VAP-GameController at SemEval-2026 Task 2: Lexicon-based and Emotion-Aware Approaches for Longitudinal Emotion Prediction

Huy M. Le[*]
MBZUAI, UAE
HuyM.Le@mbzuai.ac.ae

Phu Truong Thien[*]
University of Information Technology
Vietnam National University, Ho Chi Minh City
23521190@gm.uit.edu.vn

Trung Tran[*]
MBZUAI, UAE
Trung.Tran@mbzuai.ac.ae

Nga N. T. Nguyen
Independent Researcher
nnganguyen.work@gmail.com

Monojit Choudhury
MBZUAI, UAE
Monojit.Choudhury@mbzuai.ac.ae

Abstract

In this work, we participate in SemEval-2026 Task 2, which focuses on predicting continuous valence and arousal trajectories from longitudinal ecological essays. We study three approaches: (1) encoder-based regressors that map raw text directly to valence and arousal, (2) a lexicon-based pipeline with linguistic rules and a dual-level calibration mechanism for personalized estimation, and (3) an emotion-augmented hybrid framework that first predicts a sentence-level emotion label and then regresses valence and arousal from the sentence content together with that predicted label. On the official shared-task test set, our official submissions are DistillBERT and DistillBERT + Emotion, while we additionally report comparative results for lexicon and BGE-M3 variants on the same test set. Evaluated with Pearson correlation (r) and MAE, the experiments highlight the strength of direct encoder-based regression and the value of calibration for lexicon-based signals.

1 Introduction

Recent NLP and AI systems have achieved strong performance on many tasks (Le et al., 2025b,a, 2026a, 2023, 2026b), but modeling affect in text is often framed as predicting a single, static label per document, while real-world emotions evolve over time. Dimensional representations such as valence (pleasantness) and arousal (activation) support fine-grained tracking of these dynamics, but most continuous-emotion studies focus on short, controlled, typically multimodal interactions (e.g., RECOLA; (Ringeval et al., 2013; Schneider et al., 2025)). SemEval-2026 Task 2 (Soni et al., 2026) extends this setting to naturalistic, longitudinal, text-only data by asking systems to estimate valence–arousal variation from ecological essays, emphasizing temporal coherence and robustness beyond static emotion detection benchmarks (e.g., Muhammad et al., 2025).

We address this task on the official dataset of 5,285 texts from 182 U.S. service-industry workers, annotated with valence, arousal, and metadata. To assess generalization to unseen individuals, we enforce a user-level split such that each writer appears in exactly one partition. Our framework explores three complementary approaches: (i) direct encoder-based regressors that predict essay-level affect from raw text, (ii) a lexicon-based pipeline using NRC VAD with linguistic heuristics and a two-stage calibration scheme (global and user-specific), and (iii) an emotion-augmented hybrid method that predicts a sentence-level discrete emotion label and concatenates it with the original sentence content before regression. We evaluate using MAE and Pearson correlation, and observe that direct encoder-based regression performs best overall while calibration substantially strengthens the lexicon baseline.

Our main contributions are:

- We compare direct encoder-based regressors and a calibrated lexicon-based pipeline for longitudinal valence–arousal prediction.
- We introduce a hybrid pipeline that augments each sentence with a predicted emotion label before regression.
- We provide an empirical comparison of encoder-based, lexicon-based, and hybrid approaches, highlighting their strengths and trade-offs on SemEval-2026 Task 2.

2 Related Work

Dimensional emotion representations. A common way to model affect is to represent emotions in a low-dimensional continuous space, most prominently along *valence* (pleasantness) and *arousal* (activation), as formalized in the circumplex view

[*] Equal contribution.

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 256–262
July 3-4, 2026 ©2026 Association for Computational Linguistics

of affect (Russell, 1980). Such dimensional representations support fine-grained prediction beyond discrete emotion categories and are widely adopted in affective computing and NLP. Complementing sentence- or document-level modeling, lexical resources provide word-level affect norms in VAD space, including large-scale norms (Warriner et al., 2013) and the NRC VAD lexicon (Mohammad, 2018).

