AI4PC-Howard University and BisonAI4PC at SemEval-2026 Task 2: Fine-Tuning DistilBERT, DeBERTa and ModernBERT for Valence–Arousal Prediction and Change Estimation

Araj Shah and Utsav Shah and Saurav K. Aryal*
Howard University
Washington, DC, USA
araj.shah@bison.howard.edu
utsav.shah@bison.howard.edu
saurav.aryal@howard.edu

Abstract

We present a merged system description for two participating teams, AI4PC-Howard University and BisonAI4PC, on longitudinal valence–arousal (VA) prediction in the SemEval-2026 Task 2 text corpus. Using only the official data, we enforce user-disjoint splits to prevent leakage and evaluate three settings: essay-level VA state estimation, short-horizon VA change forecasting, and long-horizon disposition change prediction. Our best systems use DistilBERT for essay-level regression, ModernBERT-based history modeling with a GRU and a blended previous-delta baseline for short-horizon change, and pooled DeBERTa history embeddings with a compact MLP for disposition change. Reporting the best-performing official submission for each subtask, we achieve $r_{\text{comp}} = 0.665/0.468$ (valence/arousal) for Subtask 1, $r = 0.597/0.413$ for Subtask 2A, and $r = 0.046/0.348$ for Subtask 2B.

1 Introduction

Longitudinal VA modeling from text is fundamentally constrained by heterogeneity in affective expression and by the ease of user-level leakage. When training and evaluation share users, models can exploit idiosyncratic lexical and stylistic cues instead of learning signals that generalize, leading to inflated performance that does not transfer to unseen writers.

SemEval-2026 Task 2 provides a benchmark for leakage-safe longitudinal VA modeling over essays written by U.S. service-industry workers (Soni et al., 2026). We address the task with three aligned prediction views: (i) per-essay VA state estimation, (ii) short-horizon user-level VA change forecasting from recent history, and (iii) long-horizon disposition change prediction from aggregated histories.

Our approach follows two principles: (1) strict user-disjoint evaluation with inference-time-safe feature construction, and (2) compact, transparent pipelines that are easy to reproduce. In this paper, we merge two participating teams (AI4PC-Howard University and BisonAI4PC), hereafter Team 1 and Team 2, and report both teams' official submissions and leaderboard results across Subtasks 1/2A/2B. Across the two official submissions, our best-performing systems use a DistilBERT regressor for essay-level VA, ModernBERT-based representations with a GRU and a blended previous-delta baseline for short-horizon change, and pooled DeBERTa history embeddings augmented with summary features for disposition change. Our contributions are as follows:

- Leakage-controlled, user-disjoint evaluation for reproducible longitudinal VA modeling.
- Lightweight systems for Subtasks 1/2A/2B with explicit, reproducible inference pipelines.
- Official evaluation results from both teams, with analysis of how backbone choice and temporal aggregation relate to performance.

2 Background

2.1 Valence–Arousal Modeling in Longitudinal Text

Valence–arousal (VA) represents affect in a continuous two-dimensional space (Russell, 1980; Aryal and Adhikari, 2023; Aryal et al., 2023a). In longitudinal text, the goal is to model both current state and its evolution within individuals over time, which yields three common prediction views: (i) per-essay VA state estimation, (ii) short-horizon VA change forecasting from recent history, and (iii) longer-horizon disposition change from aggregated user trajectories.

*Corresponding author.

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 617–623

2.2 SemEval Task Setting, Data, and Evaluation

SemEval-2026 Task 2 benchmarks longitudinal VA modeling on essays written over time by U.S. service-industry workers, with timestamps and continuous VA annotations and user-level change targets (Soni et al., 2026). Because users contribute multiple texts, row-level splits can leak user-specific cues across training and evaluation; we therefore adopt user-disjoint splits throughout. The official metrics are correlation-based: Subtask 1 reports composite Pearson correlation (and components) for valence/arousal, while Subtasks 2A/2B report Pearson correlation for delta targets, emphasizing preservation of relative variation over absolute error.

