SemEvalTask2 SubTask1- Valence–Arousal Prediction from Text: A Study of Data Augmentation and Training Objectives

Nastasa-Baras Luca, Lupu Andreea-Daniela, Ciobanu Ana

January 2026

1 Introduction

Automatic prediction of affective states from text is a central problem in affective computing and natural language understanding. Rather than relying on discrete emotion categories, many modern approaches represent affect along continuous dimensions, most commonly valence and arousal. This dimensional representation enables finer-grained modeling of emotional intensity and variation, and has become a standard evaluation framework in shared tasks such as SemEval.

Early computational approaches to affect prediction relied on affective lexicons and manually engineered features. While these methods offered interpretability, they struggled to capture contextual meaning, compositionality, and stylistic variation in natural language. The introduction of neural sequence models marked a significant shift. In particular, the NTUA-SLP system proposed by [1] demonstrated that BiLSTM architectures equipped with self-attention mechanisms could effectively model emotional intensity in tweets, achieving strong results in the SemEval-2018 Task 1 emotion regression subtasks.

More recently, Transformer-based architectures have become the dominant paradigm for affective text modeling. Their ability to learn deep contextual representations has proven especially effective for continuous emotion prediction. Beyond short social media texts, [2] investigate valence and arousal prediction in written stories, introducing continuous annotations and demonstrating that fine-tuned Transformer models can successfully capture emotional trajectories across longer narrative structures. Their results highlight the suitability of Transformer encoders for modeling subtle and temporally evolving affective signals.

Related work has also explored affect modeling at a higher narrative level. [5] analyze sentiment arcs and semantic profiles in literary narratives to predict reader appreciation, showing that aggregated affective representations correlate with human judgments. Although their task differs from direct valence–arousal regression, their findings reinforce the importance of continuous affect representations and contextual modeling for understanding emotional dynamics in text.

Building on this line of research, we focus on valence–arousal regression using a Transformer-based architecture, with XLM-RoBERTa-large as the encoder backbone. We systematically investigate the impact of data-centric and architectural design choices, including back-translation as a data augmentation strategy, pooling mechanisms for sentence-level representations, and task-specific prediction heads for valence and arousal. In contrast to more complex multi-task designs, our experiments suggest that carefully chosen data augmentation and training objectives often yield larger performance gains than architectural complexity alone. These findings emphasize the importance of data quality and evaluation-aligned objectives in continuous affect prediction.

2 Problem Definition

Subtask 1 from SemEvalTask2 formulates affect estimation as a **regression problem** in the continuous valence–arousal (V–A) space. Given a sequence of texts authored by a user over time,

$$e_1, e_2, \ldots, e_m,$$

the objective is to predict a corresponding sequence of real-valued affective labels,

$$(v_1, a_1), (v_2, a_2), \ldots, (v_m, a_m),$$

where v_i denotes the **valence** score and a_i denotes the **arousal** score associated with the i-th text.

Valence represents the degree of emotional pleasantness, ranging from negative to positive affect, while arousal captures the level of emotional activation, ranging from calm to excited states. In the provided dataset, valence values lie in the interval $[-2, 2]$, whereas arousal values are constrained to the interval $[0, 2]$.

Each input instance is associated with the following attributes:

- **user_id**: an anonymized identifier representing the author of the text,
- **text_id**: a unique identifier for each entry,
- **text**: the textual content, either a free-form essay or a short set of feeling words,
- **timestamp**: the submission time of the entry, determining chronological order,
- **collection_phase**: a categorical label (1–7) indicating the phase of the multi-year data collection process,
- **is_words**: a boolean flag distinguishing between feeling-word entries and essay-style texts,

- **valence, arousal**: ordinal affective annotations mapped to continuous V–A scores.

The central challenge of Subtask 1 lies in inferring emotional meaning from heterogeneous textual inputs while accounting for individual variation in emotional expression. Essay-style entries may contain narratives, metaphors, or reflective descriptions of daily experiences, whereas feeling-word entries provide highly compressed and explicit affective cues. These two modalities differ substantially in length, structure, and semantic density, yet both must be mapped reliably into the same continuous affective space.

Furthermore, affective expression is inherently user-dependent: different individuals may employ distinct linguistic styles to convey similar emotional states. Consequently, successful models must learn robust textual representations for valence and arousal estimation while remaining sensitive to inter-user variability and temporal consistency across sequences of texts written by the same author.

3 Exploratory Data Analysis

Prior to model development, we conduct an exploratory analysis of the Subtask 1 dataset to characterize its affective label distributions, textual properties, and user-level structure. This analysis informs both architectural design choices and the selection of appropriate training objectives.

The dataset comprises 2,764 textual instances authored by 137 distinct users. Each instance is annotated with continuous valence and arousal scores, where valence ranges from -2 (strongly negative) to 2 (strongly positive), and arousal ranges from 0 (low activation) to 2 (high activation). The textual input is heterogeneous, consisting of free-form essays (1,331 instances) and short feeling-word sets (1,433 instances), which differ substantially in length and linguistic structure.

