VerbaNex AI at SemEval-2026 Task 2: DeBERTa for Longitudinal Valence and Arousal Prediction

Melissa Moreno Novoa, Edwin Puertas, Juan Carlos Martinez-Santos
Universidad Tecnológica de Bolívar
Cartagena, Colombia
{memoreno, epuerta,jcmartinezs}@utb.edu.co

Abstract

This paper describes our submission to SemEval 2026 Subtask 1: Longitudinal Affect Assessment, which aims to predict continuous valence and arousal scores from chronologically ordered texts. Implement two regression-based configurations built on DeBERTa fine-tuning: a contextual model and a hybrid model that incorporates normalized lexical features derived from the NRC VAD lexicon. Both systems preserve temporal ordering and apply user-level data splits to ensure generalization to unseen individuals. Results show competitive performance, with stronger valence outcomes than arousal. The integration of lexical features does not yield consistent improvements for arousal, highlighting the difficulty of modeling emotional intensity dynamics. Error analysis indicates challenges in handling implicit emotions, pragmatic ambiguity, and subtle affective shifts over time. Overall, the findings underscore the importance of combining contextual representations with structured lexical knowledge to address longitudinal variability in emotional activation.

1 Introduction

Emotions constitute psychophysiological responses that enable individuals to adapt to their environment and assign meaning to their experiences (Wu et al., 2025). They manifest in language and vary according to context, and can be described through continuous dimensions such as valence (V), associated with the degree of pleasantness or unpleasantness, and arousal (A), related to the level of activation (Moreno Novoa et al., 2025). Analyzing these dimensions is essential for understanding affective dynamics in human interaction. In this context, this paper proposes a system that predicts a pair of valence and arousal values for each chronologically ordered text entry, aiming to model emotional evolution over time.

We adopt a regression strategy based on fine-tuning DeBERTa to predict V and A from longitudinal text (Takenaka, 2025). Temporal structure and user-level separation are preserved to evaluate generalization to unseen individuals. In the base configuration, the model learns contextual representations and produces two continuous outputs. In the hybrid configuration, normalized NRC VAD (The NRC Valence, Arousal, and Dominance Lexicon)features (Mohammad, 2025) are concatenated with the transformer representation before a final prediction layer. This approach combines deep contextual signals with structured lexical affective information.

This work contributes to the study of longitudinal affect prediction by explicitly addressing the challenge of generalizing to unseen users, a critical factor in real-world emotional modeling scenarios. The proposed system demonstrates stronger performance in V estimation, while confirming the increased complexity of modeling A due to its high temporal and intra-user variability. In comparison with other participating systems, the approach achieves mid-range performance, reflecting a balanced trade-off between predictive correlation and temporal stability. Furthermore, the analysis identifies limitations associated with very short texts, implicitly expressed emotions, and subtle variations in emotional activation, providing empirical evidence of the key challenges that remain in longitudinal emotion prediction.[1]

2 Background

Affective analysis aims to automatically infer emotional states from language, taking into account not only explicitly affective words but also narrative structures and contextual patterns (Mendes and Martins, 2023). Within this framework, the dimensional model represents emotions in a continuous

[1] https://github.com/VerbaNexAI/SemEval2026

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 635–640

Field	Description
user_id	Unique identifier of the author.
text_id	Unique identifier of the text instance.
text	Raw textual content written by the user.
timestamp	Date and time when the text was produced.
collection_phase	Phase of data collection (e.g., 1–7).
is_words	Indicates whether the input corresponds to isolated words or full text.
valence	Numerical affective score representing pleasantness (range: -2 to 2).
arousal	Numerical affective score representing activation level (range: 0 to 2).

Table 1: Structure of the training data for Subtask 1: Longitudinal Affect Assessment.

Example
user_id: 25 **text_id:** 1487 **timestamp:** 2021-06-08 **valence:** 1.0 **arousal:** 0.0 **text:** I am currently feeling tired and sleepy due to lack of sleep. I woke up thinking it would be a good day, but realized my outlook was pessimistic.
user_id: 25 **text_id:** 1488 **timestamp:** 2021-06-09 **valence:** 1.0 **arousal:** 1.0 **text:** Woke up earlier than usual. It was a nice day, so I walked around and felt good about myself.
user_id: 25 **text_id:** 1489 **timestamp:** 2021-06-10 **valence:** 2.0 **arousal:** 1.0 **text:** I am feeling great. Went to work and socialized. My boss recognized my personality, and I felt emotionally stable.

