Khaleesiyali at SemEval-2026 Task 2: Lexicon-Augmented RoBERTa for Valence–Arousal Regression on Ecological Essays

Eleale Nusi Tee
Waseda University
tee.eleale@akane.waseda.jp

Abstract

This paper presents a lexicon-augmented RoBERTa system for the SemEval-2026 Task 2 valence–arousal regression challenge. The model integrates deep contextual embeddings with a 6-dimensional feature vector derived from the NRC VAD lexicon, achieving a high token coverage rate of 72.05%. Under official user-aware evaluation, the system reached a competitive average composite correlation of 0.547, significantly outperforming the ridge-regression baseline. The system demonstrated particular robustness in valence ($r = 0.656$) and achieved strong generalization to unseen users ($r_{\text{arousal}} = 0.519$). These findings indicate that lightweight lexicon-based statistics provide valuable complementary cues for longitudinal emotion modeling in modern transformer architectures.

1 Introduction

Emotions are often modeled in a continuous valence–arousal space, where valence captures the degree of positivity or negativity of an experience and arousal reflects its level of activation or intensity. This dimensional view has been widely adopted in affective computing and natural language processing (NLP), and has motivated the construction of several resources that provide valence–arousal (and dominance) ratings at the word or text level (Warriner et al., 2013; Mohammad, 2018; Buechel and Hahn, 2017). In parallel, transformer-based language models have become the de facto standard for text-based emotion detection, offering strong performance across a range of sentiment and affect recognition benchmarks (Acheampong et al., 2021; Mendes and Martins, 2023).

SemEval-2026 Task 2, Subtask 1 (Soni et al., 2026) frames dimensional emotion prediction as a valence–arousal regression problem on anonymized ecological essays. Given a short essay describing an everyday experience, the goal is to predict continuous valence and arousal scores for each text, while evaluation is carried out with user-aware metrics that combine between-user and within-user statistics. Such a setup is closely related to recent work on tracking emotion dynamics and well-being over time from free-form language (Sun et al., 2020; Teodorescu et al., 2023) and to growing interest in personalized affective models (Li et al., 2024; Han et al., 2024; Kargarandehkordi et al., 2024).

The main strategy of this system is to fine-tune a RoBERTa encoder for joint valence–arousal regression and to augment its contextual representation with lightweight, interpretable features derived from the NRC VAD lexicon (Mohammad, 2018). For each essay, word-level valence and arousal scores are summarized into a six-dimensional vector of simple statistics, which is concatenated to the pooled RoBERTa representation before prediction. This lexicon-augmented design aims to combine the strengths of deep contextual modeling with continuous affective priors, while keeping the overall architecture compact and data-efficient.

Lexicon-based sentiment and emotion analysis methods have sometimes been criticized for limited accuracy in fine-grained prediction settings, particularly when compared to end-to-end neural approaches. At the same time, such lexicon-based methods have been shown to be highly useful and robust in interdisciplinary applications and at higher levels of granularity, especially when qualitative or large-scale supervised modeling is not feasible (Öhman, 2021). In this context, the present system adopts lexicon features not as a standalone method, but as a complementary signal layered on top of a fine-tuned transformer encoder.

On the official Subtask 1 test set, the resulting RoBERTa+Lex system substantially outperforms the organizers' baseline based on ridge regression over frozen BERT embeddings. Clear gains are observed in user-aware composite correlation and

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 522–527

mean absolute error for both valence and arousal, with particularly strong improvements for arousal. In addition, a small ablation study on the development split indicates that the lexicon-augmented variant provides a modest but consistent advantage over a plain RoBERTa regressor without lexicon features. The contributions of this work are threefold: (i) a simple lexicon-augmented transformer architecture for valence–arousal regression on ecological essays, (ii) an empirical comparison against a strong linear baseline, and (iii) an analysis of the added value of lexicon-based affective statistics in a modern transformer-based setting.

2 Background

2.1 Dimensional Emotion Analysis and Resources

Dimensional approaches to emotion modeling represent affective states in a continuous space, most commonly along valence, arousal, and sometimes dominance dimensions. This paradigm has motivated the creation of several lexical and sentence-level resources that provide continuous ratings for large inventories of English words and texts. For example, norms for 13,915 English lemmas with valence, arousal, and dominance scores have been released by Warriner et al. (2013), and reliable human ratings for approximately 20,000 English words have been obtained for the NRC VAD lexicon (Mohammad, 2018). At the sentence level, EmoBank provides dimensional emotion annotations for a corpus of English documents, enabling systematic study of annotation perspectives and representation formats for valence–arousal analysis (Buechel and Hahn, 2017).

