UAlberta at SemEval-2026 Task 2: Temporal Fusion Models for Predicting Affect Over Time

Duc Ho, Khanh Bui, Daniela Teodorescu, Grzegorz Kondrak
Alberta Machine Intelligence Institute
Department of Computing Science
University of Alberta, Edmonton, Canada
{dtho,kdbui,dteodore,gkondrak}@ualberta.ca

Abstract

We describe our systems for the SemEval 2026 Task 2 on Predicting Variation in Emotional Valence and Arousal from Ecological Essays. To predict affect in a single instance, and for forecasting dispositional change, we use embeddings from a language model and a Recurrent Neural Network. To predict state changes from a previous timestep to the next, we use time-series forecasting. Our systems ranked first for forecasting dispositional change, and third for forecasting state change over time. We make our code publicly available.

1 Introduction

SemEval 2026 Task 2 on Predicting Variation in Emotional Valence and Arousal from Ecological Essays is about modeling affect using longitudinal data. Affect consists of two dimensions: valence (pleasantness–unpleasantness) and arousal (activated–calm). More specifically, the task is to predict affect scores for longitudinal data (Subtask 1), and changes in affect in future texts (Subtask 2a and 2b). While sentiment analysis is a popular NLP task, gold labels are usually determined through third-party annotations, such as crowd sourcing, without considering how the individual feels when determining emotion labels In contrast, this SemEval presents a new dataset of ecological essays and self-reports of affect in which individuals describe "how they are feeling". The data is collected over three years, which allows for longitudinal analysis of emotions (Soni et al., 2026).

The SemEval task consists of three subtasks. In Subtask 1, the input consists of sequences of text for a user, and the output is the valence and arousal score per instance. This subtask involves users whose texts have been *seen* during training, and also those whose text is not in the training data (i.e., *unseen* users). Subtask 2 focuses on forecasting changes in valence and arousal, and consists of two subtasks. The input in Subtask 2a is a sequence of t texts, and the output is the valence and arousal score for text t+1 in the sequence. For Subtask 2b, the input is likewise a sequence of t texts in group 1, and the output is the change in average valence and arousal for a future equally-sized segment of texts (group 2) for a user. We use similar systems for Subtask 1 and 2b, and a different system for Subtask 2a.

Our systems at a high-level represent the text as embeddings from a language model; we then use these representations to train a neural network to predict the affect score. For Subtask 1, we generate text embeddings for each instance using BERT (Devlin et al., 2019), and then train separate BiLSTM models (Graves and Schmidhuber, 2005) to predict the valence and arousal scores. For Subtask 2b, we use a similar system, except that we train the BiLSTM model to predict valence and arousal scores together. To predict the user's group 2 average scores, we first input the concatenated texts for a user's group 1. Then, we determine disposition change by subtracting the average valence and arousal score for group 1 from those predicted for group 2. Lastly, for Subtask 2a, we similarly use BERT to extract the representations for the sequences of text for a user. Then we train a Temporal Fusion Transformer model (Lim et al., 2021) to jointly predict the valence and arousal score for the next timestep given the previous text, valence, and arousal score.

Our systems are highly competitive in forecasting affect variation. For Subtask 2b, our system ranks first among the submitted systems on the test set, achieving correlations of 0.405 for valence, and 0.602 for arousal. For Subtask 2a, our system ranks third, achieving correlations of 0.615 for valence and 0.674 for arousal, with the average valence and arousal score close to the top performing system. In addition, our relatively light-weight system for Subtask 1 fares reasonably well on the arousal di-

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 605–611

mension, outperforming some of the higher-ranked teams. We make the code for our systems publicly available.[1]

2 Related Work

Past work predicts affect at an instance-level, generating Valence-Arousal (V-A) scores for short sentences or words using pre-trained language models (such as BERT or RoBERTa) on multilingual datasets (Mendes and Martins, 2023). In the context-aware sentiment analysis task, a hybrid model classifies emotions in social media posts and movie reviews by combining a BiLSTM with RoBERTa (Rahman et al., 2025). Their models are comparable to ours in that they combine an LSTM with a pretrained language model. A similar architecture has been used for detecting sarcasm in online texts (Pandey and Singh, 2023). However, both Rahman et al. (2025) and Pandey and Singh (2023) use output layers for classification, whereas we use them for regression.

