CSIRO-LT at SemEval-2026 Task 2: *In-the-Wild* Valence and Arousal Forecasting on Ecological Text Time Series

Jiyu Chen[1*], **Necva Bölücü**[1*], **Sarvnaz Karimi**[1], **Diego Mollá**[2,1], **Cécile Paris**[1,2]
[1]CSIRO, Australia [2]Macquarie University, Australia
firstname.lastname@csiro.au diego.molla-aliod@mq.edu.au

Abstract

Predicting emotional valence and arousal in text is challenging due to the continuous, dynamic, and context-dependent nature of emotions. The *SemEval 2026 Task 2: Predicting Variation in Emotional Valence and Arousal over Time from Ecological Essays* shared task investigates longitudinal affect prediction from real-world personal essays, including forecasting short-term state and longer-term dispositional changes. We compare Pre-trained Language Models (PLMs) and Large Language Models (LLMs) for these subtasks, examining different input representations and feature formulations. We show that sentiment-aware PLMs are most effective for continuous valence and arousal prediction, and LLMs are effective for short-term state forecasting. Modelling dispositional changes remains challenging, and none of our neural approaches surpasses a simple historical baseline approach in this setting.[1]

1 Introduction

Understanding human emotions through textual expression is a central challenge at the intersection of social science and Natural Language Processing (NLP). Emotions are inherently complex and dynamic: they vary in intensity, combine in subtle ways, and evolve over the course of a narrative. Accurately modelling these emotional dynamics is crucial for applications such as mental health support, social behaviour analysis, and the development of systems that respond appropriately to human affect (Saffar et al., 2023; Teodorescu et al., 2023; Yang and Li, 2025).

Early computational approaches often treated emotion as a discrete phenomenon, assigning texts to fixed categories such as joy, anger, or sadness (Mohammad et al., 2018; Demszky et al., 2020). Emotions can also be represented along dimensional axes: *valence*, indicating the positivity or negativity of an emotional state, and *arousal*, reflecting its intensity or activation level (Russell, 1980). The dimensional framework enables a fine-grained and psychologically grounded representation of affect.

Beyond being continuous, emotions are also inherently temporal. A single textual instance rarely provides a complete picture of an individual's emotional state. Rather, emotions develop, fluctuate, and shift in response to contextual factors. Modelling such temporal trajectories enables systems not only to analyse past affective patterns but also to anticipate future changes—an ability particularly relevant in longitudinal and predictive settings.

The *SemEval-2026 Task 2: Predicting Variation in Emotional Valence and Arousal over Time from Ecological Essays* shared task (Soni et al., 2026)[2] is part of the International Workshop on Semantic Evaluation (SemEval). Its objective is to predict continuous emotional valence and arousal trajectories within sequences of real-world personal essays. The shared task is organised into two subtasks: *Subtask 1: Longitudinal Affect Assessment* is the task of predicting continuous valence and arousal values for chronologically ordered sequences; and, *Subtask 2: Forecasting Future Variation in Affect* focuses on predicting both state and dispositional changes based on historical data.

Our team participated in both subtasks, focusing on the comparison of Pre-trained Language Models (PLMs) and Large Language Models (LLMs). While LLMs have demonstrated substantial performance gains across a wide range of NLP tasks (Brown et al., 2020; Touvron et al., 2023), their application to fine-grained, continuous emotion prediction remains underexplored. Existing studies have primarily employed LLMs in

[*]Primary authors for this work.

[1]The code for both the PLM and LLM models used in this shared task is available at https://github.com/jiyuc/semeval2026-task2.

[2]https://www.codabench.org/competitions/9963

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 160–166

zero-shot or few-shot settings for emotion-related tasks (Mendes and Martins, 2023; Muhammad et al., 2025; Fazzi et al., 2025). Notably, Becker et al. (2026) show that fine-tuning LLMs as regression models can outperform few-shot prompting approaches for emotional trajectories. Building on this insight, we investigate the effectiveness of LLM-based regression fine-tuning for modelling continuous valence and arousal trajectories and compare their performance against a PLM baseline. Our analysis aims to better understand the relative strengths and limitations of the regression fine-tuning strategy in longitudinal affect prediction tasks compared to PLMs.