2.3 Related Work

Transformer encoders are standard backbones for affective regression from text (Vaswani et al., 2017; Devlin et al., 2019; Prioleau and Aryal, 2023a; Aryal et al., 2023a; Prioleau and Aryal, 2023a; Aryal et al., 2023a; Ngueajio et al., 2025; Aryal et al., 2023b), including distilled variants such as DistilBERT (Sanh et al., 2019). We also consider stronger encoder families (DeBERTa (He et al., 2021b,a) and ModernBERT (Warner et al., 2024)) and implement fine-tuning and inference using the Transformers framework (Wolf et al., 2020). We use these encoder families as task-specific representation choices rather than assuming a single backbone is uniformly optimal across state estimation, short-horizon forecasting, and disposition-change prediction. For temporal modeling, common approaches include sequence models such as GRUs (Cho et al., 2014; Aryal et al., 2022; Prioleau and Aryal, 2023b) and simple history-based baselines for stability and sanity checks. Recent work across multilingual sentiment analysis, biomedical modeling, and multimodal behavioral prediction further highlights the importance of robust encoders and structured temporal representations for real-world affective computing tasks (Aryal et al., 2023a; Ngueajio et al., 2025; Hagos et al., 2025; Prioleau and Aryal, 2023a; Aryal et al., 2023b).

3 System Overview

We address longitudinal valence–arousal (VA) modeling with lightweight, reproducible systems trained only on the official data and evaluated under user-disjoint splits, consistent with leakage-aware evaluation practices in prior longitudinal NLP and behavioral modeling work (Aryal et al., 2023b; Johnson et al., 2025; Aryal et al., 2023a).

3.1 Problem Decomposition

We decompose VA modeling into:

1. **Essay-level state estimation**: predict continuous valence and arousal from each essay.
2. **Short-horizon change forecasting**: predict near-term user-level deltas $(\Delta v, \Delta a)$ from recent history.
3. **Long-horizon disposition-change prediction**: predict user-level disposition-change deltas from aggregated trajectory information.

3.2 Shared Principles Across Systems

Across all components, we enforce:

- **Official data only**: no external corpora or auxiliary supervision.
- **Leakage-safe evaluation**: user-disjoint validation and model selection.
- **Lightweight heads**: compact regressors on top of pretrained encoders.
- **Transparent inference**: deterministic pipelines that generate the submission files.

3.3 Essay-level VA State Estimation

Team 1 submission. Team 1 fine-tunes a ModernBERT-base encoder with a lightweight MLP head to predict (v, a). To improve optimization stability, we rescale arousal targets from $[0, 2]$ to $[-2, 2]$ during training and invert the scaling back to $[0, 2]$ at inference time before clipping to the task bounds. This aligns arousal with the valence scale during joint regression and improves loss balance during optimization. On the official evaluation, this ModernBERT variant achieved $r_{\text{comp}} = 0.631$ (valence) and 0.462 (arousal).

Team 2 submission. Team 2 fine-tunes a DistilBERT encoder with a lightweight regression head for per-essay VA prediction. Given an input essay, the system tokenizes the text and encodes it with the transformer backbone. The pooled representation is passed through a small MLP head to jointly predict (v, a). This compact DistilBERT setup was chosen for efficient end-to-end fine-tuning in the dense per-essay supervision setting.

3.4 Short-horizon VA Change Forecasting

Team 1 submission. For short-horizon forecasting, Team 1 predicts user-level state-change deltas $(\Delta v, \Delta a)$ at each forecasting anchor using only the user's observed history up to that point. Our primary approach represents each user by (i) a fixed-length window of recent ModernBERT-base text embeddings and (ii) trajectory-derived numeric features (recent deltas, time gaps, and history summaries such as means/trends). A GRU sequence regressor encodes the embedding window and fuses it with the numeric features to output $(\widehat{\Delta v}, \widehat{\Delta a})$. To improve stability, we also fit a simple linear previous-delta baseline and form the final forecast by blending the baseline and GRU predictions with fixed weights. ModernBERT embeddings were used here to provide fixed text representations for recent-history sequence modeling, where the downstream GRU operates over windows of temporally ordered essays.

Team 2 submission. Team 2 uses a per-anchor regressor that predicts $(\Delta v, \Delta a)$ by combining the current essay text with compact history statistics. Specifically, we encode the anchor text with DistilBERT (max length 512) and concatenate the [CLS] embedding with six trajectory features: current (v_t, a_t), history mean $(\bar{v}, \bar{a})$, and per-dimension trend (last–first; zero if insufficient history). An MLP regressor predicts $(\Delta v, \Delta a)$, trained with MSE on user-disjoint splits.