Within Continuous Emotion Recognition (CER), the prediction of dimensional affect (valence and arousal) is treated as a dynamic time-series problem rather than a discrete classification task. Recent advancements in CER heavily rely on multimodal, time-sensitive inputs to capture continuous emotional trajectories. For example, past work has demonstrated the effectiveness of spatio-temporal convolutional and recurrent neural networks in extracting continuous valence and arousal signals from sequential video data (Teixeira et al., 2021). Similarly, multimodal strategies integrating textual transcripts, voice tones, and facial expressions have been utilized to bridge discrete labels by mapping complex emotional expressions directly into a continuous VAD space (Jia et al., 2025). Beyond audiovisual and textual data, time-sensitive neurophysiological inputs, such as continuous EEG signals, have also been successfully leveraged to predict exact valence and arousal values over time (Galvão et al., 2021). Building upon this foundation, our methodology frames emotion prediction as affective forecasting, utilizing time-aware attention mechanisms to project the continuous trajectory of emotional states.

Recent advancements in Large Language Models (LLMs), such as GPT-4 and LLaMA, have demonstrated impressive zero-shot and few-shot capabilities across various affective computing tasks. Studies exploring the fine-grained affective processing of LLMs indicate that generative models can meaningfully interpret and generate text mapped to the Valence, Arousal, and Dominance (VAD) dimensions (Broekens et al., 2023). However, comprehensive evaluations reveal that while LLMs excel at coarse-grained sentiment classification (e.g., identifying broad positive or negative polarities), they frequently struggle to accurately quantify subtle, continuous emotional cues without extensive alignment or complex prompting strategies (Wang et al., 2024). Furthermore, because standard LLMs encode numbers as text, they are inherently discontinuous in both their encoding and decoding stages (Golkar et al., 2023). While these discrete models can learn to approximate continuous functions, treating numerical outputs as discrete semantic tokens rather than true continuous variables makes them prone to quantization errors when applied to strict regression tasks. Consequently, while LLMs represent a powerful tool for generalized emotion extraction, modeling exact affective trajectories over time often requires more specialized, deterministic architectures.

3 Methods

In this section, we present our methods for each subtask. By framing affect prediction as both a static prediction and a dynamic forecasting problem, our pipeline draws on principles of affective dynamics, emphasizing the continuous temporal evolution of emotional states.

3.1 Subtask 1 and Subtask 2b

Our method for Subtask 1 and 2b builds upon established architectures for emotion regression prediction by combining a pre-trained Transformer-based Language Model (LM) with a bidirectional Recurrent Neural Network (RNN). The LM is utilized to generate dense, high-dimensional representations of the input text. These embeddings are subsequently processed by the bidirectional RNN, which extracts contextual information from both preceding and succeeding tokens to guide the continuous prediction of affective dimensions – which is well suited to the temporal nature of the data.

We adapt the standard classification outputs seen in similar pipelines (Rahman et al., 2025; Pandey and Singh, 2023) to a regression layer for continuous V-A score prediction. Depending on the

[1] https://github.com/UAlberta-NLP/SemEval2026-EmoVA

task constraints, the model capacity is scaled to accommodate different input sequence lengths. For Subtask 1, the architecture is instantiated as independent single-target predictors for valence and arousal. Since it is subjective how different users may judge valence and arousal, we conjecture that independent models should be more robust against inconsistencies between the two dimensions. For Subtask 2b, which requires combining all texts from a user's group, the architecture is unified to predict both dimensions simultaneously.

For the RNN architecture, we chose a BiLSTM (Graves and Schmidhuber, 2005), because it extracts contextual information from both the past and future parts of the input text. The choice of the hybrid BiLSTM architecture over only the Transformer is motivated by the longitudinal nature of the texts in Task 2. For Subtask 1, the BiLSTM processes text sequences to maintain a persistent hidden state that captures the user's affective trajectory. This temporal smoothing mitigates "lexical jitters" noise from isolated high-intensity keywords that may not represent a genuine shift in the user's underlying emotional baseline. In Subtask 2b, the BiLSTM acts as a learned aggregator. While Transformer-only models often treat concatenated text as an unordered collection of tokens, the BiLSTM models the input as a sequential flow. This preserves the longitudinal narrative arc, and captures dispositional evolution more effectively than simple linear averaging of token-level features.

3.2 Subtask 2a

We propose Affective Dynamics Forecasting via Attention-Based Time-Series Modeling as our method for Subtask 2a. To address the temporal aspects, we frame the challenge through the lens of affective dynamics, which conceptualizes emotions as continuously evolving trajectories rather than isolated, static occurrences. Predicting the change in V-A scores from one timestep to the next requires a time-series forecasting approach.

