowed to make sample submissions to run through our submissions' format checker, while in the evaluation phase participants submitted their system predictions on the held out test data with the last entry considered as their official submission for ranking. The task participants did not have access to the results of their submissions until the end of the evaluation phase.

4.3 Evaluation and Baselines

Evaluation Metrics All systems are evaluated using Pearson correlation (r) and Mean Absolute Error (mae) for each affect outcome: valence and arousal. Pearson correlation is used for interpretability with a bounded measure, whereas mae captures the absolute scale. For Subtask 1 (Longitudinal Affect Assessment), evaluation captures both between-user differences and within-user temporal variation. The between-user metric computes r_{between} and mae_{between} across users by comparing the mean predicted and gold scores aggregated over each user's texts. The within-user metric computes r_{within} and mae_{within} across texts for each user and then averages the results across users. To balance both aspects, a composite correlation ($r_{\text{composite}}$), and composite mae ($mae_{\text{composite}}$) is computed by combining between- and within-user correlations and mae using Fisher's z-transformation. The Fisher's z-transformation converts correlations into a space where they are normally distributed and ensures the final composite correlation remains within the bounded correlation measure requiring high performance across both the stable traits (between-user) and dynamic states (within-user) of affect. For Subtasks 2a and 2b (Forecasting Variation in Affect), performance is evaluated at the user level using Pearson r and MAE between predicted and gold state change and disposition change values. Systems are ranked using the average of $r_{\text{composite}}$ across V & A scores for Subtask 1, and the average of r across V & A scores for Subtasks 2a and 2b each. We define each metric used for Subtask 1, computed for valence and arousal separately, as follows:

$$f_{\text{between}}(\{\hat{y}_{u,t}\}, \{y_{u,t}\}) = f\Big(\{\text{mean}_{t \in u}(\hat{y}_{u,t})\}_{u=1}^{N}, \{\text{mean}_{t \in u}(y_{u,t}))\}_{u=1}^{N}\Big) \quad (1)$$

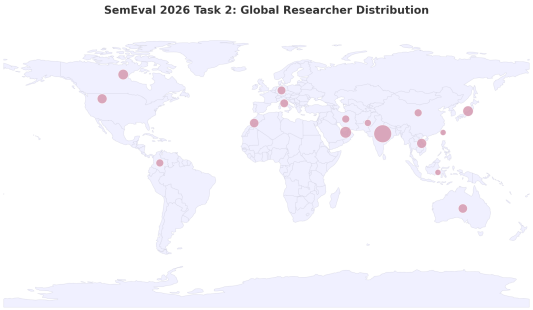

Figure 3: The official affiliations of the 90 participants who submitted a system description paper ranged across 16 countries: India, the United Arab Emirates, Canada, Japan, the United States, Vietnam, Australia, Morocco, Germany, Italy, China, Colombia, Iran, Pakistan, Indonesia, and Taiwan. Circle sizes are proportional to the number of researchers affiliated with each country.

[2] https://www.codabench.org/competitions/9963/

$$f_{\text{within}}(\{\hat{y}_{u,t}\}, \{y_{u,t}\}) = mean_{\forall u}\Big(f_u(\{\hat{y}_{u,t}\}_{t\in u}, \{y_{u,t}\}_{t\in u})\Big) \quad (2)$$

$$f_{\text{composite}} = \tanh\left(\frac{\text{arctanh}(f_{\text{within}}) + \text{arctanh}(f_{\text{between}})}{2}\right) \quad (3)$$

where f is in [pearson r, MAE], and t are the *texts* written by the user u.
And, metrics for Subtask 2a and 2b, computed for valence and arousal separately, are defined as:

$$f(\hat{y_u}, y_u) \quad (4)$$

where f is in [pearson r, MAE] and u is a user.

Baselines. We run simple random and linear baselines across the subtasks. Random baseline for subtask 1 predicts the global mean valence and arousal from the training set for every text in the test set, and for subtask 2a and 2b forecasts no state change and disposition change (i.e., zero) for all users in the test set. For linear baselines, we train L2-regularized linear regression (ridge regression) models using BERT-base-uncased mean-pooled token embeddings. *linear(BERT)* predicts valence and arousal directly from text embeddings. For forecasting affect change in Subtasks 2a and 2b, we evaluate two additional variants: *linear(BERT; previous)* uses the current text embedding together with the current valence/arousal value as features, while *linear(previous)* uses only the current valence/arousal value without text representations. For state change (2a), the target is the difference between consecutive affect values ($\Delta_i = V_{i+1} - V_i$). For disposition change (2b), users' texts are split chronologically into two halves and the target is the difference between the mean affect of the second and first halves.

5 Participating Systems and Results

We report results only for the teams that submitted a system description paper. A total of 25 teams made the rankings on the official leaderboard, while 3 additional teams were included in post-deadline leaderboard without a ranking. We discuss the results for Subtask 1 with 28 participating teams in Section 5.1, the results for Subtask 2a with 17 participating teams in Section 5.2, and the results for Subtask 2b with 13 participating teams in Section 5.3. We present a deeper analysis for *seen* and *unseen* users performance (Section 5.4), *ecological essays* and *feeling words* results (Section 5.5), methods analysis (Section 5.6), and a final discussion in Section 6.

5.1 Subtask 1: Longitudinal Affect Assessment

Table 1 reports Subtask 1 leaderboard ranking including valence and arousal $r_{composite}$ results for participating teams. The top three valence scores, NLPGROUP8 (Arthur et al., 2026) ($r = 0.688$), LAMANHNGUYEN (NGUYEN, 2026) ($r = 0.687$), and CURIOSAI (Beppu et al., 2026) ($r = 0.683$), were separated by fewer than 0.006 points and all used RoBERTa or DeBERTa backbones.

Team	Valence (V) $r_{composite}$ ↑	Arousal (A) $r_{composite}$ ↑	V&A Avg
UKP_Psycontrol	0.667	**0.554**	**0.611**
YNU	0.677	0.528	0.603
cclin	0.647	0.527	0.587
AFourP	0.679	0.466	0.573
lamanhnguyen	0.687	0.458	0.573
CSIRO-LT	0.656	0.488	0.572
CuriosAI	0.683	0.451	0.567
Bison AI4PC	0.665	0.468	0.567
UIT	0.637	0.489	0.563
mcmaster4z03	0.665	0.460	0.562
Perspicere	0.623	0.497	0.560
NLPGroup8	**0.688**	0.416	0.552
Cherish	0.596	0.505	0.551
Ajman Univ.	0.656	0.439	0.548
VerbaNex AI	0.632	0.463	0.547
IMEZO/Khaleesi	0.656	0.437	0.547
AI4PC-Howard	0.631	0.462	0.546
LexMachina	0.645	0.434	0.539
Emo-tica	0.645	0.409	0.527
AGI	0.600	0.452	0.526
EcoAffectTrack	0.663	0.373	0.518
UAlberta	0.556	0.444	0.500
NLP-FSDM	0.546	0.453	0.499
VAP-GameCtrl.	0.615	0.322	0.469
linear(BERT)	0.557	0.299	0.428
One and Only	0.527	0.315	0.421
rand	0.000	0.000	0.000
Momentum	0.638	0.455	0.547
ES4MLL	0.650	0.433	0.541
Draken	0.594	0.296	0.445

Table 1: Subtask 1 Leaderboard. Performance is reported using Pearson $r_{\text{composite}}$ for valence (V) and arousal (A), along with their average (V&A Avg). Cell colors indicate a performance heatmap (green = higher correlation, peach = lower correlation); * $p < 0.01$ $^+$ $p < 0.05$; **bold** values denote the best performance per column. Blue-shaded rows represent standard baselines. Gray-shaded rows represent post-deadline submissions and are excluded from the official rankings.

Looking more closely at what distinguished them: NLPGROUP8 constructed five independent models with different output objectives (sigmoid regression, ordinal decomposition via SoftMax, and binary threshold formulations), pooling their predictions through averaging. This ensembling approach shows that heterogeneous predictors reduce systematic error. LAMANHNGUYEN trained two DeBERTa variants with complementary methods: one regularized with strict early stopping to favour stability, the other trained longer to increase sensitivity to high-variance emotional states, with a weighted linear combination balancing bias and variance across the label distribution. CURIOSAI used a two-phase training curriculum, first on EmoBank for broadly grounded dimensional affect regression, then on the task data, achieving domain adaptation that anchored the model's prediction scale before task-specific fine-tuning.

For arousal, the rankings shifted considerably: UKP_PSYCONTROL (Hryhoryeva et al., 2026) ($r = 0.554$), YNU (Lan et al., 2026) ($r = 0.528$), and CCLIN (Lin, 2026) ($r = 0.527$) led, none of which ranked in the top five for valence. This divergence is consistent with a broader pattern in affective computing: valence and arousal differ both statistically and psycholinguistically. In this dataset, valence spans five discrete levels (-2 to $+2$) whereas arousal spans three (0 to 2), yielding lower variance for arousal. At the psycholinguistic level, valence is well-captured by lexical choices recoverable by standard language models(Warriner et al., 2013), whereas arousal more closely tracks physiological and paralinguistic intensity cues less directly encoded in written text (Gunes and Pantic, 2010). This may partly explain why approaches incorporating temporal dynamics and personal context such as UKP_PSYCONTROL's user-aware GPT-5 prompting with discrete emotion labels, and YNU's time-aware LSTM gated on inter-post intervals, benefited arousal prediction relatively more than valence. YNU also ranking 5th on valence ($r = 0.677$) suggests temporal modeling improved both dimensions, though its relative benefit is larger for the lower-variance arousal target. We note that approaches incorporating temporal dynamics and personal context may disproportionately benefit arousal because the isolated text instance carries less of the relevant signal for that dimension.

Twenty-five of the twenty-eight teams surpassed the *linear(BERT)* baseline on valence; while all but one surpassed it on arousal. Three teams fell below baseline on valence, all relying on either zero-shot LLM inference or models trained from scratch without established pre-trained representations.

5.2 Subtask 2a: Forecasting State Change (Δ_1)

Table 2 reports Subtask 2a leaderboard ranking including pearson r for valence and arousal *state change* for participating teams. The *linear (prev)* baseline that does not use linguistic features ($r = 0.615$ valence, $r = 0.670$ arousal) captured a large share of forecasting affect by forwarding each user's most recent affect values. Four teams surpassed it on valence: YNU (Lan et al., 2026) (0.692), UKP_PSYCONTROL (Hryhoryeva et al., 2026) (0.675), UIT (Phuong et al., 2026) (0.629), and CSIRO-LT (Chen et al., 2026) (0.621), while only UKP_PSYCONTROL (Hryhoryeva et al., 2026) (0.683) and UALBERTA (Ho et al., 2026) (0.674) exceeded it on arousal.