2 Methods

We approach both subtasks as supervised regression tasks comparing PLM and LLM over continuous emotional dimensions (valence and arousal).

2.1 Regression Framework

For both PLM and LLM models, the input sequence at timestep t is first encoded into a contextual representation $\mathbf{h}_t$. Depending on the subtask, additional auxiliary features $\mathbf{g}_t$ may be available (e.g., short-term state changes or phase-level summaries from longer time periods). The final representation fed to the regression head is:

$$\mathbf{z}_t = \begin{cases} \mathbf{h}_t, & \text{(text-only setting)} \\ [\mathbf{h}_t; \|; \phi(\mathbf{g}_t)], & \text{(text + auxiliary features)} \end{cases} \tag{1}$$

where $\phi(\cdot)$ denotes a learnable embedding function and $\|$ represents vector concatenation.

2.2 PLM

We build a regression model based on RoBERTa (Liu et al., 2019), initialised from the TweetNLP sentiment model (Camacho-collados et al., 2022)[3]. The original sentiment classification head is removed, and the contextualised `[CLS]` token representation from the final hidden layer is used as a generic encoding of the input sequence.

Given an input sequence at timestep t, denoted by x_t, the model produces:

$$\mathbf{h}_t = \text{RoBERTa}(x_t)[\texttt{CLS}] \in \mathbb{R}^d. \tag{2}$$

The regression prediction is computed as:

$$\hat{y}_t = \mathbf{w}^\top \mathbf{z}_t + b. \tag{3}$$

Separate models are trained for valence and arousal, resulting in two independent regression models optimised using mean squared error (MSE) loss.

Subtask 1 Each timestep is treated independently of historical information:

$$\mathbf{z}_t^{(1)} = \mathbf{h}_t. \tag{4}$$

Subtask 2a For timestep t, historical features from the preceding window of size n are incorporated. The auxiliary feature vector is defined as:

$$\mathbf{g}_t^{(2A)} = \Big[v_{t-n:t-1},\ a_{t-n:t-1},\ \Delta v_{t-n+1:t-1},\ \Delta a_{t-n+1:t-1},\ \Delta t_{t-n+1:t-1}\Big], \tag{5}$$

where v_t and aa_t denote the valence and arousal at timestep t, $\Delta v_\tau = v_\tau - v_{\tau-1}$ and $\Delta a_\tau = a_\tau - a_{\tau-1}$ for $\tau \in \{t-n+1, \ldots, t-1\}$. At inference time, all historical values $v_{t-n:t-1}$ and $a_{t-n:t-1}$ are gold standard annotations as provided in the shared task data, rather than model-predicted values.

Subtask 2b Let t^* denote the most recent timestep in the first half of a user sequence. Long-term summary statistics are computed over the preceding n timesteps $\{y_{t^*-n}, \ldots, y_{t^*-1}\}$, with $y \in \{v, a\}$:

$$\mathbf{g}^{(2B)} = \Big[\mu_n,\ \text{MSSD}_n,\ \rho_n,\ \sigma_{\text{roll},n},\ \text{Swing}_n,\ \text{Bounce/Inertia}_n\Big], \tag{6}$$

where μ_n is the mean, MSSD_n the mean squared successive difference, ρ_n the lag-1 autocorrelation, $\sigma_{\text{roll},n}$ the rolling standard deviation, and Swing_n and Bounce/Inertia_n measure volatility and directional persistence.

2.3 LLM

We fine-tune `Gemma3` (Team et al., 2025) 270M for joint valence-arousal regression[4]. The model

[3] https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment

[4] We also evaluated larger models from both the same and different families on the validation set; however, these achieved lower performance than the Gemma3 270M model in initial experiments.

consists of a frozen pre-trained LLM backbone and task-specific regression heads.