Continuous affect recognition and temporal trajectories. Continuous valence–arousal prediction has been extensively studied in controlled settings, especially with multimodal data. Benchmarks such as RECOLA (Ringeval et al., 2013) and the AVEC challenges (Ringeval et al., 2015; Valstar et al., 2016) emphasize modeling temporally coherent affect trajectories and have inspired sequence-aware architectures (e.g., recurrent models and attention-based variants). However, these datasets typically involve short interactions and controlled protocols, leaving open questions about modeling *longitudinal* affect dynamics in naturalistic, text-only narratives.

Lexicon-based, hybrid, and personalized approaches. Lexicon-driven affect estimation remains attractive due to interpretability and robustness in low-resource or domain-shift scenarios, leveraging affect norms (Warriner et al., 2013; Mohammad, 2018) and rule-based analyzers such as VADER (Hutto and Gilbert, 2014). Recent approaches integrate lexical or sentiment knowledge into neural models via feature augmentation or knowledge-aware pretraining, e.g., SentiLARE (Ke et al., 2020) and SentiBERT (Yin et al., 2020). Large-scale emotion-labeled corpora such as GoEmotions (Demszky et al., 2020) further support emotion-aware adaptation of pretrained encoders. Separately, personalization and author-conditioned modeling (e.g., incorporating user signals through attention) has shown benefits in sentiment prediction (Chen et al., 2016), aligning with the user-specific and longitudinal nature of Task 2.

3 Task and Data

3.1 Dataset

We use the SemEval-2026 Shared Task 2 dataset of 5,285 longitudinal texts (ecological essays and feeling-word entries) collected from 182 U.S. service-industry workers between 2021 and 2024. Each text is annotated with valence and arousal labels defined on bounded ordinal scales: valence $v \in \{-2, -1, 0, 1, 2\}$ and arousal $a \in \{0, 1, 2\}$. We model these targets with regression while retaining their original bounded label ranges. Each entry is also associated with metadata such as user identifier, timestamp, collection phase, and a binary flag indicating whether the `text` field is a free-form essay or an explicit list of feelings.

3.2 Task Modeling

Given a set of one or more textual entries for each user, where each entry is associated with a timestamp, the task is to predict the corresponding affective states along the two target dimensions of valence and arousal.

Formally, for each user u_i, we observe a sequence of text–timestamp pairs:

$$u_i : \{(e_{i1}, t_{i1}), (e_{i2}, t_{i2}), \ldots, (e_{in_i}, t_{in_i})\},$$

where e_{ij} represents the textual content (e.g., essays or reported feeling words), t_{ij} is the associated timestamp, and n_i denotes the number of entries for user i.

Each text e_{ij} is annotated with a label pair

$$y_{ij} = (v_{ij}, a_{ij}),$$

where valence $v_{ij} \in \{-2, -1, 0, 1, 2\}$ and arousal $a_{ij} \in \{0, 1, 2\}$. In our formulation, these bounded labels are predicted with regression heads whose outputs are constrained to the corresponding target ranges.

The goal is to develop a predictive model that estimates

$$\hat{y}_{ij} = (\hat{v}_{ij}, \hat{a}_{ij})$$

for every textual entry.

4 Methods

4.1 Lexicon-based Approach

We propose an efficient lexicon-based regression pipeline for each entry in the dataset. We use the NRC VAD Lexicon (Mohammad, 2018), which provides human ratings of valence and arousal for more than 20,000 English words. Figure 1 illustrates the overall pipeline.

Lookup: First, for each token t in an entry e that appears in the lexicon, we retrieve its valence and arousal scores and rescale them to the task ranges.

IDF weights: Inspired by (Buechel and Hahn, 2016), we compute inverse document frequency (IDF) scores on the training data restricted to the lexicon vocabulary:

$$w_{t,i} = \mathrm{idf}(t). \tag{1}$$

Following Hutto and Gilbert (2014), we apply several handcrafted rules to handle specific linguistic phenomena in the text. Degree modifiers scale the magnitude of both valence and arousal scores, nearby negations flip the valence score within a short left-context window, contrastive conjunctions such as *but* or *however* downweight the preceding clause and upweight the following one, and exclamation marks provide a small boost to arousal.

