CuriosAI at SemEval-2026 Task 2: Predicting Emotion using RoBERTa-large model

Fumika Beppu, Hiroki Takushima, Manoj Kumar Aiswariya, Daichi Yamaga, Yuki Shibata, and Takayuki Hori
SoftBank Corp.
fumika.beppu@g.softbank.co.jp

Abstract

This paper proposes a method for predicting continuous emotion dimensions, namely Valence and Arousal, from text by combining affective intermediate training with multi-task learning. The proposed approach consists of two training phases: an intermediate pretraining phase using external emotion datasets, followed by a multi-task learning phase using task-specific data. RoBERTa-large is employed as the backbone model, and independent regression heads are introduced for each subtask. Experimental results show that the proposed method achieves Pearson correlation coefficients of 0.68 for Valence and 0.45 for Arousal on Subtask 1, demonstrating stable performance, particularly in capturing inter-user differences in emotional expression.

1 Introduction

We participated in SemEval-2026 Task 2 (Soni et al., 2026), which focuses on predicting variation in emotional valence and arousal over time from ecological essays.

SemEval 2026 Task 2 focuses on predicting emotional states and their temporal dynamics from text. Unlike conventional sentiment analysis tasks that estimate emotion at a single time point, this task explicitly addresses changes in emotional states over time, making it more challenging and realistic.

In this work, we aim to improve continuous emotion prediction by first equipping the model with general affective knowledge and then jointly learning multiple related tasks. By integrating affective intermediate training with multi-task learning, the proposed method seeks to achieve more robust and stable performance across subtasks.

2 Related Work

Dimensional representations of affect, particularly the valence–arousal framework, provide a continuous alternative to discrete emotion categories and are commonly grounded in the circumplex model of affect (Russell, 1980). In NLP, several resources operationalize continuous emotion representations, including EmoBank (Buechel and Hahn, 2017), the NRC Valence–Arousal–Dominance lexicon (Mohammad, 2018), and large-scale normative ratings such as Warriner et al. (Warriner et al., 2013). Pretrained Transformer encoders such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) have become standard backbones for regression-based emotion prediction.

Our two-phase training strategy is closely related to intermediate-task fine-tuning approaches, often referred to as STILTs (Supplementary Training on Intermediate Labeled-data Tasks) (Phang et al., 2018). Prior work has shown that intermediate supervised training can improve downstream performance and stability, depending on task relatedness and data regime (Pruksachatkun et al., 2020; Chang and Lu, 2021). In contrast to typical STILTs settings that use natural language inference or classification tasks, we employ affective intermediate supervision by performing regression on Valence and Arousal using EmoBank, explicitly adapting the encoder to a continuous affective space before task-specific training.

Multi-task learning (MTL) is another established paradigm for leveraging related supervision through shared representations (Caruana, 1997). Recent analyses compare intermediate fine-tuning and MTL and discuss when each transfer strategy is preferable (Weller et al., 2022). In affective shared-task settings, including SemEval and WASSA tracks (Strapparava and Mihalcea, 2007; Mohammad et al., 2018; Tafreshi et al., 2021), systems typically rely on pretrained encoders combined with careful development-set checkpoint selection. Following this convention, our approach integrates affective intermediate training with multi-task regression for SemEval 2026 Task 2.

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 121–126

3 Proposed Method

3.1 Overall Architecture

Figure 1 illustrates the overall architecture of the proposed two-stage framework. In the first stage, RoBERTa-large is further pre-trained on regression tasks for Valence and Arousal in order to adapt it to a continuous affective space. In the second stage, the encoder obtained from the first stage is used as the initialization, and three regression tasks—Subtask 1, Subtask 2a, and Subtask 2b—are jointly trained in a multi-task learning setting.

The purpose of this two-stage design is to first enable the general-purpose language model to learn the geometric structure of the affective space, and then to adapt it to more task-specific emotion change estimation.

3.2 Pre-Fine-Tuning with EmoBank

In Phase 1, regression for Valence and Arousal was performed using the EmoBank dataset and RoBERTa-large model. The output consisted of a two-dimensional continuous vector representing Valence and Arousal, respectively.

