Fine-Tuning DistilBERT, DeBERTa and ModernBERT for Valence–Arousal Prediction and Change Estimation

Araj Shah, Utsav Shah, Saurav K Aryal, PhD

AI4PC Lab, Institute for Human-Centered Artificial Intelligence

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 (SemEval-2026 Task 2 Organizers, 2026).

We address the task with three aligned prediction views:

1. per-essay VA state estimation
2. short-horizon user-level VA change forecasting from recent history
3. long-horizon disposition change prediction from aggregated histories.

Our approach follows two principles:

1. strict user-disjoint evaluation with inference-time-safe feature construction
2. compact, transparent pipelines that are easy to reproduce.

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:

- A leakage-aware, user-disjoint protocol for longitudinal VA modeling across state and change settings.
- Lightweight systems for Subtasks 1/2A/2B with explicit, reproducible inference pipelines.
- Official evaluation results and controlled alternative variants to clarify the effect of backbone choice and temporal aggregation.

METHODS

1. Data & Preprocessing

- **Dataset:** Official SemEval-2026 Task 2 (English).
- **Protocol:** Strict **user-disjoint splits** to prevent leakage and ensure reproducibility.
- **Features:**
  - **Essay-level:** Tokenized raw text (max length 512).
  - **Short-horizon:** Fixed-length windows of recent embeddings + numeric trajectory features.
  - **Disposition:** Pooled DeBERTa history embeddings (mean/most-recent) + summary stats.

2. Model Architecture We deployed three lightweight regressors aligned with the prediction views:

- **VA State:** Fine-tuned **DistilBERT** (ModernBERT variant explored).
- **Short-horizon Change: GRU sequence regressor** + previous-delta baseline blend.
- **Disposition Change: Feed-forward MLP** over pooled history embeddings.

3. Training & Optimization

- **Optimizer:** AdamW with standard regression losses (MSE/Huber).
- **Regularization:** Dropout, gradient clipping, and early stopping.
- **Selection:** Models selected via official validation metrics under user-disjoint splits

4. Evaluation Metrics

- **Essay-level State:** Composite Pearson correlation *(r)* for Valence and Arousal.
- **Change/Disposition:** Pearson correlation *(r)* for predicted deltas against targets.