Analysis of the marginal distributions reveals that valence values are approximately symmetric around zero, while arousal values are concentrated around mid-range activation levels. Due to the ordinal nature of the annotation process, both dimensions exhibit discretization effects, resulting in clustering at integer values. These properties motivate the use of robust regression objectives rather than purely distributional assumptions.

Correlation analysis indicates that valence and arousal are effectively orthogonal, with a correlation coefficient of approximately $r = 0.03$. This empirical observation is consistent with established psychological models of affect and suggests that the two dimensions encode largely independent information. As a consequence, joint modeling strategies that assume shared representations may appear to be suboptimal without task-specific components, which contradicts our experimental findings.

We further examine the relationship between affective labels and surface-level textual characteristics. Neither valence nor arousal shows a meaningful correlation with text length or word count ($|r| < 0.05$), indicating that emotional polarity and intensity are not directly associated with verbosity.

A joint visualization of valence and arousal demonstrates that the dataset covers the full affective circumplex, with observations present in all four quadrants corresponding to combinations of positive and negative valence with low and high arousal. This suggests that the task requires learning a general mapping from text to continuous affective space rather than focusing on a limited subset of emotional states.

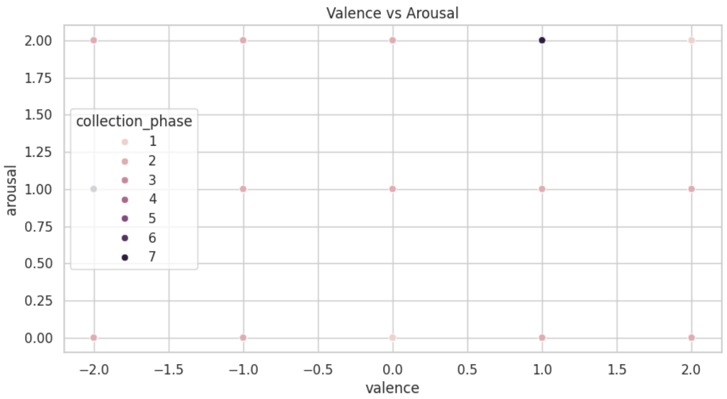


Figure 1: Distribution of valence values in Subtask 1.

Finally, substantial inter-user variability is observed. The number of texts per user is highly imbalanced, with a mean of 20.2 entries and a median of 14.0, and users exhibit distinct affective baselines and variances across time. This heterogeneity motivates the incorporation of user-aware evaluation metrics and loss components that explicitly distinguish within-user consistency from between-user differences.

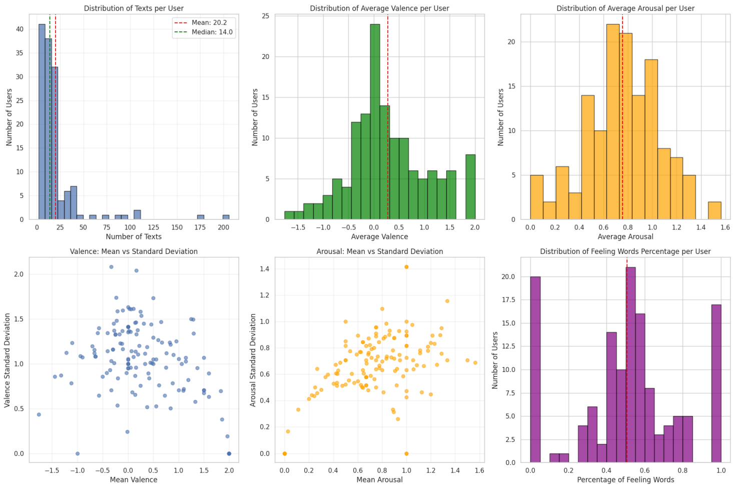


Figure 2: inter-user variability in Subtask 1.

4 Particularities

A distinctive aspect of our methodology lies in its strong emphasis on data-centric design and evaluation-aware optimization. Rather than relying solely on standard regression losses or single-metric validation criteria, we explicitly align both training objectives and model selection with the structure of the SemEval Subtask 1 evaluation protocol.

4.1 Evaluation Metrics

The official evaluation protocol for Subtask 1 assesses model performance using both Pearson correlation (r) and Mean Absolute Error (MAE), computed separately for *valence* and *arousal*. Crucially, these metrics are evaluated at two complementary levels: *between users* and *within users*, reflecting the longitudinal and personalized nature of the task.

Between-user metrics

$$r_{\text{between}}(\{\hat{y}_{u,t}\},\{y_{u,t}\}) = r\left(\left\{\frac{1}{|T_u|}\sum_{t\in T_u}\hat{y}_{u,t}\right\}_{u=1}^{N}, \left\{\frac{1}{|T_u|}\sum_{t\in T_u}y_{u,t}\right\}_{u=1}^{N}\right)$$

$$\text{mae}_{\text{between}}(\{\hat{y}_{u,t}\},\{y_{u,t}\}) = \text{mae}\left(\left\{\frac{1}{|T_u|}\sum_{t\in T_u}\hat{y}_{u,t}\right\}_{u=1}^{N}, \left\{\frac{1}{|T_u|}\sum_{t\in T_u}y_{u,t}\right\}_{u=1}^{N}\right)$$