Team	Valence (V) $r \uparrow$	Arousal (A) $r \uparrow$	V&A Avg
UKP_Psycontrol	0.675*	**0.683***	**0.679**
YNU	**0.692***	0.647*	0.670
UAlberta	0.615*	0.674*	0.645
linear(prev)	0.615*	0.670*	0.643
Ajman Univ.	0.615*	0.670*	0.642
UIT	0.629*	0.633*	0.631
CSIRO-LT	0.621*	0.477*	0.549
AI4PC-Howard	0.597*	0.413*	0.505
linear(B;p)	0.430*	0.405*	0.418
Emo-tica	0.424*	0.355+	0.390
CuriosAI	0.467*	0.275	0.371
linear(BERT)	0.290*	0.199*	0.245
Bison AI4PC	0.379*	0.085	0.232
NLPGroup8	0.152	0.126	0.139
rand	0.000	0.000	0.000
EcoAffectTrack	-0.243	-0.011	-0.127
AGI	-0.167	-0.147	-0.157
lamanhnguyen	-0.273	-0.275	-0.274
Cherish	NaN	NaN	NaN
Momentum	0.553*	0.589*	0.571
One and Only	-0.194	-0.423*	-0.308

Table 2: Subtask 2a leaderboard. Performance is reported using Pearson r for valence (V) and arousal (A) *state change*, along with their average (V&A Avg). Cell colors indicate a performance heatmap; * $p < 0.01$ + $p < 0.05$; **bold** denotes the best per column. Blue-shaded rows represent standard baselines. Gray-shaded rows represent post-deadline submissions and are excluded from the official rankings

All four teams that surpassed *linear(prev)* on valence used an explicit sequential component conditioned on temporal history: YNU fed

log-transformed inter-post intervals into a time-aware LSTM; UIT processed a sliding window (k=8) of prior texts through a Mixture-of-Experts head; CSIRO-LT concatenated a 24×1 temporal feature embedding with the [CLS] token; and UKP_PSYCONTROL incorporated the previous state change with a trainable user embedding into a feed-forward regressor. In contrast, NLPGROUP8's transformer-decoder architecture achieved the highest Subtask 1 valence score ($r = 0.688$) but scored ($r = 0.152$) lower than all baselines on state-change valence. This gap highlighted that representations optimized for cross-sectional affect assessment did not automatically transfer to temporal change prediction, where the target is an *incremental* quantity.

5.3 Subtask 2b: Forecasting Dispositional Change (Δ_{avg})

Subtask 2b required predicting the long-term average shift in a user's emotional baseline between two longitudinal segments, the most demanding subtask, with a target signal of substantially lower variance than either the instantaneous scores or the state changes. Table 3 reports Subtask 2b leaderboard ranking including pearson r for valence and arousal *disposition change* for participating teams.

UALBERTA (Ho et al., 2026) achieved the highest overall scores ($r = 0.405$ valence, $r = 0.602$ arousal), yet did not surpass the *linear(prev)* baseline on valence ($0.405 < 0.434$) and only marginally exceeded it on arousal (0.602 vs. 0.584). Eight of thirteen teams produced negative valence correlations, meaning the majority of submitted systems systematically inverted the direction of long-term dispositional shift. Only five teams yielded positive valence correlations (UALBERTA (Ho et al., 2026), NLPGROUP8 (Arthur et al., 2026), EMO-TICA (Noor and Fatima, 2026), AI4PC-HOWARD (Shah et al., 2026), and AGI (Rathva, 2026)), while LAMANHNGUYEN (NGUYEN, 2026) performed worst on both dimensions, again suggesting that momentum-based dampening is counterproductive for long-term dispositional prediction.

An indirect prediction strategy separated UALBERTA from the rest: rather than regressing a single per-user change label, they predicted mean affect for each longitudinal segment separately using a BERT + BiLSTM pipeline, then derived dispositional change by differencing predicted group means. This allowed the model to leverage the same affect-prediction signal as Subtask 1, avoiding the need to independently learn a long-term difference signal from sparse training examples. EMO-TICA similarly sidestepped direct text modeling, fitting ridge regression on user-level trajectory statistics (mean, standard deviation, trend slope, extrema, time span). Aggregation-based approaches, i.e., sequential predictions and hand-crafted trajectory features, were more robust than end-to-end neural approaches attempting to forecast the dispositional change target directly from text.

Team	Valence (V) $r \uparrow$	Arousal (A) $r \uparrow$	V&A Avg
linear(prev)	**0.434***	0.584*	**0.509**
UAlberta	0.405*	**0.602***	0.503
NLPGroup8	0.354^{+}	0.388*	0.371
Emo-tica	0.257	0.418*	0.337
AI4PC-Howard	0.046	0.348*	0.197
Ajman Univ.	-0.124	0.456*	0.166
AGI	0.086	-0.081	0.003
rand	0.000	0.000	0.000
linear(B;p)	-0.029*	0.019*	-0.005
EcoAffectTrack	-0.243	0.226	-0.009
linear(BERT)	-0.088*	0.070*	-0.009
CSIRO-LT	-0.147	0.114	-0.017
CuriosAI	-0.161	0.011	-0.075
Bison AI4PC	-0.120	-0.103	-0.111
UIT	-0.169	-0.060	-0.114
lamanhnguyen	-0.398*	-0.577*	-0.488
One and Only	-0.185	0.016	-0.084

Table 3: Subtask 2b leaderboard. Performance is reported using Pearson r for valence (V) and arousal (A) *disposition change*, along with their average (V&A Avg). Cell colors indicate a performance heatmap; * $p < 0.01$ $^{+}$ $p < 0.05$; **bold** denotes the best per column. Blue-shaded rows represent standard baselines. Gray-shaded rows represent post-deadline submissions and are excluded from the official rankings

5.4 Seen-User vs. Unseen-User Performance

Table 4 reports the r_{comp} for *seen users* (users present during training, evaluated on future texts, referring to as prospective generalization) and *unseen users* (users absent during training, referring to as cross-sectional generalization), for valence and arousal respectively. The two rightmost columns (Δ_V, Δ_A) show the difference in score between unseen and seen users for each dimension; positive values indicate better cross-sectional performance than prospective performance.

Valence: the baseline collapses for unseen users; submitted systems near-invariant or slightly worse. The *linear(BERT)* baseline achieves the highest seen-user valence of any system ($r =$

0.748), yet drops to $r = 0.562$ for unseen users ($\Delta_V = -0.186$, the largest drop in the table). A linear regression fitted on BERT features effectively learns a user-specific intercept for seen individuals, making it a strong prospective predictor; when applied to new users with no training history that advantage disappears, and it reverts to population-level predictions. Among submitted systems, 16 of the 25 show $|\Delta_V| \leq 0.03$, meaning their valence predictions were essentially equally good for seen and unseen users. AI4PC-HOWARD (Shah et al., 2026) stands out with $\Delta_V = -0.002$, the smallest drop in the table, suggesting its DeBERTa + GRU architecture generalized robustly across user types without requiring explicit seen/unseen routing. YNU (Lan et al., 2026) similarly showed near-invariance ($\Delta_V = -0.007$), suggesting its time-aware LSTM generalized via temporal text dynamics rather than user-specific signal. While ten teams show positive Δ_V (better on unseen), approaches without strong user-specific modeling often generalize as well or better to new individuals. The largest drops among submitted systems, UALBERTA (Ho et al., 2026) (-0.111) and CURIOSAI (Beppu et al., 2026) (-0.069), both relied on architectures designed for prospective prediction with access to prior user history.

Arousal: a near-universal improvement for unseen users. Performance in arousal prediction was opposite of valence. 23 of the 25 teams show a *positive* Δ_A, meaning most systems predicted arousal *better* for unseen users than for seen users. The gains are substantial: NLPGROUP8 (Arthur et al., 2026) improves by $+0.304$ (0.297 seen vs. 0.601 unseen), VERBANEX AI (Moreno et al., 2026) by $+0.269$, AFOURP (Thota et al., 2026) by $+0.246$, and LEXMACHINA (Ganguli et al., 2026) by $+0.231$. Only three systems show negative Δ_A: the *linear(BERT)* baseline (-0.184), UKP_PSYCONTROL (Hryhoryeva et al., 2026) (-0.068), and UALBERTA (Ho et al., 2026) (-0.048).

This asymmetry between valence and arousal is consistent with their distinct psycholinguistic properties. Within-person *valence* tends to track topically driven changes in a user's writing such as positive or negative events leave durable lexical traces in ongoing essays (Bolger et al., 2003b; Warriner et al., 2013), making prospective valence prediction feasible even without explicit temporal modeling. Within-person *arousal*, by contrast, reflects activation and physiological intensity which fluctuates more rapidly and is encoded less stably in written language (Gunes and Pantic, 2010). As a result, seen-user arousal prediction (requiring prospective modeling of a specific individual's activation dynamics) is harder than unseen-user arousal prediction (which benefits from the more stable between-person text-level signal that a highly activated writer differs from a calm writer).

Team	Seen Users		Unseen Users		Δ_V	Δ_A
	Val.	Aro.	Val.	Aro.		
UKP_Psycontrol	0.688	0.568	0.662	0.500	−0.026	−0.068
YNU	0.684	0.466	0.677	0.612	−0.007	+0.146
cclin	0.664	0.457	0.629	0.614	−0.035	+0.157
AFourP	0.696	0.359	0.662	0.605	−0.034	+0.246
lamanhnguyen	0.708	0.392	0.673	0.552	−0.035	+0.160
CSIRO-LT	0.650	0.395	0.673	0.598	+0.023	+0.203
CuriosAI	0.715	0.403	0.646	0.510	−0.069	+0.107
Bison AI4PC	0.664	0.404	0.673	0.553	+0.009	+0.149
UIT	0.639	0.405	0.620	0.555	−0.019	+0.150
mcmaster4z03	0.666	0.402	0.660	0.550	−0.006	+0.148
Perspicere	0.614	0.414	0.641	0.591	+0.027	+0.177
NLPGroup8	0.698	0.297	0.679	0.601	−0.019	+0.304
Cherish	0.604	0.465	0.594	0.534	−0.010	+0.069
Ajman Univ.	0.669	0.361	0.650	0.530	−0.019	+0.169
VerbaNex AI	0.606	0.356	0.664	0.625	+0.058	+0.269
IMEZO	0.648	0.386	0.675	0.519	+0.027	+0.133
AI4PC-Howard	0.633	0.415	0.631	0.506	−0.002	+0.091
LexMachina	0.636	0.343	0.669	0.574	+0.033	+0.231
Emo-tica	0.658	0.332	0.640	0.512	−0.018	+0.180
AGI	0.639	0.400	0.655	0.545	+0.016	+0.145
EcoAffectTrack	0.688	0.327	0.642	0.426	−0.046	+0.099
UAlberta	0.578	0.505	0.467	0.457	−0.111	−0.048
NLP-FSDM	0.541	0.444	0.556	0.447	+0.015	+0.003
VAP-GameCtrl.	0.615	0.264	0.618	0.400	+0.003	+0.136
linear(BERT)	0.748	0.422	0.562	0.238	−0.186	−0.184
One and Only	0.508	0.268	0.554	0.356	+0.046	+0.088

Table 4: Subtask 1 r_{comp} disaggregated by user type. Δ_V and Δ_A = unseen minus seen for valence and arousal respectively; green = better on unseen users, red = worse on unseen users. Teams sorted by Subtask 1 leaderboard order.

5.5 Essays vs. Feeling Words Performance

Table 5 (in appendix) reports r_{comp} separately for each text type: 'ecological essays' and 'feeling words', with $\Delta_V = r_{\text{words},V} - r_{\text{essays},V}$ and $\Delta_A = r_{\text{words},A} - r_{\text{essays},A}$ indicating the gain from feeling words over essays.