We use the Temporal Fusion Transformer (TFT), an attention-based forecasting architecture originally designed for complex time-series data (Lim et al., 2021; Vaswani et al., 2023). The input consists of a sequence of text embeddings and score vectors. Because the forecasting architecture does not process raw text naively, the text is first encoded into dense embeddings using a pre-trained LM. These embeddings, paired with their historical V-A scores, are fed into the forecasting module to predict the subsequent multi-dimensional affective state (valence and arousal simultaneously).

The TFT was selected because simpler time-series models, such as ARIMA or standard LSTMs, often suffer from recency bias or lack the flexibility to handle static user traits alongside dynamic mood shifts. The TFT's interpretable multi-head attention captures long-range dependencies across the user's history, ensuring the system identifies sustained dispositional shifts rather than reacting to isolated noise in the text.

4 Experimental Setup

In this section, we describe the dataset preparation and training setup.

4.1 Dataset Preparation

The task organizers provided three training datasets, one associated with each subtask. Each dataset comprises 137 users, and each user has an average of 58.7 texts. To preserve the sequential structure of each user's data, we split the dataset into training and development sets by user ID. Each user's entire sequence is allocated to either the training or development set, with 80% of the users assigned to the training set. The development set was held out to validate the trained model.

The datasets for Subtask 2 already include the state change and disposition change values. For Subtask 2a, the final text in each user's sequence is excluded from the training data, as no subsequent text is available to calculate the state change. For Subtask 2b, each text in a user's sequence belongs to either Group 1 or Group 2 based on its recording time. Because our goal is to calculate the change in average valence and arousal between groups, we dropped all group 2 texts for each user during training.

4.2 Training Setup

For all subtasks, the BERT-based models that are used to generate embeddings are frozen. A complete list of hyperparameters for all models can be found in the Appendix A.

For Subtask 1, the organizers provided the test dataset consisting of 1,737 longitudinal texts. We use DistilBERT to generate embeddings, as it is less resource-intensive for inference than other BERT models. After hyperparameter finetuning on the development set, the best parameters are

	V/A	Ours		Baseline	
		r	MAE	r	MAE
1	V	0.556	0.734	0.557	0.743
	A	0.444	0.427	0.299	0.459
2a	V	0.615	1.208	0.615	1.168
	A	0.674	0.635	0.670	0.638
2b	V	0.405	0.635	0.434	0.406
	A	0.602	0.261	0.584	0.286

Table 1: Pearson r correlation and MAE for the three subtasks.

Subtask	Dimension	r	MAE
1	Valence	0.528	0.733
	Arousal	0.444	0.427
2a	Valence	0.623	1.252
	Arousal	0.668	1.137
2b	Valence	0.379	0.672
	Arousal	0.614	0.483

Table 2: Our average results over 5 runs.

trained for 300 epochs, using an embedding token size of 128. Afterward, using these parameters, we retrain the models on the original training set to produce our final system.

For Subtask 2a, we drop all test users whose IDs appear in the test set, and train our final model only on the remaining users. The test users' sequences are then used to infer the next valence-arousal score, and state change is calculated as the prediction minus the last valence-arousal score observed in the sequence from the training data. For this Subtask, we use BERT's last-layer token embeddings to incorporate textual information, since TFT is not a text-processing model. Unlike in other subtasks, we use PCA on the 768-dimensional BERT embeddings to reduce them to 32 dimensions. This is because feeding raw 768-dimensional BERT embeddings into the TFT's Variable Selection Network (a component in the TFT) across a temporal sequence significantly increases the risk of overfitting and computational latency. The PCA is also trained only on the training set. Our optimal setup uses 1 training epoch.

We use a similar setup to train the model for Subtask 2b. We exclude any users whose IDs appeared in the test set. For the remaining users, we concatenate all of their Group 1 texts, separating them with a special `[SEP]` token, and average their corresponding valence and arousal scores. To evaluate our model in the test set, we concatenate the Group 1 and Group 2 texts for each user. We then average their valence and arousal scores of Group 1 and Group 2 to establish what we call the "original average". Next, we train our model to predict the average V-A of the second group. Finally, we subtract these predicted scores from the original average to calculate the disposition change. For our text input, to handle long texts produced by sequence concatenation, we use BERT's last token embeddings, which have a dimension of 768. The number of training epochs is set to 3.

