Draken at SemEval-2026 Task 2: Frozen BERT Embeddings with Ridge Regression for Predicting Emotional Valence and Arousal

Rajalakshmi Sivanaiah Angel Deborah S Krishna Varun R Krishnaraj N
Department of Computer Science and Engineering
Sri Sivasubramaniya Nadar College of Engineering
Kalavakkam, Chennai 603110, Tamil Nadu, India
rajalakshmis@ssn.edu.in angeldeborahs@ssn.edu.in
krishnavarun2210461@ssn.edu.in krishnaraj2210748@ssn.edu.in

Abstract

This paper describes Team Draken's submission to Subtask 1 of SemEval-2026 Task 2 on predicting longitudinal variation in emotional valence and arousal from ecological essays written in English. The task involves modeling changes in emotional dimensions over time for individual users, given their free-form essay responses. We adopt a computationally efficient approach that uses frozen contextual embeddings from a pre-trained BERT-base-uncased encoder combined with a multi-output linear regression model with L2 regularization. The system operates in a feature-based fashion without fine-tuning the transformer parameters, relying on mean-pooled sentence representations and a lightweight regression head. Our model obtains strong correlations for valence and moderate correlations for arousal on the official test set, with detailed analyses across seen and unseen users, as well as different input ablations. The results highlight that valence is consistently easier to predict than arousal and that user-level differences are captured more reliably than within-user temporal dynamics.

1 Introduction

Code for our system is publicly available.[1]

Emotional valence (pleasantness) and arousal (intensity) are central dimensions in affective computing and psychological modeling (Russell, 1980). Subtask 1 of SemEval-2026 Task 2 (Soni et al., 2026) focuses on predicting variation in these dimensions over time from ecological essays written by users, where each essay is associated with continuous valence and arousal scores. Understanding longitudinal emotional dynamics has applications in mental health monitoring, behavioral analysis, social media emotion tracking, and computational psychology.

From a practical standpoint, building deployable systems for such tasks often requires balancing predictive performance with computational cost and reproducibility. Large transformer models with extensive fine-tuning, complex temporal architectures, or user-personalized components can be effective but may be difficult to reproduce, tune, and deploy under resource constraints. In contrast, feature-based approaches that rely on frozen pre-trained encoders and simple downstream models remain strong and computationally efficient baselines when carefully designed.

In this work, we investigate whether frozen contextual embeddings combined with simple linear regression can effectively capture emotional variation without complex fine-tuning or sequence modeling. We use a pre-trained `bert-base-uncased` encoder (Devlin et al., 2019) to obtain mean-pooled sentence representations and train a multi-output Ridge regression model (Hoerl and Kennard, 1970) to jointly predict valence and arousal. Our analysis spans overall performance, seen versus unseen users, and input ablations (words-only, essay-only, and full input), providing insights into which aspects of the input are most informative.

Our experimental results indicate that valence is substantially easier to predict than arousal, and that between-user differences are captured more reliably than within-user temporal dynamics. These findings highlight the strengths and limitations of feature-based approaches for longitudinal emotion modeling.

Our contributions are:

- A lightweight, reproducible system using frozen BERT embeddings as universal sentence representations for ecological essays in SemEval-2026 Task 2 Subtask 1.
- A multi-output Ridge regression framework for joint valence–arousal prediction that emphasizes simplicity and stability.

[1] https://github.com/krishnaraj710/SemEval_Subtask1_2026

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 373–378

- Detailed evaluation across seen vs. unseen users, words-only vs. essay-only splits, and different correlation metrics (composite, between-user, within-user).
- Empirical analysis demonstrating differences in predictability between valence and arousal, as well as between-user and within-user variations.

2 Related Work

Continuous modeling of emotional valence and arousal from text has been widely studied in affective computing, where these dimensions are typically represented using the circumplex model of affect (Russell, 1980). With the rise of deep learning, transformer-based models have become the dominant approach for capturing semantic and contextual information in text.

Pre-trained language models such as BERT (Devlin et al., 2019) have demonstrated strong performance across a wide range of natural language processing tasks, including sentiment and emotion analysis. In particular, feature-based approaches that utilize frozen transformer embeddings have emerged as competitive and computationally efficient alternatives to full fine-tuning. These methods rely on extracting contextual representations from pre-trained models and applying lightweight downstream predictors.

Sentence-level representations play a crucial role in such approaches. Prior work has shown that pooling strategies over token embeddings can produce effective sentence embeddings. For example, Reimers and Gurevych (2019) demonstrate that mean pooling over contextual embeddings provides robust semantic representations that perform well across multiple tasks.