Aggregation: We then use the weighted-average formulation of (Buechel and Hahn, 2016) to predict the final regression values:

$$\hat{v}(e) = \frac{\sum_{t,i} w_{t,i}\, v_{t,i}}{\sum_{t,i} w_{t,i}} \tag{2}$$

$$\hat{a}(e) = \frac{\sum_{t,i} w_{t,i}\, a_{t,i}}{\sum_{t,i} w_{t,i}}. \tag{3}$$

Calibration: Because the lexicon is not tailored to this dataset, we apply a linear calibration step to better match the task labels. For $y \in \{v, a\}$, we compute

$$\tilde{y} = \alpha_0 + \alpha_1\, \hat{y}. \tag{4}$$

We further apply user-specific bias and scaling for seen users to balance personalization and generalization:

$$\tilde{y}^{(u)} = \beta_0^{(u)} + \beta_1^{(u)}(\alpha_0 + \alpha_1 \hat{y}). \tag{5}$$

4.2 Encoder-based Regression

In the direct regression route (Figure 2-top), raw text is fed directly into a fine-tuned encoder-based regressor to obtain continuous valence and arousal scores. Concretely, we use either BGE-M3 (Chen et al., 2024) or DistillBERT (Mavdol, 2025), with the latter initialized from a model pretrained on synthesized VA data. Both models are trained on labeled text–label pairs to map each entry to (v, a) without any intermediate emotion prediction.

We adapt each encoder to the task by adding two separate linear heads on top of the shared representation $\mathbf{h}(x)$: one for valence and one for arousal. Each head maps the embedding to a scalar with an activation chosen to respect the natural range of its affective dimension:

$$\hat{v}(x) = 2 \cdot \tanh\left(\mathbf{w}_v^\top \mathbf{h}(x) + b_v\right), \tag{6}$$

$$\hat{a}(x) = 2 \cdot \sigma\left(\mathbf{w}_a^\top \mathbf{h}(x) + b_a\right), \tag{7}$$

where $\sigma(\cdot)$ is the sigmoid, yielding $\hat{v}(x) \in [-2, 2]$ and $\hat{a}(x) \in [0, 2]$.

This direct route is simple and data-efficient when the text already contains unambiguous affective cues or when the task setting ($is\_words = True$) focuses on short, emotion-laden words or phrases.

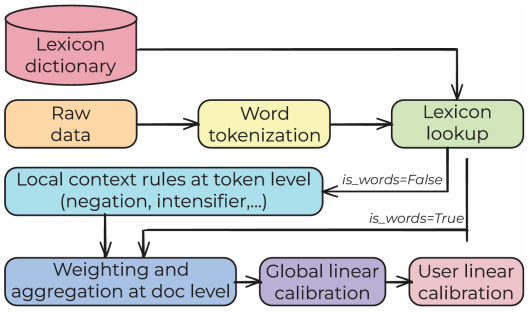

Figure 1: Flow of our lexicon-based approach.

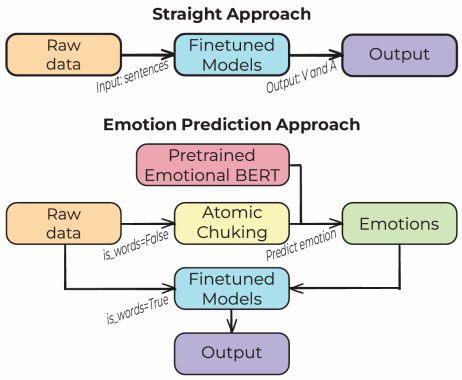

Figure 2: Overview of the two routes: (top) direct regression from raw text; (bottom) sentence-level emotion augmentation with a pretrained emotion recognizer, with a direct bypass for $is\_words = True$.

4.3 Emotion-Augmented Regression

The emotion-augmented route (Figure 2-bottom) adds an explicit sentence-level emotion label before regression. For entries with $is\_words = False$, we split the text into sentences using punctuation-based boundaries (periods, question marks, and exclamation marks). Each sentence s_k is passed to the pretrained emotion recognizer[1], and we keep the single most probable label z_k among the 27

[1] SamLowe/roberta-base-go_emotions

emotion labels predicted by the model. We then concatenate the original sentence content with its predicted emotion label to form the downstream input $x'_k = s_k \,\|\, z_k$. The same encoder-based regressor from Section 4.2 is applied to each augmented sentence, producing sentence-level predictions $(\hat{v}_k, \hat{a}_k)$. Final document-level scores are computed by mean aggregation:

$$\hat{v}(e) = \frac{1}{m} \sum_{k=1}^{m} \hat{v}_k, \tag{8}$$

$$\hat{a}(e) = \frac{1}{m} \sum_{k=1}^{m} \hat{a}_k, \tag{9}$$

where m is the number of sentences in entry e.

