McMaster NLP at SemEval-2026 Task 2: A Lightweight Multi-Feature System for Predicting Emotional Valence and Arousal over Time

Hongyi Zhang and Daniel Hu and Allison Lahnala
McMaster University, Faculty of Engineering
Department of Computing and Software
{zhanh279,hus44,lahnalaa}@mcmaster.ca

Abstract

We present a lightweight, feature-based regression system for predicting **valence** (pleasantness) and **arousal** (activation) from longitudinal language data. The language data ranges from longer free-form ecological essays to short affect-word, organized by user and time, reflecting natural variation in affective expression and experience. Our approach combines three complementary signals: (i) sentence-level semantic embeddings, (ii) psycholinguistic category features capturing affect- and function-related word usage, (iii) similarity measures between the language data with archetypal sentences, and (iv) trainable user-embeddings to account for between-user differences. The resulting feature vector is passed to a multi-layer perceptron trained to jointly predict valence and arousal. Our design provides a strong and interpretable baseline by making it possible to isolate the contribution of semantic, psycholinguistic, similarity, and user-specific signals. We further analyze our model's predictions to identify which feature groups are most informative and where errors are concentrated across users and input types.

1 Introduction

Affect, along the dimensions of valence and arousal, is closely tied to day-to-day well-being, stress, and mental-health (Whitehead and Bergeman, 2013). Language is a valuable medium for studying these states because it externalizes internal thoughts, emotions, and social experience in a form that can be analyzed at scale (Zhang et al., 2022). Prior work in psychology, computational linguistics, and NLP has shown that everyday language patterns can reveal psychologically meaningful information; for instance, textual signals can support the detection and monitoring of mental-health-relevant phenomena such as depression and distress (Eichstaedt et al., 2018). This motivates language-based affect modeling as a complement to traditional self-report or clinician-administered assessment, especially in settings where frequent, low-burden measurement is desirable (Tausczik and Pennebaker, 2010).

Within this broader context, SemEval-2026 Task 2 (Soni et al., 2026) frames affect as a dynamic, within-person process rather than a single static label. In **Subtask 1**, the goal is to predict continuous **valence** and **arousal** scores from repeated text entries over time. This longitudinal setup, organized by user and time with both longer free-text reflections (ecological essays) and shorter "feeling word" entries, offers a more naturalistic view of emotional expression than isolated prompts or single-post snapshots (Doherty et al., 2020). Reliable models of such trajectories through language could support personalized, low-burden well-being tracking where repeated assessment fatigue is a concern.

The task is grounded in the **circumplex model of affect**, which represents affective states in a 2D space with orthogonal axes for valence and arousal (Russell, 1980). Unlike one-dimensional sentiment polarity ("positive vs. negative"), this 2D representation distinguishes, for example, *high-arousal negative* states (e.g., anger/anxiety) from *low-arousal negative* states (e.g., sadness/fatigue), important for affective computing and mental-health–adjacent analysis where activation level can be as informative as polarity (Posner et al., 2005).

Prior work suggests that strong models often combine **contextual semantic representations** with **affect-informed features** (e.g., lexicon/psycholinguistic resources) (Verma and Tiwary, 2017), as we discuss in §2. As such, we encode each entry using (1) **sentence embeddings** to capture semantic content, (2) **psycholinguistic category counts** using Linguistic Inquiry and Word Count (LIWC) to capture affect- and function-related word usage, and (3) **lexicon-based similarity features** that compare an entry to words

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 686–698
July 3-4, 2026 ©2026 Association for Computational Linguistics

with known valence and arousal norms, thereby measuring how strongly it aligns with high- or low-valence/arousal prototypes. To account for between-person variation, as users can differ in their individual baseline affect, we integrate a learned **user embedding** that captures user-specific tendencies in affect expression. We concatenate these features and train a multi-layer perceptron regressor to jointly predict valence and arousal.

2 Background and Related work

Language provides a rich behavioral trace of how people think, feel, and cope, making natural language analysis a central tool for studying psychological processes at scale. Boyd and Schwartz (2021) situate modern NLP-for-psychology within a longer tradition of verbal behavior research, emphasizing that language features can capture both stable individual differences (e.g., trait-like patterns) and time-varying signals (e.g., shifts in affect). In mental health contexts, this motivates **language-based mental health assessments (LBAs)**: systems that infer clinically relevant constructs (e.g., distress, depression risk, well-being) from everyday language including social media posts, narratives, and self-reflective writing (Pennebaker, 2018).

A key challenge for LBAs is ensuring **robustness across contexts**. Mangalik et al. (2024) explicitly target this by evaluating whether language-based mental health signals persist across time and geography, demonstrating how temporal and geographic shifts can break naive assumptions of stationarity. This concern aligns with broader work in robust NLP, which identifies distribution shifts, spurious correlations, and domain change can degrade model behavior (Omar et al., 2022). In longitudinal affect modeling, these robustness issues are amplified: texts from the same individual are not i.i.d., and affect is expected to fluctuate with life events and context. This framing is directly relevant to SemEval 2026 Task 2, which is to model valence and arousal over time from ecologically collected entries.