Given an input sequence at timestep t, denoted by x_t, the LLM backbone produces contextualised hidden representations:

$$\mathbf{H}_t = \text{LLM}(x_t) \in \mathbb{R}^{L \times d}, \tag{7}$$

where L is the sequence length and d the hidden dimension. A masked mean pooling operation produces a global representation:

$$\mathbf{h}_t = \frac{\sum_{i=1}^{L} m_i \mathbf{H}_{t,i}}{\sum_{i=1}^{L} m_i}. \tag{8}$$

The regression module consists of two parallel projection heads operating on the pooled representation $\mathbf{h}_t$:

$$\begin{aligned} \hat{v}_t &= 2 \cdot \tanh(\mathbf{W}_v \mathbf{h}_t + b_v), \\ \hat{a}_t &= 2 \cdot \sigma(\mathbf{W}_a \mathbf{h}_t + b_a), \end{aligned} \tag{9}$$

where $\hat{v}_t \in [-2, 2]$ and $\hat{a}_t \in [0, 2]$. The model is trained jointly using MSE loss over both dimensions.

Subtask 1 Each timestep is treated independently without historical information:

$$\mathbf{z}_t^{(1)} = \mathbf{h}_t. \tag{10}$$

Subtask 2a For timestep t, short-term affective dynamics from the preceding window of size n are incorporated as:

$$\begin{aligned} \mathbf{g}_t^{(2A)} = \Big[&(v_{t-n}, a_{t-n}, \Delta v_{t-n}, \Delta a_{t-n}), \ldots, \\ &(v_{t-1}, a_{t-1}, \Delta v_{t-1}, \Delta a_{t-1}) \Big], \end{aligned} \tag{11}$$

where $\Delta v_t = v_t - v_{t-1}$ and $\Delta a_t = a_t - a_{t-1}$; the trajectory and most recent state (v_{t-1}, a_{t-1}) are encoded to predict (v_t, a_t). At inference time, all historical values $v_{t-n:t-1}$ and $a_{t-n:t-1}$ are gold standard annotations as provided in the shared task data, rather than model-predicted values.

Subtask 2b For each user, phase-level mean valence and arousal are computed. Let $\mu_v^{(k)}$ and $\mu_a^{(k)}$ denote the average valence and arousal in phase k. The target is the dispositional change between consecutive phases:

$$\mathbf{g}^{(2B)} = \left[\mu_v^{(k+1)} - \mu_v^{(k)},\ \mu_a^{(k+1)} - \mu_a^{(k)} \right]. \tag{12}$$

3 Experimental Setup

3.1 Shared Task Dataset

The dataset used in the shared task consists of 5,285 texts ("ecological-essays") written by 182 authors, collected over multiple years (2021-2024). The training data contains 2,764 texts from 137 users. For Subtask 1, the evaluation data consists of 1,737 texts written by 91 users. For Subtask 2, the test data contains the records for the same users in the training data.

Dataset Augmentation Assigning emotion words to feelings is a psychologically rich process, shaped by how past experiences influence current emotional perception (Soni et al., 2026). Free-form textual descriptions and explicit feeling words, therefore, capture complementary aspects of affective experience: text conveys contextual and narrative information, while feeling words provide concise and direct emotional labels. The shared task dataset contains either free-form textual description (e.g., *I am generally feeling good this morning.*) or, in some instances, explicitly feeling words (indicated by `is_words = True`) (e.g., *Congested, Sick, Tired, Relaxed, Mellow*). To ensure that both types of aspects are available for every sample, we apply controlled data augmentation to generate the missing type of aspect. Specifically, when `is_words = False`, we generate the corresponding feeling words from the textual description. Conversely, when `is_words = True`, we generate a textual description from the provided feeling words.

Augmentation is applied to both the training and test sets of the shared task. As a result, each instance in the augmented sets contains both a complete free-form textual description field and a complete feeling field. Augmentation is performed using GPT-OSS:20B (OpenAI, 2025) (`temperature=0.0`, `num_ctx=2000`, and `think=low`).