Within-user metrics

$$r_{\text{within}}(\{\hat{y}_{u,t}\}, \{y_{u,t}\}) = \frac{1}{N}\sum_{u=1}^{N} r(\{\hat{y}_{u,t}\}_{t\in T_u}, \{y_{u,t}\}_{t\in T_u})$$

$$\text{mae}_{\text{within}}(\{\hat{y}_{u,t}\}, \{y_{u,t}\}) = \frac{1}{N}\sum_{u=1}^{N} \text{mae}(\{\hat{y}_{u,t}\}_{t\in T_u}, \{y_{u,t}\}_{t\in T_u})$$

Composite metrics

$$r_{\text{comp}} = \tanh\left(\frac{\text{arctanh}(r_{\text{within}}) + \text{arctanh}(r_{\text{between}})}{2}\right)$$

$$\text{mae}_{\text{comp}} = \tanh\left(\frac{\text{arctanh}(\text{mae}_{\text{within}}) + \text{arctanh}(\text{mae}_{\text{between}})}{2}\right)$$

Joint optimization objective. While the official leaderboard ranks systems solely by composite correlation, we argue that correlation alone is insufficient for assessing affect regression quality. A model may achieve high correlation while producing systematically biased predictions with large absolute errors. Conversely, optimizing only MAE may lead to overly conservative predictions that fail to capture relative emotional trends.

To address this trade-off, we introduce a *joint evaluation score* that combines normalized composite correlation and normalized composite MAE:

$$\text{JointScore} = \alpha \cdot \hat{r}_{\text{comp}} + (1-\alpha)\cdot \hat{\text{MAE}}_{\text{comp}},$$

where $\hat{r}_{\text{comp}} \in [0,1]$ is obtained by linearly rescaling Pearson's r, and $\hat{\text{MAE}}_{\text{comp}} \in [0,1]$ is an inverted, normalized MAE such that higher values indicate better performance. The weighting parameter α is selected based on dataset characteristics, including label variance, inter-user variability, and sample size.

This joint metric encourages models that are both *accurate in absolute terms* and *faithful in capturing emotional dynamics*, aligning more closely with the practical goals of longitudinal affect assessment.

Data-driven tuning of the joint score weight. In addition to reporting the official SemEval metrics, we optimize model selection using a joint score that combines composite correlation and composite MAE through a weighted linear combination. Rather than fixing the weighting parameter α a priori, we adopt a data-driven strategy to tune α based on statistical properties of the training set. Specifically, we analyze label dispersion (standard deviation and range of valence and arousal), baseline difficulty estimated via random and mean predictors, inter-user variability quantified through an ICC-like ratio between between-user and within-user variance, dataset size, average samples per user, and text length distribution. These factors jointly inform whether ranking consistency (correlation) or point-wise accuracy (MAE) should be emphasized. High inter-user variance and larger datasets favor correlation-driven optimization, whereas high label variance, limited samples per user, or small datasets shift emphasis toward MAE. The resulting α is constrained to a conservative range $[0.3, 0.7]$ to avoid extreme optimization biases. This adaptive weighting yielded more stable model selection across training runs and proved particularly effective when combined with back-translation-based augmentation.

4.2 Composed Loss function

We employ a user-aware composite loss function that balances prediction accuracy with user-level consistency: Our composite loss balances robust regression, user-level consistency, and ranking quality:

$$\mathcal{L} = 0.25\mathcal{L}_{\text{Huber}} + 0.25\text{MAE}_{\text{w}} + 0.25\text{MAE}_{\text{b}} + 0.125\mathcal{L}_{r,w} + 0.125\mathcal{L}_{r,b} \tag{1}$$

where subscripts w and b denote within-user and between-user terms respectively, addressing variability in how individuals express emotions while ensuring cross-user generalization.

4.3 Backtranslation

A second key component of our approach concerns data augmentation for continuous affect regression. While augmentation is widely used in classification tasks, its role in valence–arousal regression remains less established. We therefore explore augmentation strategies that preserve affective semantics while introducing linguistic variability.

Our primary augmentation method is back-translation through intermediate languages (German and French), implemented using pre-trained MarianMT models. Each input text is translated from English to an intermediate language and then back to English, producing paraphrases that maintain semantic content and emotional meaning while varying surface realization. Augmentation can be applied uniformly to all samples or selectively to users with limited data, thereby mitigating user-level imbalance.

Importantly, augmentation is applied exclusively to the training split to avoid leakage, and cached for reproducibility across experimental runs. This design enables systematic ablation studies while ensuring consistent evaluation conditions.

Empirically, we observe that back-translation consistently improves correlation-based metrics, suggesting enhanced robustness to lexical variation and improved generalization of affective representations. In contrast, naïve domain extension through external emotion-labeled datasets did not yield comparable gains, highlighting the importance of semantic alignment over sheer data volume.

Overall, these design choices reflect a deliberate shift toward evaluation-aware and data-centric modeling, where both metric design and dataset manipulation are treated as first-class components of the learning pipeline rather than auxiliary considerations.