These resources have established valence–arousal representations as a standard choice in text-based emotion analysis and have supported a wide range of downstream applications. In particular, word-level lexicons can be summarized into aggregate features for longer spans of text, while sentence-level corpora provide training data for supervised regression models. The present system builds directly on this line of work by leveraging the NRC VAD lexicon to extract continuous affective statistics that are combined with transformer-based representations.

2.2 Lexicon-Based and Transformer-Based Emotion Modeling

Lexicon-based sentiment and emotion analysis methods remain widely used, especially in applied and interdisciplinary research settings such as digital humanities and computational social science (Öhman, 2021). Although such methods may underperform machine learning models in some fine-grained prediction tasks, it has been argued that their validity should be judged in terms of usefulness and interpretability rather than raw accuracy alone (Öhman, 2021). Lexicon features are easy to compute, domain-agnostic, and directly grounded in human ratings, which makes them attractive as complementary signals even in neural architectures.

In parallel, transformer models—and BERT-style architectures in particular—have become the dominant approach for text-based emotion detection. A large body of work has demonstrated the effectiveness of BERT and related models for emotion classification and intensity regression across multiple benchmarks and languages (Acheampong et al., 2021). Recent studies have also explored dimensional valence–arousal prediction with multilingual pre-trained transformers, showing that such models can achieve strong performance in cross-lingual and cross-domain settings (Mendes and Martins, 2023). Advances in representation learning for affect include continuous adversarial training schemes that seek to improve robustness and generalization in affective recognition tasks (Son et al., 2025), as well as richer exploitation of social and contextual signals (Park et al., 2018).

Within this landscape, the present work adopts a hybrid strategy: a RoBERTa encoder is fine-tuned end-to-end for valence–arousal regression, and a small set of lexicon-derived features is concatenated to the pooled transformer representation. This design exploits the strong contextual modeling capacity of transformers while retaining the interpretability and robustness of lexicon-based affective priors. The architecture is deliberately kept simple, avoiding more complex adversarial or multi-task setups, in order to focus on the incremental value of lexicon augmentation in a standard fine-tuning regime.

2.3 Emotion Dynamics and User-Aware Evaluation

The ecological essay setting of the shared task is closely related to recent work that uses everyday language as a window into emotional well-being and mental health. Studies on the "language of well-being" have shown that fluctuations in self-reported emotion experience can be tracked through everyday speech and written reflections (Sun et al., 2020). More recently, measures of emotion dynamics derived from text have been proposed as potential linguistic biosocial markers for mental health, emphasizing the importance of longitudinal patterns rather than isolated judgments (Teodorescu et al., 2023). In parallel, the affective computing community has increasingly investigated personalized versus generalized affective models, including comparisons of user-specific and population-level approaches in wearable-sensor and multimodal settings (Li et al., 2024; Han et al., 2024; Kargarandehkordi et al., 2024).

The evaluation protocol of SemEval-2026 Task 2, Subtask 1 reflects these concerns by combining between-user and within-user statistics into composite metrics for valence and arousal. Correlation and error are assessed both across users (based on user-level mean predictions) and within users (based on per-user series of texts), and then aggregated into composite scores. This user-aware evaluation emphasizes consistency of predictions at both the population level and the individual trajectory level, aligning the task with broader efforts to model emotion dynamics and personalization. The RoBERTa+Lex system described in this paper is designed and analyzed under this evaluation framework, with particular attention to how transformer fine-tuning and lexicon augmentation impact user-aware composite performance.

3 System Overview

The main strategy of this system is to fine-tune a RoBERTa (Liu et al., 2019) encoder for valence–arousal regression while augmenting it with lightweight affective features derived from the NRC VAD lexicon (Mohammad, 2025). In this setup, the transformer provides contextualized sentence representations, and the lexicon contributes summary statistics that encode prior knowledge about the affective properties of individual words. The combination is implemented as a single regression model (*RoBERTa+Lex*) that predicts continuous valence and arousal scores jointly.