The most consistent finding is an arousal advantage for feeling words: All 25 submitted systems show a positive Δ_A, with a median gain of $+0.19$ arousal points. This is striking given an essay averages 58 words and most 'feeling words' are 3–5 words (with few instances having multiple

word phrases as a single 'feeling words', refer Appendix Table 7)). The pattern mirrors the arousal asymmetry observed for unseen users (Section 5.4): since arousal reflects general activation and intensity; emotion terms such as *anxious*, *exhausted*, or *energised* carry strong physiological connotations that are directly recoverable from the label lexicon, even by models that never see the full essay context, consistent with prior work on brief versus extended natural language response formats for psychological assessment (Gu et al., 2025).

Team	Essays		Feeling Words		Δ_V	Δ_A
	Val.	Aro.	Val.	Aro.		
UKP_Psycontrol	0.685	0.500	0.658	0.575	−0.027	+0.075
YNU	0.673	0.419	0.683	0.602	+0.010	+0.183
cclin	0.632	0.422	0.662	0.620	+0.030	+0.198
lamanhnguyen	0.659	0.358	0.677	0.578	+0.018	+0.220
AFourP	0.679	0.357	0.670	0.566	−0.009	+0.209
CSIRO-LT	0.623	0.388	0.673	0.598	+0.050	+0.210
CuriosAI	0.653	0.403	0.669	0.524	+0.016	+0.121
Bison AI4PC	0.647	0.333	0.665	0.568	+0.018	+0.235
UIT	0.586	0.397	0.693	0.526	+0.107	+0.129
mcmaster4z03	0.653	0.365	0.658	0.542	+0.005	+0.177
Perspicere	0.619	0.370	0.649	0.569	+0.030	+0.199
NLPGroup8	0.666	0.335	0.690	0.574	+0.024	+0.239
Cherish	0.618	0.380	0.586	0.576	−0.032	+0.196
Ajman Univ.	0.618	0.326	0.648	0.528	+0.030	+0.202
VerbaNex AI	0.599	0.437	0.684	0.515	+0.085	+0.078
IMEZO	0.645	0.395	0.662	0.573	+0.017	+0.178
AI4PC-Howard	0.631	0.343	0.650	0.540	+0.019	+0.197
LexMachina	0.627	0.307	0.655	0.572	+0.028	+0.265
Emo-tica	0.602	0.313	0.669	0.582	+0.067	+0.269
AGI	0.599	0.370	0.600	0.554	+0.001	+0.184
EcoAffectTrack	0.644	0.332	0.667	0.539	+0.023	+0.207
UAlberta	0.517	0.339	0.555	0.400	+0.038	+0.061
NLP-FSDM	0.513	0.359	0.572	0.516	+0.059	+0.157
VAP-GameCtrl.	0.581	0.278	0.646	0.318	+0.065	+0.040
linear(BERT)	0.546	0.395	0.470	0.158	−0.076	−0.237
One and Only	0.421	0.229	0.633	0.426	+0.212	+0.197

Table 5: Subtask 1 r_{comp} disaggregated by text type. Δ_V and Δ_A = feeling words minus essays per dimension; green = feeling words better, red = essays better. Teams sorted by Subtask 1 leaderboard order.

For valence, 22 of 25 teams show a positive Δ_V, but gains are modest (median +0.02), and three teams (UKP_PSYCONTROL (Hryhoryeva et al., 2026), AFOURP (Thota et al., 2026), CHERISH (Parahita, 2026)) along with our baseline (*linear(BERT)*) do *better* on essays for valence. Valence is well-encoded in rich lexical content: longer narratives provide syntactic negation, contextual framing, and sentiment-bearing sentences that a few terse emotion words cannot replicate (Warriner et al., 2013). Team ONE AND ONLY (Dinh, 2026) provides the clearest contrast: their zero-shot GPT-5 prompting gains $\Delta_V = +0.212$ from feeling words, the largest in the table, consistent with LLMs being very effective at interpreting short, emotionally direct labels but struggling to extract valence from unstructured narrative.

5.6 Methodological Analysis and Takeaways

Handling Temporal Structure and Extra-Linguistic Features. The dominant approach was to process individual posts independently through a Transformer encoder with a regression head, treating each post as an isolated regression instance. This proved sufficient for better performance on Subtask 1's instantaneous affect assessment but inadequate for the forecasting subtasks, where predicting change inherently requires modeling dynamics across time. Among teams that added an explicit sequential component (LSTM, GRU, or Temporal Fusion Transformer), Subtask 2a performance was consistently higher than architecturally similar teams that relied only on Transformer representations. Only a minority exploited the provided `timestamp` column; most treated temporal order implicitly through chronological sorting. Teams that encoded inter-post intervals as explicit features clustered above the *linear(prev)* baseline on Subtask 2a, while those without timestamp features clustered at or below it.

User Representations. User representation was a common strategy for Subtask 1 but was less uniformly helpful for the forecasting subtasks. A fixed-dimensional trainable user embedding concatenated with the text representation was the basic approach for most teams that outperformed the baseline on Subtask 1. As Table 4 illustrates, performance differences between seen and unseen users depended critically on how teams handled the cold-start problem: systems with explicit fallback mechanisms for unseen users (a dedicated `UNK` embedding, user-agnostic prompting, or a data-conditioned gate) tended to show smaller seen/unseen gaps than those relying on a fixed global fallback. LEXMACHINA's (Ganguli et al., 2026) adversarial approach (DANN) took a structurally different route, explicitly penalizing the encoding of user identity during arousal training to prevent models from collapsing to each person's mean score rather than modeling within-person dynamics.

Loss Functions. Standard MSE was the dominant training objective, but several teams adopted losses more closely aligned with the Pearson-correlation evaluation metric. NLPGROUP8 used a 90% Pearson R / 10% MSE composite, directly

optimizing the leaderboard criterion. AGI implemented a correlation-first phased training schedule with explicit variance-preservation terms to prevent prediction collapse, a problem they identified as particularly acute for arousal, where lower label variance can lead models to default to near-constant predictions. EcoAffectTrack (Kumar and Joshi, 2026), UIT (Phuong et al., 2026), and Ajman University (Jumakhan et al., 2026) used Concordance Correlation Coefficient (CCC) loss, which jointly penalizes deviations in mean, variance, and correlation.

6 Discussion

Across the three subtasks, several patterns emerged. Most teams fine-tuned pre-trained encoders (typically RoBERTa or DeBERTa) with regression heads resulting in similar performance for top Subtask 1 valence assessment systems. However, strong instantaneous affect prediction did not translate into better temporal change forecasts, where systems incorporating explicit temporal modeling performed well. Interestingly, a simple linear baseline using only current affect scores with no linguistic features proved to be a stronger forecaster of variation in affect than most systems, emphasizing the inherent challenge of the subtasks. Additionally, arousal assessments proved to be harder than valence assessments for all systems, consistent with previous work (Soni et al., 2025a), although generalization patterns differed with valence predictions slightly worse on unseen users while arousal predictions often improved. Further, feeling words reflected consistently stronger signals than essays for affect assessment, particularly for arousal, suggesting physiological intensity information may be directly recoverable from the label lexicon (Warriner et al., 2013), but at the same time highlighting the self-reported nature of the labels: feeling words likely mimic the rating scale more closely than essays do (Nilsson et al., 2025). We encourage future work to explore this distinction further, particularly for psychological states less amenable to self-report, such as clinical depression.

7 Conclusion

We introduced the SemEval-2026 Task 2 shared task on predicting variation in affective valence and arousal from longitudinal, ecologically embedded language. The task is built on a multi-year dataset of self-reported affect paired with chronological essays and feeling words written by U.S. service-industry workers. By framing affect prediction both as an assessment problem (Subtask 1) and a forecasting problem (Subtask 2), the task encouraged models to move beyond static sentiment classification/estimation toward capturing within-person affective dynamics over time. The task attracted substantial participation from the community, with over 200 participants on Codabench and 90 participants forming 28 teams submitting system description papers. Across systems, transformer-based models dominated affect assessment performance, while forecasting tasks highlighted the importance of explicit temporal modeling and the inherent challenge of predicting longer-term dispositional change.

Beyond benchmarking model performance, the shared task surfaces several broader insights about language and affect. First, representations optimized for cross-sectional affect prediction do not necessarily transfer to temporal forecasting tasks. Second, the asymmetry between valence and arousal prediction reflects both psycholinguistic properties of language and the statistical structure of the labels. Third, brief feeling-word responses often provide stronger signals for arousal than longer essays, suggesting that concise self-descriptions can capture affective intensity more directly than narrative text. We hope that the release of this dataset and the analyses presented here will stimulate further research at the intersection of NLP, psychology, and affective science.

Ethical Considerations

The "Data Science and Alcohol Consumption Study" to collect ‘ecological essays’ and ‘feeling words’ dataset and assess affect underwent ethical review by central IRB from the University of Pennsylvania with IRB agreement (IRB2019-00678) from Stony Brook University. Participants in the study were paid ∼$20 on signing up and ∼$1 per completed response (typically taking 1 minute). We secured further approval from IRB to consent our existing participants, using additional consent form via Qualtrics, to publicly share their anonymized daily survey responses for research purposes. We clearly explained to participants how their written essays would be anonymized, by removing any usernames, person names, email addresses, and URLs, and that, if they would give permission, we will publicly share the anonymized

data. Although users were not asked to enter any identifiable information in this data, we additionally removed all named entities (named persons, places, organizations, or things) or contact information (phone numbers, addresses, URLs) from the written responses via an automated named entity and contact recognizer. We also manually looked over to verify anonymization for all data to be released.

While our shared task predicts emotions from ecological essays and feeling words, such automated systems should not be mistaken for objective measures of emotional state. Language is inherently subjective and shaped by personal, social, and cultural contexts, meaning that similar emotions may be expressed differently across individuals. Ignoring this variability risks oversimplifying or misrepresenting emotions, making it essential to interpret predictions with caution and a clear awareness of these limitations. More on ethics for AI tasks is provided by Mohammad (2022).

Acknowledgments

This work was supported in part by a grant from the NIH-NIAAA (Data Science for Unhealthy Drinking; R01 AA028032), a grant from the CDC/NIOSH (U01 OH012476), and a grant from the NIH/NIMH (CREATE: Center for Advancing Therapy with AI; P50 MH 139450) . The conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, by any government organization or the U.S. Government.

References

Omar AlZoubi, Saja Khaled Tawalbeh, and AL-Smadi Mohammad. 2022. Affect detection from arabic tweets using ensemble and deep learning techniques. *Journal of King Saud University-Computer and Information Sciences*, 34(6):2529–2539.

Troy Arthur, Aidan Kelley, and Sierra Reschke. 2026. NLPGroup8 at SemEval-2026 Task 2: Diverse Ensembles and Hierarchical Transformers for Emotional State Prediction. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*.

Nabiha Asghar. 2016. Yelp dataset challenge: Review rating prediction. *arXiv preprint arXiv:1605.05362*.

Abdessamad Benlahbib, Zouhir Essalmani, Achraf Boumhidi, Anass Fahfouh, and Hamza Alami. 2026. NLP-FSDM at SemEval-2026 Task 2: Temporal Smoothing and CCC-MAE Optimization for Balanced Longitudinal Affect Assessment. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*.