Table 2: Examples of longitudinal entries with continuous valence and arousal labels.

space defined by valence and arousal, allowing the capture of intensity, variability, and emotional transitions (Pólya and Csertő, 2023). This continuous representation is particularly suitable for longitudinal settings, as it enables modeling gradual affective trajectories rather than discrete emotional categories.

This work addresses Subtask 1: Longitudinal Affect Assessment at SemEval 2026. The task consists of modeling emotional dynamics in chronological sequences of texts produced by different users (Soni et al., 2026). For each text within a sequence, the system must predict a continuous pair of valence and arousal, which formulates the problem as a multi-output regression task.

The input consists of temporally organized texts, including short essays and isolated sentiment words, in English. The output consists of a pair of continuous values (valence, arousal) for each instance. Evaluation additionally considers the model's ability to generalize to unseen users.

In this context, the dataset comprises 5,285 longitudinal texts produced by 182 authors in the service sector in the United States between 2021 and 2024. Table 1 presents the structure of the training data, including textual content, temporal metadata, and continuous affective labels. Each instance corresponds to a longitudinal observation associated with a specific user. Table 2 illustrates representative examples of chronologically ordered entries and their corresponding valence and arousal ranges.

Recent studies have shown that integrating Transformer-based models with dimensional affect representations, such as NRC VAD, strengthens the analysis of emotional intensity and affective changes in text and conversation (Moreno Novoa et al., 2025) (Lim et al., 2025). Furthermore, the expansion of the NRC VAD Lexicon has consolidated the use of valence, arousal, and dominance as a continuous and reliable framework for affect modeling (Garcia et al., 2024; Suresh et al., 2024; Ghosh et al., 2023). In line with these advances, our study adopts a longitudinal regression-based perspective to model continuous emotional trajectories.

3 System Overview

This section describes the algorithmic and modeling decisions adopted for Subtask 1. Two clearly differentiated configurations were implemented: a model based exclusively on contextual representations (System A) and a hybrid model integrating

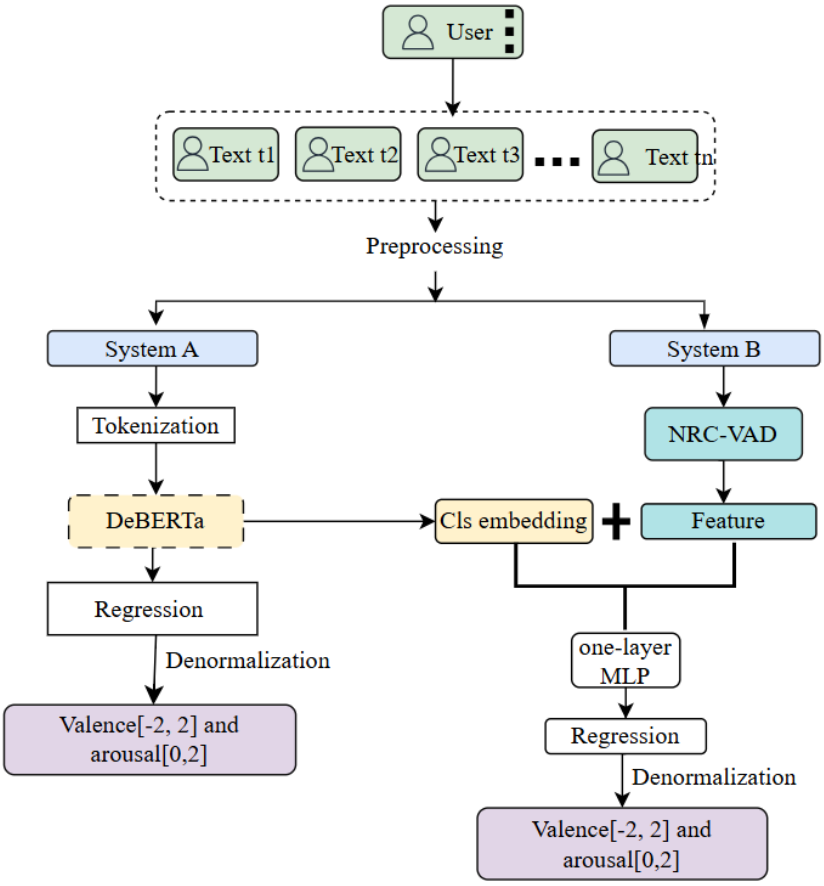

Figure 1: Overall processing pipeline.

contextual information with external lexical knowledge (System B). Both systems were designed to ensure generalization to unseen users, training stability, and robust prediction of valence and arousal under a repeated-observation evaluation setting. Figure 1 illustrates the overall processing pipeline.