If the input consists of isolated words or short emotion phrases ($is\_words = True$), we bypass sentence splitting and emotion prediction and feed the original content directly into the regressor, since the text itself already serves as an explicit emotion cue.

5 Experiments and Results

5.1 Experimental Setup

Hardware Resources. All experiments are conducted on GPU-enabled cloud environments using NVIDIA T4 GPUs on Google Colab and Kaggle.

Model configuration. For reproducibility, we fix all random seeds and enable deterministic CUDNN behavior. All Transformer backbone parameters and task-specific regression heads are fully fine-tuned, with multiple training runs performed to ensure stable and robust performance for the valence–arousal prediction task. The main hyperparameters are summarized in Table 2.

5.2 Data Splitting

To faithfully simulate the official shared-task setting, we construct data splits that are both user-aware and time-aware.

User partitioning We first partition users into seen and unseen groups. Specifically, we randomly sample 30% of users as unseen users, while the remaining 70% are treated as seen users.

Time-based splitting For each seen user, we sort all samples by their timestamps to preserve chronological order. We then assign the earliest 70% of samples to the training set, and allocate the remaining samples to the validation set. For unseen users, no samples are used for training; their data are reserved for validation.

Official data usage All splitting procedures are conducted on the official training data, while the final evaluation reported in Table 1 is performed on the separate official test set provided by the competition.

5.3 Evaluation Metrics

To quantify prediction quality for bounded valence and arousal targets modeled with regression, we report two standard regression metrics: MAE and Pearson correlation (r).

Mean Absolute Error (MAE) measures the average magnitude of the deviations between predicted and true labels, disregarding their direction:

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^{N} \left| y_i - \hat{y}_i \right|. \tag{10}$$

Pearson Correlation Coefficient (r) evaluates the linear correlation between the predicted values and the ground truth:

$$r = \frac{\sum_{i=1}^{N} (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{N} (y_i - \bar{y})^2} \sqrt{\sum_{i=1}^{N} (\hat{y}_i - \bar{\hat{y}})^2}}. \tag{11}$$

5.4 Main Results

Table 1 summarizes results on the official shared-task test set. The first block lists our two official submissions (DistillBERT and DistillBERT + Emotion), while the remaining rows are additional comparison systems evaluated on the same test set. We compare (i) lexicon-based pipelines with calibration, (ii) direct encoder-based regressors, and (iii) emotion-augmented hybrid variants that append predicted sentence-level emotion labels before regression.

Lexicon baselines. The uncalibrated lexicon baseline is a relatively weak reference point ($\bar{r} = 0.2140$, $\overline{\text{MAE}} = 0.8060$), indicating that raw lexicon aggregation alone does not align well with the label distribution and writing style of ecological essays. Adding *global calibration* substantially reduces error ($\overline{\text{MAE}}$ from 0.8060 to 0.6331), and the *global + user calibration* further improves both dimensions, reaching $\bar{r} = 0.3555$ and $\overline{\text{MAE}} = 0.5788$. This confirms that even simple linear recalibration can correct systematic bias and partially account for individual differences.

Model	$r_v \uparrow$	$\text{MAE}_v \downarrow$	$r_a \uparrow$	$\text{MAE}_a \downarrow$	$\bar{r} \uparrow$	$\overline{\text{MAE}} \downarrow$
Official shared-task submissions						
DistillBERT	0.6151	0.6288	0.3220	0.3959	0.4686	0.5124
DistillBERT + Emotion	0.5818	0.6635	0.1829	0.4385	0.3824	0.5510
Additional comparisons on the official test set						
Lexicon (no calibration)	0.2598	0.8761	0.1681	0.7358	0.2140	0.8060
Lexicon (global calibration)	0.3611	0.8098	0.1129	0.4564	0.2370	0.6331
Lexicon (global + user calibration)	0.4186	0.7498	0.2924	0.4077	0.3555	0.5788
BGE-M3	**0.6306**	**0.6183**	**0.4389**	**0.3463**	**0.5348**	**0.4823**
BGE-M3 + Emotion	0.5511	0.6798	0.3419	0.3905	0.4465	0.5352

Table 1: Performance on the official shared-task test set. DistillBERT and DistillBERT + Emotion are our two official leaderboard submissions; the remaining systems are additional comparison results on the same official test set. r_v / MAE_v and r_a / MAE_a denote Pearson correlation (r) and Mean Absolute Error for valence and arousal, respectively.