Fumika Beppu, Hiroki Takushima, Aiswariya Kumar Manoj, Daichi Yamaga, Yuki Shibata, and Takayuki Hori. 2026. CuriosAI at SemEval-2026 Task 2: Predicting Emotion using RoBERTa-large model. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*.

Niall Bolger, Angelina Davis, and Eshkol Rafaeli. 2003a. Diary methods: Capturing life as it is lived. *Annual Review of Psychology*, 54:579–616.

Niall Bolger, Angelina Davis, and Eshkol Rafaeli. 2003b. Diary methods: Capturing life as it is lived. *Annual review of psychology*, 54(1):579–616.

John T. Cacioppo and Wendi L. Gardner. 1999. Emotion. *Annual Review of Psychology*, 50(1):191–214.

Jiyu Chen, necva bölücü, Sarvnaz Karimi, Diego Molla, and Cecile L. Paris. 2026. CSIRO-LT at SemEval-2026 Task 2: In-the-Wild Valence and Arousal Forecasting on Ecological Text Time Series. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*.

G. Coppersmith, C. Hilland, O. Frieder, and R. Leary. 2017. Scalable mental health analysis in the clinical whitespace via natural language processing. In *2017 IEEE EMBS International Conference on Biomedical Health Informatics (BHI)*, pages 393–396.

E. Diener. 2000. Subjective well-being: The science of happiness and a proposal for a national index. *The American Psychologist*, 55(1):34–43.

Nam Dinh. 2026. One and Only at SemEval-2026 Task 2: Evaluating Zero-Shot Autonomous LLM Agents and Heuristic Proxies in Ecological Affect Forecasting. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*.

Johannes C Eichstaedt, H. Andrew Schwartz, Margaret L Kern, Gregory Park, Darwin R Labarthe, Raina M Merchant, Sneha Jha, Megha Agrawal, Lukasz A Dziurzynski, Maarten Sap, Christopher Weeg, Emily E Larson, Lyle H Ungar, and Martin E P Seligman. 2015. Psychological language on Twitter predicts county-level heart disease mortality. *Psychological Science*, 26(2):159–169.

Adithya V Ganesan, Vasudha Varadarajan, Juhi Mittal, Shashanka Subrahmanya, Matthew Matero, Nikita Soni, Sharath Chandra Guntuku, Johannes Eichstaedt, and H Andrew Schwartz. 2022. Wwbp-sqt-lite: Multi-level models and difference embeddings for moments of change identification in mental health forums. In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*, pages 251–258.

Somdev Ganguli, Vibhan Dutta, Romit Datta, Amit Barman, and Sudip Kumar Naskar. 2026. LexMachina at 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)*.

Debanjan Ghosh, Kai North, Ekaterina Kochmar, Mamoru Komachi, and Marcos Zampieri, editors. 2026. *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*. Association for Computational Linguistics, San Diego, United States.

Zhuojun Gu, Katarina Kjell, H. Andrew Schwartz, and Oscar Kjell. 2025. Natural language response formats for assessing depression and worry with large language models: A sequential evaluation with model pre-registration. *Assessment*.

Hatice Gunes and Maja Pantic. 2010. Automatic, dimensional and continuous emotion recognition. *International Journal of Synthetic Emotions (IJSE)*, 1(1):68–99.

Duc Ho, Khanh Dao Duy Bui, Daniela Teodorescu, and Grzegorz Kondrak. 2026. UAlberta at SemEval-2026 Task 2: Temporal Fusion Models for Predicting Affect Over Time. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*.

Darya Hryhoryeva, Amaia Zurinaga, Hamidreza Jamalabadi, and Iryna Gurevych. 2026. UKP_Psycontrol at SemEval-2026 Task 2: Modeling Valence and Arousal Dynamics from Text. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*.

Haseebullah Jumakhan, Soud Assad, Seyed Abdullah, and Mahmoud Al-Ayyoub. 2026. Ajman University at SemEval-2026 Task 2: Overcoming Scale Collapse in Temporal Emotion Modeling via Residual Learning. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*.

Diya Satish Kumar and Om Sujal Joshi. 2026. EcoAffectTrack at SemEval-2026 Task 2: A Hierarchical DeBERTa-Transformer Framework with CCC Optimization for Longitudinal Affect Modeling. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*.

Xin Lan, Jin Wang, and Xuejie Zhang. 2026. YNU-HPCC at SemEval-2026 Task 2: Contrastive Calibration and Temporal Modeling for Continuous Valence-Arousal Prediction. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*.

Huy Le, Truong Thien Phu, Trung Tran, Nga Nguyen, and Monojit Choudhury. 2026. VAP-GameController at SemEval-2026 Task 2: Lexical-based and Emotion-Aware Approaches for Longtitudinal Emotion Prediction. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*.

Jing-Jun Lin. 2026. cclin at SemEval-2026 Task 2 : SLM-Enhanced Lightweight Multi-BERT Ensemble for Longitudinal Affect Assessment. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*.

Andrea Lolli, Chiara Lunazzi, Riccardo Coppola, and Flavio Giobergia. 2026. ES4MLL at SemEval-2026 Task 2: Set Attention Aggregation and Recurrent Temporal Modeling for Longitudinal Affect Prediction. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*.

Siddharth Mangalik, Johannes C Eichstaedt, Salvatore Giorgi, Jihu Mun, Farhan Ahmed, Gilvir Gill, Adithya V. Ganesan, Shashanka Subrahmanya, Nikita Soni, Sean AP Clouston, and 1 others. 2024. Robust language-based mental health assessments in time and space through social media. *NPJ Digital Medicine*, 7(1):109.

Matthew Matero, Nikita Soni, Niranjan Balasubramanian, and H Andrew Schwartz. 2021. Melt: Message-level transformer with masked document representations as pre-training for stance detection. In *Findings of the association for computational linguistics: EMNLP 2021*, pages 2959–2966.

Julian John McAuley and Jure Leskovec. 2013. From amateurs to connoisseurs: modeling the evolution of user expertise through online reviews. In *Proceedings of the 22nd international conference on World Wide Web*, pages 897–908.

Saif Mohammad. 2022. Ethics sheets for AI tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8368–8379, Dublin, Ireland. Association for Computational Linguistics.

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.

Melissa Andrea Moreno, Juan Carlos Martinez Santos, and Edwin Puertas. 2026. VerbaNex AI at SemEval-2026 Task 2: DeBERTa for Longitudinal Valence and Arousal Prediction. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*.

Supriya S. Nadiger, SUNIL SAUMYA, Rahul Pujari, VEERESH S. HIREMATH, Kiran A. Chikaraddi, and Anoop U. Kadkol. 2026. Momentum at SemEval-2026 Task 2: LongVA-RoBERTa, a transformer-Based Longitudinal Valence and Arousal Modeling. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*.

LAM ANH NGUYEN. 2026. lamanhnguyen at SemEval-2026 Task 2: Uncovering Lexical Bias and Momentum Lag in Longitudinal Emotion Prediction using Multi-task DeBERTa. In *Proceedings of the*

20th International Workshop on Semantic Evaluation (SemEval-2026).

August Håkan Nilsson, Ryan L. Boyd, Adithya V. Ganesan, Oscar Kjell, Syeda Mahwish, Haitao Huang, Richard N. Rosenthal, Lyle Ungar, and H. Andrew Schwartz. 2025. Language-based assessments for experienced well-being: Accuracy and external validity across behaviors, traits, and states. Manuscript under review.

Sadia Noor and Mehwish Fatima. 2026. Emo-tica at SemEval-2026 Task 2: Trait?State Affect Forecaster for Longitudinal Valence and Arousal. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*.

Cetta Reswara Parahita. 2026. Cherish at SemEval-2026 Task 2: Enhancing RoBERTa-Based Models for Emotional Valence and Arousal Prediction in Ecological Essays with Personalized PLoRA and Temporal Embeddings. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*.

Son The Phuong, My Thuy Tra Ngo, Tri Minh Dao, and Duc-Vu Nguyen. 2026. CITD@UIT at SemEval-2026 Task 2: Temporal Mixture-of-Experts for Longitudinal Valence and Arousal Prediction from Ecological Essays. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*.

Jonathan Posner, James A Russell, and Bradley S Peterson. 2005. The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Developmental Psychopathology*, 17(3):715–734.

Daniel Preoţiuc-Pietro, H. Andrew Schwartz, Gregory Park, Johannes Eichstaedt, Margaret Kern, Lyle Ungar, and Elisabeth Shulman. 2016. Modelling valence and arousal in Facebook posts. In *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 9–15, San Diego, California. Association for Computational Linguistics.

Harsh Rathva. 2026. "AGI" Team at 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)*.

James A. Russell and Lisa Feldman Barrett. 1999. Core affect, prototypical emotional episodes, and other things called emotion: Dissecting the elephant. *Journal of Personality and Social Psychology*, 76(5):805–819. Place: US.

Araj Shah, Utsav Shah, and Saurav K. Aryal. 2026. AI4PC-Howard University at SemEval-2026 Task 2: Fine-Tuning DistilBERT, DeBERTa and ModernBERT for Valence?Arousal Prediction and Change Estimation. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*.

Khushboo Singh, Vasudha Varadarajan, Adithya V. Ganesan, August Håkan Nilsson, Nikita Soni, Syeda Mahwish, Pranav Chitale, Ryan L. Boyd, Lyle Ungar, Richard N. Rosenthal, and H. Andrew Schwartz. 2025. Systematic evaluation of auto-encoding and large language model representations for capturing author states and traits. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 18955–18973, Vienna, Austria. Association for Computational Linguistics.

Rajalakshmi Sivanaiah, Angel Deborah S, Krishna Varun R, and Krishnaraj N. 2026. Draken at SemEval-2026 Task 2: Frozen BERT Embeddings with Ridge Regression for Predicting Emotional Valence and Arousal. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*.

Nikita Soni, Niranjan Balasubramanian, H Andrew Schwartz, and Dirk Hovy. 2024a. Comparing pretrained human language models: Is it better with human context as groups, individual traits, or both? In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 316–328.

Nikita Soni, Pranav Chitale, Khushboo Singh, Niranjan Balasubramanian, and H. Andrew Schwartz. 2025a. 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, Dhruv Vijay Kunjadiya, Pratham Piyush Shah, Dikshya Mohanty, H Andrew Schwartz, and Niranjan Balasubramanian. 2026. Addressing the ecological fallacy in larger lms with human context. *arXiv preprint arXiv:2603.05928*.

Nikita Soni, Matthew Matero, Niranjan Balasubramanian, and H. Andrew Schwartz. 2022. Human language modeling. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 622–636, Dublin, Ireland. Association for Computational Linguistics.

Nikita Soni, August Håkan Nilsson, Syeda Mahwish, Vasudha Varadarajan, H Andrew Schwartz, and Ryan L Boyd. 2025b. Who we are, where we are: Mental health at the intersection of person, situation, and large language models. In *Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2025)*, pages 300–313.

Nikita Soni, H. Andrew Schwartz, João Sedoc, and Niranjan Balasubramanian. 2024b. Large human language models: A need and the challenges. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8631–8646, Mexico City, Mexico. Association for Computational Linguistics.

Eleale Nusi Tee. 2026. Khaleesiyali at SemEval-2026 Task 2: Lexicon-Augmented RoBERTa for Valence?Arousal Regression on Ecological Essays. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*.