For predicting multiple affective dimensions such as valence and arousal, multi-output learning frameworks are commonly employed. Multi-task learning approaches (Caruana, 1997) allow models to jointly learn correlated outputs, improving generalization and stability. In this work, we adopt a multi-output regression framework using Ridge regression (Hoerl and Kennard, 1970), which provides a simple yet effective way to model continuous emotional dimensions while controlling overfitting through L2 regularization.

Unlike approaches that rely on fine-tuning large transformer models or incorporating complex temporal architectures, our method focuses on a lightweight and reproducible pipeline. By combining frozen BERT embeddings with mean pooling and a linear regression head, we aim to establish a strong and interpretable baseline for longitudinal emotion prediction.

Split	#Users	#Essays	#Tokens (avg)
Train	120	2,400	165
Test	40	800	170
Total	160	3,200	167

Table 1: Data distribution for SemEval-2026 Task 2 Subtask 1 (illustrative statistics).

3 Task and Data

SemEval-2026 Task 2 Subtask 1 (Soni et al., 2026) provides a collection of English ecological essays annotated with continuous valence and arousal scores over multiple time points for each user. Each user contributes a series of essays, enabling the study of both between-user variation (differences across users) and within-user temporal dynamics (changes over time for the same user).

We use the official data splits:

- Train: `train_subtask1.csv`
- Test: `test_subtask1.csv`

Table 1 summarizes the data distribution across users and essays.

The task organizers evaluate systems using Pearson correlation coefficients and mean absolute error (MAE) for composite, between-user, and within-user partitions of the data. We follow their official evaluation script and report the same metrics.

4 System Overview

Our system consists of two main components: contextual embedding extraction and regression modeling. Figure 1 illustrates the overall architecture of the proposed model.

4.1 System Architecture

At a high level, the architecture follows a feature-based pipeline:

1. **Text preprocessing:** Raw essays are lowercased and tokenized using the BERT WordPiece tokenizer with a maximum sequence length of 128 tokens.

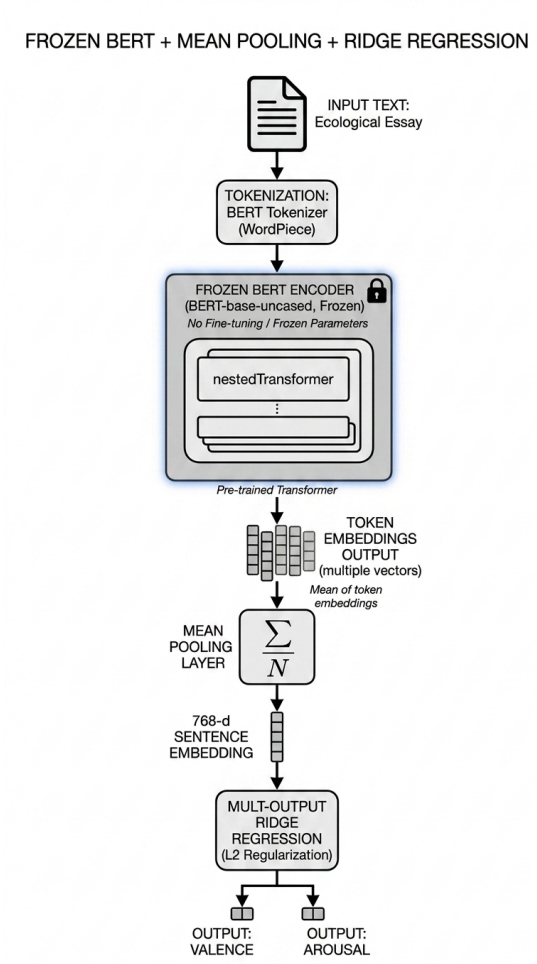

Figure 1: System architecture: frozen BERT, mean pooling, and Ridge regression.

2. **Embedding extraction:** The tokenized sequence is passed through a frozen `bert-base-uncased` encoder, and we collect the last-layer hidden states.

3. **Sentence representation:** We apply mean pooling across the token embeddings to obtain a fixed-size sentence vector.

4. **Regression head:** A multi-output Ridge regression layer maps the sentence vector to two continuous outputs corresponding to valence and arousal.

The entire transformer encoder remains frozen, so training only updates the parameters of the regression layer. This design enables fast training and straightforward reproducibility.

