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

Somdev Ganguli, Vibhan Dutta, Romit Datta, Amit Barman and Sudip Kumar Naskar
Department of Computer Science and Engineering
Jadavpur University, Kolkata, India
{somdevg, vibhand, romitd}.cse.ug@jadavpuruniversity.in
{amitbarman811, sudip.naskar}@gmail.com

Abstract

Tracking emotional dynamics like valence and arousal is critical for understanding users' affective baselines in ecological text. However, encoder models often struggle to distinguish stable user traits from dynamic shifts, leading to poor generalization. This paper presents LexMachina, our system for SemEval-2026 Task 2, addressing "domain shift" and "regression to the mean." LexMachina utilizes a DeBERTa-v3-Base backbone with a bifurcated strategy: post-hoc Isotonic Regression for valence calibration and a Domain Adversarial Neural Network (DANN) to mitigate user-bias in arousal. LexMachina achieved composite scores of $r = 0.645$ (Valence) and $r = 0.434$ (Arousal), demonstrating that adversarial disentanglement effectively captures nuances in longitudinal affective data.

1 Introduction

Affective quantification traditionally relies on the Valence-Arousal space (Russell, 1980). In this work, we address SemEval-2026 Task 2 (Subtask 1) (Soni et al., 2026) by predicting continuous valence-arousal trajectories from longitudinal essays. A primary challenge in this domain is annotator idiosyncrasy and label noise. We propose a multi-stage pipeline built on DeBERTa-v3-base (He et al., 2020) embeddings. To ensure signal integrity, we implemented an Out-of-Fold (OOF) Data Sanitation Protocol, purging 9% of the training data exhibiting extreme deviation ($\delta > 1.5$) from model consensus. Our modeling approach is bifurcated: Valence is refined via post-hoc Isotonic Regression to fix calibration drift, while Arousal is modeled through a Domain Adversarial Neural Network (DANN) to disentangle user-specific biases from universal emotional signals.

2 Background and Related Work

2.1 Related Work

Transformer-based architectures have set benchmarks in mapping text to continuous Valence-Arousal space (Mendes and Martins, 2023). Recent innovations include integrating Valence-Arousal-Dominance (VAD) features into attention mechanisms (Moreno et al., 2025) and utilizing ordinal classification to minimize perceptual errors (Mitsios et al., 2024). For low-resource or domain-specific settings, strategies like Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) and multi-teacher pseudo-labeling have proven effective for cross-domain adaptation (Jian et al., 2025). Our work builds on these prior studies by applying adversarial disentanglement to longitudinal user data.

2.2 Task Definition and Evaluation

The task requires mapping a user's chronological stream $S = \{e_1, e_2, \ldots, e_m\}$ to bivariate outputs (v_i, a_i) where $v \in [-2, 2]$ and $a \in [0, 2]$. Evaluation relies on a Composite Correlation (Soni et al., 2026) (r_{comp}), aggregating inter-user traits ($r_{between}$) and temporal fluctuations (r_{within}) via the Fisher-Z transformation (Fisher, 1915):

$$r_{comp} = \tanh\left(\frac{1}{2}\Big(\operatorname{arctanh}(r_{between}) + \operatorname{arctanh}(r_{within})\Big)\right) \quad (1)$$

2.3 Dataset Description

The dataset consists of longitudinal essays and feeling words. Our partition strategy highlights the stylistic gap addressed by our adversarial approach (cf. Table 1).