The prompt used to generate text for the given feeling words is as follows:

```
You are an emotion analysis assistant.

Given the following feelings: {FEELINGS}

Generate a short, descriptive text (2-3
sentences) summarising the emotional state of
a person with these feelings, considering
both valence (positive/negative) and arousal
(high/low) dimensions.
```

```
Example input: Calm, Tired
Example output: "The person feels relaxed and
at ease, but also a bit drained and low on
energy."
```

The prompt used to generate feeling words for the given text is as follows:

```
Extract the main feelings or emotions from
    the following text.

Return them as a simple comma-separated list
with short phrases (1-3 words each).

Text: {TEXT}

Example output: Calm, Tired, Slightly anxious
```

3.2 Evaluation Metrics

For both subtasks, we follow the evaluation metrics specified by the shared task organisers.[5] The metrics are Pearson's correlation coefficient (r) and Mean Absolute Error (MAE) for each outcome: valence and arousal.

3.3 Hyperparameters

PLM: All PLM-based models use a maximum sequence length of 512 and are optimised with AdamW. Hyperparameters are selected via cross-validation on the training data. We intentionally adopt different configurations across subtasks to reflect their varying learning dynamics and feature complexity.

- *Subtask 1.* Hyperparameters are selected via cross-validation (5-fold) on the training set. Learning rate 2×10^{-5}, weight decay 0.01, batch size 16, 4 epochs, and dropout $p = 0.2$ in the regression head. The final model is trained on the full training data using the selected configuration.
- *Subtask 2a.* We use 5-fold user-level cross-validation and select the checkpoint with the lowest validation MSE ($3^{rd} fold$). The model is trained with a learning rate 2×10^{-5}, weight decay 0.01, batch size 16, and 5 epochs, incorporating historical features from the previous $n = 5$ posts.
- *Subtask 2b.* We use 5-fold user-level cross-validation and select the checkpoint with the lowest validation MSE ($3^{rd} fold$). The batch size remains 16 with weight decay 0.01, while the learning rate is increased to 1×10^{-3}. Models are trained for 5 epochs with a maximum sequence length of 512. Longer-term history is incorporated using $n = 10$ previous posts.

[5] https://github.com/semeval2026task2/EmotionValArouTimeVariation2026

Team	Valence (V)	Arousal (A)	V&A Avg.
Subtask 1			
Max	0.667	0.554	0.611
Avg.	0.635	0;440	0.538
Min	0.581	0.256	0.418
CSIRO-LT	0.656	0.488	0.572
linear(BERT)	0.557	0.299	0.428
rand	0.000	0.000	0.000
Subtask 2a			
Max	0.675	0.683	0.679
Avg.	0.370	0.329	0.350
Min	-0.272	-0.274	-0.273
CSIRO-LT	0.621	0.477	0.545
linear (BERT; prev)	0.430	0.405	0.418
linear (BERT)	0.290	0.199	0.245
rand	0.000	0.000	0.000
Subtask 2b			
Max	0.405	0.602	0.503
Avg.	0.017	0.179	0.098
Min	-0.398	-0.576	-0.487
CSIRO-LT	-0.147	0.114	-0.016
rand	0.000	0.000	0.000
linear (prev)	0.434	0.584	0.509
linear (BERT; prev)	-0.029	0.019	-0.005
linear (BERT)	-0.088	0.07	-0.009

Table 1: Leaderboard of SemEval 2026 Task 2. For each subtask, we report the Max, Min and Avg. r scores of all official submissions alongside our system (CSIRO-LT). The baseline systems are those under the dotted lines.[6]

LLM: For all subtasks, the training data is split into training and validation partitions based on user identities. The best model is selected based on the R^2 score, which jointly penalises bias and variance and showed more stable model ranking than Pearsons's r or MAE during pilot experiments on the validation set. Hyperparameter tuning for each subtask was performed using the validation set.