3.1 Task Formulation

Subtask 1 of SemEval-2026 Task 2 is formulated as a valence–arousal regression problem on longitudinal ecological essays. Given an input essay x (an anonymized English text describing everyday experiences), the goal is to predict two continuous scores: *valence* and *arousal*. While the shared task provides longitudinal data, this system treats the problem as a generalized multi-output regression task under an i.i.d. assumption. Each text is modeled independently to establish a robust performance baseline for a static sentence-level regressor using a shared encoder and a two-dimensional regression head.

Formally, for an input essay x, a RoBERTa-based encoder first produces a contextual representation $h(x)$. A regression head then predicts

$$\hat{\mathbf{y}} = (\hat{v}, \hat{a}) = f_\theta(h(x)) \in R^2, \tag{1}$$

where $\hat{v}$ and $\hat{a}$ denote the predicted valence and arousal, respectively, and f_θ is a linear projection parameterized by θ. Model parameters are optimized using a mean squared error (MSE) loss over both dimensions.

3.2 Lexicon-Augmented RoBERTa Architecture

The final system, RoBERTa+Lex, builds on `roberta-base` (Liu et al., 2019) as the text encoder. Given a tokenized essay x, RoBERTa produces a sequence of hidden states

$$\mathbf{H} \in R^{T \times d}, \tag{2}$$

where T is the sequence length and d is the hidden size. Following standard practice, the hidden state of the first special token (the RoBERTa "`[CLS]`"–equivalent) is used as a sentence representation:

$$h = \mathbf{H}_{[\text{CLS}]} \in \mathbf{R}^d. \tag{3}$$

To inject explicit affective knowledge, a 6-dimensional feature vector $f_{\text{lex}}(x) \in \mathbf{R}^6$ is computed from the NRC VAD Lexicon. Essays are first tokenized into word tokens using a simple regex over lowercase text, and each token is matched against the NRC VAD entries (restricted to single-word items). For all matched tokens, the following summary statistics are computed:

- mean, standard deviation, and maximum of valence;

- mean, standard deviation, and maximum of arousal.

Although the NRC VAD lexicon includes dominance, it was excluded from $f_{\text{lex}}(x)$ to keep the feature space compact and focus on the task's valence and arousal regression targets. If an essay has no tokens covered by the lexicon, $f_{\text{lex}}(x)$ is set to the zero vector.

The contextual representation and the lexicon features are then concatenated:

$$z = [\, h \,;\, f_{\text{lex}}(x) \,] \in \mathbf{R}^{d+6}, \tag{4}$$

and passed to a linear regression head:

$$\hat{\mathbf{y}} = Wz + \mathbf{b}, \quad \hat{\mathbf{y}} \in \mathbf{R}^2, \tag{5}$$

where W and $\mathbf{b}$ are learned jointly with the RoBERTa encoder. In effect, the model learns to combine sentence-level semantics captured by RoBERTa with lexicon-derived statistics that summarize the distribution of affective word scores in each essay.

During development, an ablated configuration without lexicon features (a plain RoBERTa regressor, M1) was trained for comparison. As shown in Table 1, the lexicon-augmented variant (M2) achieved higher composite correlation for arousal, although a small decrease in valence correlation and a modest increase in valence composite MAE were observed. Despite this trade-off, the RoBERTa+Lex configuration was selected as the final system for submission. This decision was motivated by the objective of enhancing performance in the arousal dimension, which is recognized as a significantly more challenging task in task-based affect assessment than valence. Furthermore, a unified multi-output architecture was prioritized over a decoupled approach (i.e., seperate models for each dimension) to ensure a more compact and data-efficient architecture that effectively leverages a shared contextual representation.

4 Experimental Setup

4.1 Dataset and Splits

The dataset for Subtask 1 consists of 5,285 anonymized ecological essays and feeling-word texts written in English by 182 users, each annotated with longitudinal valence and arousal scores. After preprocessing, the official training portion contains 2,764 essay–label pairs. For model development, this portion is split into:

Model	$\mathbf{r}^{\mathbf{v}}_{\text{comp}}$	$\mathbf{r}^{\mathbf{a}}_{\text{comp}}$	$\text{MAE}^{\mathbf{v}}_{\text{comp}}$	$\text{MAE}^{\mathbf{a}}_{\text{comp}}$
M1	0.708	0.500	0.512	0.388
M2	0.693	0.545	0.552	0.418

Table 1: Development-set performance of a M1: plain RoBERTa regressor and the M2: lexicon-augmented RoBERTa+Lex model. Scores are composite correlation (r_{comp}) and composite mean absolute error (MAE_{comp}) for valence (v) and arousal (a). Higher r_{comp} and lower MAE_{comp} are better.