3.5 Long-horizon Disposition-change Prediction

Team 1 submission. For disposition-change prediction, Team 1 represents each user via pooled DeBERTa-v3-base embeddings computed over their history. We combine the pooled text representation with normalized summary features derived from the user's trajectory and apply a compact feed-forward MLP regressor to predict disposition-change deltas. DeBERTa was used in this setting to provide stronger pooled user-history representations for longer-horizon aggregation.

Team 2 submission. Team 2 predicts disposition-change deltas with a sequence model that encodes each text using DistilBERT and feeds the ordered [CLS] sequence into a biLSTM (Hochreiter and Schmidhuber, 1997; Schuster and Paliwal, 1997), concatenated with observed-segment VA summary statistics, followed by an MLP regressor for disposition-change deltas.

3.6 Outputs

Our systems output continuous predictions for each setting: essay-level $(\hat{v}, \hat{a})$ per essay, short-horizon $(\widehat{\Delta v}, \widehat{\Delta a})$ per forecasting user, and disposition-change $(\widehat{\Delta v}_{\text{disp}}, \widehat{\Delta a}_{\text{disp}})$ per forecasting user. Each view is implemented as an explicit inference pipeline that reads the official inputs and writes the required submission CSV.

4 Experimental Setup

4.1 Data and Splits

We use only the official SemEval-2026 Task 2 releases. All texts are in English. For development and model selection, we enforce user-disjoint evaluation to prevent leakage across longitudinal histories. Specifically, we rely on frozen unseen-user splits provided in our repository (seeded for reproducibility) and keep the split fixed across runs for comparable ablations and tuning.

4.2 Preprocessing and Representations

For essay-level prediction, raw essays are tokenized with the corresponding pretrained tokenizer and truncated/padded to the model maximum length. The submitted Subtask 1 DistilBERT encoder uses max length 512; the embedding-based Subtask 2A and Subtask 2B pipelines use max length 256, and the Team 2 per-anchor Subtask 2A DistilBERT regressor also uses max length 512. For the change-forecasting and disposition-change settings, we additionally construct user-level histories:

- **Short-horizon forecasting**: Team 1 represents each user with the most recent $k = 32$ text embeddings and compact trajectory-derived numeric features. Histories shorter than 32 are left-padded with zero embedding rows, and true sequence lengths are passed to the GRU so padded positions are ignored. Irregular intervals are represented using `dt_prev_seconds`, the elapsed time in seconds from the previous timestamp.

- **Disposition-change prediction**: users are represented by pooled history embeddings augmented with normalized trajectory summary statistics; timestamps are used for ordering histories, but no explicit time-gap feature is used.

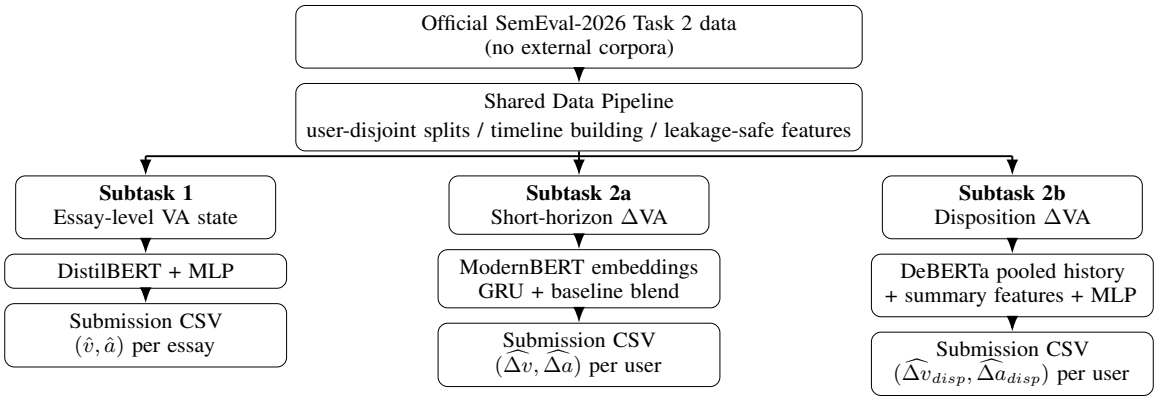

Figure 1: Unified pipeline for longitudinal valence–arousal (VA) modeling, decomposed into three prediction views with leakage-safe user-disjoint evaluation.