System A employs DeBERTa-v3-base as the core model, implemented through the Hugging Face Transformers AutoModelForSequenceClassification interface and configured for two-output regression. The model uses the default sequence-level pooled representation internally provided by the architecture, followed by a shared regression head producing a 2-dimensional output vector that jointly predicts valence and arousal, without imposing sequential dependency between targets. As a design decision, texts were chronologically ordered by user and timestamp to preserve the repeated-observation structure of the dataset.

The training and validation split was performed at the user level using an 80/20 partition, preventing texts from the same individual from appearing in both sets. This strategy is consistent with the evaluation setting of the task, which assesses both between-person and within-person performance, and reduces the risk of overestimating generalization. Each post was processed as an independent training instance, and mini-batches contained samples from multiple users; no sequential backpropagation or user-specific recurrence was applied. Historical posts from the same user were not concatenated or used as additional context during training or inference. During training, original labels were normalized to a common range to stabilize optimization. Final predictions were subsequently transformed back to the original task scale: valence in [-2,2] and arousal in [0,2].

System B extends this configuration by integrating the NRC VAD as an additional external resource. For each text, lexical features were extracted based on unigram, bigram, and trigram matches, including mean VAD scores, coverage proportion, number of matches, and total token count. Features were normalized independently: VAD scores were linearly mapped from $[0, 1]$ to $[-1, 1]$; coverage was preserved in $[0, 1]$; and matched counts and token counts were transformed using $\log(1 + x)$ followed by z-score standardiza-

tion computed exclusively on the training set to prevent information leakage. Architecturally, the hybrid model concatenates the contextual representation from the base encoder with the normalized lexical feature vector, which is processed by a feed-forward layer producing a shared 2-dimensional output vector for joint prediction of valence and arousal. The model was optimized using mean squared error loss under the same fine-tuning configuration described for System A.

Recent advances in affective analysis have incorporated Transformer-based models to improve emotion and empathy detection in text. In particular, architectures such as DeBERTa have shown competitive results in emotion prediction and conversational emotion modeling tasks (Furniturewala and Jaidka, 2024; Liang et al., 2024). However, most existing approaches remain limited to predictions at the individual text or conversational turn level, without considering repeated observations across users or generalization to unseen individuals, which reveals an important limitation of current approaches.

A central challenge of the task lies in its repeated-observation evaluation setting. The objective is not merely to predict isolated values but to capture intra-user variation while maintaining consistent inter-individual distinctions. Evaluation therefore considers both between-person and within-person correlations, allowing assessment of profile differentiation and sensitivity to user-level fluctuations. Another challenge concerns semantic heterogeneity: some texts contain explicit emotional cues, while others convey affect implicitly. System A relies entirely on contextual inference, whereas System B reinforces explicit affective signals through structured lexical information. Together, these configurations explore complementary strategies for affect prediction under user-level repeated observations.

4 Experimental Setup

Texts were chronologically ordered before processing to preserve the repeated-observation structure of the dataset. Data were split at the user level into 80:20 proportions for training and validation, preventing overlap between partitions, while the test set remained fully isolated for final prediction. Each post was processed as an independent training instance, and mini-batches contained samples from multiple users; no sequential backpropagation or user-specific recurrence was applied. Valence and arousal labels were normalized to the interval [-1,1] during training and transformed back to their original ranges after inference. In the hybrid model, lexical features were standardized using statistics computed exclusively from the training set to prevent information leakage.

The model was fine-tuned for 3 epochs using the AdamW optimizer with a learning rate of 2×10^{-5}, batch size 8, weight decay 0.01, and a linear learning-rate scheduler. The architecture combines contextual representations from the base model with six normalized lexical features through a neural network with a single hidden layer of 256 units and ReLU activation. Tokenization was performed with a maximum sequence length of 256 tokens. Experiments were implemented using PyTorch and Hugging Face Transformers. Model selection was based on validation mean absolute error (MAE), and evaluation followed the official protocol using between-person correlation, within-person correlation, composite score, and mean absolute error.