Mean Squared Error (MSELoss) was used as the loss function. The final model was selected based on the lowest mean MAE (mae_mean) of Valence and Arousal on the validation set. An early stopping mechanism was introduced to prevent overfitting.

Through this phase, the model acquires the ability to map textual representations into a continuous affective space.

3.3 Multi-Task Learning

In Phase 2, the encoder obtained in Phase 1 (roberta_emotion_finetuned_es) was used for initialization, and three subtasks were trained jointly.

In Subtask 1, the model takes text and user ID as input and regresses Valence and Arousal. In Subtask 2a, the model predicts Valence and Arousal corresponding to state change from the text and user ID. In Subtask 2b, the input consists of multi-segment text along with user ID, segment-level features, and essay-level features, and the model regresses disposition change.

The user ID is learned as a 32-dimensional embedding vector and integrated with the text encoder output. In Subtask 2b, segment-level features (32 dimensions) and essay-level features (64 dimensions) are further incorporated, and contextual integration is performed through Transformer layers (4 layers, 8 attention heads, feed-forward dimension of 3072). The dropout rate was set to 0.05.

The loss function is defined as the weighted sum of the MSELoss for each subtask:

$$L = \lambda_1 L_{\text{sub1}} + \lambda_{2a} L_{2a} + \lambda_{2b} L_{2b}$$

In this study, all weighting coefficients λ were set to 1.0.

4 Experiments

4.1 Datasets

Our training procedure consists of two phases, each relying on different datasets.

For the Pre-Fine-Tuning phase, we used the publicly available EmoBank dataset. EmoBank is a large-scale corpus annotated with continuous emotion ratings based on the Valence–Arousal–Dominance (VAD) framework. The dataset contains 10,062 English sentences drawn from diverse textual genres, including news headlines, fiction, blogs, and essays. Each sentence is annotated with real-valued scores for Valence, Arousal, and Dominance on a continuous scale.

In this study, we utilized the Valence and Arousal annotations to train the model as a regression task in a two-dimensional affective space. The original train/dev/test splits provided by the dataset were followed. This intermediate training phase enables the language model to adapt to continuous affective representations before being exposed to the task-specific data.

In the second Multi-Task Learning phase, we used the official dataset provided for SemEval 2026 Task 2. The dataset consists of user-generated textual data annotated for continuous emotional states and emotion change across multiple subtasks (Subtask 1, Subtask 2a, and Subtask 2b). Following the shared task setup, we trained the model using the released training data and evaluated model checkpoints on the provided development set.

4.2 Implementation Details

All experiments were implemented using PyTorch and the HuggingFace Transformers library. We used roberta-large as the backbone model. The encoder obtained from Phase 1 was used to initialize the multi-task model in Phase 2.

In Phase 1, training was conducted for up to 50 epochs with a batch size of 8 and a maximum sequence length of 128 tokens. We used the AdamW optimizer with a learning rate of 1e-5 and a weight

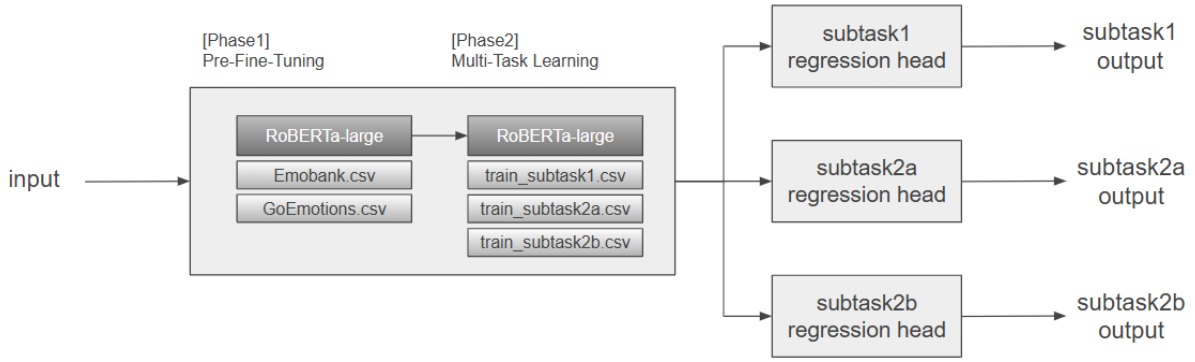

Figure 1: Overview of the proposed two-stage framework. In Phase 1, RoBERTa-large is adapted to the continuous affective space via intermediate regression training on Valence and Arousal. In Phase 2, the encoder is fine-tuned in a multi-task learning setting with three task-specific regression heads for Subtask 1, Subtask 2a, and Subtask 2b.