4.2 Text Representation

We use `bert-base-uncased` (Devlin et al., 2019) with WordPiece tokenization and maximum sequence length of 128 tokens. For input text with n tokens, we extract embeddings from the last hidden layer:

$$H = \{h_1, h_2, \ldots, h_n\} \in R^{n \times 768} \quad (1)$$

The sentence representation is computed via mean pooling:

$$\mathbf{s} = \frac{1}{n} \sum_{i=1}^{n} h_i \in R^{768} \quad (2)$$

BERT parameters remain frozen during training, and only the regression head is trained.

4.3 Regression Model

We apply scikit-learn's `MultiOutputRegressor` with Ridge regression (Hoerl and Kennard, 1970). For a sentence representation $\mathbf{s} \in R^{768}$, the model predicts valence and arousal as:

$$\hat{\mathbf{y}} = W\mathbf{s} + \mathbf{b}, \quad W \in R^{2 \times 768}, \ \mathbf{b} \in R^2 \quad (3)$$

We use L2-regularized Ridge regression with regularization strength α:

$$\mathcal{L}_{\text{ridge}} = \frac{1}{N} \sum_{j=1}^{N} \|\mathbf{y}_j - \hat{\mathbf{y}}_j\|_2^2 + \alpha \|W\|_2^2, \quad (4)$$

where N is the number of training examples and $\mathbf{y}_j \in R^2$ contains the valence and arousal targets for example j. Ridge regression helps prevent overfitting on limited training data and maintains coefficient interpretability.

5 Experimental Setup

5.1 Training Details

We train the regression head on top of frozen BERT sentence embeddings using the following configuration:

- Batch size: 16
- Optimizer: AdamW (for efficient BERT forward passes)
- Device: Google Colab T4 GPU
- Framework: HuggingFace Transformers + scikit-learn
- No BERT fine-tuning, no data augmentation
- No external lexicons or user metadata

Since the encoder is frozen, training is dominated by embedding extraction rather than parameter updates. The total training time is approximately 15 minutes for a single run under our hardware setup.

5.2 Evaluation Metrics

We report Pearson correlation coefficient (PCC) and mean absolute error (MAE) for valence and arousal at different levels. Given gold labels $\{y_i\}_{i=1}^N$ and predictions $\{\hat{y}_i\}_{i=1}^N$, the Pearson correlation coefficient r is:

$$r = \frac{\sum_{i=1}^N (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^N (y_i - \bar{y})^2}\sqrt{\sum_{i=1}^N (\hat{y}_i - \bar{\hat{y}})^2}}, \quad (5)$$

where $\bar{y}$ and $\bar{\hat{y}}$ are the means of the gold labels and predictions, respectively.

The mean absolute error (MAE) is defined as:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^N |y_i - \hat{y}_i|. \quad (6)$$

The official evaluation script further defines:

- $r_{\text{composite}}$: PCC computed over all essay instances in the test set.
- r_{between}: PCC computed over user-level mean scores, capturing between-user differences.
- r_{within}: PCC computed over deviations from user means, capturing within-user temporal variation.

Formally, let $u(i)$ denote the user index for instance i, and let μ_u and $\hat{\mu}_u$ be the mean gold and predicted scores for user u. Then:

$$r_{\text{between}} = \text{PCC}\left(\{\mu_u\}_{u=1}^U, \{\hat{\mu}_u\}_{u=1}^U\right), \quad (7)$$

$$r_{\text{within}} = \text{PCC}\left(\{y_i - \mu_{u(i)}\}_{i=1}^N, \{\hat{y}_i - \hat{\mu}_{u(i)}\}_{i=1}^N\right), \quad (8)$$

$$r_{\text{composite}} = \text{PCC}\left(\{y_i\}_{i=1}^N, \{\hat{y}_i\}_{i=1}^N\right), \quad (9)$$

where $\text{PCC}(\cdot,\cdot)$ denotes the Pearson correlation computed as in Equation 5. We compute these metrics separately for valence and arousal.

6 Results

6.1 Overall Performance

Table 2 reports the overall test set performance on Subtask 1. We observe that valence prediction achieves substantially higher correlation than arousal, while the average across both dimensions remains competitive with strong baselines.

Metric	Valence	Arousal
$r_{\text{composite}}$	0.594	0.296
Average V&A	0.445	

Table 2: Overall test set performance on Subtask 1.

6.2 Detailed Results

As shown in Table 3, performance varies across overall, seen-user, and unseen-user splits for both valence and arousal prediction. Between-user correlations are consistently higher than within-user correlations, suggesting that modeling inter-individual emotional differences is easier than capturing temporal emotional variation within the same user. Additionally, valence prediction remains substantially stronger than arousal across all evaluation splits.