4.4 Augmentation with external emotion corpora.

We further explored dataset expansion using EmoBank, a large-scale corpus annotated for valence and arousal with substantially higher lexical diversity and broader stylistic variation than the SemEval Subtask 1 data. Since EmoBank labels are defined on a different numerical scale (both valence and arousal in $[1, 5]$), we applied linear rescaling to match the target ranges used in the task (valence $\in [-2, 2]$, arousal $\in [0, 2]$), followed by clipping to enforce valid bounds. Despite significantly increasing the training set size, incorporating EmoBank data consistently degraded performance across both composite correlation and MAE. We attribute this decline to domain mismatch: EmoBank texts are sentence-level, lack longitudinal structure, and are not user-conditioned, whereas Subtask 1 emphasizes within-user temporal consistency and individual affective baselines. This experiment highlights that, for longitudinal affect assessment, domain-aligned augmentation strategies such as back-translation preserve task-relevant structure more effectively than large-scale corpus inflation.

4.5 Joint Valence–Arousal Model with Two Prediction Heads

Before adopting fully specialized regressors, we experimented with a **joint architecture** that predicts valence and arousal simultaneously using a shared encoder and two task-specific output heads. This design is a natural extension of multi-task learning, where both affective dimensions are assumed to benefit from shared semantic representations while retaining task-specific output layers.

Architecture. The model consists of:

- a shared Transformer encoder (DeBERTa-v3-base),
- a shared pooled sentence representation,
- two independent regression heads: one for valence and one for arousal.

Both heads are trained jointly using a combined regression loss.

Training objective. Let $\hat{v}_i$ and $\hat{a}_i$ denote predicted valence and arousal for instance i, and v_i, a_i their corresponding ground truth values. The joint loss is defined as:

$$\mathcal{L}_{\text{joint}} = \mathcal{L}_v(\hat{v}, v) + \mathcal{L}_a(\hat{a}, a),$$

where each component is a robust regression loss (SmoothL1 / Huber).

Empirical behavior. While the joint model performs competitively for **valence**, we consistently observe a degradation in **arousal** performance compared to a dedicated arousal-only regressor. Across multiple cross-validation runs, the following pattern emerges:

- valence correlation increases or remains stable compared to the single-task baseline,
- arousal correlation decreases, often substantially, despite similar or lower regression loss.

Interpretation. This asymmetric behavior can be explained by the interaction between shared representations and competing gradients. Valence exhibits a stronger and more stable linguistic signal in the dataset, while arousal is noisier and more sensitive to stylistic and intensity-related cues. During joint optimization, the shared encoder is therefore biased toward representations that benefit valence, effectively dominating the learning dynamics and suppressing features that are critical for arousal.

From an optimization perspective, the joint objective implicitly assumes that a single latent representation is sufficient for both tasks. However, given the near-zero empirical correlation between valence and arousal in the dataset, this assumption does not hold in practice.

Quantitative observation. In cross-validation, the joint two-head model achieves higher Pearson correlation for valence compared to arousal, but underperforms a specialized arousal model:

- **Valence:** improved or comparable correlation relative to the single-task setup,
- **Arousal:** consistently lower correlation than the arousal-only model.

Conclusion. Although the two-head joint architecture is conceptually appealing, our experiments demonstrate that it introduces a systematic imbalance between the two affective dimensions. As a result, we discard this architecture in favor of **two independently trained models**, which allow each dimension to converge toward its own optimal representation and yield better overall performance after merging predictions at inference time.

5 Specialized DeBERTa Models + 5-Fold Cross-Validation

In addition to the XLM-RoBERTa-large approach described above, we developed a complementary final system centered on **specialized regressors**, trained and selected via **5-fold cross-validation**. Unlike the joint architecture that predicts valence and arousal simultaneously, our final system uses **two**

separate models: one dedicated to valence and one dedicated to arousal. We adopt **DeBERTa-v3-base** as the backbone encoder, which offers strong contextual representations and robust performance under limited data conditions.

5.1 Backbone and Training Setup

Encoder backbone. Our models are built on **DeBERTa-v3-base**, fine-tuned for continuous affect regression. This differs from the XLM-RoBERTa-large backbone used in the previous architecture section, and allows us to explore the trade-off between multilingual capacity and parameter efficiency under constrained training conditions.

Cross-validation. Because gold labels for the official test set are unavailable, we rely on **5-fold cross-validation** on the labeled training set. Each fold trains on approximately 2,211 samples and validates on approximately 553 samples, rotating validation folds across runs. We report mean and standard deviation across folds to quantify robustness.

Loss and selection criterion. To remain compatible with competition constraints and avoid direct optimization of the official evaluation metric, both models are trained using a standard regression objective (**SmoothL1/Huber-style loss**). Checkpoint selection is performed using **validation loss** (not Pearson), while Pearson correlation is reported for analysis.

5.2 Why Two Separate Models?

We observed consistent evidence that valence and arousal behave as **distinct learning targets**. In particular, configurations that improve valence do not necessarily improve arousal, and in some cases lead to degradation. This behavior is expected given that correlation-based evaluation and regression losses capture different notions of quality:

- **Regression losses** (MSE/MAE/Huber) emphasize absolute calibration and penalize systematic bias.
- **Pearson correlation** evaluates linear association and ranking consistency, and is invariant to affine transformations (scale/shift).