5 Results

Table 1 presents the official results for the three subtasks, as well as the best performing baseline from the organizers. Our average results over 5 runs are in Table 2. We report Pearson r correlation and Mean Absolute Error (MAE), the official metrics for the task. For Subtask 1, the composite r and MAE are presented.

5.1 Official Results

Overall, our proposed pipeline demonstrates highly competitive performance across the evaluation, establishing specific strengths in predicting arousal and modeling temporal affective dynamics. In Subtask 1, while our composite r valence is comparable to the baseline provided by the task organizers, our composite r arousal is substantially better, validating our decision to train independent, specialized models for the two dimensions.

Our system particularly excels in the more complex challenges. For Subtask 2a, our results significantly exceed the baseline, ranking third on the leaderboard. This strong comparative performance suggests that framing emotion shift as a time-series forecasting problem is an effective approach. We placed first on the leaderboard for both valence and arousal correlations in Subtask 2b, demonstrating that our integration of contextual language model embeddings with a recurrent neural network is exceptionally robust for extracting affective signals from long texts.

5.2 Error Analysis

To better understand the limitations of our model for Subtask 1, we conducted a quantitative and qualitative error analysis focusing on the magnitude of continuous prediction errors. Given our selection of DistilBERT, which we hypothesized would be

well-suited for shorter sequence lengths, we investigated whether our MAE correlated with the word count of the input texts. A Pearson correlation analysis revealed no meaningful linear relationship between sequence length and absolute error for either valence ($r = 0.057$) or arousal ($r = 0.068$). This indicates that our model handles the dataset's varying text lengths robustly; prediction decay is not caused by input truncation or a lack of contextual length, but is rather driven by the semantic and affective complexity of the texts themselves.

Qualitative analysis of the predictions with the highest absolute error revealed that our model struggles with lexical polarity imbalance and the contextual weighting of psychological states. Because the model relies entirely on the emergent properties of the DistilBERT embeddings, it frequently defaults to a superficial aggregation of token-level sentiment. For example, when presented with texts containing a high frequency of negative physical descriptors alongside a single strong positive psychological state (e.g., *"Tired, Hungry, Motivated, Lightheaded, Sore"*) the model predicted a strongly negative valence (-2). However, the gold label for this text was strongly positive (2). This discrepancy indicates that the model's attention mechanism allows the sheer quantity of negative tokens to overwhelm the qualitative importance of the positive emotional anchor. In this case it appears that the human annotator prioritized the psychological drive over physical discomfort when assessing overall valence, a nuance the current baseline fails to capture.

Regarding the conjecture that independent models should be more robust to inconsistencies between the two dimensions mentioned in Section 3.1, we analyzed the orthogonality of the absolute prediction errors in our model. A Pearson correlation analysis revealed a near-zero correlation between valence errors and arousal errors ($r = 0.058$). This demonstrates that our independent predictors fail orthogonally: the semantic complexity of predicting one dimension does not bleed into and artificially degrade the prediction of the other. Furthermore, we evaluated performance on atypical emotional expressions by isolating the top 20% of texts with the highest V-A discrepancy. These highly inconsistent pairs inherently pose a more difficult predictive task, exhibiting higher MAE ($V = 1.08$, $A = 0.78$) compared to consistent pairs ($V = 0.82$, $A = 0.52$); however, the decoupled architecture ensures that these performance decays remain localized to the challenging dimension. This confirms that separating the prediction heads effectively insulates the model from cascading failures caused by the subjectivity and inconsistency inherent in multidimensional human affect.

6 Conclusion

We propose a system for SemEval 2026 Task 2, focused on the continuous prediction of valence and arousal across both static and temporal contexts. By framing emotion prediction as a dynamic time-series forecasting challenge grounded in affective dynamics rather than merely as a static extraction task, the proposed architectures demonstrate remarkable robustness. The integration of contextual language models with bidirectional recurrent neural networks is highly effective for processing long-form text, achieving first place on the leaderboard for Subtask 2b. Similarly, our temporal forecasting approach for Subtask 2a yields competitive results, validating the use of multi-horizon attention mechanisms for tracking shifts in affect. Future work could incorporate data-augmentation techniques to better handle data scarcity and improve model training.