Historically, LBAs often relied on feature engineering and classical representations. For example, Cohen et al. (2008) adapted latent semantic analysis to better capture clinically meaningful concepts in psychiatric narrative, illustrating both the promise and limitations of early distributional methods. Since then, the field has increasingly incorporated pre-trained transformer models or large language models (LLMs). A systematic review by Guo et al. (2024) summarizes the expanding landscape of LLM-based mental health applications, highlighting both opportunities (e.g., strong general language understanding, flexible adaptation) and risks (e.g., safety, evaluation gaps, overconfidence). Complementing this, Vu et al. (2024) propose adapting transformer-based LLMs to reflect psychological attributes (e.g., traits, personality, mental health), pointing toward personalization and trait-aware modeling as a route to better alignment with human-relevant constructs. This idea resonates with longitudinal settings where stable between-person differences exist alongside within-person change.

3 Task and Dataset

The objective of this work is to model variations in an individual's affect over time (Subtask 1: Longitudinal Affect Assessment) over a dataset of ecological essays and feeling words written by U.S. service-industry workers. Aside from the writings, participants provided self-reports of their affect measured as valence and arousal ratings shown in Table 1. The collection took place between 2021-2024 and features longitudinal writings and affect reports per participant. This enables the task of predicting an individual's affect provided language and self-reports of prior entries. The task is structured as a multi-output regression problem. For each text item, the model predicts two continuous values for valence and arousal. Predictions are evaluated independently for each dimension.

The dataset exhibits two main properties that shape modeling choices. First, it is **longitudinal and user-indexed**: each `user_id` contributes multiple entries over time, which introduces strong within-user dependence and makes naive i.i.d. assumptions unrealistic. As a result, systems benefit from capturing **stable individual baselines** (some authors systematically report higher/lower valence or arousal) while still remaining sensitive to **short-term affective variation** across successive entries.

Second, the input is **heterogeneous**. Instances include both free-form ecological essays and short "feeling word" lists (flagged by `is_words`). These modalities differ in length, discourse structure, and lexical explicitness: essays may express affect indirectly via events, appraisal, and narrative framing, whereas word lists often contain direct affect de-

scriptors but limited context. This heterogeneity increases variance in the mapping from text to affect and can reward feature sets that are robust across both long and short inputs (e.g., semantic embeddings paired with lexicon- or category-based cues).

Data cleaning. We remove instances with missing values in any required fields, including `user_id`, `text_id`, `text`, `valence`, or `arousal`. After filtering, entries are sorted by `user_id`, timestamp, and `text_id` to ensure a consistent temporal ordering for each user's longitudinal timeline. No additional text normalization is applied at this stage, as affective markers may be expressed through higher-level lexical variation.

4 System Description

Our system is a lightweight, feature-based regression model designed to predict valence and arousal for each text independently, while accounting for stable between-user differences common in longitudinal affect data. Figure 1 provides an overview of the architecture.

Semantic embeddings To capture the linguistic meaning of each entry, we encode text using `all-MiniLM-L6-v2`, a pre-trained Transformer encoder from the `sentence-transformers` library (Reimers and Gurevych, 2019), producing a fixed-dimensional dense semantic embedding.

Psycholinguistic features Psycholinguistic word usage has been consistently linked to emotional states and mental health outcomes (Tausczik and Pennebaker, 2010; Boyd and Schwartz, 2021). We extract features using LIWC (Linguistic Inquiry and Word Count) (Pennebaker et al., 2015), which maps words to affective, cognitive, and functional linguistic categories. For each entry, we compute LIWC category frequencies, providing interpretable signals related to emotional tone, self-focus, and discourse style that complement dense semantic embeddings. In the submitted system these LIWC counts were used as-is, without length normalization or feature scaling. In a followup refinement we additionally (i) length-normalize each category by dividing by the post token length (appending $\log(1+\text{num\_tokens})$ as an explicit length feature) and (ii) standardize the resulting features with a `StandardScaler` fit on the training set.[1]

[1]LIWC features were neither length-normalized nor scaled in the submitted system; both steps were added only in post-submission refinement.

Item-score similarity features We further extract task-specific features by measuring the similarity between each text entry and a small set of predefined valence and arousal archetype sentences. The archetypes were generated using a large language model informed by the PHQ-9 item score framework, designed to reflect low, neutral, and high levels of affect. Sentence embeddings are computed using the `mxbai-embed-large-v1` model (Lee et al., 2024; Li and Li, 2023), and similarity scores to each archetype are encoded directly as model features.