5 Result

The results indicate that both systems achieve competitive performance according to the official evaluation metrics. In Experiment 1 (System A), the text-only configuration shows strong valence performance, with a composite correlation of 0.6316 and a between-person correlation of 0.6895, as reported in Table 3. Performance in arousal is more moderate, with a composite correlation of 0.4629, although inter- and intra-user metrics remain consistent. These findings suggest that the model effectively captures global differences among individuals but struggles to model temporal variations within users.

Experiment 2 (System B), which integrates NRC VAD lexical features, yields a slight improvement in valence, reaching a composite correlation of 0.6364 and an improved within-person correlation of 0.6092 (Table 3). This pattern indicates that explicit affective signals help capture individual emotional fluctuations. However, performance in arousal decreases compared to the contextual baseline, with a composite correlation of 0.2197, suggesting that lexical information does not consistently benefit this dimension and may introduce noise when activation is expressed implicitly.

Comparison with the best performing team (Ta-

System A (Text-only DeBERTa)						
Dimension	r_{comp}	r_{between}	r_{within}	MAE_{comp}	$\text{MAE}_{\text{between}}$	$\text{MAE}_{\text{within}}$
Valence	0.6316	0.6895	0.5657	0.6581	0.4656	0.7885
Arousal	0.4629	0.5238	0.3973	0.3447	0.2462	0.4362
System B (Hybrid Transformer + NRC-VAD)						
Valence	0.6364	0.6621	0.6092	0.3389	0.2371	0.4407
Arousal	0.2197	0.1947	0.2444	0.3958	0.2454	0.5462

Table 3: Results of Systems A, B.

Team	Valence r_{comp}	Arousal r_{comp}	V&A Avg
UKP_Psycontrol (Best)	0.667	0.554	0.611
VerbaNex AI	0.632	0.463	0.547

Table 4: Comparison with the best-performing team on Subtask 1. Higher values indicate better performance.

ble 4) contextualizes our results. Our system achieves a composite valence correlation of 0.632 compared to 0.667 for the leading team, and 0.463 versus 0.554 in arousal. The overall V and A average of 0.547 places the system below the top submission but within a competitive range. The primary margin for improvement lies in modeling arousal, particularly in cases involving implicit emotional intensity.

From an ablation perspective, the comparison between Systems A and B shows that lexical integration improves valence consistency, particularly at the intra-user level, but does not enhance arousal. Error analysis reveals that false negatives in arousal frequently correspond to texts that express fatigue, apathy, or mixed states, without explicit lexical markers. False positives in valence often occur in cases of irony or affective ambivalence. Additionally, the dataset exhibits greater stability in valence than in arousal, whose intra-user variability complicates consistent modeling of emotional change over time. Overall, the system demonstrates robust valence modeling and competitive performance, with clear opportunities to improve its capture of activation dynamics.

6 Conclusions

Results indicate that emotional valence can be modeled with relative stability across users. In contrast, emotional arousal remains inherently challenging due to its strong intra-individual variability and its frequently implicit linguistic realization. The findings suggest that the performance limitations observed in arousal prediction are not solely attributable to model capacity, but rather to the temporal and contextual nature of emotional activation itself.

A key contribution of this work is providing empirical evidence that preserving chronological structure and enforcing user-level generalization fundamentally alter affect-prediction behavior. The results highlight that longitudinal emotion modeling should be approached as a trajectory learning problem, in which emotional states evolve rather than emerging from isolated textual signals.

Overall, this study provides methodological insights for future research in affective computing by demonstrating the need for temporally aware architectures and evaluation strategies that can model evolving emotional patterns in real-world settings. The proposed framework provides a competitive and reproducible foundation for studying continuous affect dynamics under realistic longitudinal conditions.

7 Future Work

As a line of future work, we propose explicitly modeling the temporal dynamics of the arousal dimension, which exhibited higher intra-user variability and lower composite correlation across experiments. Future research will investigate sequence-aware architectures, including recurrent neural networks, temporal attention mechanisms, and transformer-based models designed to capture dependencies across repeated user observations. Rather than predicting texts independently, these approaches may model affective variation as a continuous temporal process.