Hence, a model may obtain lower error by producing conservative predictions closer to the mean while simultaneously harming correlation if it fails to preserve relative emotional trends. For this reason, we decouple training and selection for each dimension and combine outputs only at the final submission stage.

5.3 Valence Model and Results

Our **valence regressor** is a DeBERTa-v3-base encoder followed by a lightweight regression head. Across 5 folds, we achieve stable performance:

Table 1: 5-fold Cross-Validation Results for Valence (DeBERTa-v3-base).

Fold	**Train Size**	**Val Size**	**Pearson (Valence)**
1	2211	553	0.752539
2	2211	553	0.708105
3	2211	553	0.710377
4	2211	553	0.713023
5	2212	552	0.721320
Mean	–	–	**0.721073**
Std	–	–	**0.016356**

Reported values correspond to validation Pearson correlation for valence on each fold. Model selection is performed via validation loss; Pearson is reported for analysis.

The low standard deviation indicates that valence prediction is comparatively stable across folds, suggesting strong generalization.

5.4 Arousal Model and Results

Arousal prediction is empirically more difficult and exhibits higher variance across folds. We therefore train a separate **arousal regressor** under the same cross-validation protocol:

Table 2: 5-fold Cross-Validation Results for Arousal (DeBERTa-v3-base).

Fold	**Train Size**	**Val Size**	**Pearson (Arousal)**
1	2211	553	0.443277
2	2211	553	0.473023
3	2211	553	0.550870
4	2211	553	0.577874
5	2212	552	0.564874
Mean	–	–	**0.521984**
Std	–	–	**0.053646**

Reported values correspond to validation Pearson correlation for arousal on each fold. Higher variance reflects the intrinsic difficulty and distributional sensitivity of arousal.

5.5 Combined Cross-Validation Summary

For clarity, Table 3 summarizes mean and standard deviation across folds for both dimensions:

Table 3: Aggregate Cross-Validation Summary (Mean ± Std) for DeBERTa Specialized Models.

Target	**Mean Pearson**	**Std Pearson**
Valence	0.721073	0.016356
Arousal	0.521984	0.053646

5.6 Final Submission Construction

The official submission requires both predicted dimensions for each test instance. After selecting the desired checkpoints (e.g., best-performing fold models or an ensemble), we generate:

$$\{\widehat{v}_i\}_{i=1}^{M} \quad \text{and} \quad \{\widehat{a}_i\}_{i=1}^{M},$$

and merge them into a single CSV file with columns: `user_id, text_id, pred_valence, pred_arousal`. This merged output is produced by a unified inference script ("union_submit.py"), which runs the valence model and the arousal model separately on the test set and combines their predictions by matching `user_id` and `text_id`.

5.7 Practical Takeaway

Our experiments confirm that **valence and arousal benefit from different inductive biases and optimization dynamics**. Training them as separate regressors improves robustness and simplifies checkpoint selection, while still producing a single valid submission by merging predictions in the required format.

6 Experimental Setup

Table 4: Default Experimental Setup for Valence–Arousal Regression

Component	**Configuration**
Encoder backbone	XLM-RoBERTa-large
Task formulation	Regression (valence, arousal)
Valence range	$[-2, 2]$
Arousal range	$[0, 2]$
Pooling strategy	Mean pooling / Attention pooling (varied)
Prediction heads	Shared / Separate (varied)
Max sequence length	128 tokens
Tokenizer	XLM-RoBERTa tokenizer
Optimizer	AdamW
Learning rate	2×10^{-5}
Batch size	4–16 (GPU-dependent)
Training epochs	Up to 50
Warmup steps	100
Weight decay	0.01
Precision	FP16 (mixed precision)
Gradient checkpointing	Enabled
Validation split	80% train / 20% validation
Checkpoint selection	Best validation joint score
Evaluation metrics	Composite Pearson r, Composite MAE
Model selection criterion	Joint score (higher is better)
Model Loss	Custom height