Numeric trajectory features are standardized with feature-wise mean/std normalization, with statistics fit only on the training split and then applied unchanged to validation and forecasting users.

4.3 Models and Training

We train lightweight regressors aligned with the three prediction views:

- **Essay-level VA state**: Team 2 uses a fine-tuned DistilBERT regressor, while Team 1 uses a ModernBERT fine-tuning variant.
- **Short-horizon change**: Team 1 uses a GRU sequence regressor over recent embeddings with a previous-delta baseline blend, while Team 2 uses a per-anchor DistilBERT+history-feature MLP.
- **Disposition-change**: Team 1 uses a feed-forward MLP over pooled DeBERTa history embeddings with normalized trajectory features, while Team 2 uses a DistilBERT-based biLSTM.

All models are optimized with AdamW (Loshchilov and Hutter, 2019) (built on Adam (Kingma and Ba, 2015)) and selected by the validation primary score under user-disjoint splits. For Subtask 1, the submitted Team 2 DistilBERT system uses AdamW with two parameter groups: encoder learning rate $5\text{e}{-6}$ and regression-head learning rate $1\text{e}{-5}$, both with weight decay 0.01. It uses batch size 16, dropout 0.1 in the MLP head, max length 512, early-stopping patience 3, and a linear warmup/decay schedule with warmup over 10% of total steps. For Subtask 2A, the GRU sequence model uses learning rate $5\text{e}{-4}$, weight decay 0.01, batch size 64, and up to 10 epochs with patience 2; final predictions blend the GRU and linear previous-delta baseline as $0.40/0.60$ for valence and $0.10/0.90$ for arousal. For Subtask 2B, the pooled-history MLP uses learning rate $1\text{e}{-3}$, weight decay $1\text{e}{-4}$, batch size 256, and 30 fixed epochs, with the best checkpoint selected by validation primary score. Models use standard regression losses (MSE or SmoothL1/Huber (Huber, 1964)); transformer backbones follow standard components such as layer normalization (Ba et al., 2016) and GELU activations (Hendrycks and Gimpel, 2016).

4.4 Evaluation Metrics

We report the official evaluation metrics for each prediction view. For essay-level state estimation, we report the official composite Pearson correlation for valence and arousal. For short-horizon and disposition-change prediction, we report Pearson correlation for predicted deltas on the corresponding targets. Unless otherwise noted, results are computed on the official evaluation split with the prescribed scoring scripts.

4.5 Reproducibility

To support reproducibility, we (i) fix random seeds, (ii) use frozen user-disjoint splits for development, (iii) log run configurations and artifacts, and (iv) provide trained checkpoints and step-by-step inference instructions for each task setting. When training uses a fixed epoch budget, the released checkpoint is selected by the best validation primary score across epochs. All submissions are generated by running the provided inference scripts from the repository root with the included artifacts placed at documented paths.

5 Results

5.1 Official Evaluation Performance

Table 1 reports the official evaluation results for the two merged teams' public leaderboard submissions (Team 1 = AI4PC-Howard University; Team 2 = BisonAI4PC) across Subtasks 1/2A/2B. For Subtask 1 (essay-level state estimation), we report the official composite correlation r_{comp} together with its between-user and within-user components, and the corresponding MAE variants. For Subtasks 2A and 2B (user-level change prediction), we report Pearson r and MAE for the valence and arousal deltas. We include the organizers' official baselines for context (linear(BERT) for Subtask 1 and linear(prev) for Subtasks 2A/2B).

6 Discussion

Our results show a consistent pattern across the three prediction views: essay-level valence–arousal state estimation is reliable for lightweight transformer regressors, while user-level change prediction—especially arousal and longer-horizon shifts—remains difficult under the available supervision and aggregation choices. We interpret these trends using both teams' official submissions under the same user-disjoint split protocol.

6.1 Why essay-level state estimation is strongest

For per-essay state prediction, the text often contains direct cues correlated with affect (e.g., appraisals, stressors, relief). Fine-tuning a compact encoder with a small regression head is sufficient to learn stable mappings from text to continuous valence and arousal. This setting also benefits from dense supervision (one label per essay) and avoids compounding uncertainty from forecasting. Team 1's ModernBERT system and Team 2's DistilBERT system show similar essay-level behavior, suggesting that backbone choice does not fundamentally change the difficulty profile for this view.