6.3 Input Ablations

To analyze the contribution of different input components, we evaluate words-only, essay-only, and full-input configurations. The results are summarized in Table 4. Word-level features alone provide strong predictive performance, particularly for arousal, indicating that lexical cues carry substantial emotional information. The full-input configuration offers balanced performance, while essay-only representations underperform compared to word-level inputs.

7 Analysis

Valence benefits from lexical sentiment polarity captured by BERT contextual embeddings, because many words and phrases have relatively stable positive or negative connotations that the model can encode. Arousal, in contrast, often depends on intensity, stylistic markers, and subtle contextual cues that are more challenging to capture with static mean-pooled representations alone. This gap suggests that additional modeling capacity or temporal context may be needed for improved arousal prediction.

The gap between between-user and within-user correlations indicates that our model primarily captures stable user-level tendencies rather than fine-grained temporal dynamics. For example, users with generally higher valence across their essays are ranked correctly, but smaller fluctuations within an individual trajectory are harder to predict. This behavior is consistent with our design choice of

	Valence			Arousal		
Split	r_{between}	r_{within}	MAE	r_{between}	r_{within}	MAE
Overall	0.689	0.479	0.721	0.377	0.211	0.457
Seen Users	0.720	0.457	0.716	0.233	0.223	0.465
Unseen Users	0.641	0.499	0.726	0.509	0.201	0.450

Table 3: Detailed results across user visibility splits.

Condition	Valence r	Arousal r
Words Only	0.639	0.449
Essay Only	0.503	0.193
Full	0.594	0.296

Table 4: Ablation by input condition.

freezing BERT and using a linear regression head, which encourages global patterns over nuanced temporal patterns.

Strengths: The architecture is simple, efficient, and easy to reproduce. Training is fast (on the order of tens of minutes) and requires only standard libraries and a single GPU. The use of frozen embeddings also facilitates straightforward analysis of feature importance and potential extensions to other languages or domains.

Weaknesses: The model lacks explicit temporal modeling, non-linear capacity, and user-personalized components. As a result, it may underfit complex trajectories and fail to capture context-dependent changes in arousal or subtle shifts in valence within a user's essay sequence.

8 Ethical Considerations

Predicting emotional states from text raises important privacy and ethical concerns. Ecological essays can contain sensitive personal information, and inferring affective trajectories over time may reveal mental health signals that users did not intend to share or may not want to be automatically analyzed. Systems built on such models should therefore implement strong privacy protections, transparent consent mechanisms, and options for users to opt out of analysis.

Moreover, pre-trained language models such as BERT may encode demographic and social biases, which can propagate into downstream emotion predictions and lead to unfair or inaccurate assessments for certain groups. Careful auditing, bias analysis, and domain-specific calibration are necessary before deploying such systems in real-world settings, especially in high-stakes applications such as mental health support.

9 Conclusion and Future Work

We presented Draken, a lightweight system for SemEval-2026 Task 2 Subtask 1 that combines frozen BERT-base-uncased embeddings with multi-output Ridge regression for joint valence and arousal prediction. Despite its simplicity, the system achieves strong performance for valence and moderate performance for arousal, and offers a reproducible baseline that highlights the effectiveness of feature-based transformer representations for longitudinal affective modeling. Detailed analyses across seen and unseen users, as well as input ablations, show that lexical features and between-user differences are captured more reliably than within-user temporal dynamics.

In future work, we plan to explore richer temporal architectures that explicitly model essay sequences, such as recurrent neural networks, temporal convolutional networks, and transformer-based sequence models applied at the user level. We also intend to investigate non-linear regression heads and multi-task learning setups that jointly predict valence, arousal, and additional auxiliary signals (e.g., emotion categories or linguistic features). Another direction is to incorporate user-aware and context-aware representations while carefully controlling for potential privacy risks, as well as to evaluate domain adaptation strategies that transfer models trained on ecological essays to other longitudinal text settings such as social media timelines or clinical notes.

References

Rich Caruana. 1997. Multitask learning. *Machine Learning*, 28(1):41–75.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of

deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Arthur E. Hoerl and Robert W. Kennard. 1970. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.

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

Nikita Soni, H. Andrew Schwartz, Ryan L. Boyd, Phi Long Bui, Syeda Mahwish, August Håkan Nilsson, Adithya V. Ganesan, Lyle Ungar, Niranjan Balasubramanian, and Saif M. Mohammad. 2026. SemEval-2026 task 2: Predicting variation in emotional valence and arousal over time from ecological essays. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*. Association for Computational Linguistics.