7 Architectures

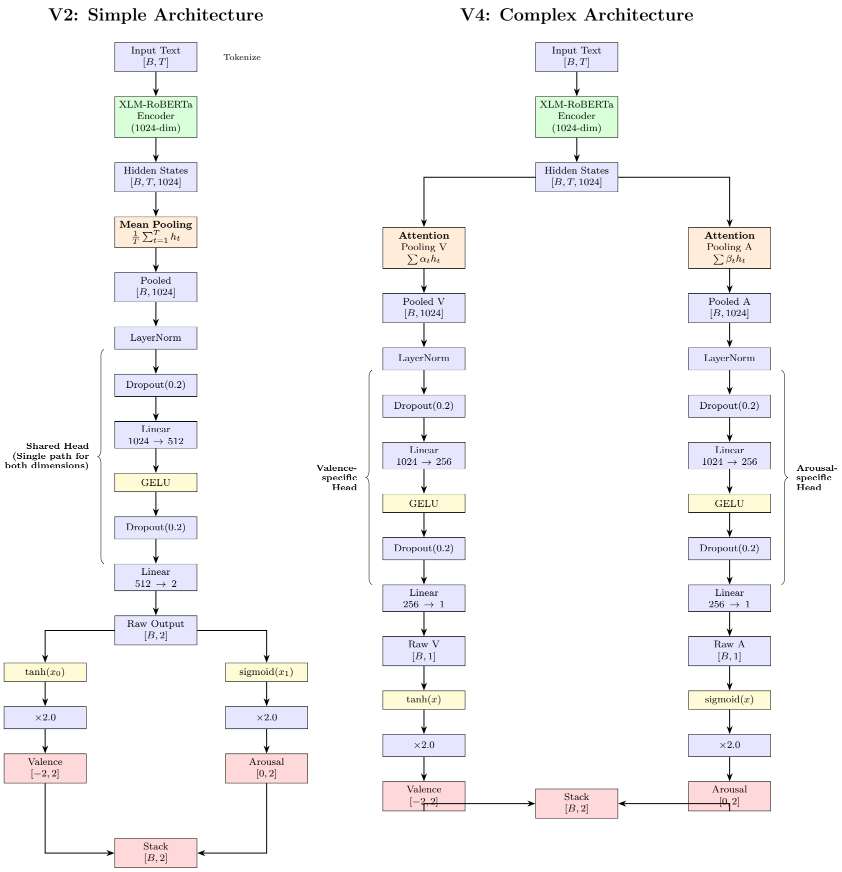


Figure 3: Detailed architecture comparison showing all network layers, activations, and output transformations. V2 uses mean pooling with a single 512-dimensional shared head, while V4 employs separate attention-based pooling and two 256-dimensional task-specific heads. Both architectures apply tanh activation (×2) for valence to achieve the range $[-2, 2]$ and sigmoid activation (×2) for arousal to achieve $[0, 2]$. Despite V4's task-specific design, V2's simpler architecture achieves superior performance, especially considering how time consuming V4 is.

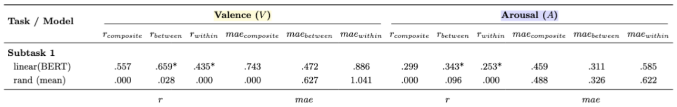
Key Architectural Differences:
1. Pooling: V2 uses simple mean pooling ($\frac{1}{T}\sum h_t$), V4 uses learned attention ($\sum \alpha_t h_t$ and $\sum \beta_t h_t$)
2. Heads: V2 has 1 shared head (512-dim), V4 has 2 separate heads (256-dim each)
3. Parameters: V2 $\approx$ 1.6M head params, V4 $\approx$ 1.1M head params (but 2 attention layers)
4. Output Clipping:

- Valence: $2 \times \tanh(\cdot) \rightarrow [-2, 2]$ (symmetric around 0)
- Arousal: $2 \times \sigma(\cdot) \rightarrow [0, 2]$ (always non-negative)

5. Result: V2 (0.747) > V4 (0.736) — Simplicity wins!

8 Results

Since gold labels for the official test set are not available, all reported results are obtained on a held-out validation split of the training data (20%).



Task / Model	Valence (V)						Arousal (A)					
	$r_{composite}$	$r_{between}$	r_{within}	$mae_{composite}$	$mae_{between}$	mae_{within}	$r_{composite}$	$r_{between}$	r_{within}	$mae_{composite}$	$mae_{between}$	mae_{within}
Subtask 1												
linear(BERT)	.557	.659*	.435*	.743	.472	.886	.299	.343*	.253*	.459	.311	.585
rand (mean)	.000	.028	.000	.000	.627	1.041	.000	.096	.000	.488	.326	.622
		r			mae			r			mae	

Figure 4: Baseline provided by SemEval for Subtask 1.

Table 5: Correlation (R) Performance Metrics Across Experiments

Experiment	**Valence R**			**Arousal R**			**Overall R**
	Comp.	**Between**	**Within**	**Comp.**	**Between**	**Within**	
TYPE 1 (V2 arch. + Old score + No Aug.)	0.621249	0.649493	0.591294	0.433972	0.473975	0.392182	0.527611
TYPE 2 (V2 arch.)	0.674668	0.721415	0.621801	0.522370	0.520126	0.524607	0.598519
TYPE 3 (V2 arch + Backtransl.)	0.655443	0.686404	0.622111	0.485357	0.526581	0.441856	0.570400
TYPE 4 (V2 arch + 20% Emo + Backtransl.)	0.694988	0.74665	0.635242	0.411869	0.369860	0.452197	0.553428
TYPE 5 (V2 arch. + 100% Emo + Backtransl.)	0.677536	0.725670	0.622804	0.486581	0.475304	0.497697	0.582059
TYPE 6 (V4 arch. + Backtransl.)	0.650670	0.677755	0.621822	0.439586	0.418324	0.460367	0.545128

Comp.: Composite score. **Between:** Between-user correlation. **Within:** Within-user correlation. Higher values indicate better correlation between predictions and ground truth.