Shrika SP Thota, Lakshmi Priya Swaminatha Rao, Shivaanee SK, Thirumurugan RA, Vishal Muralidharan, and Dhannya Santhakumari Madhavan. 2026. AFourP at 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)*.

Debra Trampe, Jordi Quoidbach, and Maxime Taquet. 2015. Emotions in everyday life. *PLOS ONE*, 10(12):1–15.

Enrica Troiano, Sebastian Padó, and Roman Klinger. 2019. Crowdsourcing and validating event-focused emotion corpora for German and English. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4005–4011, Florence, Italy. Association for Computational Linguistics.

Vasudha Varadarajan, Allison Lahnala, Adithya V Ganesan, Gourab Dey, Siddharth Mangalik, Ana-Maria Bucur, Nikita Soni, Rajath Rao, Kevin Lanning, Isabella Vallejo, and 1 others. 2024. Archetypes and entropy: Theory-driven extraction of evidence for suicide risk. In *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*, pages 278–291.

Vera Vine, Ryan L. Boyd, and James W. Pennebaker. 2020. Natural emotion vocabularies as windows on distress and well-being. *Nature Communications*, 11(4525):1–9. Number: 1.

Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. 2013. Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior research methods*, 45(4):1191–1207.

D Watson. 2000. *Mood and temperament*. Emotions and social behavior. Guilford Press, New York.

Kamyar Moradian Zehab, Mohammad Sadegh Poulaei, and Nasser Mozayani. 2026. Perspicere at SemEval-2026 Task 2: Transfer Learning for Valence and Arousal Prediction in Ecological Essays. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*.

Hongyi Zhang, Daniel Hu, and Allison Claire Lahnala. 2026. McMaster NLP at SemEval-2026 Task 2: A Lightweight Multi-Feature System for Predicting Emotional Valence and Arousal over Time. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*.

A Appendix

The appendix tables consist of further details on the dataset, the participating systems, and further elaborated results. Another minor point to note in the results analysis is that a majority of submissions (15 of 25) exhibited small but statistically significant negative correlations between absolute error and document temporal position ($r \in$ [-.05, -.10]; Benjamini-Hochberg corrected, $p < .05$), indicating that models achieved lower error on documents later in an individual's timeline despite being trained at the document level. We leave systematic investigation of this finding to future work.

User-ID	Text-ID	Text	Timestamp	Collection Phase	Is Words	Valence	Arousal
162	4138	I'm a bit on edge, my stomach feels a little tight, but I also feel somewhat content after talking with my mom. I am kind of on the verge of tears, actually, but I can't totally pinpoint why. I'm feeling conflicting emotions, and I'm excited to go to sleep tonight.	2024-11-11 22:03:30	7	FALSE	-1	1
162	4152	Stressed, Unsure, Excited, Tired, Hungry	2024-11-17 17:01:16	7	TRUE	0	1
16	869	Feeling pretty good and refreshed after my weekend and heading into work today. Hoping that the ORG draft being in town will be lucrative this weekend for me, but hopefully nothing too crazy! Lol. Feeling calm mellow ready to work!	2022-04-29 15:01:00	3	FALSE	2	0

Table 6: The structure of train dataset for both Subtask 1 and Subtask 2 including selected rows from the data as examples. Train data for Subtask 2 also included state change and disposition change labels for each user.

Statistics	Numbers per User			Number of Words per Instance		
	Texts	Essays	Feeling-Words	Text	Essay	Feeling-Words
mean	72.83	53.09	48.28	31.83	58.96	4.98
std	72.33	60.98	58.85	32.72	26.20	1.34
min	3.00	1.00	1.00	3.00	9.00	3.00
25%	25.00	13.00	13.00	5.00	47.00	5.00
50%	35.00	18.00	18.00	11.00	52.00	5.00
75%	152.00	117.00	42.00	52.00	60.00	5.00
max	215.00	168.00	177.00	230.00	230.00	38.00

Table 7: Full Dataset Statistics: Summary of text counts and word distributions for the full dataset, consisting of 5285 texts from 182 authors, of which 2628 are 'ecological essays' and 2657 are 'feeling words'

Statistics	Numbers per User			Number of Words per Instance		
	Texts	Essays	Feeling-Words	Overall Text	Essay	Feeling-Word
mean	58.66	40.27	42.03	30.30	57.45	5.08
std	61.70	52.23	54.89	31.68	25.67	1.51
min	2.00	1.00	1.00	3.00	9.00	3.00
25%	16.00	9.00	9.00	5.00	46.00	5.00
50%	31.00	14.00	16.00	7.00	51.00	5.00
75%	97.00	63.00	61.00	51.00	59.00	5.00
max	206.00	168.00	177.00	212.00	212.00	38.00

Table 8: Training Data Statistics: Summary of text counts and word distributions for the training data, consisting of 2764 texts from 137 authors, of which 1331 are 'ecological essays' and 1433 are 'feeling words'

Statistics	Numbers per User			Number of Words per Instance		
	Texts	Essays	Feeling-Words	Overall Text	Essay	Feeling-Word
count	1737.00	807.00	930.00	1737.00	807.00	930.00
mean	35.98	23.89	30.68	30.37	59.71	4.91
std	30.29	31.71	34.34	32.60	26.05	1.12
min	1.00	1.00	1.00	3.00	33.00	3.00
25%	17.00	9.00	9.00	5.00	47.00	5.00
50%	28.00	13.00	15.00	6.00	53.00	5.00
75%	36.00	18.00	23.00	52.00	61.00	5.00
max	107.00	105.00	107.00	230.00	230.00	16.00

Table 9: Subtask 1 Test Data Statistics: Summary of text counts and word distributions for the Subtask 1 held-out test data, consisting of 1737 texts from 91 authors, of which 807 are 'ecological essays' and 930 are 'feeling words'

Team	Core Approach	Val.	Aro.	Avg.
rand	*Baseline*	*0.000*	*0.000*	*0.000*
linear(BERT)	*Baseline*	*0.557*	*0.299*	*0.428*
UKP_PSYCONTROL	GPT-5 prompting, seen/unseen-aware strategies	0.667	**0.554**	**0.611**
YNU	RoBERTa + contrastive learning + prompt reformulation	0.677	0.528	0.603
CCLIN	LLM (Mistral/Qwen LoRA) + multi-BERT ensemble	0.647	0.527	0.587
LAMANHNGUYEN	DeBERTa-v3-base dual-variant ensemble + momentum post-proc.	0.687	0.458	0.573
AFOURP	RoBERTa-base + linear regression head	0.679	0.466	0.573
CSIRO-LT	RoBERTa-Twitter, temporal feature diffusion	0.656	0.488	0.572
CURIOSAI	RoBERTa-large, EmoBank pre-training + multi-task FT	0.683	0.451	0.567
BISON AI4PC	DistilBERT + MLP, user-based train/val splits	0.665	0.468	0.567
UIT	RoBERTa-Cardiff + Mixture-of-Experts + CCC loss	0.637	0.489	0.563
MCMASTER4Z03	Sentence embeds. + LIWC features + seed word similarity + user embeds + MLP	0.665	0.460	0.563
PERSPICERE	Jasper frozen embeddings (Matryoshka) + SVM	0.623	0.497	0.560
NLPGROUP8	RoBERTa ensemble (5 heads, diverse loss objectives)	**0.688**	0.416	0.552
CHERISH	RoBERTa/BERT + PLoRa personalisation	0.596	0.505	0.551
AJMAN UNIV.	DistilBERT + BiLSTM + gated user embeddings	0.656	0.439	0.548
VERBANEX AI	RoBERTa-base + NRC VAD lexicon	0.632	0.463	0.548
IMEZO	RoBERTa-base + NRC VAD lexicon augmentation	0.656	0.437	0.547
AI4PC-HOWARD	DeBERTa / ModernBERT fine-tuning + GRU	0.631	0.462	0.547
LEXMACHINA	DeBERTa-v3 + DANN adversarial head (arousal)	0.645	0.434	0.540
EMO-TICA	DistilBERT Trait–State model + learned user embeds.	0.645	0.409	0.527
AGI	RoBERTa-large + unidirectional GRU + inertia gate	0.600	0.452	0.526
ECOAFFECTTRACK	DeBERTa-v3-base + CCC loss	0.663	0.373	0.518
UALBERTA	BERT + BiLSTM, temporal sequence modeling	0.556	0.444	0.500
NLP-FSDM	ModernBERT + temporal smoothing + ensemble	0.546	0.453	0.500
VAP-GAMECTRL.	NRC VAD lexicon + LLM + time-aware fusion	0.615	0.322	0.469
ONE AND ONLY	GPT-5 zero-shot + lexicon, no fine-tuning	0.527	0.315	0.421

Table 10: Subtask 1 overall results: Pearson r_{comp} for valence (Val.), arousal (Aro.), and their average (Avg.). **Bold** = best per column. Gray rows = baselines). See Table 4 for seen/unseen breakdown.

Team	Core Approach	Val.	Aro.	Avg.
rand	*Baseline*	0.000	0.000	*0.000*
linear(BERT)	*Baseline*	0.290*	0.199*	*0.245*
linear(prev)	*Baseline*	0.615*	0.670*	*0.643*
linear(BERT;prev)	*Baseline*	0.430*	0.405*	*0.418*
UKP_PSYCONTROL	RoBERTa-base + user embed. + MLP regressor	0.675*	**0.683***	0.679
YNU	RoBERTa + time-aware LSTM ($\log \Delta t$ gating)	**0.692***	0.647*	0.670
UALBERTA	BERT + Temporal Fusion Transformer (TFT)	0.615*	0.674*	0.645
AJMAN UNIV.	DeBERTa-v3-base + "Megaphone" MLP + CCC loss	0.615*	0.670*	0.643
UIT	RoBERTa-Cardiff + MoE (sliding window k=8)	0.629*	0.633*	0.631
CSIRO-LT	RoBERTa + temporal feature diffusion (5 posts)	0.621*	0.477*	0.549
AI4PC-HOWARD	DeBERTa + GRU sequence encoder	0.597*	0.413*	0.505
EMO-TICA	Ridge (val.) + LightGBM (aro.) + temporal stats	0.424*	0.355	0.390
CURIOSAI	RoBERTa-large multi-task (no explicit history)	0.467*	0.275	0.371
BISON AI4PC	DistilBERT + history stats (mean, trend)	0.379*	0.085	0.232
NLPGROUP8	RoBERTa + transformer decoder (sliding window)	0.152	0.126	0.139
ECOAFFECTTRACK	DeBERTa-v3 LSTM (frozen enc.) + instance norm.	−0.243	−0.011	−0.127
AGI	RoBERTa-large + GRU + zero-inflated Δ model	−0.167	−0.147	−0.157
LAMANHNGUYEN	DeBERTa ensemble + momentum dampening	−0.273	−0.275	−0.274
CHERISH	RoBERTa/BERT + PLoRa personalisation + GRU	NaN	NaN	NaN

Table 11: Subtask 2a results: Pearson r for state change valence and arousal, sorted by average (Avg.) of the two dimensions. Gray rows = baselines. * $p < 0.01$. **Bold** = best per column.