- **train**: 2,211 instances,
- **development**: 553 instances.

The remaining 1,737 instances form the official test set, whose labels are held out by the organizers. Both M1 (RoBERTa) and M2 (RoBERTa+Lex) are trained on the train split and tuned on the development split. The test set is used only for final evaluation by the task organizers.

4.2 Preprocessing

Text processing. The HuggingFace `RobertaTokenizer` is used for subword tokenization. During training and development, the maximum sequence length is set to 128 tokens, with longer essays truncated. For final test-time inference, the maximum length is increased to 256 tokens to reduce truncation for longer inputs.

Lexicon feature extraction. For (RoBERTa+Lex), the 6-dimensional NRC VAD feature vector $f_{\text{lex}}(x)$ is computed for each essay using the following steps:

1. The essay is tokenized into word tokens with a regex applied to lowercase text.
2. Each token is looked up in the NRC VAD lexicon (restricted to single-word entries).
3. All available valence and arousal scores for matched tokens are collected.
4. Mean, standard deviation, and maximum values are computed separately for valence and arousal.

Analysis of the training data indicates that approximately 72.05% of all tokens were successfully matched against the lexicon, and the feature vector $f_{\text{lex}}(x)$ resulted in a zero vector for 0.00% of the essays. These features are concatenated to the RoBERTa representation as described in Section 3.2 and are used only in M2.

Model	r^{v}_{comp}	r^{a}_{comp}	$\bar{r}_{comp}$	MAE^{v}_{comp}	MAE^{a}_{comp}	$\overline{MAE}_{comp}$
linear(BERT)	0.557	0.299	0.428	0.743	0.459	0.601
RoBERTa+Lex	0.656	0.438	0.547	0.653	0.411	0.532

Table 2: Official Subtask 1 test-set results for the task baseline (linear(BERT)) and the submitted system (RoBERTa+Lex)

Category	r^{v}_{comp}	r^{a}_{comp}	$\bar{r}_{comp}$	MAE^{v}_{comp}	MAE^{a}_{comp}	$\overline{MAE}_{comp}$
All Users	0.656	0.437	0.547	0.653	0.411	0.532
User Visibility						
Seen Users	0.648	0.386	0.517	0.653	0.415	0.534
Unseen Users	0.675	0.519	0.597	0.656	0.408	0.532
Text Type						
Words Only	0.662	0.573	0.618	0.636	0.384	0.510
Essay Only	0.645	0.395	0.520	0.672	0.445	0.559

Table 3: RoBERTa+Lex performance analysis for Subtask 1, by user visibility and text category.

4.3 Training Configuration

All models share the same training configuration. The encoder is `roberta-base` from HuggingFace Transformers, and a mean squared error (MSE) loss is optimized over the two regression outputs (valence and arousal). Optimization is performed with AdamW and a fixed learning rate of 2×10^{-5}. The batch size is set to 16 for training and development, and increased to 32 for final test-time inference. Training is run for up to 10 epochs, with early stopping patience of 2 epochs based on the mean of valence and arousal composite scores on the development set. A linear learning rate scheduler with no warmup steps is applied, and all experiments are conducted on a single NVIDIA L4 GPU.

5 Results

Table 2 reports the official scores released by the SemEval-2026 Task 2 organizers for Subtask 1. The submitted system, which is a lexicon-augmented RoBERTa regressor (RoBERTa+Lex), is compared against the shared baseline from the task overview, which trains a ridge regression model on averaged BERT token embeddings (*linear(BERT)*).

Following the official evaluation protocol, the table shows the user-aware composite correlation $r_{\text{composite}}$ and composite mean absolute error $\text{MAE}_{\text{composite}}$ for both valence and arousal, along with their simple averages $\bar{r}_{\text{comp}}$ and $\overline{\text{MAE}}_{\text{comp}}$ across the two dimensions.