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 553–560

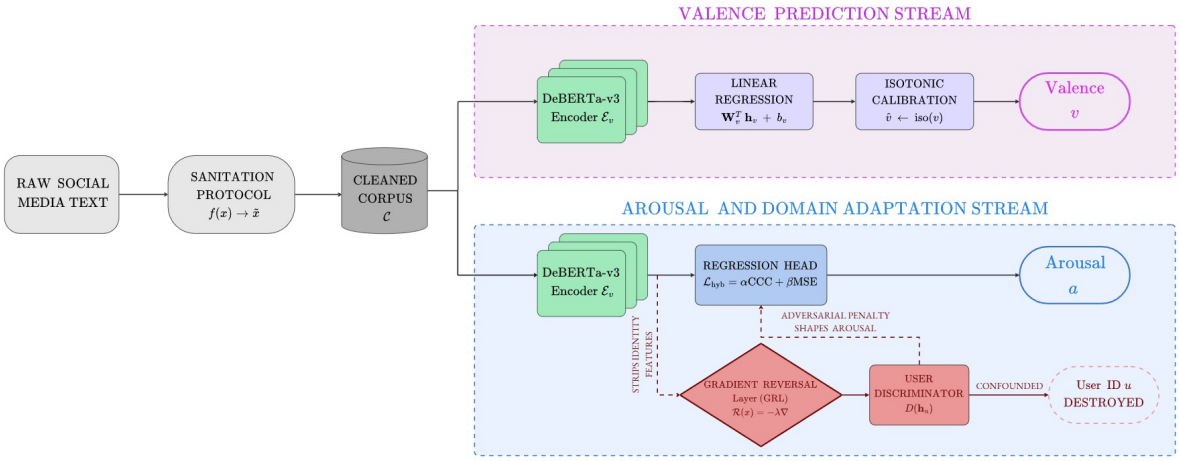

Figure 1: System Architecture. The pipeline splits into two parallel streams after the DeBERTa-v3 backbone. The Valence stream (top) utilizes a standard regression head followed by Isotonic Calibration to map predictions to the target distribution. The Arousal stream (bottom) employs a Domain-Adversarial Neural Network (DANN) with a Gradient Reversal Layer (GRL) to enforce user-invariant feature learning, mitigating the generalization gap for unseen subjects.

Partition	User Type	# Users	# Texts	Val ($\mu \pm \sigma$)	Aro ($\mu \pm \sigma$)
Train	Seen	137	2,487	0.22 ± 1.29	0.75 ± 0.76
Dev	Seen	92	277	0.22 ± 1.30	0.75 ± 0.74
Test	**Unseen**	91	1,737	*Hidden*	*Hidden*

Table 1: Dataset Statistics.

3 System Overview

We propose a dual-stream architecture designed to address the distinct generalization challenges of the 'Seen' versus 'Unseen' user cohorts. While both streams utilize a shared linguistic backbone, they employ divergent optimization strategies: a direct regression approach with post-hoc calibration for Valence, and a Domain-Adversarial framework for Arousal.

3.1 Data Sanitation Protocol

Preliminary error analysis revealed a subset of samples with noisy labels where ground truth annotation sharply contradicted the semantic polarity of the text. To address this, we implemented an automated sanitation strategy grounded in the principle of pruning training sets to remove instances that impede representation learning (Angelova et al., 2005).

We utilized a 3-Fold OOF Cross-Validation strategy to identify these outliers. Samples with an absolute error exceeding $\delta > 1.5$ were identified as poisonous and discarded, removing approximately 9% of the training data. After the official test labels were released prior to paper submission, we also ran post-hoc forensic analysis on the test set. Table 2 illustrates examples of extreme test-label noise observed in that post-release analysis.

Test Set Essay Excerpt	True Val.	Pred Val.	δ
"I have had one shitty night... I hate life sometimes... I'm very sad and angry."	2.00	-1.39	3.39
"Tired , Annoyed , Furious , Boredom , Upset"	2.00	-1.39	3.39
"Tired , Depress , Sad , Conflicted , Trapped"	2.00	-1.39	3.39

Table 2: Gold Standard Annotation Errors in the Released Official Test Labels. Post-release forensic error analysis ($\delta > 3.0$) reveals instances where our model correctly predicted extreme negative Valence, but was heavily penalized by apparently inverted ground-truth labels.