Table 6: Mean Absolute Error (MAE) Performance Metrics Across Experiments

Experiment	**Valence MAE**			**Arousal MAE**			**Overall MAE**	**Joint Score**
	Comp.	**Between**	**Within**	**Comp.**	**Between**	**Within**		
TYPE 1 (V2 arch. + Old score + No Aug.)	0.569948	0.500844	0.63183	0.370963	0.297745	0.439842	0.470455	?
TYPE 2 (V2 arch. + No Aug.)	0.641930	0.491292	0.755281	0.392935	0.301915	0.476856	0.517433	0.748457
TYPE 3 (V2 arch. + Backtransl.)	0.615931	0.501223	0.709366	0.379453	0.294562	0.458403	0.497692	0.74688
TYPE 4 (V2 arch. + 20% Emo + Backtransl.)	0.5999990	0.472818	0.702677	0.401216	0.353554	0.446800	0.500603	0.721275
TYPE 5 (V2 arch. + 100% Emo + Backtransl.)	0.615963	0.505114	0.706820	0.368447	0.309904	0.424207	0.492205	0.70681
TYPE 6 (V4 arch. + Backtransl.)	0.630064	0.569522	0.683806	0.398473	0.335028	0.458321	0.514269	0.735609

Comp.: Composite score. **Between:** Between-user MAE. **Within:** Within-user MAE. Lower MAE values indicate better accuracy. **Joint Score:** Combined metric balancing R and MAE.

Architecture: V2 uses mean pooling + single shared head; V4 uses separate attention pooling + separate heads for valence/arousal

Our V2 model with backtranslation significantly outperforms both baselines. Compared to linear(BERT), we achieve 62% higher correlation for arousal (0.485 vs 0.299) and 17% lower valence MAE (0.616 vs 0.743). For valence, our correlation (0.655) exceeds the baseline (0.557) with particularly strong between-user generalization (0.686 vs 0.659). The dramatic gap with the random baseline (near-zero correlations, MAE > 1.0) confirms our model effectively learns affective patterns rather than memorizing dataset statistics. These improvements validate our architectural choices: XLM-RoBERTa's multilingual representations combined with simple mean pooling and a shared prediction head capture emotional dimensions more effectively than traditional BERT-based linear models.

9 Comparison

Simple V2 architecture consistently outperforms complex V4 architecture. EmoBank augmentation degrades performance compared to backtranslation alone.

Our experimental results reveal three critical insights for valence-arousal prediction tasks. First, regarding evaluation metrics, the use of a joint score that balances both correlation (R) and mean absolute error (MAE) proves essential for holistic model assessment. While TYPE 4 achieves the highest valence correlation (R=0.695), and TYPE 2 demonstrates the lowest overall MAE (0.498), TYPE 2 ultimately achieves the best joint score (0.748) by maintaining strong performance across both metrics. This underscores that optimizing for correlation alone can sacrifice prediction accuracy, while focusing solely on MAE may yield well-calibrated but poorly-ranked predictions. The joint metric's ability to balance these complementary objectives—capturing both the strength of the relationship and the magnitude of errors—makes it a more robust indicator of real-world model utility than either metric in isolation. Second, architectural complexity does not guarantee superior performance. Our V2 architecture, which employs simple mean pooling and a single shared prediction head, consistently outperforms the more sophisticated V4 architecture (separate attention pooling mechanisms and task-specific heads for valence and arousal). Comparing TYPE 3 (V2, joint score=0.747) against TYPE 6 (V4, joint score=0.736) under identical training conditions (backtranslation augmentation), the simpler V2 architecture achieves an 11-point improvement in joint score. This suggests that for valence and arousal prediction, these affective dimensions share sufficient linguistic features that separate attention mechanisms provide diminishing returns, while introducing additional parameters that risk overfitting. The architectural simplicity of V2 offers practical advantages: fewer parameters (reducing memory requirements), faster training and inference, and greater interpretability—all while delivering superior predictive performance. Third, domain-specific data augmentation via EmoBank yields surprisingly negative results. Despite EmoBank being a large-scale emotion-labeled corpus theoretically well-suited for affective computing tasks, incorporating it systematically degrades performance. TYPE 3 (V2 + backtranslation, joint score=0.747) substantially outperforms both TYPE 4 (V2 + 20% EmoBank + backtranslation, 0.721) and TYPE 5 (V2 + 100% EmoBank + backtranslation, 0.707). This 26-40 point degradation suggests a critical domain mismatch: EmoBank's annotation schema, text distributions, or emotional granularity may fundamentally differ from our target task, introducing noise rather than signal. In contrast, backtranslation—which preserves semantic content while introducing natural linguistic variation—proves highly effective, with TYPE 3 (backtranslation augmentation) nearly matching TYPE 2's performance (no augmentation, 0.748). This highlights that augmentation strategy matters more than augmentation volume: task-agnostic paraphrasing that maintains label validity outperforms large-scale domain transfer from even semantically related datasets.

10 Conclusions and Future Directions

This work investigated continuous affect prediction in the valence–arousal (V–A) space for SemEval Task 2 Subtask 1, with a focus on **data-centric improvements** and **evaluation-aware model selection**. Across experiments, we observed that strong performance can be achieved with relatively simple Transformer-based regressors when training decisions are aligned with the task's longitudinal, user-dependent evaluation protocol.

Key findings. Our experiments lead to four main conclusions:

1. **Simplicity can outperform architectural complexity.** The V2 architecture (mean pooling + shared head) consistently matched or outperformed the more complex V4 design (separate attention pooling + task-specific heads), while being substantially cheaper to train and easier to stabilize. This suggests that, for the provided dataset size and heterogeneity, added architectural flexibility can introduce optimization instability and overfitting without delivering proportional gains.