In particular, temporal transformers and attention-based models may enable the system to capture gradual emotional transitions and contextual accumulation across consecutive observations.

The effectiveness of these approaches could be quantitatively evaluated through trajectory-level metrics, such as temporal consistency, reconstruction of intra-user variance, and prediction accuracy over affective transitions between consecutive timestamps. These measures would allow assessment of whether models successfully capture dynamic changes in arousal rather than static affective signals.

Additionally, future work will explore discourse-level and pragmatic features associated with implicit emotional activation, including contextual intensity markers, narrative cues, and indirect expressions of affect. Since arousal is often conveyed implicitly, incorporating higher-level linguistic representations may improve sensitivity to latent emotional activation.

Finally, the hybrid framework could be extended through adaptive lexical representations and user-aware learning strategies, such as meta-learning or personalized embeddings, to enhance generalization to unseen individuals. These directions aim to advance affect prediction under repeated-observation settings toward more adaptive and robust real-world monitoring systems.

Acknowledgments

The authors express their gratitude to the Call 933 "Training in National Doctorates with a Territorial, Ethnic and Gender Focus in the Framework of the Mission Policy — 2023" of the Ministry of Science, Technology and Innovation (Minciencia). In addition, we thank the team of the Artificial Intelligence Laboratory VerbaNex [2], affiliated with the UTB, for their contributions to this project.

References

Shaz Furniturewala and Kokil Jaidka. 2024. Turn-level empathy prediction using psychological indicators. *Preprint*, arXiv:2407.08607.

Santiago Garcia, Elizabeth Martinez, Juan Cuadrado, Juan Carlos Martinez-Santos, and Edwin Puertas. 2024. VerbaNexAI lab at SemEval-2024 task 10: Emotion recognition and reasoning in mixed-coded conversations based on an NRC VAD approach. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1332–1338, Mexico City, Mexico. Association for Computational Linguistics.

Soumitra Ghosh, Asif Ekbal, and Pushpak Bhattacharyya. 2023. Vad-assisted multitask transformer framework for emotion recognition and intensity prediction on suicide notes. *Information Processing & Management*, 60:103234.

Chenyi Liang, Jin Wang, and Xuejie Zhang. 2024. YNU-HPCC at SemEval-2024 task10: Pre-trained language model for emotion discovery and reasoning its flip in conversation. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 777–784, Mexico City, Mexico. Association for Computational Linguistics.

Dongha Lim, Kangwon Lee, Junhui Jo, Hyeonji Lim, Hyeongchan Bae, and Changgu Kang. 2025. Web-based platform for quantitative depression risk prediction via vad regression on korean text and multi-anchor distance scoring. *Applied Sciences*, 15:10170.

Gonçalo Azevedo Mendes and Bruno Martins. 2023. Quantifying valence and arousal in text with multilingual pre-trained transformers. In *Advances in Information Retrieval*, pages 84–100, Cham. Springer Nature Switzerland.

Saif M. Mohammad. 2025. Nrc vad lexicon v2: Norms for valence, arousal, and dominance for over 55k english terms. *Preprint*, arXiv:2503.23547.

Melissa Moreno Novoa, Edwin Puertas, and Juan Carlos Martinez-Santos. 2025. UTBNLP at Semeval-2025 task 11: Predicting emotion intensity with BERT and VAD-informed attention. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 1217–1222, Vienna, Austria. Association for Computational Linguistics.

Tibor Pólya and István Csertő. 2023. Emotion recognition based on the structure of narratives. *Electronics*, 12(4).

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.

Tharun Suresh, Ayan Sengupta, Md Shad Akhtar, and Tanmoy Chakraborty. 2024. A comprehensive understanding of code-mixed language semantics using hierarchical transformer. *IEEE Transactions on Computational Social Systems*, 11:4139–4148.

Yoichi Takenaka. 2025. Performance evaluation of emotion classification in japanese using roberta and deberta. *Preprint*, arXiv:2505.00013.

Haoqi Wu, Daicong Li, Zhenan Chen, Xiaolan Tang, and Guangyu Wang. 2025. Linking forests, coasts, and people: social media insights into sentiment and wellness perceptions in china's nature reserves. *Trees, Forests and People*, 22:101068.

[2] https://github.com/VerbaNexAI/SemEval2026/blob/main/SemEval2026_task2.ipynb