Team	Core Approach	Val.	Aro.	Avg.
rand	*Baseline*	0.000	0.000	*0.000*
linear(BERT)	*Baseline*	−0.088*	0.070*	*-0.009*
linear(prev)	*Baseline*	0.434*	0.584*	*0.509*
linear(BERT;prev)	*Baseline*	−0.029*	0.019*	*-0.005*
UALBERTA	BERT + BiLSTM, group-mean difference strategy	**0.405***	**0.602***	**0.503**
NLPGROUP8	RoBERTa context encoder + transformer decoder	0.354	0.388*	0.371
EMO-TICA	Ridge regression on affect trajectory statistics	0.257	0.418*	0.338
AI4PC-HOWARD	DeBERTa + MLP on pooled embeddings	0.046	0.348	0.197
AJMAN UNIV.	DeBERTa-v3-large + Siamese difference pooling	−0.124	0.456*	0.166
AGI	RoBERTa-large + GRU + time-weighted EMA	0.086	−0.081	0.003
ECOAFFECTTRACK	Ridge on DeBERTa-v3 user-profile embeddings	−0.243	0.226	−0.009
CSIRO-LT	RoBERTa + temporal feature diffusion (15 posts)	−0.147	0.114	−0.017
CURIOSAI	RoBERTa-large multi-task	−0.161	0.011	−0.075
BISON AI4PC	DistilBERT + BiLSTM + affect statistics	−0.120	−0.103	−0.112
UIT	RoBERTa-Cardiff + MoE + dual-sequence input	−0.169	−0.060	−0.115
LAMANHNGUYEN	DeBERTa ensemble + temporal smoothing	−0.398*	−0.577*	−0.488

Table 12: Subtask 2b results: Pearson r for dispositional change in valence and arousal, sorted by average (Avg.) of the two dimensions. Gray rows = baselines. * $p < 0.01$. **Bold** = best per column.

Team	Subtask 1 (r_{comp})		Subtask 2a (r)		Subtask 2b (r)	
	Val.	Aro.	Val.	Aro.	Val.	Aro.
UKP_Psycontrol (Hryhoryeva et al., 2026)	0.667	**0.554**	0.675*	**0.683***	-	-
YNU (Lan et al., 2026)	0.677	0.528	**0.692***	0.647*	-	-
cclin (Lin, 2026)	0.647	0.527	-	-	-	-
AFourP (Thota et al., 2026)	0.679	0.466	-	-	-	-
lamanhnguyen (NGUYEN, 2026)	0.687	0.458	-0.273	-0.275	-0.398*	-0.577*
CSIRO-LT (Chen et al., 2026)	0.656	0.488	0.621*	0.477*	-0.147	0.114
CuriosAI (Beppu et al., 2026)	0.683	0.451	0.467*	0.275	-0.161	0.011
Bison AI4PC (Shah et al., 2026)	0.665	0.468	0.379*	0.085	-0.120	-0.103
UIT (Phuong et al., 2026)	0.637	0.489	0.629*	0.633*	-0.169	-0.060
mcmaster4z03 (Zhang et al., 2026)	0.665	0.460	-	-	-	-
Perspicere (Zehab et al., 2026)	0.623	0.497	-	-	-	-
NLPGroup8 (Arthur et al., 2026)	**0.688**	0.416	0.152	0.126	0.354^{+}	0.388*
Cherish (Parahita, 2026)	0.596	0.505	NaN	NaN	-	-
Ajman Univ. (Jumakhan et al., 2026)	0.656	0.439	0.615*	0.670*	-0.124	0.456*
VerbaNex AI (Moreno et al., 2026)	0.632	0.463	-	-	-	-
IMEZO/Khaleesi (Tee, 2026)	0.656	0.437	-	-	-	-
AI4PC-Howard (Shah et al., 2026)	0.631	0.462	0.597*	0.413*	0.046	0.348*
LexMachina (Ganguli et al., 2026)	0.645	0.434	-	-	-	-
Emo-tica (Noor and Fatima, 2026)	0.645	0.409	0.424*	0.355^{+}	0.257	0.418*
AGI (Rathva, 2026)	0.600	0.452	-0.167	-0.147	0.086	-0.081
EcoAffectTrack (Kumar and Joshi, 2026)	0.663	0.373	-0.243	-0.011	-0.243	0.226
UAlberta (Ho et al., 2026)	0.556	0.444	0.615*	0.674*	0.405*	**0.602***
NLP-FSDM (Benlahbib et al., 2026)	0.546	0.453	-	-	-	-
VAP-GameCtrl. (Le et al., 2026)	0.615	0.322	-	-	-	-
linear(BERT)	0.557	0.299	0.290*	0.199*	-0.088*	0.070*
One and Only (Dinh, 2026)	0.527	0.315	-	-	-	-
rand	0.000	0.000	0.000	0.000	0.000	0.000
linear(prev)	-	-	0.615*	0.670*	**0.434***	0.584*
linear(B;p)	-	-	0.430*	0.405*	-0.029*	0.019*
Momentum (Nadiger et al., 2026)	0.638	0.455	0.553*	0.589*	-	-
ES4MLL (Lolli et al., 2026)	0.650	0.433	-	-	-	-
Draken (Sivanaiah et al., 2026)	0.594	0.296	-	-	-	-
One and Only (Dinh, 2026)	-	-	-0.194	-0.423*	-0.185	0.016

Table 13: Main Metrics across all subtasks. Performance is reported using Pearson r_{comp} for Subtask 1 and Pearson r for Subtasks 2a and 2b. Heatmap colors represent relative performance within each metric; * $p < 0.01$ $^{+}$ $p < 0.05$; **bold** values denote the best performance per column. Blue-shaded rows represent standard baselines. Gray-shaded rows represent post-deadline submissions and are excluded from the official rankings

Team	Valence (V)						Arousal (A)					
	r_{comp}	r_{bet}	r_{wit}	mae_{comp}	mae_{bet}	mae_{wit}	r_{comp}	r_{bet}	r_{wit}	mae_{comp}	mae_{bet}	mae_{wit}
UKP_Psycontrol	0.667	0.761*	0.546*	0.595	0.402	0.738	**0.554**	0.701*	0.363*	0.345	0.210	0.467
YNU	0.677	0.741*	0.601*	0.606	0.404	0.751	0.528	0.606*	0.439*	0.367	0.231	0.488
cclin	0.647	0.695*	0.593*	0.653	0.453	0.790	0.527	0.611*	0.430*	0.365	0.226	0.489
AFourP	0.679	0.749*	0.595*	0.633	0.414	0.782	0.466	0.526*	0.402*	0.395	0.253	0.520
lamanhnguyen	0.687	0.779*	0.567*	0.639	0.388	0.801	0.458	0.523*	0.388*	0.395	0.243	0.528
CSIRO-LT	0.656	0.721*	0.580*	0.654	0.798	0.440	0.488	0.547*	0.425*	0.401	0.253	0.531
CuriosAI	0.683	0.788*	0.541*	0.622	0.383	0.783	0.451	0.505*	0.393*	0.404	0.257	0.533
Bison AI4PC	0.665	0.744*	0.569*	0.633	0.419	0.780	0.468	0.540*	0.389*	0.395	0.257	0.518
UIT	0.637	0.713*	0.546*	0.694	0.466	0.835	0.489	0.600*	0.359*	0.408	0.258	0.539
mcmaster4z03	0.665	0.763*	0.536*	0.625	0.392	0.782	0.460	0.555*	0.353*	0.399	0.253	0.528
Perspicere	0.623	0.682*	0.555*	0.656	0.445	0.798	0.497	0.590*	0.392*	0.403	0.254	0.533
NLPGroup8	**0.688**	0.777*	0.572*	NaN	1.890	1.925	0.416	0.455*	0.376*	0.395	0.250	0.523
Cherish	0.596	0.648*	0.538*	NaN	0.614	1.021	0.505	0.616*	0.375*	0.361	0.234	0.477
Ajman University	0.656	0.726*	0.573*	0.637	0.424	0.783	0.439	0.528	0.000*	0.443	0.338	0.536
VerbaNex AI	0.632	0.689*	0.566*	0.656	0.466	0.789	0.463	0.524	0.000*	0.345	0.246	0.436
IMEZO /Khaleesiyali	0.656	0.747*	0.542*	0.653	0.423	0.804	0.437	0.476*	0.397*	0.411	0.273	0.533
AI4PC - Howard Uni	0.631	0.701*	0.548*	0.670	0.443	0.817	0.462	0.511*	0.410*	0.416	0.296	0.523
LexMachina	0.645	0.712*	0.567*	0.652	0.437	0.796	0.434	0.461*	0.406*	0.421	0.303	0.527
Emo-tica	0.645	0.705*	0.577*	0.685	0.472	0.822	0.409	0.430*	0.388*	0.407	0.275	0.524
AGI	0.600	0.599*	0.600*	0.738	0.573	0.846	0.452	0.488*	0.416*	0.413	0.277	0.533
EcoAffectTrack	0.663	0.723*	0.593*	0.714	0.479	0.854	0.373	0.348*	0.398*	0.473	0.317	0.605
UAlberta	0.556	0.671*	0.415*	0.734	0.470	0.878	0.444	0.596*	0.262*	0.427	0.264	0.566
NLP-FSDM	0.546	0.660*	0.408*	0.756	0.504	0.889	0.453	0.588*	0.293*	0.405	0.252	0.537
VAP-GameController	0.615	0.691*	0.525*	0.629	0.464	0.751	0.322	0.385*	0.256*	0.396	0.284	0.497
linear(BERT)	0.557	0.659*	0.435*	0.743	0.472	0.886	0.299	0.343*	0.253*	0.459	0.311	0.585
One and Only	0.527	0.617*	0.423*	0.738	0.490	0.876	0.315	0.335*	0.001*	0.873	0.844	0.897
rand	0.000	0.028*	0.000	0.000	0.627	1.041	0.000	0.096*	0.000	0.488	0.326	0.622
Momentum	0.638	0.715*	0.546*	0.645	0.416	0.798	0.455	0.507*	0.000*	0.400	0.256	0.526
ES4MLL	0.650	0.727*	0.555*	0.630	0.418	0.776	0.433	0.480*	0.384*	0.428	0.292	0.548
Draken	0.594	0.689*	0.479*	0.721	0.446	0.871	0.296	0.377*	0.211*	0.457	0.282	0.602

Table 14: Detailed performance metrics for Subtask 1, disaggregated by valence (V) and arousal (A). For each dimension, we report the composite correlation (r_{comp}), between-user correlation (r_{bet}), within-user correlation (r_{wit}), and their corresponding Mean Absolute Error (MAE) values. * $p < 0.01$ $^+$ $p < 0.05$; Blue-shaded rows represent standard baselines. Gray-shaded rows represent post-deadline submissions and are excluded from the official rankings.