decay of 0.01. A linear learning rate scheduler with a warmup ratio of 0.08 was applied. Gradient clipping was performed with a maximum norm of 1.0 to ensure stable optimization.

The loss function in Phase 2 was defined as the weighted sum of the Mean Squared Error (MSE) losses for the three subtasks. All weighting coefficients were set to 1.0. The development set ratio was set to 0.2, and model checkpoints were evaluated at the end of each epoch. The random seed was fixed at 42 for reproducibility. Early stopping was applied if no improvement in the model selection metric was observed for 7 consecutive epochs.

For model selection, we used the sum of the mean Pearson correlation coefficients across the three subtasks (sum pearson_mean) as the primary validation metric. The checkpoint that achieved the highest validation sum pearson_mean was selected as the final model.

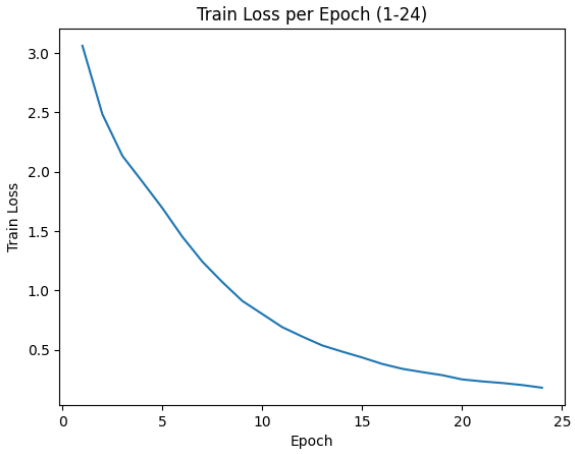

Figure 2: Training loss across 24 epochs. The loss decreases rapidly in the early stage and gradually converges after epoch 15, indicating stable optimization.

4.3 Training Dynamics

To analyze the optimization behavior of our model, we trained it and monitored the training loss at each epoch. Figure 2 illustrates the trajectory of the training loss throughout the training process in Phase 2.

As shown in Figure 2, the loss decreases sharply during the early stage of training (epochs 1–5), indicating that the model rapidly captures the underlying patterns of the dataset. This rapid reduction suggests that the model benefits from effective parameter initialization and stable gradient updates.

After approximately epoch 6, the rate of decrease becomes more gradual, reflecting a transition from coarse-grained learning to fine-grained parameter adjustment. From epoch 15 onward, the loss curve shows clear signs of convergence, with only marginal improvements observed in subsequent epochs.

Importantly, no instability, divergence, or oscillatory behavior was observed during training. The loss decreases monotonically across epochs, suggesting that the chosen optimization settings (learning rate, batch size, and optimizer configuration) are well-suited to the task. The smooth convergence pattern also indicates that the model does not suffer from optimization difficulties such as exploding gradients.

4.4 Training Result

To better understand how our model learns across subtasks, we track performance during training by evaluating the model checkpoint at each epoch and computing the mean Pearson correlation for each subtask. Figure 3 shows the evolution of the subtask-wise Pearson mean over 24 epochs in Phase2.