To check whether the purged samples came from specific users, we compared observed per-user purge counts with expected counts under a uniform error-rate assumption. The results are summarized in Table 3. The chi-squared test was not statistically significant, and correlation tests also showed no meaningful link between purge rate and user-level affective profile. This suggests that the

Algorithm 1 Adversarial Training Loop for Arousal Stream

Require: Sanitized Corpus $\mathcal{S}_{clean}$, Max Epochs E, Total Steps T
Require: Feature Extractor G_f, Regressor G_y, Discriminator G_d
Initialize network parameters $\theta_f, \theta_y, \theta_d$
$t \leftarrow 0$ ▷ Current step counter
for $epoch = 1$ **to** E **do**
  for each batch (X, Y, U) **in** $\mathcal{S}_{clean}$ **do**
    $p \leftarrow \frac{t}{T}$ ▷ Training progress $p \in [0, 1]$
    $\lambda \leftarrow \frac{2}{1+\exp(-10 \cdot p)} - 1$ ▷ Dynamic GRL schedule
    Forward Pass:
    $H \leftarrow G_f(X; \theta_f)$ ▷ Extract latent features
    $\hat{Y} \leftarrow G_y(H; \theta_y)$ ▷ Predict Arousal
    $\hat{U} \leftarrow G_d(H; \theta_d)$ ▷ Predict User ID
    Compute Losses:
    $\mathcal{L}_{task} \leftarrow \alpha \text{MSE}(Y, \hat{Y}) + (1 - \alpha)\text{CCC}(Y, \hat{Y})$
    $\mathcal{L}_{domain} \leftarrow \text{CrossEntropy}(U, \hat{U})$
    Backward Pass (with Gradient Reversal):
    $\nabla_{\theta_y} \leftarrow \frac{\partial \mathcal{L}_{task}}{\partial \theta_y}$
    $\nabla_{\theta_d} \leftarrow \frac{\partial \mathcal{L}_{domain}}{\partial \theta_d}$
    $\nabla_{\theta_f} \leftarrow \frac{\partial \mathcal{L}_{task}}{\partial \theta_f} - \lambda \frac{\partial \mathcal{L}_{domain}}{\partial \theta_f}$ ▷ Invert domain gradient
    Update Parameters:
    $\theta_y \leftarrow \theta_y - \eta \nabla_{\theta_y}$
    $\theta_d \leftarrow \theta_d - \eta \nabla_{\theta_d}$
    $\theta_f \leftarrow \theta_f - \eta \nabla_{\theta_f}$
    $t \leftarrow t + 1$
  end for
end for

removed samples reflect random annotation noise, not consistent bias from particular users.

The pruning threshold $\delta = 1.5$ was set before model training, based on label-scale structure (Valence $\in \{-2, -1, 0, 1, 2\}$; Arousal $\in \{0, 1, 2\}$). Since adjacent labels are one unit apart, $\delta = 1.5$ separates ordinary one-step disagreement from larger polarity inversions. We did not tune δ on downstream performance to avoid information leakage between sanitation and training.

Test	Statistic	Interpretation
Chi-squared goodness-of-fit	$p = 0.076$	No significant user-level concentration
Purge rate vs. mean valence	$r = 0.062,\ p = 0.47$	No significant correlation
Purge rate vs. mean arousal	$r = -0.163,\ p = 0.06$	No significant correlation
Purge rate vs. valence std. dev.	$r = -0.136,\ p = 0.11$	No significant correlation

Table 3: Evidence that purged samples are not user-specific.

3.2 Linguistic Backbone

For feature extraction, we utilize DeBERTa-v3-base, grounded in the foundational bidirectional Transformer attention mechanism (Vaswani et al., 2017). DeBERTa-v3 was selected specifically for its disentangled attention mechanism and superior pre-training via ELECTRA-style discriminative objectives (Clark et al., 2020), which provide a highly efficient weight initialization for downstream affective tasks.

We extract the contextualized word embeddings from the final hidden layer and aggregate them using Max-Pooling (alongside the `[CLS]` token) to capture the most salient features of the input sequence x_i, where $h_i \in \mathbb{R}^{768}$ serves as the latent affective representation.

$$h_i = \text{MaxPool}(\text{DeBERTa}(x_i)) \quad (2)$$

3.3 Valence Stream: Regression & Calibration

The Valence stream projects the latent vector h_i through a standard regression head followed by a linear layer. Modern deep networks are frequently miscalibrated, exhibiting a conservative bias where predictions cluster around the mean (Guo et al., 2017).