We used a batch size of 4, with a maximum sequence length of 512. The Gemma3 270M backbone was kept frozen during training; only the task specific regression heads were updated. The model was fine-tuned for 5 epochs using a learning rate of $5e-5$ and L2 weight decay of $1e-4$ to reduce overfitting. For features in Subtask 2a, the historical features from the previous $n = 5$ posts are used.

4 Experimental Results

Our official submission (CISRO-LT) is the PLM (RoBERTa) model. The leaderboard of the shared

	Valence (V)						Arousal (A)					
Task / Model	r_c (↑)	r_b (↑)	r_w (↑)	mae_c (↓)	mae_b (↓)	mae_w (↓)	r_c (↑)	r_b (↑)	r_w (↑)	mae_c (↓)	mae_b (↓)	mae_w (↓)
Subtask 1												
UKP_Psycontrol	0.667	0.761*	0.546*	0.595	0.402	0.738	0.554	0.701*	0.363*	0.345	0.210	0.467
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
rand (mean)	0.000	0.028	0.000	0.000	0.627	1.041	0.000	0.096	0.000	0.488	0.326	0.622
PLM (RoBERTa)	0.656	0.721*	0.580*	0.654	0.440	0.799	0.488	0.547*	0.425*	0.401	0.253	0.531
LLM (original)	0.561	0.635*	0.447*	0.765	0.525	0.890	0.405	0.418*	0.374*	0.435	0.286	0.552
LLM (feeling)	0.598	0.670*	0.516*	0.743	0.505	0.876	0.433	0.457*	0.408*	0.420	0.275	0.547
LLM (text)	0.542	0.604*	0.473*	0.770	0.564	0.886	0.391	0.415*	0.367*	0.427	0.283	0.553
LLM (feeling + text)	0.618	0.698*	0.523*	0.704	0.450	0.840	0.356	0.378*	0.334*	0.437	0.282	0.570
Subtask 2a												
	r (↑)			mae (↓)			r (↑)			mae (↓)		
YNU	0.692*			1.074			0.647*			0.641		
linear (BERT)	0.290*			1.294			0.199*			0.744		
linear (BERT; prev)	0.430*			1.251			0.405*			0.708		
linear (prev)	0.615*			1.168			0.670*			0.638		
rand (zero)	0.000			1.261			0.000			0.696		
PLM (RoBERTa)	0.621*			1.190			0.477*			0.740		
LLM (original)	0.660*			1.195			0.668*			0.629		
LLM (feeling)	0.659*			1.126			0.677*			0.628		
LLM (text)	0.681*			1.105			0.678*			0.632		
LLM (feeling + text)	0.669*			1.121			0.684*			0.626		
Subtask 2b												
linear (prev)	0.434*			0.406			0.584*			0.286		
UAlberta	0.405*			0.635			0.602*			0.261		
rand (zero)	0.000			0.417			0.000			0.296		
linear (BERT; prev)	-0.029			0.436			0.019			0.305		
linear (BERT)	-0.088			0.438			0.070			0.303		
PLM (RoBERTa)	-0.147			0.593			0.114			0.373		
LLM (original)	-0.080			0.502			0.394*			0.345		
LLM (feeling)	0.063			0.558			-0.024			1.268		
LLM (text)	-0.081			0.548			-0.036			0.432		
LLM (feeling + text)	0.093			0.622			0.199			0.340		

Table 2: Effectiveness of baseline systems (linear models and rand), of ranked 1^{st} systems (UKP_Psycontrol, YNU, and UAlberta), and of our approaches (PLM and LLM) on the test set for all subtasks. The results are provided by the task organisers. * indicates $p < 0.01$ (significantly different from 0). Note: $r_c = r_{composite}$, $r_b = r_{between}$, $r_w = r_{within}$, $mae_c = mae_{composite}$, $mae_b = mae_{between}$, and $mae_w = mae_{within}$.

task is given in Table 1.