User embeddings Affective self-reports may vary systematically between persons (e.g., consistently higher or lower reported affect). Comparing the variance of the users' mean ratings to the total variance in the training data, we estimate that 36.6% of valence variance and 17.5% of arousal variance are attributable to between-user differences. To account for this, we learn a trainable 32-dimensional embedding for each `user_id`. These user embeddings are initialized at random near-zero values and concatenated with text-extracted features. The user embedding dimensions are updated during training via backpropagation jointly with the other features. This allows the model to represent between-user variation in reported affect while still modeling within-user variation across time. At inference time, users present in the training set reuse their learned embedding, while unseen users are assigned a fixed zero vector of the same dimensionality, so that predictions for cold-start users rely solely on the text-derived features.

Regression model We use a multi-layer perceptron (MLP) regressor with ReLU activations and dropout. The input is the concatenation of all text-derived features and the user embedding. The MLP consists of three hidden layers (512, 256, 128) followed by a 2-unit output layer producing valence and arousal predictions jointly. Models are trained with mean squared error (MSE) loss for both targets, and optimized using Adam (Kingma and Ba, 2015) with a learning rate of 10^{-4}.

5 Experimental Setup

The experiments are performed on the official train/test splits provided for the shared task. We further split the training data 80/20 into train and validation sets at the instance level using a fixed random seed, stratifying by neither user nor label to reflect a within-user evaluation scenario where

the goal is to model longitudinal affective variation while maintaining stable individual baselines. This validation set is used for hyperparameter tuning and early stopping. Performance is evaluated independently for valence and arousal using Mean Squared Error (MSE) and Mean Absolute Error (MAE). The MLP is trained with MSE loss, optimized using Adam (Kingma and Ba, 2015) with learning rate=10^{-4}, batch size=32, and dropout rate=0.2. The input is the concatenation of all text-derived features and the user embedding. Training runs for up to 100 epochs with early stopping patience of 5, based on validation loss.

6 Submission Results

Our system achieved a composite Pearson correlation of $r = 0.665$ for valence and $r = 0.460$ for arousal, with an overall V&A average of 0.562, ranking **10th out of 29** submitted systems (including the random baseline). Across participating teams, valence predictions consistently yield higher correlations than arousal, a pattern also observed in our system. This trend is evident across nearly all submissions (Figure 2), suggesting that the lower performance on arousal is not system-specific but reflects inherent differences between affect dimensions. Valence is often expressed more directly through explicit lexical cues, whereas arousal captures activation intensity that is conveyed more implicitly through contextual, stylistic, or tonal signals (Bradley and Lang, 1994). As a result, arousal may be inherently more challenging to model using surface-level lexical and semantic representations. Additional metrics provided by the organizers further indicate model behavior. For our system, performance remains relatively stable across seen and unseen users (valence: $r = 0.666$ seen vs. 0.660 unseen; arousal: 0.402 seen vs. 0.550 unseen), suggesting reasonable generalization to new users. In the words-only setting, performance is close to the full evaluation (valence $r = 0.658$, arousal $r = 0.542$). In contrast, essay-only inputs show slightly reduced correlations (valence $r = 0.653$, arousal $r = 0.365$), particularly for arousal, supporting the hypothesis that longer narrative text introduces additional semantic variability that is harder for feature-based regression models to capture.

Overall, these results reinforce two observations: (1) valence appears more robustly encoded in lexical signals across systems and evaluation splits, and (2) arousal prediction remains more sensitive to user variability and discourse structure, highlighting an intrinsic modeling challenge rather than a system-specific weakness.

7 Ablation Study

To better understand the contribution of each feature component,we conduct a series of ablation experiments on the validation set. Starting from the full model (semantic embeddings, LIWC features, similarity-based features, feeling representation features, structural indicators, and user embeddings), we remove or keep individual components to evaluate their impact on performance. The analysis contains composite, between-user, and within-user correlations, along with MAE metrics Table 3. To clarify feature contributions for the more difficult arousal dimension, Table 4 summarizes selected ablations relative to the full model.

Overall performance The best overall performance in terms of $r_{\text{composite}}$ for valence is achieved by the full model (0.756), while for arousal the full model also remains competitive (0.586). Removing major components consistently reduces performance, particularly for valence. This pattern indicates that the majority of engineered features contribute meaningfully for model's prediction, showing that valence is strongly supported by lexical-semantic signals. In contrast, arousal appears less sensitive to individual feature removals, potentially reflecting a more diffuse or context-dependent signal.

Effect of user embeddings User embeddings have a substantial role in capturing stable individual patterns. Removing user embeddings decreases valence composite correlation from 0.756 to 0.730 and arousal from 0.586 to 0.522. The reduction is especially significant in between-user correlation, indicating that user-specific representations primarily contribute in inter-individual variation rather than within-user fluctuations. This shows the importance of modelling individuals' behaviour in longitudinal settings.