Table 2: Hyperparameters for encoder-based models.

Hyperparameter	Value
Output heads	Scaled Valence and Arousal heads
Training objective	RMSE over (v, a) pairs
Optimizer	AdamW
Weight decay	0.01
Learning rate	2×10^{-5}
Training epochs	100
LR scheduler	Linear (warm-up ratio 0.1)
Mixed precision	FP16
Early stopping	15

Encoder-based models. Neural encoders clearly outperform lexicon-only methods. Among them, BGE-M3 achieves the best overall performance with $\bar{r} = 0.5348$ and $\overline{\text{MAE}} = 0.4823$, and it yields the strongest arousal correlation ($r_a = 0.4389$). DistillBERT is competitive ($\bar{r} = 0.4686$, $\overline{\text{MAE}} = 0.5124$) but consistently behind BGE-M3, especially on arousal ($r_a = 0.3220$). Overall, the best model improves average correlation by $+0.1793$ over the strongest lexicon baseline (0.5348 vs. 0.3555) and reduces $\overline{\text{MAE}}$ by 0.0965 (0.4823 vs. 0.5788).

Effect of sentence-level emotion augmentation. Adding predicted sentence-level emotion labels (+*Emotion*) degrades performance for both backbones: DistillBERT drops from $\bar{r} = 0.4686$ to 0.3824 and BGE-M3 drops from 0.5348 to 0.4465, with corresponding $\overline{\text{MAE}}$ increases. This suggests that a single predicted emotion label per sentence is too coarse as an auxiliary signal for this task and that errors from the emotion classifier propagate to the downstream regressor.

5.5 Discussion

Why does VA-pretrained DistillBERT underperform BGE-M3? DistillBERT is pretrained for valence–arousal prediction on VA-style resources (Mavdol, 2025), which would intuitively make it a strong fit for this task. However, our results show that it is consistently outperformed by the more recent BGE-M3 encoder (Chen et al., 2024), especially on arousal (0.3220 vs. 0.4389). We hypothesize two main reasons: (i) *domain and objective mismatch*–VA pretraining typically relies on shorter texts and more explicit affect cues, while ecological essays often express affect implicitly through events, appraisal, and discourse context; and (ii) *representation strength*–BGE-M3 is a modern, general-purpose multilingual encoder trained with large-scale objectives (self-knowledge distillation and multi-function embedding learning), which may transfer better to long-form, naturalistic narratives despite not being specialized solely for VA regression.

Compute-efficient encoders are still highly competitive. All of our models are encoder-based with fewer than 1B parameters, selected to fit limited GPU resources. Despite these constraints, our best system reaches $\bar{r} = 0.5348$, which is close to the strongest scores observed on the public leaderboard (around 0.611 at the time of evaluation). This indicates that *compact encoder-only approaches* remain a strong and practical choice for longitudinal affect modeling: they offer a favorable accuracy–efficiency trade-off and can be trained/reproduced with modest hardware, while leaving room for future gains from larger backbones or more explicit user-/time-aware sequence modeling.

6 Conclusion

We presented our submission to SemEval-2026 Task 2 on predicting continuous valence and arousal from longitudinal ecological essays. Our study showed that direct encoder-based regression consistently outperformed lexicon-only methods, while calibration substantially strengthened the lexicon baseline. The results also show that sub-1B encoder-based models remain highly competitive under limited compute. We also found that sentence-level emotion augmentation degraded performance for both backbones, and that a VA-pretrained DistillBERT model still lagged behind the more modern BGE-M3 encoder (notably on arousal), suggesting that representation strength and transfer robustness matter more than VA-specific pretraining alone in this naturalistic setting.

Acknowledgments

The authors gratefully acknowledge the support of Mohamed bin Zayed University of Artificial Intelligence for this research.

References

Sven Buechel and Udo Hahn. 2016. Emotion analysis as a regression problem — dimensional models and their implications on emotion representation and metrical evaluation.