To address this, we applied **Post-Hoc Isotonic Calibration** (Zadrozny and Elkan, 2002). We also evaluated parametric alternatives like Platt Scaling (Platt, 1999) to fix the calibration drift, but opted for the non-parametric route to better map the cumulative distribution function (CDF) of the validation set:

$$\hat{v}_{final} = f_{iso}(\hat{v}_{raw}) \tag{3}$$

3.4 Arousal Stream: Domain-Adversarial Neural Network (DANN)

For Arousal, we identified a critical User Generalization Gap. In the official test split, **60.2% of the total text volume** (1,045 out of 1,737 texts) was written by users not present in the training set. This heavy skew toward previously unseen users meant that standard supervision models, which often overfit to the writing styles of "Seen" users, would fail to generalize.

To mitigate this, we implemented a **DANN** (Ajakan et al., 2014; Ganin et al., 2016). The architecture consists of three components: (1) The shared Feature Extractor (G_f), (2) A Label Predictor (G_y) using Hybrid Loss, and (3) A User Discriminator (G_d) attempting to guess the User ID from h_i. By introducing a Gradient Reversal Layer (GRL), we create a 'min-max game' in the latent space (Tzeng et al., 2017), forcing the model to strip stylistic identity markers and learn user-invariant representations of emotional intensity. The complete step-by-step optimization procedure for this adversarial min-max game is formalized in Algorithm 1.

4 Experimental Setup

4.1 Implementation Details

The source code for LexMachina is available on GitHub.[1] All models were implemented using PyTorch and the HuggingFace library, with training conducted on a single NVIDIA Tesla T4 GPU. Inputs were tokenized via `microsoft/deberta-v3-base` and truncated to a maximum length of 128 tokens to accommodate the average essay length (cf. Table 1).

We utilized the AdamW optimizer with a learning rate of 2×10^{-5} and weight decay of 0.01. To maintain backbone stability, we applied a linear learning rate scheduler with a 10% warmup phase. The Arousal DANN was trained for 4 epochs—an early-stopping threshold empirically determined to allow the adversarial User Discriminator sufficient time to converge, while strictly preventing the shared feature extractor from overfitting to the 'seen' user domain. The independent Valence regressor achieved optimal convergence in 3 epochs. Both streams utilized a consistent batch size of 16. For the DANN, the hybrid task loss was balanced with $\alpha = 0.5$, and the adversarial weight λ was dynamically scaled from 0 to 1 as detailed in Algorithm 1.

We used a constrained tuning strategy to keep the system reproducible. Most hyperparameters were fixed from standard practice or hardware limits, and only two of the hyperparameters were tuned on a validation split. Table 4 details our system hyperparameters. The final setup remained stable across runs, and the full 3-seed pipeline took about 35–75 minutes on one Tesla T4 GPU.

Component	Values / Setting	Selection Basis
Learning rate	$\{1e^{-5}, 2e^{-5}, 5e^{-5}\} \rightarrow 2e^{-5}$	Tuned on validation split
Hybrid-loss balance (α)	$\{0.3, 0.5, 0.7\} \rightarrow 0.5$	Tuned on validation split
Optimizer	AdamW, weight decay 0.01	Standard transformer default
GRL schedule	$\lambda = \frac{2}{1+\exp(-10p)} - 1$	Canonical DANN schedule
Batch size	16	Hardware limit (Tesla T4)
Max length	128 tokens	Dataset length profile + efficiency
Arousal / Valence epochs	4/3	Validation convergence trend

Table 4: Hyperparameter selection summary.

4.2 Ensemble Strategy

To reduce output variance and improve generalization stability (Dietterich, 2000), our final submission utilized a 3-seed ensemble (Seeds 42, 43, 44). Models were trained from scratch for each seed, and the final predictions represent the arithmetic mean of these independent runs.

5 Results and Analysis

5.1 Official Results and Cohort Analysis

Table 5 presents the performance of the LexMachina architecture across the official evaluation criteria. Our system achieved an overall composite correlation (r_c) of **0.645** for Valence and **0.434** for Arousal.

For **Valence**, performance remained remarkably stable across both cohorts ($r_c = 0.636$ for Seen, 0.669 for Unseen). This consistency validates our use of DeBERTa paired with post-hoc Isotonic Calibration, which successfully mapped texts to the continuous space without overfitting to specific user histories.