Effect of semantic embeddings and similarity features Semantic embeddings and similarity features provide complementary signals. Removing similarity features ("No Sims") reduces valence from 0.756 to 0.735, while removing contextual embeddings also leads to measurable decreases. Models only rely on embeddings or similarity features perform substantially worse than the the mod-

els have them together, suggesting that these features capture distinct but complementary aspects of affect expression. Embeddings likely encode contextual semantics, while similarity features provide alignment with affective prototypes, and their combination yields more robust predictions.

Effect of LIWC features Removing LIWC features produces only a minor decrease in performance (valence: 0.756 $\rightarrow$ 0.753; arousal: 0.586 $\rightarrow$ 0.599). This indicates that psycholinguistic features contribute minor gains. Their limited standalone impact suggests that while function-word and category distributions correlate with affect, they are less discriminative than contextual semantic representations.

Within- vs. between-user behavior Across nearly all ablation studies, between-user correlations are higher than within-user correlations. This pattern indicates that modeling stable individual baselines is substantially easier than capturing longitudinal affective fluctuations.

8 Error Analysis

We perform error analysis along three linguistic dimensions. First, we compute Pearson correlations between POS tag[2] frequencies and absolute prediction error, applying Bonferroni correction across 17 POS categories ($p < .003$). We categorize test items by error magnitude into good (MAE < 0.3), medium ($0.3 - 0.6$), and poor (> 0.6) groups. Second, we compare mean POS ratios across these groups.

Third, since inputs are either narrative essays or affect-word lists, we compare error distributions between these types to see how input format relates to performance. All analyses are conducted separately for valence and arousal to account for their distinct affective dynamics. Results are shown in Figure 3.

Prediction errors vary systematically with both linguistic structure and input modality. For arousal, size POS categories correlate significantly ($p < .003$) with absolute error: more particles, pronouns, adpositions, and auxiliaries correlate with higher error, while adjectives and punctuation correlate with lower error. Interestingly, POS composition is not meaningfully associated with *true* arousal ratings; it may relate more to lexical semantics than syntactic distribution. The model may predict arousal better when text contains more descriptive (ADJ) or structured (PUNCT) content, though further fine-grained analysis is needed. For valence, only adpositions correlate significantly with absolute error, weakly ($r = 0.07$). Unlike arousal, self-reported valence does correlate with POS composition: higher valence is associated with more nouns, determiners, and adpositions, and lower valence with more adjectives and punctuation.

Second, stratification by prediction quality indicates that linguistic structure contributes to model performance. Harder instances tend to contain distinct more pronouns, verbs, and nouns compared to easier ones. This could suggest that affect may be conveyed implicitly through event descriptions or relational context rather than explicit emotion terms, making it harder to capture with surface-level features (Zhou et al., 2021).

Third, modality comparisons show pretty similar error distributions between essays and affect-word lists, though essays exhibit a marginally heavier tail for arousal errors, suggesting that narrative text introduces slightly more variability for that dimension. No clear modality effect is observed for valence.

Overall, linguistic structure and input modality meaningfully influence prediction difficulty, even when using strong semantic representations and user embeddings.

9 Conclusion

In this paper, we presented McMaster NLP's system for SemEval-2026 Task 2, Subtask 1: a lightweight multi-feature regression model for predicting longitudinal variation in emotional valence and arousal from ecologically collected text. Our approach combines sentence embeddings, psycholinguistics features, lexicon-based similarity features, and user embeddings in a joint MLP regressor, allowing the model to capture both entry-level linguistic signals and stable between-user differences in a simple, interpretable framework.

Our results show that this feature-based approach provides a strong baseline for the shared task, with valence predicted more reliably than arousal. Our ablation study further indicates that valence is strongly supported by lexical-semantic signals, while arousal appears less sensitive to individual feature removals. Error analysis further shows that prediction difficulty varies by both lin-

[2]POS tags are extracted with spaCy 3.8.11 using the `en_core_web_sm` v3.8.0 model and its Universal POS tag set (17 categories).

guistic structure and input modality: longer, more narrative essay responses tend to be harder to model than shorter feeling-word entries, and texts with certain syntactic patterns are associated with larger errors. In particular, arousal prediction is more sensitive to narrative and syntactic complexity, while valence relies more on lexical-semantic cues.

Overall, these findings suggest that relatively lightweight and interpretable architectures can still perform effectively on longitudinal affect prediction, especially when they combine contextual text representations with psychologically motivated features and user-level modeling. In future work, we aim to better capture implicit affect in longer narrative writing, explore stronger temporal modeling of user trajectories, and refine task-specific affect features to improve robustness across input types and affect dimensions. Our code is publicly available at https://github.com/McMasterNLP/SemEval-2026.

Limitations

Several limitations should be acknowledged. First, the affect archetypes used in our feature design were generated through prompted language modeling rather than constructed by clinical or psychological experts. Although the archetypes were designed to reflect established valence–arousal definitions (Bradley and Lang, 1994), expert-created prototypes grounded in psychological theory may yield more precise and theoretically aligned representations.