Huimin Chen, Maosong Sun, Cunchao Tu, Yankai Lin, and Zhiyuan Liu. 2016. Neural sentiment classification with user and product attention. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1650–1659, Austin, Texas. Association for Computational Linguistics.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *Preprint*, arXiv:2402.03216.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online. Association for Computational Linguistics.

Clayton J. Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 8, pages 216–225.

Pei Ke, Haozhe Ji, Siyang Liu, Xiaoyan Zhu, and Minlie Huang. 2020. SentiLARE: Sentiment-aware language representation learning with linguistic knowledge. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6975–6988, Online. Association for Computational Linguistics.

Huy M. Le, Vy T. Luong, and Ngoc Hoang Luong. 2023. Data augmentation with large language models for vietnamese abstractive text summarization. In *2023 International Conference on Multimedia Analysis and Pattern Recognition (MAPR)*, pages 1–6.

Huy M. Le, Dat Tien Nguyen, Phuc Binh Nguyen, Gia Bao Le Tran, Phu Truong Thien, Cuong Dinh, Minh Nguyen, Nga Nguyen, Thuy T. N. Nguyen, Huy Gia Ngo, Tan Nhat Nguyen, and Binh T. Nguyen. 2026a. Fusionista2.0: Efficiency retrieval system for large-scale datasets. In *MultiMedia Modeling*, pages 167–175, Singapore. Springer Nature Singapore.

Huy M. Le, Dat Tien Nguyen, Ngan T. T. Vo, Tuan D. Q. Nguyen, Nguyen Le Binh, Duy Minh Ho Nguyen, Daniel Sonntag, Lizi Liao, and Binh T. Nguyen. 2026b. Reinforce trustworthiness in multimodal emotional support system. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(37):31474–31482.

Huy M. Le, Dat Nguyen Tien, Khang Le Duy, Tuan Nguyen Dang Quang, Toan Nguyen Khanh, Tuyen Nguyen, and Binh T. Nguyen. 2025a. Fustar: Divide and conquer query in video retrieval system. In *Information and Communication Technology*, pages 92–105, Singapore. Springer Nature Singapore.

Huy M. Le, Dat Nguyen Tien, Khang Le Duy, Tuan Nguyen Dang Quang, Nguyen Khanh Toan, Tuyen Nguyen, and Binh T. Nguyen. 2025b. Fusionista: Fusion of 3-d information of video in retrieval system. In *MultiMedia Modeling*, pages 278–285, Singapore. Springer Nature Singapore.

Mavdol. 2025. Valence and arousal annotations for interactive characters. https://huggingface.co/Mavdol/NPC-Valence-Arousal-Prediction.

Saif M. 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.

Shamsuddeen Hassan Muhammad, Nedjma Ousidhoum, Idris Abdulmumin, Seid Muhie Yimam, and Jan Philip Wahle. 2025. Semeval-2025 task 11: Bridging the gap in text-based emotion detection. *CoRR*, abs/2503.07269.

Fabien Ringeval, Björn Schuller, Michel Valstar, Roddy Cowie, and Maja Pantic. 2015. AVEC 2015 – the 5th international audio/visual emotion challenge and

workshop. In *Proceedings of the 23rd ACM International Conference on Multimedia*, pages 1335–1336, Brisbane, Australia. ACM.

Fabien Ringeval, Andreas Sonderegger, Jürgen Sauer, and Denis Lalanne. 2013. Introducing the recola multimodal corpus of remote collaborative and affective interactions. pages 1–8.

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

Helen Schneider, Svetlana Pavlitska, Helen Gremmelmaier, and Marius Zöllner. 2025. Datasets for valence and arousal inference: A survey. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2025, Nashville, TN, USA, June 11-15, 2025*, pages 5657–5664. Computer Vision Foundation / IEEE.

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.

Michel Valstar, Jonathan Gratch, Björn Schuller, Fabien Ringeval, Denis Lalanne, Mercedes Torres Torres, Stefan Scherer, Giota Stratou, Roddy Cowie, and Maja Pantic. 2016. AVEC 2016 – depression, mood, and emotion recognition workshop and challenge. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, pages 3–10, Amsterdam, The Netherlands. ACM.

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.

Da Yin, Tao Meng, and Kai-Wei Chang. 2020. SentiBERT: A transferable transformer-based architecture for compositional sentiment semantics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3695–3706, Online. Association for Computational Linguistics.