Team	Valence (V)		Arousal (A)	
	$r \uparrow$	$mae \downarrow$	$r \uparrow$	$mae \downarrow$
UKP_Psycontrol	0.675*	1.118	**0.683***	0.641
YNU	**0.692***	1.074	0.647*	0.641
UAlberta	0.615*	1.208	0.674*	0.635
linear(prev)	0.615*	1.168	0.670*	0.638
Ajman Univ.	0.615*	1.582	0.670*	0.636
Uni IT	0.629*	1.141	0.633*	0.689
CSIRO-LT	0.621*	1.190	0.477*	0.740
AI4PC-Howard	0.597*	1.180	0.413*	0.720
linear(B;p)	0.430*	1.251	0.405*	0.708
Emo-tica	0.424*	1.297	0.355+	0.842
CuriosAI	0.467*	1.220	0.275	0.890
linear(BERT)	0.290*	1.294	0.199*	0.744
Bison AI4PC	0.379*	1.202	0.085	0.767
NLPGroup8	0.152	1.420	0.126	0.838
rand	0.000	1.261	0.000	0.696
EcoAffectTrack	-0.243	1.342	-0.011	0.806
AGI	-0.167	1.515	-0.147	0.844
lamanhnguyen	-0.273	1.322	-0.275	0.736
Cherish	NaN	1.565	NaN	2.130
Momentum	0.553*	1.139	0.589*	0.698
One and Only	-0.194	1.398	-0.423*	0.818

Table 15: Detailed performance metrics for Subtask 2a, focusing on the prediction of emotional state changes. We report Pearson correlation (r) and Mean Absolute Error (mae) for both valence (V) and arousal (A) dimensions. * $p < 0.01$ $^+$ $p < 0.05$; Blue-shaded rows represent standard baselines. Gray-shaded rows represent post-deadline submissions and are excluded from the official rankings

tabcolsep 3pt

Team	Valence (V)		Arousal (A)	
	$r \uparrow$	$mae \downarrow$	$r \uparrow$	$mae \downarrow$
linear(prev)	**0.434***	0.406	0.584*	0.286
UAlberta	0.405*	0.635	**0.602***	0.261
NLPGroup8	0.354+	0.425	0.388*	0.393
Emo-tica	0.257	0.461	0.418*	0.298
AI4PC-Howard	0.046	0.419	0.348*	0.292
Ajman Univ.	-0.124	0.623	0.456*	0.291
AGI	0.086	0.633	-0.081	0.363
rand	0.000	0.417	0.000	0.296
linear(B;p)	-0.029*	0.436	0.019*	0.305
EcoAffectTrack	-0.243	0.419	0.226	0.292
linear(BERT)	-0.088*	0.438	0.070*	0.303
CSIRO-LT	-0.147	0.593	0.114	0.373
CuriosAI	-0.161	0.795	0.011	0.429
Bison AI4PC	-0.120	0.424	-0.103	0.296
Uni IT	-0.169	0.723	-0.060	0.568
lamanhnguyen	-0.398*	0.431	-0.577*	0.309
One and Only	-0.185	0.899	0.016	0.483

Table 16: Detailed performance metrics for Subtask 2b, focusing on the prediction of emotional state changes. We report Pearson correlation (r) and Mean Absolute Error (mae) for both valence (V) and arousal (A) dimensions. * $p < 0.01$ $^+$ $p < 0.05$; Blue-shaded rows represent standard baselines. Gray-shaded rows represent post-deadline submissions and are excluded from the official rankings

Team	Valence (V)						Arousal (A)					
	r_{comp}	r_{bet}	r_{wit}	mae_{comp}	mae_{bet}	mae_{wit}	r_{comp}	r_{bet}	r_{wit}	mae_{comp}	mae_{bet}	mae_{wit}
UKP_Psycontrol	0.688	0.787*	0.553*	0.370	0.634	0.514	**0.568**	0.753*	0.301*	0.300	0.196	0.398
YNU	0.684	0.767*	0.578*	0.430	0.736	0.605	0.466	0.546*	0.377*	0.377	0.257	0.486
cclin	0.664	0.738*	0.575*	0.473	0.779	0.652	0.457	0.552*	0.350*	0.376	0.246	0.492
AFourP	0.696	0.781*	0.587*	0.443	0.762	0.628	0.359	0.400*	0.316*	0.419	0.292	0.532
lamanhnguyen	0.708	0.804*	0.575*	0.410	0.762	0.616	0.392	0.424*	0.359*	0.409	0.282	0.522
CSIRO-LT	0.650	0.735*	0.545*	0.686	0.810	0.502	0.395	0.428*	0.360*	0.429	0.301	0.542
CuriosAI	0.715	0.832*	0.535*	0.371	0.717	0.569	0.403	0.432*	0.373*	0.413	0.291	0.521
Bison AI4PC	0.664	0.771*	0.520*	0.453	0.772	0.640	0.404	0.502*	0.295*	0.405	0.280	0.517
UIT	0.639	0.740*	0.510*	0.563	0.819	0.714	0.405	0.508*	0.291*	0.421	0.308	0.523
mcmaster4z03	0.666	0.791*	0.488*	0.638	0.423	0.785	0.402	0.492*	0.304*	0.398	0.268	0.514
Perspicere	0.614	0.709*	0.496*	0.476	0.801	0.670	0.414	0.478*	0.346*	0.415	0.293	0.523
NLPGroup8	0.698	0.812*	0.533*	NaN	1.890	1.827	0.297	0.306*	0.287*	0.436	0.314	0.543
Cherish	0.604	0.695*	0.495*	0.574	0.986	0.916	0.465	0.596*	0.311*	0.378	0.258	0.487
Ajman University	0.669	0.736*	0.590*	0.651	0.473	0.778	0.361	0.464	0.001*	0.428	0.325	0.521
VerbaNex AI	0.606	0.716*	0.467*	0.771	0.621	0.866	0.356	0.436*	0.029*	0.395	0.326	0.460
IMEZO /Khaleesiyali	0.648	0.758*	0.501*	0.454	0.789	0.653	0.386	0.408*	0.365*	0.415	0.290	0.525
AI4PC - Howard Uni	0.633	0.725*	0.519*	0.505	0.829	0.702	0.415	0.444*	0.385*	0.407	0.295	0.508
LexMachina	0.636	0.718*	0.537*	0.491	0.805	0.678	0.343	0.323*	0.363*	0.448	0.345	0.541
Emo-tica	0.658	0.732*	0.569*	0.543	0.823	0.710	0.332	0.383*	0.278*	0.415	0.288	0.529
AGI	0.639	0.687*	0.587*	0.499	0.790	0.670	0.400	0.438*	0.360*	0.436	0.319	0.540
EcoAffectTrack	0.688	0.768*	0.586*	0.467	0.803	0.668	0.327	0.314*	0.340*	0.495	0.351	0.616
UAlberta	0.578	0.770*	0.290*	0.520	0.845	0.720	0.505	0.692*	0.255*	0.413	0.295	0.518
NLP-FSDM	0.541	0.673*	0.375*	0.522	0.861	0.735	0.444	0.571*	0.297*	0.396	0.262	0.514
VAP-GameController	0.615	0.710*	0.498*	0.503	0.730	0.630	0.264	0.323*	0.203*	0.406	0.306	0.496
linear(BERT)	**0.748**	0.588*	0.851*	0.598	0.766	0.354	0.422	0.592	0.215	0.408	0.268	0.531
One and Only	0.508	0.595*	0.409*	0.782	0.578	0.894	0.268	0.265	0.270*	0.820	0.795	0.843
ES4MLL	0.597	0.707*	0.459*	0.752	0.571	0.863	0.304	0.311*	0.297*	0.446	0.340	0.541
Momentun	0.595	0.704*	0.459*	0.756	0.582	0.863	0.268	0.331*	0.106*	0.462	0.343	0.566
Draken	0.605	0.720*	0.457*	0.716	0.470	0.859	0.228	0.233*	0.223*	0.465	0.321	0.588

Table 17: Detailed performance metrics for Subtask 1 (Seen Users), focusing on the prediction of emotional state changes. We report Pearson correlation (r) and Mean Absolute Error (mae) for both valence (V) and arousal (A) dimensions. * $p < 0.01$ $^+$ $p < 0.05$; Blue-shaded rows represent standard baselines. Gray-shaded rows represent post-deadline submissions and are excluded from the official rankings

Team	Valence (V)						Arousal (A)					
	r_{comp}	r_{bet}	r_{wit}	mae_{comp}	mae_{bet}	mae_{wit}	r_{comp}	r_{bet}	r_{wit}	mae_{comp}	mae_{bet}	mae_{wit}
UKP_Psycontrol	0.662	0.757*	0.540*	0.434	0.843	0.691	0.500	0.574*	0.417*	0.393	0.225	0.539
YNU	0.677	0.726*	0.622*	0.377	0.767	0.607	0.612	0.708*	0.494*	0.356	0.205	0.490
cclin	0.629	0.648*	0.609*	0.431	0.801	0.654	**0.614**	0.707*	0.499*	0.355	0.206	0.487
AFourP	0.662	0.714*	0.601*	0.385	0.802	0.638	0.605	0.708*	0.476*	0.370	0.214	0.508
lamanhnguyen	0.673	0.761*	0.560*	0.366	0.842	0.667	0.552	0.667*	0.412*	0.380	0.203	0.533
CSIRO-LT	0.673	0.727*	0.610*	0.622	0.786	0.376	0.598	0.694*	0.482*	0.372	0.203	0.519
CuriosAI	0.646	0.728*	0.546*	0.394	0.850	0.684	0.510	0.598*	0.409*	0.396	0.223	0.545
Bison AI4PC	0.673	0.727*	0.612*	0.384	0.788	0.626	0.553	0.626*	0.470*	0.385	0.233	0.519
UIT	0.620	0.666*	0.570*	0.436	0.833	0.682	0.555	0.690*	0.381*	0.400	0.232	0.545
mcmaster4z03	0.660	0.729*	0.579*	0.611	0.360	0.779	0.550	0.674*	0.396*	0.401	0.237	0.543
Perspicere	0.641	0.673*	0.608*	0.413	0.795	0.642	0.591	0.715*	0.432*	0.391	0.215	0.543
NLPGroup8	**0.679**	0.741*	0.607*	1.986	2.026	NaN	0.601	0.717*	0.452*	0.354	0.185	0.503
Cherish	0.594	0.611*	0.576*	0.654	1.057	NaN	0.534	0.624*	0.430*	0.344	0.210	0.466
Ajman University	0.650	0.726*	0.558*	0.623	0.374	0.789	0.530	0.625	0.000*	0.458	0.352	0.553
VerbaNex AI	0.664	0.704*	0.619*	0.568	0.384	0.709	0.625	0.755	0.000*	0.317	0.199	0.425
IMEZO /Khaleesiyali	0.675	0.754*	0.578*	0.392	0.820	0.656	0.519	0.603*	0.425*	0.408	0.256	0.541
AI4PC - Howard Uni	0.631	0.680*	0.575*	0.380	0.804	0.638	0.506	0.573*	0.432*	0.426	0.298	0.539
LexMachina	0.669	0.734*	0.593*	0.382	0.788	0.625	0.574	0.681*	0.443*	0.394	0.260	0.513
Emo-tica	0.640	0.691*	0.583*	0.399	0.822	0.660	0.512	0.540*	0.483*	0.398	0.261	0.519
AGI	0.655	0.694*	0.613*	0.649	0.902	0.811	0.545	0.617*	0.464*	0.390	0.235	0.525
EcoAffectTrack	0.642	0.681*	0.599*	0.491	0.906	0.770	0.426	0.404*	0.448*	0.451	0.282	0.593
UAlberta	0.467	0.493*	0.439*	0.476	0.876	0.734	0.457	0.629*	0.243*	0.431	0.227	0.599
NLP-FSDM	0.556	0.657*	0.436*	0.486	0.918	0.783	0.447	0.582*	0.289*	0.414	0.242	0.561
VAP-GameController	0.618	0.678*	0.550*	0.425	0.773	0.629	0.400	0.490*	0.303*	0.386	0.261	0.498
linear(BERT)	0.562	0.494*	0.624*	0.772	0.924	0.411	0.238	0.258*	0.218	0.488	0.309	0.633
One and Only	0.554	0.654*	0.434*	0.692	0.400	0.856	0.356	0.395	0.007*	0.928	0.894	0.952
ES4MLL	0.678	0.747*	0.595*	0.574	0.361	0.730	0.483	0.541*	0.420*	0.428	0.289	0.550
Momentun	0.645	0.692*	0.592*	0.595	0.350	0.764	0.482	0.514*	0.004*	0.383	0.222	0.524
Draken	0.574	0.641*	0.499*	0.726	0.422	0.884	0.365	0.509*	0.201*	0.450	0.243	0.617