Second, the gold labels are based on self-reported affect, which is inherently subjective and context-dependent. Self-reports may vary across individuals due to differences in interpretation, mood, personality, or reporting style, introducing noise that cannot be fully captured by textual features alone. This subjectivity may be particularly pronounced for arousal, which reflects activation intensity rather than clear evaluative polarity. Future work could integrate additional personalized attributes beyond the learned user embeddings to investigate these between-person variations.

Third, our approach primarily relies on engineered linguistic and statistical features combined with a shallow regression architecture. While this provides interpretability and efficiency, such feature-based models may struggle to capture complex contextual, discourse-level, and temporal dynamics—especially for arousal, which may depend on subtle stylistic or pragmatic cues. Future work could explore deeper neural architectures (e.g., contextual transformers with temporal modeling or multimodal representations) to better capture these higher-order affective signals.

Third, our approach primarily relies on engineered linguistic and statistical features combined with a shallow regression architecture. While this provides interpretability and efficiency, such feature-based models may struggle to capture complex contextual, discourse-level, and temporal dynamics—especially for arousal, which may depend on subtle stylistic or pragmatic cues. In particular, although the shared task defines the input as a chronologically ordered sequence of texts per user, our model predicts valence and arousal for each text independently: the user embedding only encodes a static, person-level baseline and does not represent dynamic emotional trajectories. Future work could replace the MLP head with a sequential module e.g., a LSTM or a lightweight temporal Transformer, to explicitly model within-user dynamics, and more broadly explore deeper neural architectures (e.g., contextual transformers with temporal modeling or multimodal representations) to better capture these higher-order affective signals.

Overall, these limitations suggest that improvements may come from (1) expert-informed affect representations, (2) more robust handling of label subjectivity, and (3) deeper contextual modeling approaches tailored to dynamic emotional expression.

Ethical Considerations

NLP systems must be designed with careful attention to social context, stigma, and harm. Barries like stigma affects who produces mental-health–related language and the content they disclose, and under what conditions—introducing selection effects into the data Bharadwaj et al. (2017). Additionally, NLP systems can encode and amplify societal power imbalances; Blodgett et al. (2020) argue that "bias" in NLP is not merely a technical artifact but is connected to histories of marginalization and unequal impacts, motivating evaluation that goes beyond average accuracy to consider who is harmed by errors and why. In affect and mental-health modeling, these concerns translate into questions about whether models systematically misread distress for some groups or

linguistic styles, and whether deployment would differentially affect already-stigmatized communities.

References

Prashant Bharadwaj, Mallesh M. Pai, and Agne Suziedelyte. 2017. Mental health stigma. *Economics Letters*, 159:57–60.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476. Association for Computational Linguistics.

Ryan L. Boyd and H. Andrew Schwartz. 2021. Natural language analysis and the psychology of verbal behavior: The past, present, and future states of the field. *Journal of Language and Social Psychology*, 40(1):21–41.

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

Trevor Cohen, Benjamin Blatter, and Vimla L. Patel. 2008. Simulating expert clinical comprehension: Adapting latent semantic analysis to accurately extract clinical concepts from psychiatric narrative. *Journal of Biomedical Informatics*, 41(6):1070–1087.

Kevin Doherty, Andreas Balaskas, and Gavin Doherty. 2020. The design of ecological momentary assessment technologies. *Interacting with Computers*, 32(3):257–278.

J. C. Eichstaedt, R. J. Smith, R. M. Merchant, L. H. Ungar, P. Crutchley, D. Preoţiuc-Pietro, D. A. Asch, and H. A. Schwartz. 2018. Facebook language predicts depression in medical records. *Proceedings of the National Academy of Sciences of the United States of America*, 115(44):11203–11208. Epub 2018 Oct 15.

Zhijun Guo, Alvina Lai, Johan H. Thygesen, Joseph Farrington, Thomas Keen, and Kezhi Li. 2024. Large language models for mental health applications: Systematic review. *JMIR Mental Health*, 11:e57400.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Sean Lee, Aamir Shakir, Darius Koenig, and Julius Lipp. 2024. Open source strikes bread - new fluffy embeddings model.

Xianming Li and Jing Li. 2023. Angle-optimized text embeddings. *arXiv preprint arXiv:2309.12871*.

Siddharth Mangalik, Johannes C. Eichstaedt, Salvatore Giorgi, Jihu Mun, Farhan Ahmed, Gilvir Gill, Adithya V. Ganesan, Shashanka Subrahmanya, Nikita Soni, Sean A. P. Clouston, and H. Andrew Schwartz. 2024. Robust language-based mental health assessments in time and space through social media. *npj Digital Medicine*, 7:109.

Marwan Omar, Soohyeon Choi, DaeHun Nyang, and David Mohaisen. 2022. Robust natural language processing: Recent advances, challenges, and future directions. *IEEE Access*, 10:86038–86056.

James W. Pennebaker. 2018. Expressive writing in psychological science. *Perspectives on Psychological Science*, 13(2):226–229. Epub 2017 Oct 9.