Limitations

Despite these strong overall results, we foresee several clear avenues for future improvement. Since our system was trained strictly on the dataset provided by the task organizers, we employed no data augmentation techniques, nor did we integrate external affective resources. In the future, we will consider incorporating external knowledge bases, such as the NRC-VAD lexicon (Mohammad, 2025), to explicitly guide the training process. Furthermore, we intend to introduce user-specific embeddings to capture individualized affective baselines and expressive habits. Learning these user-level patterns could help the model contextualize emotional expressions more accurately, specifically mitigating the problem of negative tokens overwhelming a user's underlying psychological anchor. The static emotion prediction capabilities could be strengthened by injecting psychological knowledge and personalized context, as well as experimenting with a broader range of language models and data augmentation strategies. Although we note that in any personalized system and sentiment analysis work, ethical considerations should be taken into account (Mohammad, 2022).

Acknowledgments

We thank Ning Shi and David Basil for their advice on our systems, methodology, and writing.

This research was supported by the Alberta Machine Intelligence Institute (Amii) and the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

Joost Broekens, Bernhard Hilpert, Suzan Verberne, Kim Baraka, Patrick Gebhard, and Aske Plaat. 2023. Fine-grained affective processing capabilities emerging from large language models. In *11th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Filipe Galvão, Soraia M. Alarcão, and Manuel J. Fonseca. 2021. Predicting exact valence and arousal values from eeg. *Sensors*, 21(10).

Siavash Golkar, Mariel Pettee, Michael Eickenberg, Alberto Bietti, Miles Cranmer, Geraud Krawezik, Francois Lanusse, Michael McCabe, Ruben Ohana, Liam Parker, and 1 others. 2023. xval: a continuous numerical tokenization for scientific language models. *arXiv e-prints*, pages arXiv–2310.

A. Graves and J. Schmidhuber. 2005. Framewise phoneme classification with bidirectional lstm networks. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, volume 4, pages 2047–2052 vol. 4.

Jiehui Jia, Huan Zhang, and Jinhua Liang. 2025. Bridging discrete and continuous: A multimodal strategy for complex emotion detection. In *2025 IEEE 35th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6.

Bryan Lim, Sercan Ö. Arık, Nicolas Loeff, and Tomas Pfister. 2021. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 37(4):1748–1764.

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. 2022. Ethics sheet for automatic emotion recognition and sentiment analysis. *Computational Linguistics*, 48(2):239–278.

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

Rajnish Pandey and Jyoti Prakash Singh. 2023. Bert-lstm model for sarcasm detection in code-mixed social media posts. *Journal of Intelligent Information Systems*, 60(1):235–254.

Md Mostafizer Rahman, Ariful Islam Shiplu, Yutaka Watanobe, and Md Ashad Alam. 2025. Robertabilstm: A context-aware hybrid model for sentiment analysis. *IEEE Transactions on Emerging Topics in Computational Intelligence*, pages 1–18.

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.

Thomas Teixeira, Éric Granger, and Alessandro Lameiras Koerich. 2021. Continuous emotion recognition with spatiotemporal convolutional neural networks. *Applied Sciences*, 11(24).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need. *Preprint*, arXiv:1706.03762.

Zengzhi Wang, Qiming Xie, Yi Feng, Zixiang Ding, Zinong Yang, and Rui Xia. 2024. Is chatgpt a good sentiment analyzer? a preliminary study. *Preprint*, arXiv:2304.04339.

A Hyperparameters

A.1 Training Hyperparameter

Hyperparameter	Value
LM fine-tuned	Frozen
Learning rate	1×10^{-5}
Optimizer	AdamW
Batch size	8
Dropout rate	0.4
LM Hidden sizes	LM 768
BiLSTM Hidden sizes	256
Loss function	L1Loss (MAE)
Random seed	51

Table 3: Training hyperparameters for Subtask 1 (Transformer LM + BiLSTM).

Hyperparameter	Value
LM fine-tuned	Frozen
Learning rate	0.03
Batch size	64
Attention head size	2
Optimizer	AdamW
Batch size	8
Dropout rate	0.1
Hidden size	32
Gradient clip	0.1
Loss function	L1Loss (MAE)
Random seed	Default

Table 4: Training hyperparameters for Subtask 2a (TFT).

Hyperparameter	Value
LM fine-tuned	Frozen
Learning rate	1×10^{-5}
Optimizer	AdamW
Batch size	8
Dropout rate	0.1
LM Hidden sizes	LM 768
BiLSTM Hidden sizes	128
Loss function	L1Loss (MAE)
Random seed	Default

Table 5: Training hyperparameters for Subtask 2b (Transformer LM + BiLSTM).