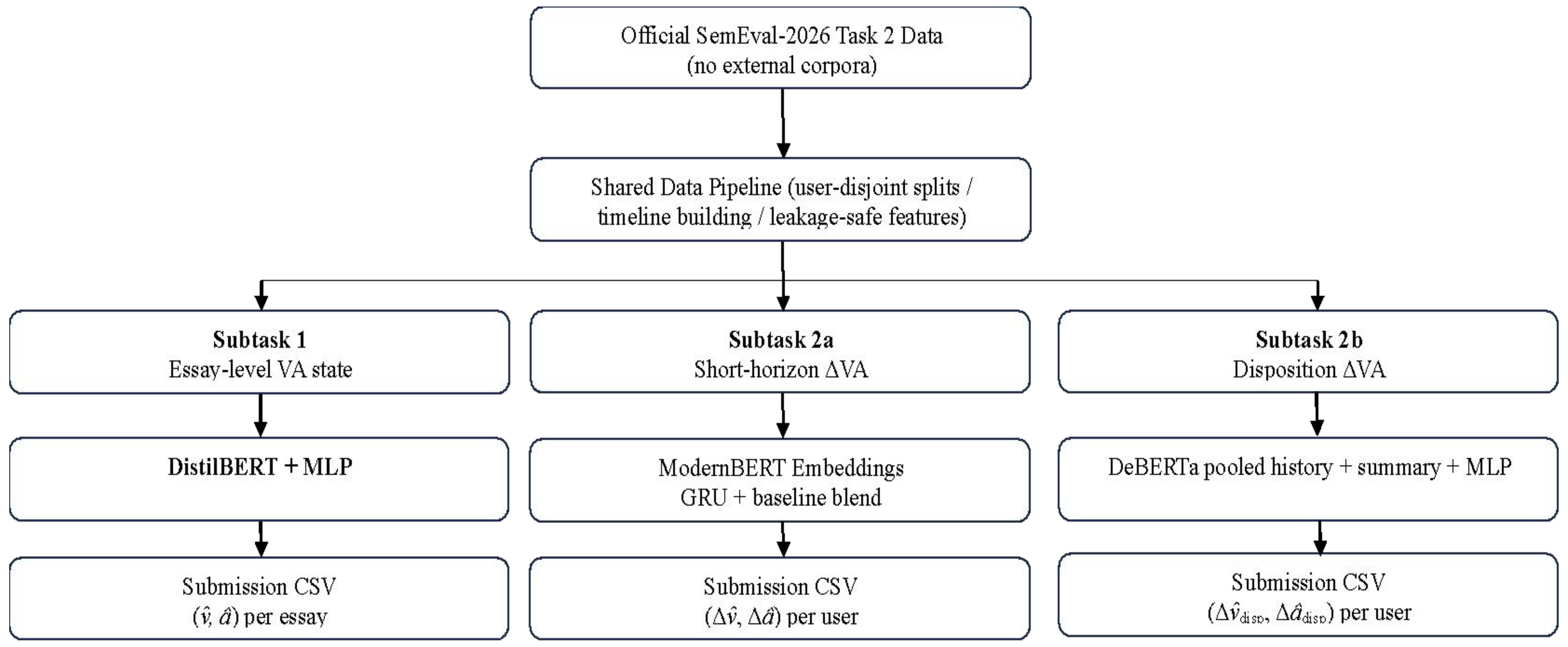

Figure: Unified pipeline for longitudinal valence-arousal (VA) modeling, decomposed, into three prediction views with leakage-safe user-disjoint evaluation

RESULTS

Subtask 1 Model	Valence (V)						Arousal (A)					
	r_{comp}	$r_{between}$	r_{within}	mae_{comp}	$mae_{between}$	mae_{within}	r_{comp}	$r_{between}$	r_{within}	mae_{comp}	$mae_{between}$	mae_{within}
Team 1 (ModerBERT)	0.631	0.701	0.548	0.67	0.443	0.817	0.462	0.511	0.41	0.416	0.296	0.523
Team 2 (DistilBERT)	0.665	0.744	0.569	0.633	0.419	0.78	0.468	0.54	0.389	0.395	0.257	0.518
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
Subtask 2A Model	**Valence (V)**						**Arousal (A)**					
	r					mae	r					mae
Team 1 (ModernBERT seq + blend)	0.597					1.18	0.413					0.72
Team 2 (DistilBERT + history MLP)	0.379					1.202	0.085					0.767
Organizers (baseline; linear(prev))	0.615					1.168	0.67					0.638
Subtask 2B Model	**Valence (V)**						**Arousal (A)**					
	r					mae	r					mae
Team 1 (DeBERTa pooled + MLP)	0.046					0.419	0.348					0.292
Team 2 (DistilBERT biLSTM)	-0.12					0.424	-0.103					0.296
Organizers (baseline; linear(prev))	0.434					0.406	0.584					0.286

Table: Official evaluation results for the submitted systems from two merged teams (Team 1 and Team 2). 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. Baselines are the organizers' provided systems (linear(BERT) for Subtask 1; linear(prev) for Subtasks 2A/2B).

RESULTS

Table above 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 rcomp 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).

CONCLUSION & FUTURE DIRECTION

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 address three prediction views: essay-level VA estimation, shorthorizon 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. Future research will explore stronger time-aware sequence objectives, uncertainty calibration for arousal changes, and improved user-history aggregation. We also aim to implement multitask coupling between state and change models to enhance signal sharing while maintaining leakage-safe, reproducible evaluation.

REFERENCES

1. Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer normalization. Preprint, arXiv:1607.06450.
2. SemEval-2026 Task 2 Organizers. 2026. Semeval-2026 task 2: Longitudinal valence–arousal modeling and change prediction. Task description paper / overview. Placeholder citation to be replaced with the official BibTeX when released during camera-ready
3. 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).
4. Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021a. DeBERTaV3: Improving DeBERTa using electrastyle pre-training with gradient-disentangled embedding sharing. Preprint, arXiv:2111.09543.
5. Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021b. DeBERTa: Decodingenhanced BERT with disentangled attention.
6. Dan Hendrycks and Kevin Gimpel. 2016. Gaussian error linear units (GELUs). Preprint, arXiv:1606.08415.
7. Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. Neural Computation, 9(8):1735– 1780.

ACKNOWLEDGEMENTS

This research project was supported in part by the Office of Naval Research grant N00014-22-1-2714. The work is solely the responsibility of the authors and does not necessarily represent the official view of the Office of Naval Research.