James W Pennebaker, Roger J Booth, Ryan L Boyd, and Martha E Francis. 2015. Linguistic inquiry and word count: Liwc2015.

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. *Development and Psychopathology*, 17(3):715–734.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

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

Nikita Soni, H. Andrew Schwartz, Ryan L. Boyd, Phi Long Bui, Syeda Mahwish, August Håkan Nilsson, Adithya V Ganesan, Lyle Ungar, Niranjan Balasubramanian, and Saif M. Mohammad. 2026. SemEval-2026 task 2: Predicting variation in emotional valence and arousal over time from ecological essays. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*. Association for Computational Linguistics.

Yla R. Tausczik and James W. Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1):24–54.

G. K. Verma and U. S. Tiwary. 2017. Affect representation and recognition in 3d continuous valence–arousal–dominance space. *Multimedia Tools and Applications*, 76:2159–2183.

Huy Vu, Huy Anh Nguyen, Adithya V. Ganesan, Swanie Juhng, Oscar N. E. Kjell, Joao Sedoc, Margaret L. Kern, Ryan L. Boyd, Lyle Ungar, H. Andrew Schwartz, and Johannes C. Eichstaedt. 2024. Psychadapter: Adapting llm transformers to reflect traits, personality and mental health.

B. R. Whitehead and C. S. Bergeman. 2013. Self-reported health bias: the role of daily affective valence and arousal. *Psychology & Health*, 28(7):784–799. Epub 2013 Jan 21.

T. Zhang, A. M. Schoene, S. Ji, and et al. 2022. Natural language processing applied to mental illness detection: a narrative review. *npj Digital Medicine*, 5:46.

Deyu Zhou, Jianan Wang, Linhai Zhang, and Yulan He. 2021. Implicit sentiment analysis with event-centered text representation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6884–6893, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Scale	Valence Rating	Arousal Rating
5	Pleasant	Excited
4	Pleased	Wide-awake
3	Neutral	Neutral
2	Unsatisfied	Dull
1	Unpleasant	Calm

Table 1: Five-point scale used for self-reported valence and arousal ratings. Valence ranges from unpleasant (1) to pleasant (5), while arousal ranges from calm (1) to excited (5), reflecting the two-dimensional affect framework (Bradley and Lang, 1994).

A System details

To construct task-specific archetype sentences for valence and arousal, we used a large language model (OpenAI ChatGPT 5.1) with few-shot prompting, informed by the PHQ-9 depression questionnaire [3]. The final prompt and resulting archetypes are shown in Tables 2 and 5, respectively. Archetypes were generated once and fixed across all experiments.

B Additional Figures

[3]PHQ-9 archetypes are from prior work

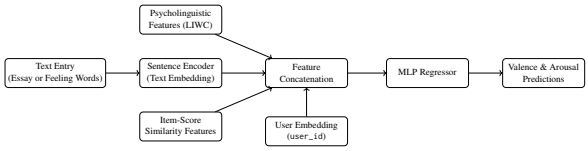

Figure 1: System overview for Subtask 1. Each textual entry is encoded using semantic embeddings, psycholinguistic features, and item-score similarity features. A trainable user embedding is concatenated to model stable individual differences. The combined representation is passed to an MLP regressor to jointly predict valence and arousal.

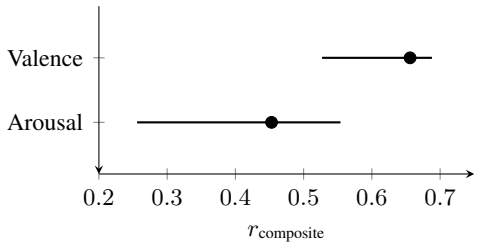

Figure 2: Min–median–max performance ranges across teams for Valence and Arousal in Subtask 1. Horizontal segments indicate min–max ranges; dots indicate medians.

Few-shot Prompt for Valence and Arousal Archetype Generation

You are given a five-point scale for Valence (emotional pleasantness) and Arousal (level of activation). Your task is to generate five short archetypal self-report statements for each dimension, corresponding to scale values 1 (lowest) to 5 (highest).
Each archetype should:

- Be written in first person
- Sound like a natural self-report sentence
- Clearly reflect the emotional intensity implied by the scale level
- Be one sentence per archetype

Example (Style Reference):
Little interest or pleasure in doing things
Feeling down, depressed, or hopeless
Trouble falling or staying asleep, or sleeping too much

I haven't been wanting to do anything.
I've been feeling really low lately.
I can't sleep properly—either too little or too much.

Scale Definitions:

Valence scale:
5 – Pleasant
4 – Pleased
3 – Neutral
2 – Unsatisfied
1 – Unpleasant

Arousal scale:
5 – Excited
4 – Wide-awake
3 – Neutral
2 – Dull
1 – Calm

Task:
Generate five Valence archetype self-report statements, ordered from 1 (Unpleasant) to 5 (Pleasant), and five Arousal archetype self-report statements, ordered from 1 (Calm) to 5 (Excited).