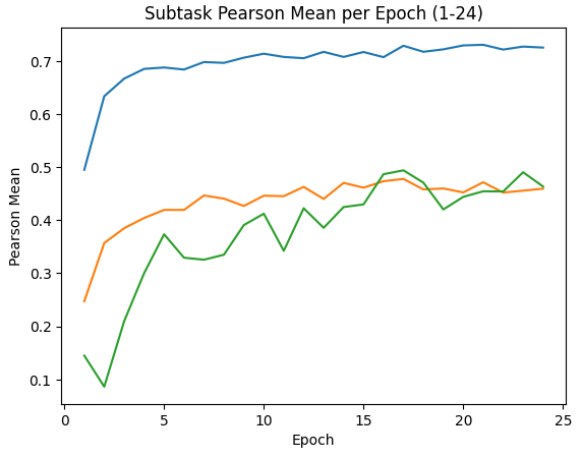

Figure 3: Subtask-wise mean Pearson correlation per epoch (1–24). Subtask 1 converges quickly and stabilizes after 10 epochs, while Subtasks 2 and 3 improve more gradually; Subtask 3 shows larger fluctuations and peaks later, motivating checkpoint selection based on development performance.

Figure 3 indicates that all subtasks benefit from training, but the learning dynamics differ substantially across them. The top curve (Subtask 1) improves very rapidly in the early stage: the Pearson mean rises sharply from around 0.50 at epoch 1 to approximately 0.65 by epoch 3, and then continues to increase more gradually, reaching a plateau around 0.72–0.74 after roughly epochs 10–12. This pattern suggests that the model quickly acquires the key signals required for Subtask 1 and that later epochs mainly provide incremental refinements rather than major gains. The stability of the curve after convergence also implies that the optimization process is well-behaved for this subtask, with minimal epoch-to-epoch variance.

In contrast, the middle curve (Subtask 2) exhibits a slower, more incremental improvement. Performance increases from roughly 0.25 at epoch 1 to around 0.40 by epoch 4–5, followed by a gradual climb toward approximately 0.45–0.47. Compared to Subtask 1, the improvements are smaller and the plateau appears earlier, indicating that Subtask 2 may be intrinsically more challenging or may require additional modeling capacity or task-specific supervision to further improve. Nevertheless, the consistently upward trend suggests that the model continues to extract useful information throughout training without signs of deterioration.

The bottom curve (Subtask 3) shows the most pronounced non-monotonic behavior. After an initially low score (around 0.10–0.15), the metric improves substantially during the first several epochs and continues to increase, but with visible fluctuations across epochs. The curve peaks around epochs 16–17 (approximately 0.48–0.50), followed by a temporary decrease and partial recovery toward the end of training. Such oscillations can occur when the subtask signal is weaker, the effective dataset size is smaller, or the optimization landscape is more sensitive to stochastic updates. Importantly, however, the overall trajectory remains positive, and the model reaches performance comparable to Subtask 2 in later epochs, suggesting that the model eventually learns representations that generalize reasonably well for this subtask.

Overall, the subtask-wise curves provide two practical insights for model selection. First, Subtask 1 converges early and remains stable, so additional epochs beyond the plateau yield limited returns. Second, Subtask 3 peaks later and is more volatile, implying that selecting the best checkpoint purely based on the final epoch may be suboptimal. In our experiments, we therefore select the final submission checkpoint based on the best development-set performance (aggregated across subtasks / or according to the official selection criterion), rather than relying on the last epoch. This strategy is consistent with the observation that different subtasks reach their optimal performance at different training stages.

5 Evaluation Results

Table 1 summarizes the main evaluation metrics for each subtask on the official test set.

For Subtask 1, the model achieved Pearson correlations of 0.6832 for Valence and 0.4508 for Arousal, with MAE values of 0.6219 and 0.4043, respectively. The relatively high correlation and low MAE indicate that the system effectively captures overall emotional tendencies, particularly for Valence prediction.

In Subtask 2a, the correlations decreased to 0.4675 (Valence) and 0.2748 (Arousal), accompanied by higher MAE values of 1.2196 and 0.8898.

Table 1: Comparison with official baseline on SemEval 2026 Task 2. Pearson correlation (r, higher is better) and MAE (lower is better).

Model	Subtask	Valence r	Valence MAE	Arousal r	Arousal MAE
Baseline	Subtask 1	0.557	0.743	0.299	0.459
	Subtask 2a	0.290	1.294	0.199	0.744
	Subtask 2b	-0.088	0.438	0.070	0.303
Ours	Subtask 1	**0.6832**	0.6219	**0.4508**	0.4043
	Subtask 2a	**0.4675**	1.2196	**0.2748**	0.8898
	Subtask 2b	-0.1610	0.7948	0.0109	0.4290

Compared to Subtask 1, this performance drop suggests that estimating state change is more challenging than direct emotion regression, as it requires modeling nuanced variations rather than static affective states.