Table 18: Detailed performance metrics for Subtask 1 (Unseen Users), focusing on the prediction of emotional state changes. We report Pearson correlation (r) and Mean Absolute Error (mae) for both valence (V) and arousal (A) dimensions. * $p < 0.01$ $^+$ $p < 0.05$; Blue-shaded rows represent standard baselines. Gray-shaded rows represent post-deadline submissions and are excluded from the official rankings

Team	Valence (V)						Arousal (A)					
	r_{comp}	r_{bet}	r_{wit}	mae_{comp}	mae_{bet}	mae_{wit}	r_{comp}	r_{bet}	r_{wit}	mae_{comp}	mae_{bet}	mae_{wit}
UKP_Psycontrol	0.658	0.758*	0.527*	0.437	0.739	0.610	0.575	0.680*	0.446*	0.354	0.255	0.447
YNU	0.683	0.765*	0.581*	0.400	0.723	0.584	0.602	0.660*	0.537*	0.351	0.245	0.448
cclin	0.662	0.718*	0.597*	0.446	0.751	0.622	**0.620**	0.689*	0.540*	0.348	0.234	0.453
AFourP	0.670	0.742*	0.581*	0.418	0.754	0.613	0.566	0.609*	0.519*	0.377	0.262	0.481
lamanhnguyen	0.677	0.765*	0.565*	0.434	0.801	0.655	0.578	0.636*	0.515*	0.366	0.257	0.466
CSIRO-LT	0.673	0.727*	0.610*	0.622	0.786	0.376	0.598	0.694*	0.482*	0.372	0.203	0.519
CuriosAI	0.669	0.775*	0.525*	0.400	0.756	0.608	0.524	0.569*	0.477*	0.402	0.298	0.497
Bison AI4PC	0.665	0.740*	0.575*	0.433	0.753	0.618	0.568	0.629*	0.499*	0.377	0.260	0.483
UIT	**0.693**	0.783*	0.576*	0.513	0.732	0.635	0.526	0.574*	0.476*	0.414	0.322	0.498
mcmaster4z03	0.658	0.758*	0.529*	0.610	0.421	0.748	0.542	0.621*	0.453*	0.383	0.255	0.498
Perspicere	0.649	0.740*	0.535*	0.436	0.754	0.620	0.569	0.623*	0.510*	0.396	0.267	0.511
NLPGroup8	0.690	0.790*	0.555*	1.953	1.980	NaN	0.574	0.629*	0.514*	0.361	0.256	0.457
Cherish	0.586	0.662*	0.498*	0.647	1.020	NaN	0.576	0.653*	0.487*	0.329	0.244	0.409
Ajman University	0.648	0.731*	0.545*	0.616	0.427	0.754	0.528	0.597*	0.451*	0.464	0.386	0.535
VerbaNex AI	0.684	0.788*	0.544*	0.615	0.490	0.715	0.515	0.515*	0.515*	0.357	0.311	0.402
IMEZO /Khaleesiyali	0.662	0.752*	0.547*	0.442	0.774	0.636	0.573	0.617*	0.525*	0.384	0.288	0.473
AI4PC - Howard Uni	0.650	0.751*	0.518*	0.438	0.769	0.632	0.540	0.572*	0.506*	0.419	0.317	0.512
LexMachina	0.655	0.730*	0.563*	0.430	0.732	0.602	0.572	0.631*	0.507*	0.402	0.321	0.478
Emo-tica	0.669	0.738*	0.586*	0.419	0.739	0.603	0.582	0.638*	0.520*	0.389	0.292	0.479
AGI	0.600	0.609*	0.592*	0.587	0.821	0.725	0.554	0.575*	0.531*	0.399	0.279	0.507
EcoAffectTrack	0.667	0.737*	0.582*	0.505	0.833	0.705	0.539	0.577*	0.498*	0.483	0.381	0.573
UAlberta	0.555	0.704*	0.359*	0.541	0.831	0.716	0.400	0.397*	0.403*	0.478	0.371	0.573
NLP-FSDM	0.572	0.728*	0.359*	0.489	0.853	0.716	0.516	0.653*	0.345*	0.394	0.252	0.520
VAP-GameController	0.646	0.720*	0.558*	0.477	0.683	0.590	0.318	0.381*	0.253*	0.418	0.314	0.512
linear(BERT)	0.470	0.357*	0.570*	0.823	0.918	0.641	0.158	0.124	0.191	0.487	0.384	0.578
One and Only	0.633	0.737*	0.500*	0.634	0.438	0.773	0.426	0.444*	0.408*	NaN	0.980	1.049
ES4MLL	0.726	0.809*	0.613*	0.575	0.447	0.680	0.516	0.570*	0.457*	0.436	0.336	0.525
Momentun	0.685	0.770*	0.577*	0.607	0.447	0.730	0.511	0.517*	0.504*	0.410	0.314	0.498
Draken	0.639	0.756*	0.482*	0.661	0.411	0.818	0.449	0.540*	0.347*	0.432	0.288	0.557

Table 19: Detailed performance metrics for Subtask 1 (Feeling Words Only), focusing on the prediction of emotional state changes. We report Pearson correlation (r) and Mean Absolute Error (mae) for both valence (V) and arousal (A) dimensions. * $p < 0.01$ $^+$ $p < 0.05$; Blue-shaded rows represent standard baselines. Gray-shaded rows represent post-deadline submissions and are excluded from the official rankings

Team	Valence (V)						Arousal (A)					
	r_{comp}	r_{bet}	r_{wit}	mae_{comp}	mae_{bet}	mae_{wit}	r_{comp}	r_{bet}	r_{wit}	mae_{comp}	mae_{bet}	mae_{wit}
UKP_Psycontrol	**0.685**	0.781*	0.559*	0.421	0.706	0.581	**0.500**	0.651*	0.311*	0.353	0.225	0.469
YNU	0.673	0.722*	0.618*	0.439	0.747	0.616	0.419	0.450*	0.388*	0.408	0.281	0.521
cclin	0.632	0.698*	0.556*	0.474	0.776	0.650	0.422	0.472*	0.369*	0.403	0.274	0.519
AFourP	0.679	0.753*	0.588*	0.451	0.776	0.642	0.357	0.374*	0.340*	0.443	0.311	0.559
lamanhnguyen	0.659	0.732*	0.572*	0.440	0.776	0.638	0.358	0.381*	0.335*	0.457	0.307	0.585
CSIRO-LT	0.623	0.653*	0.592*	0.704	0.819	0.534	0.388	0.413*	0.362*	0.440	0.302	0.560
CuriosAI	0.653	0.724*	0.567*	0.472	0.793	0.662	0.403	0.468*	0.334*	0.423	0.280	0.547
Bison AI4PC	0.647	0.705*	0.582*	0.478	0.785	0.658	0.333	0.339*	0.326*	0.444	0.310	0.561
UIT	0.586	0.643*	0.522*	0.585	0.868	0.760	0.397	0.565*	0.199*	0.465	0.349	0.567
mcmaster4z03	0.653	0.737*	0.549*	0.637	0.450	0.770	0.365	0.442*	0.283*	0.432	0.294	0.553
Perspicere	0.619	0.654*	0.580*	0.486	0.791	0.666	0.370	0.434*	0.303*	0.433	0.284	0.562
NLPGroup8	0.666	0.736*	0.583*	1.868	1.907	NaN	0.335	0.343*	0.326*	0.450	0.313	0.569
Cherish	0.618	0.664*	0.566*	0.630	1.015	NaN	0.380	0.451*	0.304*	0.423	0.306	0.527
Ajman University	0.618	0.690*	0.534*	0.660	0.476	0.789	0.326	0.350*	0.301*	0.450	0.335	0.551
VerbaNex AI	0.599	0.631*	0.566*	0.669	0.542	0.767	0.437	0.523*	0.342*	0.385	0.326	0.442
IMEZO /Khaleesiyali	0.645	0.731*	0.539*	0.466	0.809	0.672	0.395	0.442*	0.346*	0.445	0.301	0.569
AI4PC - Howard Uni	0.631	0.677*	0.580*	0.506	0.807	0.685	0.343	0.330*	0.356*	0.443	0.320	0.551
LexMachina	0.627	0.665*	0.586*	0.498	0.823	0.694	0.307	0.315*	0.298*	0.468	0.338	0.580
Emo-tica	0.602	0.652*	0.547*	0.563	0.858	0.745	0.313	0.296*	0.330*	0.437	0.298	0.558
AGI	0.599	0.606*	0.592*	0.619	0.862	0.767	0.370	0.364*	0.375*	0.438	0.309	0.552
EcoAffectTrack	0.644	0.701*	0.580*	0.516	0.864	0.735	0.332	0.274*	0.388*	0.499	0.356	0.619
UAlberta	0.517	0.522*	0.512	0.592	0.859	0.755	0.339	0.517*	0.133	0.473	0.356	0.576
NLP-FSDM	0.513	0.573*	0.447*	0.573	0.902	0.789	0.359	0.457*	0.253	0.438	0.290	0.565
VAP-GameController	0.581	0.647*	0.506*	0.488	0.777	0.656	0.278	0.279*	0.277*	0.403	0.298	0.499
linear(BERT)	0.546	0.451*	0.629*	0.710	0.820	0.548	0.395	0.395*	0.395*	0.468	0.358	0.566
One and Only	0.421	0.423*	0.420*	0.837	0.624	0.934	0.229	0.187	0.269*	0.765	0.749	0.780
Momentun	0.637	0.671*	0.600*	0.671	0.525	0.779	0.488	0.559*	0.410*	0.419	0.324	0.505
ES4MLL	0.611	0.659*	0.558*	0.654	0.504	0.766	0.360	0.314*	0.405*	0.490	0.398	0.573
Draken	0.503	0.560*	0.440*	0.799	0.558	0.916	0.193	0.244*	0.142*	0.518	0.361	0.646

Table 20: Detailed performance metrics for Subtask 1 (Essays Only), focusing on the prediction of emotional state changes. We report Pearson correlation (r) and Mean Absolute Error (mae) for both valence (V) and arousal (A) dimensions. * $p < 0.01$ $^+$ $p < 0.05$; Blue-shaded rows represent standard baselines. Gray-shaded rows represent post-deadline submissions and are excluded from the official rankings