[1] https://github.com/NoviceDev92/SemEval-2026-Task2

Evaluation Slice	Valence			Arousal		
	r_c	r_b	r_w	r_c	r_b	r_w
Overall System	**0.645**	**0.712**	**0.567**	**0.434**	**0.461**	**0.406**
User Generalization Breakdown						
Seen Users	0.636	0.718	0.537	0.343	0.323	0.363
Unseen Users	0.669	0.734	0.593	**0.574**	**0.681**	**0.443**
Modality Breakdown						
Words Only	0.655	0.730	0.563	0.572	0.631	0.507
Essay Only	0.627	0.665	0.586	0.307	0.315	0.298

Table 5: Official Subtask 1 Results. Performance broken down by user cohort and input modality. Metrics are Pearson correlations: Composite (r_c), Between-User (r_b), and Within-User (r_w).

For **Arousal**, the cohort breakdown reveals a striking validation of our Domain-Adversarial Neural Network (DANN). Designed to explicitly optimize for zero-shot generalization by stripping user identity markers, the DANN yielded an exceptional Unseen Arousal score of $r_c = 0.574$. However, this adversarial disentanglement introduced a stark generalization trade-off. By penalizing the encoding of user-specific traits, the model ignored the predictive historical baselines of known individuals, causing the Seen Arousal score to drop to $r_c = 0.343$.

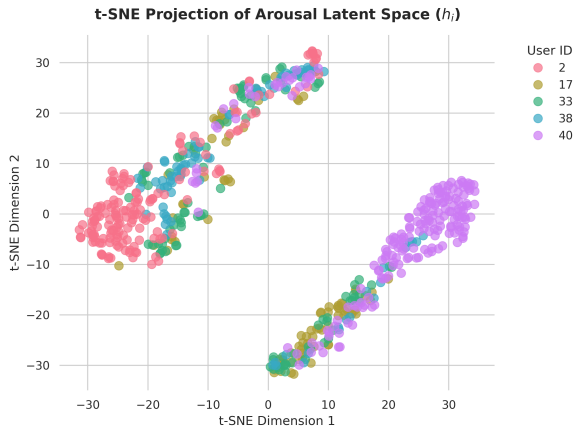

Figure 2: t-SNE Visualization. A two-dimensional projection of the Arousal stream's latent space (h_i). The high degree of overlap among different user clusters visually confirms that the DANN successfully forced the network to learn user-agnostic affective representations.

To visually confirm this adversarial effect, we present a t-SNE projection of the Arousal stream's latent representations (cf. Figure 2). The plot illustrates a highly overlapping, user-agnostic feature space, confirming that the Gradient Reversal Layer successfully stripped stylistic identity markers from the embeddings.

5.2 Modality and Ablation Analysis

Modality analysis showed a clear pattern; Valence remained stable across both input types, while Arousal was much stronger in feeling words than in essays.

Ablation results are shown in Table 6. For Valence, removing OOF sanitation caused the largest drop, showing that label-noise filtering is essential. Removing isotonic calibration also reduced performance, confirming that post-hoc calibration helps align predictions on unseen data. For Arousal, disabling DANN ($\lambda = 0$) caused a major generalization failure: both unseen-user correlation and overall composite score dropped sharply.

Setting	Valence	Arousal
Full system (official)	0.645	0.434
Words only (modality)	0.655	0.572
Essay only (modality)	0.627	0.307
Without OOF sanitation	0.456	–
Without isotonic calibration	0.576	–
Without DANN ($\lambda = 0$)	–	0.260
Without DANN: unseen r_b	–	0.070

Table 6: Modality and ablation results (r_c unless noted).

6 Conclusion

We presented LexMachina, a dual-stream framework for longitudinal affect assessment. Anchored by a disentangled DeBERTa-v3 backbone, our architecture effectively navigates the complexities of dynamic emotional shifts in ecological text. By implementing a rigorous OOF data sanitation protocol, we insulated our representation learning from severe label noise and semantic inversions (Northcutt et al., 2021). Furthermore, our bifurcated optimization strategy yielded highly competitive correlations, particularly in the Arousal dimension. Through adversarial disentanglement (Ganin et al., 2016), we demonstrated that stripping stylistic user biases allows for the extraction of universal, subject-agnostic linguistic markers of emotional intensity from unseen populations.