Subtask 2b showed a negative correlation for Valence (-0.1610) and near-zero correlation for Arousal (0.0109), with MAE values of 0.7948 and 0.4290. Although the absolute errors are not extremely large, the low and negative correlations indicate that the model fails to reliably capture the direction of disposition change. This result highlights the substantial difficulty of predicting longer-term or dispositional affective shifts.

Overall, the results are consistent with the trends observed in the released baselines. Subtask 1 remains the strongest component of the system, while Subtask 2b appears substantially more challenging, particularly for Valence prediction.

6 Discussion

The pre-fine-tuning on continuous emotion regression using EmoBank appears to have been effective in adapting the model to the affective space. In particular, the high correlation observed in Subtask 1 likely reflects the benefit of this pre-training phase.

Furthermore, the multi-task learning framework, in which the three subtasks were trained jointly, is presumed to have facilitated representation sharing across related tasks, leading to more stable learning. However, given the difficulty of state change estimation, future work may benefit from incorporating models that explicitly handle temporal structure or from introducing contrastive learning approaches.

7 Conclusion

In this study, we proposed a two-stage training framework that combines emotion regression pre-fine-tuning using EmoBank with multi-task learning across three subtasks. The proposed model, based on RoBERTa-large, demonstrated effectiveness in continuous emotion estimation and emotion change estimation tasks.

Future work will focus on further improving dynamic emotion estimation performance through the incorporation of temporal modeling and structured user information.

Limitations

This study has several limitations. First, the proposed two-stage framework was evaluated only on the SemEval 2026 Task 2 dataset, and its generalizability to other affective computing benchmarks remains unverified. Second, while the intermediate regression training encourages adaptation to the Valence–Arousal space, we did not conduct ablation studies to isolate the precise contribution of each component. Third, the performance on Subtask 2b indicates that the model struggles to capture long-term or dispositional affective changes, suggesting that additional mechanisms for modeling temporal dynamics may be necessary. Future work will investigate more robust approaches to affective change modeling and evaluate the framework across diverse datasets.

Acknowledgments

We thank the organizers of SemEval 2026 Task 2 for providing the dataset and evaluation framework. We also appreciate the valuable discussions within our research team that contributed to this work.

References

Sven Buechel and Udo Hahn. 2017. EmoBank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 578–585, Valencia, Spain. Association for Computational Linguistics.

Rich Caruana. 1997. Multitask learning. *Machine Learning*, 28(1):41–75.

Ting-Yun Chang and Chi-Jen Lu. 2021. Rethinking why intermediate-task fine-tuning works. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 706–713, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Saif Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 174–184, Melbourne, Australia. Association for Computational Linguistics.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. SemEval-2018 task 1: Affect in tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 1–17, New Orleans, Louisiana. Association for Computational Linguistics.

Jason Phang, Thibault Févry, and Samuel R. Bowman. 2018. Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks. *arXiv preprint arXiv:1811.01088*.

Yada Pruksachatkun, Jason Phang, Haokun Liu, Phu Mon Htut, Xiaoyi Zhang, Richard Yuanzhe Pang, Clara Vania, Katharina Kann, and Samuel R. Bowman. 2020. Intermediate-task transfer learning with pretrained language models: When and why does it work? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5231–5247, Online. 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.

Carlo Strapparava and Rada Mihalcea. 2007. SemEval-2007 task 14: Affective text. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 70–74, Prague, Czech Republic. Association for Computational Linguistics.

Shabnam Tafreshi, Orphee De Clercq, Valentin Barriere, Sven Buechel, João Sedoc, and Alexandra Balahur. 2021. WASSA 2021 shared task: Predicting empathy and emotion in reaction to news stories. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 92–104, Online. Association for Computational Linguistics.

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

Orion Weller, Kevin Seppi, and Matt Gardner. 2022. When to use multi-task learning vs intermediate fine-tuning for pre-trained encoder transfer learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 272–282, Dublin, Ireland. Association for Computational Linguistics.