6.2 Short-horizon change forecasting is sensitive to arousal

Short-horizon forecasting requires predicting a *delta* from limited history, where targets can be small/noisy and arousal changes are often less explicit than valence changes. Our blended approach reflects a practical trade-off: a simple previous-delta baseline stabilizes predictions, while the GRU sequence regressor captures non-linear temporal effects from recent embeddings and trajectory features. The remaining gap suggests that arousal dynamics may require richer context, better handling of time gaps, or objectives emphasizing within-user consistency. Team 2's per-anchor DistilBERT + history-feature regressor underperformed Team 1's sequence + blend system, reinforcing the benefit of explicitly modeling recent temporal context. The below-baseline results in Subtasks 2A/2B suggest that simple temporal baselines remain strong when affect trajectories are low-variance, sparse, or irregularly sampled. In these cases, learned models can amplify noise in small delta targets, especially for arousal, while previous-value or linear-history baselines provide stable estimates.

6.3 Disposition-change prediction remains challenging

Disposition-change prediction is the hardest view. Long-horizon shifts may not be well captured by simple pooling plus coarse summary statistics: pooling can wash out rare but informative signals, and supervision is sparse relative to variability in user histories. More expressive aggregation (e.g., attention over events, time-aware pooling, or hierarchical modeling) may be needed, as shown in prior multimodal and biomedical temporal modeling studies (Hagos et al., 2025; Aryal et al., 2023c; Ngueajio et al., 2025), to capture long-term structure without sacrificing reproducibility. Team 2's sequence-based DistilBERT biLSTM did not outperform Team 1's pooled-history DeBERTa model, suggesting that aggregation choices and summary features may be more important than sequence-encoder complexity in this comparison.

6.4 Reproducibility considerations

A central goal of our systems is transparency and ease of reproduction. We enforce user-disjoint splits to prevent leakage and keep evaluation comparable, and we provide deterministic inference pipelines and frozen artifacts where applicable.

6.5 Limitations and future directions

This work focuses on lightweight baselines and simple temporal modeling choices. Future work could explore (i) stronger time-aware sequence objectives, (ii) uncertainty-aware training and calibration for arousal changes, (iii) improved user-history aggregation for long-horizon disposition shifts, and (iv) multitask coupling between state estimation and change prediction to share signal across views,

Subtask 1 model	Valence (V)						Arousal (A)						Avg.
	r_{comp}	r_{between}	r_{within}	mae_{comp}	$\text{mae}_{\text{between}}$	$\text{mae}_{\text{within}}$	r_{comp}	r_{between}	r_{within}	mae_{comp}	$\text{mae}_{\text{between}}$	$\text{mae}_{\text{within}}$	r
Team 1 (ModernBERT)	0.631	0.701	0.548	0.670	0.443	0.817	0.462	0.511	0.410	0.416	0.296	0.523	0.547
Team 2 (DistilBERT)	0.665	0.744	0.569	0.633	0.419	0.780	0.468	0.540	0.389	0.395	0.257	0.518	0.567
Organizers (baseline; linear(BERT))	0.557	0.659	0.435	0.743	0.472	0.886	0.299	0.343	0.253	0.459	0.311	0.585	0.428

Subtask 2A model	Valence (V)		Arousal (A)		Avg.
	r	mae	r	mae	r
Team 1 (ModernBERT seq + blend)	0.597	1.180	0.413	0.720	0.505
Team 2 (DistilBERT + history MLP)	0.379	1.202	0.085	0.767	0.232
Organizers (baseline; linear(prev))	0.615	1.168	0.670	0.638	0.643

Subtask 2B model	Valence (V)		Arousal (A)		Avg.
	r	mae	r	mae	r
Team 1 (DeBERTa pooled + MLP)	0.046	0.419	0.348	0.292	0.197
Team 2 (DistilBERT biLSTM)	−0.120	0.424	−0.103	0.296	−0.112
Organizers (baseline; linear(prev))	0.434	0.406	0.584	0.286	0.509

Table 1: Official evaluation results for the submitted systems from two merged teams: Team 1 = AI4PC-Howard University and Team 2 = BisonAI4PC. Subtask 1 reports the official composite correlation r_{comp} and its between/within components, along with the corresponding MAE variants. Subtasks 2A and 2B report Pearson r and MAE for delta targets. Avg. r denotes the mean of the valence and arousal correlation scores for the primary correlation metric in each subtask. Baselines are the organizers' provided systems (linear(BERT) for Subtask 1; linear(prev) for Subtasks 2A/2B).

while preserving leakage-safe, reproducible evaluation. Because backbone choice and aggregation strategy vary together across systems, these comparisons are informative but do not isolate the independent effect of each encoder family.