The RoBERTa+Lex system substantially improves over the baseline across all main metrics. In terms of composite correlation, the baseline reaches $r_{\text{composite}} = 0.557$ for valence and 0.299 for arousal, whereas RoBERTa+Lex increases these to 0.656 and 0.438, respectively. This corresponds to absolute gains of approximately $+0.10$ for valence and $+0.14$ for arousal. Composite MAE also decreases from 0.743 to 0.653 for valence and from 0.459 to 0.411 for arousal, indicating more accurate predictions on both affective dimensions. Overall, the average composite correlation improves from 0.428 to 0.547, while the average composite MAE drops from 0.601 to 0.532.

A granular performance breakdown for the system is provided in Table 3. For the arousal dimension, a composite correlation of 0.519 is achieved for unseen users, compared to 0.386 for seen users. Performance on *Words-only* entries reaches 0.662 for valence and 0.573 for arousal, while *Essay-only* performance is 0.643 and 0.369.

6 Conclusion

This paper presented a lexicon-augmented RoBERTa system for valence–arousal regression in SemEval-2026 Task 2. By fusing contextual embeddings with NRC VAD priors—which covered 72.05% of tokens—the system achieved a competitive 0.547 average composite correlation. The model significantly outperformed the 0.428 linear baseline and demonstrated robust zero-shot generalization to unseen users ($r_{\text{arousal}} = 0.519$). These results confirm that integrating lightweight affective priors with fine-tuned transformers provides a critical refinement for modeling longitudinal emotional trajectories in everyday ecological language.

Limitations and Future Works

Several key limitations are acknowledged despite the competitive results of the system. A primary constraint involves the adoption of an i.i.d. assumption, which treats each ecological essay as an independent event. By disregarding user identity and chronological text sequences, the model operates as a static sentence-level regressor rather than a longitudinal tracker. While this design ensures robustness for unseen users, attaining an arousal correlation of 0.519, the lack of personalization likely resulted in a performance ceiling for the "Seen" user split 0.386. The absence of user-specific adaptation mechanisms limited the ability to calibrate to unique individual neutral baselines, which is identified as a factor disproportionately affecting within-user correlation for seen individuals. Furthermore, a performance gap remains in the arousal dimension compared to valence. Methodologically, the system relies on a simple feature aggregation scheme for lexicon data. Future work will focus on implementing user-adaptive layers and exploring sophisticated sequential modeling techniques, such as Recurrent Neural Networks (RNNs), to more effectively track affective trajectories over time.

References

Francisca Adoma Acheampong, Henry Nunoo-Mensah, and Wenyu Chen. 2021. Transformer models for text-based emotion detection: a review of bert-based approaches. *Artificial Intelligence Review*, 54(8):5789–5829.

Sven Buechel and Udo Hahn. 2017. Emobank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In *Conference of the European Chapter of the Association for Computational Linguistics*.

Yunjo Han, Panyu Zhang, Minseo Park, and Uichin Lee. 2024. Systematic evaluation of personalized deep learning models for affect recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(4):1–35.

Ali Kargarandehkordi, Matti Kaisti, and Peter Washington. 2024. Personalization of affective models using classical machine learning: a feasibility study. *Applied Sciences*, 14(4):1337.

Joe Li, Peter Washington, and 1 others. 2024. A comparison of personalized and generalized approaches to emotion recognition using consumer wearable devices: Machine learning study. *JMIR AI*, 3(1):e52171.

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. *arXiv preprint arXiv:1907.11692*.

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. 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.

Saif M. Mohammad. 2025. Nrc vad lexicon v2: Norms for valence, arousal, and dominance for over 55k english terms. *arXiv preprint arXiv:2503.23547*.

Emily Öhman. 2021. The validity of lexicon-based sentiment analysis in interdisciplinary research. In *Proceedings of the Workshop on Natural Language Processing for Digital Humanities*, pages 7–12, NIT Silchar, India. NLP Association of India (NLPAI).

Ji Ho Park, Peng Xu, and Pascale Fung. 2018. Plusemo2vec at semeval-2018 task 1: Exploiting emotion knowledge from emoji and# hashtags. *arXiv preprint arXiv:1804.08280*.

Seungah Son, Andres Saurez, and Dongsoo Har. 2025. Continuous adversarial text representation learning for affective recognition. In *2025 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, pages 0433–0438. 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.

Jessie Sun, H Andrew Schwartz, Youngseo Son, Margaret L Kern, and Simine Vazire. 2020. The language of well-being: Tracking fluctuations in emotion experience through everyday speech. *Journal of Personality and Social Psychology*, 118(2):364.

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.

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.