Table 2: Prompt used to generate archetypal self-report statements for valence and arousal dimensions using a large language model.

	Valence (V)						Arousal (A)						Average	
Configuration	$r_{\text{comp}}\uparrow$	$r_{\text{btw}}\uparrow$	$r_{\text{win}}\uparrow$	$\text{mae}_{\text{comp}}\downarrow$	$\text{mae}_{\text{btw}}\downarrow$	$\text{mae}_{\text{win}}\downarrow$	$r_{\text{comp}}\uparrow$	$r_{\text{btw}}\uparrow$	$r_{\text{win}}\uparrow$	$\text{mae}_{\text{comp}}\downarrow$	$\text{mae}_{\text{btw}}\downarrow$	$\text{mae}_{\text{win}}\downarrow$	$\bar{r}_{\text{comp}}\uparrow$	$\overline{\text{mae}}_{\text{comp}}\downarrow$
No LIWC	0.752	0.755	0.708	0.652	0.471	0.532	0.599	0.571	0.473	0.482	0.334	0.392	**0.676**	0.567
All features (full)	0.756	0.770	0.707	0.660	0.454	0.548	0.586	0.588	0.447	0.484	0.322	0.398	0.671	0.572
No feeling_rep	0.762	0.793	0.709	0.648	0.438	0.546	0.576	0.595	0.440	0.492	0.322	0.398	0.669	0.570
Sims + User Emb	0.762	0.744	0.709	0.624	0.470	0.528	0.572	0.466	0.430	0.499	0.356	0.403	0.667	**0.562**
Embedding + User Emb	0.751	0.786	0.701	0.646	0.459	0.529	0.581	0.575	0.419	0.492	0.323	0.405	0.666	0.569
No is_words	0.751	0.769	0.701	0.663	0.457	0.552	0.572	0.583	0.427	0.493	0.327	0.402	0.662	0.578
No token_num	0.750	0.764	0.702	0.666	0.462	0.552	0.573	0.580	0.431	0.493	0.327	0.401	0.662	0.579
No Embedding	0.753	0.794	0.694	0.665	0.443	0.564	0.550	0.503	0.414	0.500	0.346	0.408	0.652	0.583
No Sims	0.735	0.748	0.687	0.681	0.494	0.559	0.553	0.572	0.398	0.504	0.329	0.411	0.644	0.593
No User Embedding	0.730	0.762	0.694	0.695	0.480	0.557	0.522	0.534	0.410	0.533	0.351	0.411	0.626	0.614
Embedding only	0.697	0.699	0.675	0.727	0.537	0.557	0.536	0.539	0.408	0.530	0.347	0.412	0.617	0.629
LIWC + User Emb	0.693	0.700	0.631	0.721	0.521	0.596	0.513	0.500	0.353	0.529	0.363	0.425	0.603	0.625
Sims only	0.699	0.707	0.665	0.720	0.508	0.572	0.487	0.389	0.412	0.549	0.381	0.407	0.593	0.634
feeling_rep + User Emb	0.596	0.623	0.526	0.777	0.611	0.630	0.520	0.446	0.361	0.529	0.374	0.428	0.558	0.653
LIWC only	0.649	0.636	0.620	0.783	0.570	0.607	0.419	0.414	0.287	0.572	0.377	0.443	0.534	0.677
feeling_rep only	0.540	0.518	0.526	0.835	0.667	0.632	0.412	0.406	0.342	0.571	0.380	0.436	0.476	0.703
User Embedding only	0.430	0.527	≈0	0.885	0.639	0.739	0.416	0.292	≈0	0.575	0.399	0.453	0.423	0.730
is_words only	−0.010	−0.030	−0.040	1.042	0.787	0.746	0.045	−0.108	−0.106	0.664	0.428	0.458	0.017	0.853
token_num only	−0.027	−0.055	−0.049	1.044	0.791	0.750	0.011	−0.018	−0.128	0.667	0.426	0.465	−0.008	0.855

Table 3: Ablation Results

Table 4: Arousal-focused summary of selected ablations from Table 3. Deltas are measured relative to the full model. It shows that user embeddings have the largest impact on arousal performance, while semantic embeddings and similarity features also provide clear gains. In contrast, LIWC has a limited and non-uniform marginal effect, suggesting partial overlap with stronger contextual representations.