7 Future Work

Future refinements will focus on architectural efficiency for both streams. For Valence, we will investigate replacing post-hoc calibration with a custom Attention Head optimized to amplify high-polarity

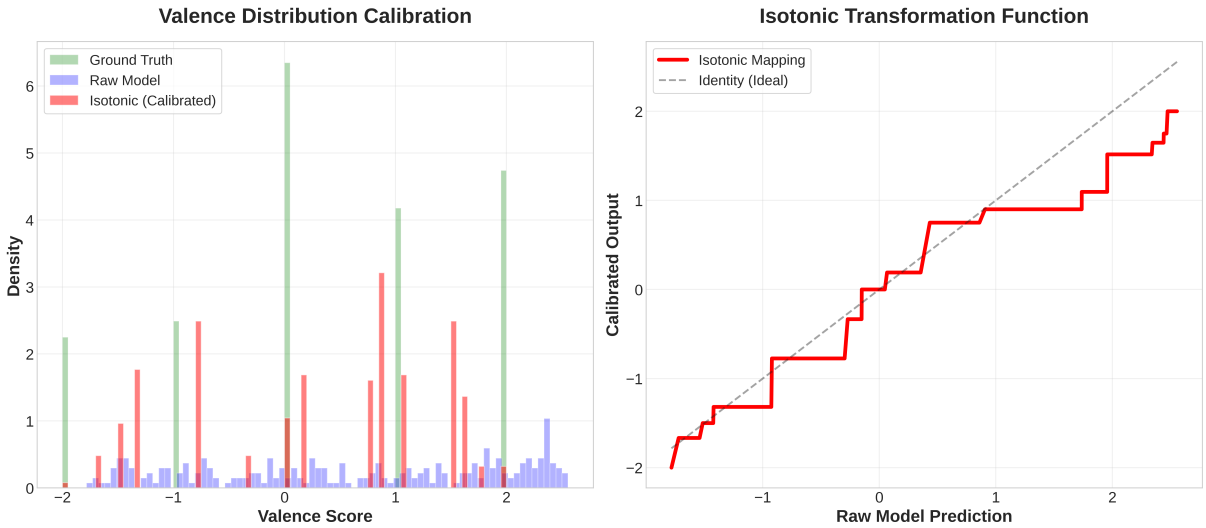

Figure 3: Valence Forensic Analysis. Visualization of the quantization artifacts ("staircase effect") introduced by the Post-Hoc Isotonic Calibration.

affective cues, solving conservative bias during the forward pass without artifact-inducing functions. For Arousal, we intend to transition to Instance-Weighted Domain Adaptation. By adopting a probabilistic perspective of generalization (Wilson and Izmailov, 2020), we propose scaling the adversarial penalty (λ) inversely to a user's historical 'reliability'. This dynamic gating would allow the network to seamlessly interpolate between zero-shot generalization for strangers and personalized fine-tuning for known subjects.

8 Limitations

While our framework exhibits strong performance, it introduces two critical architectural trade-offs.

First, the strict decoupling of user identity enforced by the DANN in the Arousal stream imposes a severe penalty on *seen* subjects. By actively unlearning stylistic idiosyncrasies to accommodate the 60.2% unseen test volume, the model discards the predictive historical baselines of known individuals. This results in a pronounced performance disparity, where unseen users significantly outperform known users on the intensity axis.

Second, our reliance on Post-Hoc Isotonic Calibration to correct the Valence regression head's "regression to the mean" introduced detrimental quantization artifacts (cf. Figure 3). Because isotonic regression maps continuous outputs to a discrete step function, it artificially discretized the continuous output space. This manifested as a visible 'staircase' effect in our error analysis, occasionally triggering severe error spikes when predicting nuanced emotional states that fell between the learned calibration steps.