We present the detailed evaluation results of our systems in Table 2. The table includes the baseline results for each subtask as provided by the organisers, the best-performing systems (ranked 1^{st} on the leaderboard), in addition to our systems for comparison. For LLM based approaches, we evaluate four input conditions: *LLM (original)* uses the raw shared task input as provided (either text or feeling words, whichever was available for a given instance); *LLM (feeling)* uses augmented feeling words, where missing feeling fields are generated from the available text; *LLM (text)* uses augmented text, where missing text fields are generated from the available feeling words; and *LLM (feeling + text)* combines both augmented fields.

Subtask 1. Our PLM (RoBERTa) approach outperforms both the baseline linear (BERT) model and our LLM-based approach. Fine-tuning a sentiment-aware RoBERTa backbone provides a clear advantage for predicting valence and arousal. For the LLM-based approach, we compare three aspects: text, feeling, and a combined representation (text concatenated with feeling words) (see 3.1 for data augmentation for details).

Compared to the LLM (original), all augmented variants generally show improvement across both dimensions. The LLM (feeling) outperforms LLM (text) for both valence and arousal, suggesting that concise, emotion-focused representations help the model capture affective signals while reducing contextual noise. Combining text and feeling words (LLM (feeling+text)) further improves valence prediction, showing that integrating both contextual and emotion-focused information benefits the modelling of positive–negative fluctuations. For arousal, however, the combined input does not outperform either feeling words or text alone, suggesting that concise emotion signals or raw textual context are more informative than the full combined input for predicting arousal. The performance of the PLM over the linear (BERT) baseline likely also stems from its sentiment-aware pretraining, which aligns well with the affective cues captured by the feeling inputs of the LLM.

Subtask 2a. Our systems consistently outperform the baseline linear (BERT) model for both valence and arousal. Compared to the PLM-based approaches, the LLM-based models demonstrate stronger predictive performance overall. For valence, LLM (text) achieves the highest scores, while for arousal, LLM (feeling+text) achieves better performance than LLM (text). Both suggest that preserving the full textual context is advantageous for this task.

Interestingly, this trend contrasts with the findings from Subtask 1. While feeling representations were beneficial in the previous setting, here, using the text input either alone or combined with feeling expressions consistently leads to better performance for LLM-based models. This shows that Subtask 2a likely requires richer contextual and discourse-level information that may be partially lost when the input is compressed into a simplified emotional summary.

Subtask 2b. Subtask 2b proves to be substantially more challenging, as evidenced by the generally lower and in some cases negative correlation scores across models. Here, the simpler linear (prev) baseline model, which relies on previous signals, outperforms all neural submissions, including ours. We attribute this to several factors. First, dispositional change is defined at the phase level, which substantially reduces the number of training instances per user and limits the ability of data-hungry neural models to generalise. Second, the target—shift in mean affect across phases—is only weakly coupled to the surface features of any single essay, making local textual representations insufficient predictors. Third, strong autocorrelation in affect trajectories means that prior affect levels are highly informative, giving the linear baseline a structural advantage that text-based models cannot easily overcome without explicit temporal modelling.

Nonetheless, for arousal, LLM (original) stands out as the only and the best-performing among our LLM-based models for *arousal*. The mixed results across input representations, with LLM (feeling+text) outperforming LLM (text) and LLM (feeling) in some instances and vice versa, suggest that no single static input formulation might be optimal for this task, and that instance-level adaptation of the input representation may be necessary to better capture the signals underlying dispositional affect change.

5 Conclusions

We participated in the *SemEval 2026 Task 2: Predicting Variation in Emotional Valence and Arousal over Time from Ecological Essays* shared task, which included tracks for continuous valence and arousal prediction (Subtask 1) and forecasting state and dispositional changes based on historical data (Subtask 2). For both subtasks, we approached them as regression tasks and compared the performance of PLMs and LLMs. We found that sentiment-aware PLM performs well in Subtask 1 but is less effective than LLM in Subtask 2a, which involves forecasting the state. For Subtask 2b, we could not achieve a stable model.