7 Conclusion

We presented lightweight, reproducible systems for longitudinal valence–arousal modeling on the official SemEval-2026 Task 2 data under leakage-aware, user-disjoint evaluation. We addressed three prediction views: essay-level VA estimation, short-horizon VA change forecasting, and longer-horizon disposition-change prediction. This paper merges two participating teams (Team 1 and Team 2) and reports both teams' official submissions and public leaderboard results across Subtasks 1/2A/2B. We hope these artifacts and baselines support future work on reliable longitudinal emotion modeling.

References

Saurav K Aryal, Howard Prioleau, and Surakshya Aryal. 2023a. Sentiment analysis across multiple african languages: A current benchmark. *arXiv preprint arXiv:2310.14120*.

Saurav K Aryal, Howard Prioleau, Surakshya Aryal, and Gloria Washington. 2023b. Baselining performance for multilingual codeswitching sentiment classification. *Journal of Computing Sciences in Colleges*, 39(3):337–346.

Saurav K Aryal, Howard Prioleau, and Legand Burge. 2022. Acoustic-linguistic features for modeling neurological task score in alzheimer's. In *Pacific Symposium on Biocomputing 2023: Kohala Coast, Hawaii, USA, 3–7 January 2023*, pages 335–346.

Saurav K Aryal, Ujjawal Shah, Howard Prioleau, and Legand Burge. 2023c. Ensembling and modeling approaches for enhancing alzheimer's disease scoring and severity assessment. In *2023 International Conference on Computational Science and Computational Intelligence (CSCI)*, pages 1364–1370. IEEE.

Saurav Keshari Aryal and Gaurav Adhikari. 2023. Evaluating impact of emoticons and pre-processing on sentiment classification of translated african tweets.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer normalization. *Preprint*, arXiv:1607.06450.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

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 (NAACL-HLT)*.

Desta Haileselassie Hagos, Saurav Keshari Aryal, Patrick Ymele-Leki, and Legand L Burge. 2025. Ai-driven multimodal colorimetric analytics for biomedical and behavioral health diagnostics. *Computational and structural biotechnology journal*, 27:2219–2232.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021a. DeBERTaV3: Improving DeBERTa using electra-style pre-training with gradient-disentangled embedding sharing. *Preprint*, arXiv:2111.09543.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021b. DeBERTa: Decoding-enhanced BERT with disentangled attention.

Dan Hendrycks and Kevin Gimpel. 2016. Gaussian error linear units (GELUs). *Preprint*, arXiv:1606.08415.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Peter J. Huber. 1964. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101.

Jazette Johnson, Lucretia Williams, Jaye Nias, Saurav K Aryal, and Gloria Washington. 2025. Centering black voices: Lessons learned and reflections from a large-scale aave data collection at a historically black university. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–7.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *Preprint*, arXiv:1412.6980.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*.

Mikel K Ngueajio, Saurav Aryal, Marcellin Atemkeng, Gloria Washington, and Danda Rawat. 2025. Decoding fake news and hate speech: A survey of explainable ai techniques. *ACM Computing Surveys*, 57(7):1–37.

Howard Prioleau and Saurav K Aryal. 2023a. Benchmarking current state-of-the-art transformer models on token level language identification and language pair identification. In *2023 International Conference on Computational Science and Computational Intelligence (CSCI)*, pages 193–199. IEEE.

Howard Prioleau and Saurav Keshari Aryal. 2023b. Feature importance analysis for mini mental status score prediction in alzheimer's disease.

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

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *Preprint*, arXiv:1910.01108.

Mike Schuster and Kuldip K. Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.

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.

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

Benjamin Warner, Alex Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, and 1 others. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *Preprint*, arXiv:2412.13663.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clément Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (EMNLP)*.

A Appendix

A.1 Code and Reproducibility

Our code and reproduction instructions are available at: `https://github.com/arajshah/semeval-2-emotion-dynamics`