9 Ethical Considerations

The longitudinal modeling of affective trajectories from personal essays necessitates a rigorous ethical framework. We identify and address the following concerns:

Privacy via Adversarial Disentanglement: A primary risk in longitudinal assessment is the inadvertent extraction of identifiable stylistic signatures. Our **DANN architecture** addresses this through a technical safeguard; by utilizing a Gradient Reversal Layer, we enforce user-invariant feature learning.

This process intentionally strips the model of its ability to associate affective features with specific stylistic identities, ensuring that latent representations capture universal emotional markers rather than identifiable individual traits. This promotes a privacy-by-design approach that prevents the model from being utilized for personal identification.

Linguistic vs. Clinical Boundary: We emphasize that the outputs of the LexMachina framework represent a *linguistic portrayal* of emotion and must not be conflated with clinical diagnosis. The predicted trajectories reflect expressed affect in text and are not a substitute for an individual's internal psychological state or a formal mental health assessment. Such systems should only be utilized under professional psychiatric oversight to avoid the risks of automated misdiagnosis.

Consent and Affective Privacy: We advocate

for the Right to Affective Privacy, asserting that emotional tracking should never be utilized for unauthorized surveillance or profiling by third parties. The deployment of affective modeling must be contingent upon the explicit, informed consent of the individual, ensuring that the technology is utilized exclusively for the user's well-being and not for external monitoring or evaluation.

Acknowledgments

We sincerely thank the anonymous reviewers for their insightful feedback, which significantly improved the final version of this manuscript. We also extend our deepest gratitude to the shared task organizers for their hard work in curating the dataset and facilitating this evaluation.

References

Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, and Mario Marchand. 2014. Domain-adversarial neural networks. *Computing Research Repository, arXiv:1412.4446*.

Anelia Angelova, Yaser Abu-Mostafa, and Pietro Perona. 2005. Pruning training sets for learning of object categories. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 494–501. IEEE.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*. ArXiv:2003.10555.

Thomas G. Dietterich. 2000. Ensemble methods in machine learning. In *Multiple Classifier Systems*, pages 1–15, Berlin, Heidelberg. Springer Berlin Heidelberg.

R. A. Fisher. 1915. Frequency distribution of the values of the correlation coefficients in samples from an indefinitely large population. *Biometrika*, 10(4):507–521.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 1321–1330. JMLR.org.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. DeBERTa: Decoding-enhanced BERT with disentangled attention. *CoRR*, abs/2006.03654. ArXiv:2006.03654.

Yi-Min Jian, An Yu Hsiao, and Shih-Hung Wu. 2025. CYUT-NLP at ROCLING-2025 shared task: Valence–arousal prediction in physicians' texts using BERT, RAG, and multi-teacher pseudo-labeling. In *Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025)*, pages 381–389, National Taiwan University, Taipei City, Taiwan. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Gonçalo Azevedo Mendes and Bruno Martins. 2023. Quantifying valence and arousal in text with multilingual pre-trained transformers. In *Advances in Information Retrieval. ECIR 2023*, volume 13980 of *Lecture Notes in Computer Science*, pages 84–100. Springer, Cham.

Michail Mitsios, Georgios Vamvoukakis, Georgia Maniati, Nikolaos Ellinas, Georgios Dimitriou, Konstantinos Markopoulos, Panos Kakoulidis, Alexandra Vioni, Myrsini Christidou, Junkwang Oh, Aimilios Chalamandaris, and Pirros Tsiakoulis. 2024. Improved text emotion prediction using combined valence and arousal ordinal classification. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 808–813, Mexico City, Mexico. Association for Computational Linguistics.

Melissa Moreno, Juan Martinez-Santos, and Edwin Puertas. 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.

Curtis G Northcutt, Anish Athalye, and Jonas Mueller. 2021. Pervasive label errors in test sets destabilize machine learning benchmarks. *arXiv preprint arXiv:2103.14749*. ArXiv:2103.14749.

John C. Platt. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*, 10(3):61–74.

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.

Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. 2017. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2962–2971.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

Andrew G Wilson and Pavel Izmailov. 2020. Bayesian deep learning and a probabilistic perspective of generalization. *Advances in Neural Information Processing Systems*, 33:4697–4708.

Bianca Zadrozny and Charles Elkan. 2002. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 694–699. ACM.