Limitations and Future Work

Our comparison of PLMs and LLMs is limited by differences in input features and model initialisation, which means observed performance gaps may reflect representational choices as much as architectural ones. We tested only a single PLM RoBERTa and a single LLM (Gemma3 270M), leaving open questions about the sensitivity of results to model scale and family. We did not perform controlled experiments comparing PLMs with and without sentiment-aware pretraining, nor did we evaluate PLMs and LLMs on the same feature sets. Our approaches treat each timestep with static encoders and do not model temporal dependencies explicitly; sequence-aware architectures such as LSTMs or temporal transformers were not explored, which may explain part of the performance gap in Subtask 2b. Finally, we did not systematically evaluate the sensitivity of results to the history window size n and to the LLM prompt design for data augmentation.

References

Jonas Becker, Liang-Chih Yu, Shamsuddeen Hassan Muhammad, Jan Philip Wahle, Terry Ruas, Idris Abdulmumin, Lung-Hao Lee, Nelson Odhiambo, Lilian Wanzare, Wen-Ni Liu, Tzu-Mi Lin, Zhe-Yu Xu, Ying-Lung Lin, Jin Wang, Maryam Ibrahim Mukhtar, Bela Gipp, and Saif M. Mohammad. 2026. Dimstance: Multilingual datasets for dimensional stance analysis. *Preprint*, arXiv:2601.21483.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu,

Clemens Winter, and 12 others. 2020. Language models are few-shot learners.

Jose Camacho-collados, Kiamehr Rezaee, Talayeh Riahi, Asahi Ushio, Daniel Loureiro, Dimosthenis Antypas, Joanne Boisson, Luis Espinosa Anke, Fangyu Liu, Eugenio Martínez Cámara, and 1 others. 2022. TweetNLP: Cutting-Edge Natural Language Processing for Social Media. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–49, Abu Dhabi, UAE. Association for Computational Linguistics.

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.

Gino Franco Fazzi, Julie Skoven Hinge, Stefan Heinrich, and Paolo Burelli. 2025. Don't Get Too Excited - Eliciting Emotions in LLMs. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '25. Association for Computing Machinery.

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.

Gonçalo Azevedo Mendes and Bruno Martins. 2023. Quantifying valence and arousal in text with multilingual pre-trained transformers. In *European Conference on Information Retrieval*, pages 84–100. Springer.

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. Association for Computational Linguistics.

Shamsuddeen Hassan Muhammad, Nedjma Ousidhoum, Idris Abdulmumin, Jan Philip Wahle, Terry Ruas, Meriem Beloucif, Christine de Kock, Nirmal Surange, Daniela Teodorescu, Ibrahim Said Ahmad, David Ifeoluwa Adelani, Alham Fikri Aji, Felermino D. M. A. Ali, Ilseyar Alimova, Vladimir Araujo, Nikolay Babakov, Naomi Baes, Ana-Maria Bucur, Andiswa Bukula, and 29 others. 2025. BRIGHTER: BRIdging the Gap in Human-Annotated Textual Emotion Recognition Datasets for 28 Languages. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8895–8916, Vienna, Austria. Association for Computational Linguistics.

OpenAI. 2025. gpt-oss-120b gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161.

Alieh Hajizadeh Saffar, Tiffany Katharine Mann, and Bahadorreza Ofoghi. 2023. Textual emotion detection in health: Advances and applications. *Journal of Biomedical Informatics*, 137:104258.

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.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 Technical Report. *Preprint*, arXiv:2503.19786.

Daniela Teodorescu, Tiffany Cheng, Alona Fyshe, and Saif Mohammad. 2023. Language and mental health: Measures of emotion dynamics from text as linguistic biosocial markers. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3117–3133. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models.

Xingwei Yang and Guang Li. 2025. Psychological and behavioral insights from social media users: Natural language processing–based quantitative study on mental well-being. *JMIR Formative Research*, 9.