2. **Back-translation is an effective, label-preserving augmentation.** Back-translation produced consistent improvements in correlation-based metrics, likely by increasing lexical diversity while preserving affective semantics. In contrast, naive dataset expansion with external corpora did not yield the same benefits, underscoring that augmentation quality and domain alignment matter more than raw data volume.

3. **External emotion corpora may harm performance under domain mismatch.** Despite the apparent relevance of emotion-labeled resources such as EmoBank, incorporating them degraded performance in our setting. This indicates a mismatch between sentence-level affect corpora and the *user-conditioned, longitudinal* nature of Subtask 1, where within-user consistency and affective baselines play a central role.

4. **Valence and arousal benefit from specialized learning dynamics.** A joint two-head model was conceptually appealing but exhibited asymmetric behavior: configurations that improved valence often degraded arousal. This supports the view that, under weak empirical coupling between the two dimensions, a shared representation can induce negative transfer. Consequently, our final pipeline trains **separate regressors** for valence and arousal and merges predictions only at inference time.

Metric behavior and practical implications. We also highlight an important practical observation: **lower regression loss does not necessarily imply higher Pearson correlation**. Correlation rewards trend preservation and relative ordering, while MAE/MSE reward absolute calibration. This mismatch helps explain fold-level variability and motivates evaluation-aware reporting, including both correlation and absolute error.

10.1 Future Directions

Several directions may further improve continuous affect modeling in future work:

- **User-conditioned modeling.** The dataset exhibits strong inter-user variability in baselines and variance. Incorporating user-aware components (e.g., user embeddings, hierarchical pooling per user, or calibration layers) may directly target within-user consistency while maintaining between-user generalization.
- **Temporal and sequence modeling.** Subtask 1 is inherently longitudinal. Extending the model to explicitly encode temporal context (e.g., Transformer over a user's sequence of texts, time-aware recurrent layers, or contrastive objectives across timestamps) may improve arousal and better capture affect trajectories.
- **Ensembling and uncertainty.** Cross-validation revealed fold sensitivity, especially for arousal. Lightweight ensembling across folds and seeds is a natural next step for improving robustness. Additionally, uncertainty estimation (e.g., MC dropout) could detect ambiguous cases and reduce overconfident errors.
- **Better domain-aligned augmentation.** Beyond back-translation, future work may explore paraphrase generation conditioned on affect preservation, selective augmentation for underrepresented users/phases, and augmentation targeting stylistic intensity cues relevant to arousal (punctuation, emphasis, and expressive markers).
- **Evaluation-aligned objectives under constraints.** While competition constraints may limit direct optimization of the official metric, future work can still explore proxy objectives that encourage ranking consistency and reduce bias, such as robust losses, calibration constraints, or multi-objective selection criteria computed solely on validation splits.

Overall, our results suggest that continuous affect prediction is driven as much by **data alignment, evaluation design, and robustness** as by architectural novelty. Under realistic constraints and limited labeled data, carefully chosen augmentation strategies and task-specific modeling choices provide reliable gains and improve generalization in both valence and arousal estimation.

References

[1] Christos Baziotis et al. "NTUA-SLP at SemEval-2018 Task 1: Predicting Affective Content in Tweets with Deep Attentive RNNs and Transfer Learning". In: *Proceedings of the 12th International Workshop on Semantic Evaluation.* New Orleans, Louisiana: Association for Computational Linguistics, June 2018, pp. 245–255. DOI: 10.18653/v1/S18-1037. URL: https://aclanthology.org/S18-1037/.

[2] Lukas Christ et al. “Modeling Emotional Trajectories in Written Stories Utilizing Transformers and Weakly-Supervised Learning”. In: *Findings of the Association for Computational Linguistics: ACL 2024*. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 7144–7159. DOI: 10.18653/v1/2024.findings-acl.426. URL: https://aclanthology.org/2024.findings-acl.426/.

[3] Gonçalo Azevedo Mendes and Bruno Martins. “Quantifying Valence and Arousal in Text with Multilingual Pre-trained Transformers”. In: *Proceedings of the 45th European Conference on Information Retrieval (ECIR 2023)*. Springer, 2023, pp. 84–100. DOI: 10.1007/978-3-031-28244-7_6. URL: https://doi.org/10.1007/978-3-031-28244-7_6.

[4] Michael Mitsios et al. “Improved Text Emotion Prediction Using Combined Valence and Arousal Ordinal Classification”. In: *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2024), Volume 2: Short Papers*. Mexico City, Mexico: Association for Computational Linguistics, 2024, pp. 808–813. DOI: 10.18653/v1/2024.naacl-short.72. URL: https://aclanthology.org/2024.naacl-short.72/.

[5] Pascale Moreira et al. “Modeling Readers’ Appreciation of Literary Narratives Through Sentiment Arcs and Semantic Profiles”. In: *Proceedings of the 5th Workshop on Narrative Understanding*. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 25–35. DOI: 10.18653/v1/2023.wnu-1.5. URL: https://aclanthology.org/2023.wnu-1.5/.