Configuration	r_{comp}	Δr_{comp}	r_{btw}	Δr_{btw}	r_{win}	Δr_{win}	MAE_{comp}	$\Delta\text{MAE}_{\text{comp}}$
All features (full)	0.586	—	0.588	—	0.447	—	0.484	—
No LIWC	0.599	+0.013	0.571	-0.017	0.473	+0.026	0.482	-0.002
No feeling_rep	0.576	-0.010	0.595	+0.007	0.440	-0.007	0.492	+0.008
No Embedding	0.550	-0.036	0.503	-0.085	0.414	-0.033	0.500	+0.016
No Sims	0.553	-0.033	0.572	-0.016	0.398	-0.049	0.504	+0.020
No User Embedding	0.522	-0.064	0.534	-0.054	0.410	-0.037	0.533	+0.049

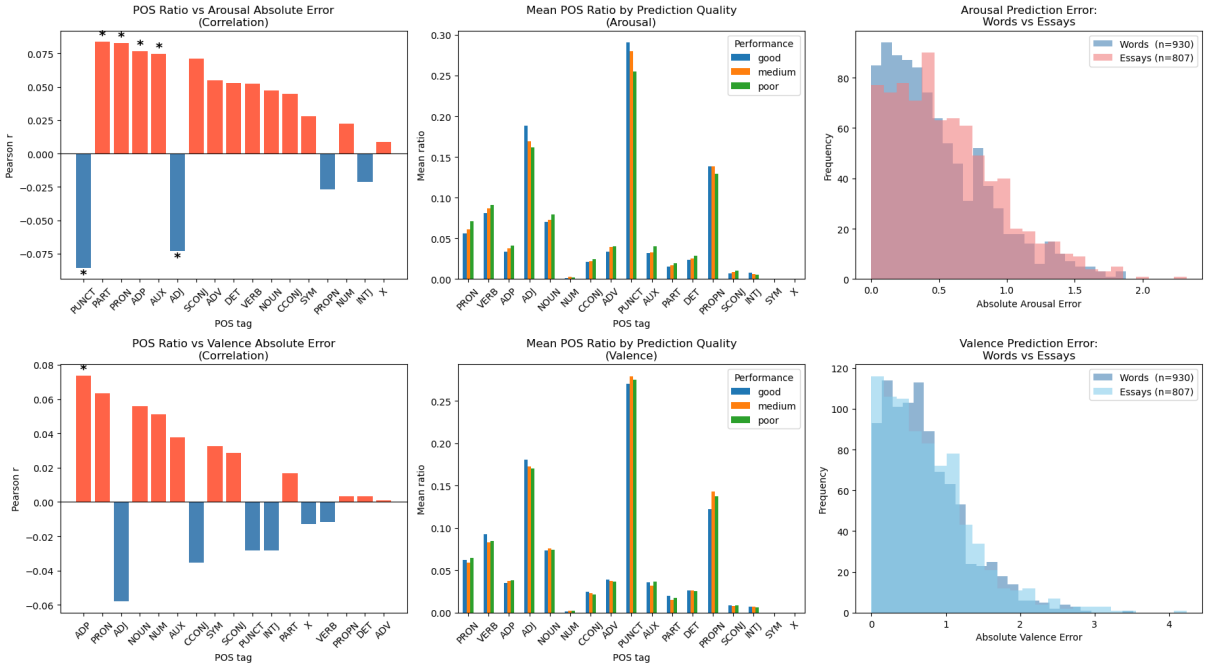

Figure 3: Error analysis across linguistic dimensions. Top row: Arousal predictions. Bottom row: Valence predictions. Left: POS ratio correlations with absolute error (Pearson r). The $*$ denotes statistical significance with p-value < 0.003 (Bonferroni correction applied for 17 POS categories). Middle: Mean POS distributions by prediction quality (good/medium/poor). Right: Error distribution by input modality (words vs. essays).

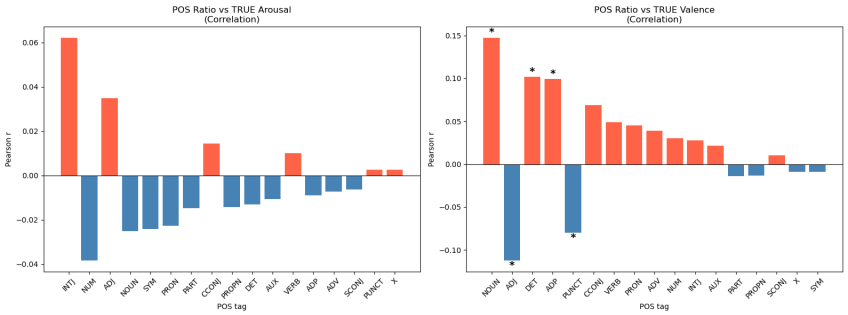

Figure 4: Correlations between POS category usage and self-reported Arousal and Valence scores.

Valence	**Arousal**
I feel genuinely good about this, and it puts me in a positive mood.	I feel really energized and excited right now.
I'm fairly happy with how things turned out.	I feel very alert and focused.
I don't feel strongly one way or the other about this.	My energy level feels pretty normal.
I'm not very happy with this, and it didn't meet my expectations.	Everything feels slow and uninteresting right now.
This makes me feel bad and uncomfortable.	I feel calm and relaxed.

Table 5: Valence and Arousal scale items.