CS 445
PROJECT REPORT

SEMEVAL 2026 TASK 2: PREDICTING VARIATION IN EMOTIONAL VALENCE AND AROUSAL OVER TIME FROM ECOLOGICAL ESSAYS

Atacan Dilber - 31037
İlhan Sertelli - 30567
Mehmet Egehan Pala - 31089
Ece Gülkanat - 31983
Oğuzhan Güzelgün - 28160

1. Introduction

This project addresses the task of predicting emotional valence and arousal over time from ecological essays, a problem that is important for understanding emotional dynamics in naturalistic and real-world text. Unlike traditional emotion classification, which relies on fixed emotion categories, this task focuses on modeling emotions on continuous dimensions, making the prediction problem more challenging but also more informative. To address this task, we adopt a robust natural language processing approach based on the RoBERTa architecture. Our method uses a RoBERTa-based regression model with a split-head design that allows valence and arousal to be predicted independently while still sharing a common textual representation. The model is trained using Concordance Correlation Coefficient Loss (CCCLoss), which directly optimizes the agreement between predicted and true continuous values and is well suited for emotion prediction tasks. To further improve model performance, Optuna is employed for automatic hyperparameter optimization, exploring different configurations of learning rate, weight decay, dropout probability, arousal loss weight, layer-wise decay, and parameter freezing strategies. Model robustness and generalization are evaluated using K-Fold Cross-Validation. Using this methodology, the final ensemble model achieves strong performance on the development set, reaching an average Composite Score of 0.7753 (0.8399 for valence and 0.7106 for arousal) and an average Global Pearson correlation of 0.8071 (0.8424 for valence and 0.7718 for arousal), demonstrating the effectiveness of the proposed RoBERTa-based framework for continuous emotion prediction from textual data.

2. Related Work

Dimensional emotion recognition has gained increasing attention, as it allows emotions to be represented on continuous valence and arousal scales. Early approaches primarily relied on emotion lexicons and traditional regression techniques, which were limited in their ability to capture contextual and compositional meaning in text. More recent studies have shown that pre-trained Transformer models considerably improve performance in predicting continuous emotional dimensions. In particular, multilingual and monolingual Transformer-based encoders have been shown to outperform earlier neural approaches across diverse datasets, while consistently exhibiting higher prediction accuracy for valence than for arousal due to the more implicit nature of arousal-related linguistic cues [2].

Beyond model architecture, several works emphasize the importance of jointly modeling emotional dimensions and incorporating affective structure. Xie et al. propose a multi-dimensional relation model that captures dependencies between valence and arousal, demonstrating that shared representations with dimension-specific outputs can improve regression performance [5]. Similarly, Paz-Arbaizar et al. adopt a Transformer-based framework for emotion forecasting and show that modeling temporal and contextual dependencies is beneficial for continuous emotion prediction [4]. These findings support the use of architectures that balance shared semantic representations with specialized prediction components.

Recent work has further explored alternative formulations of emotion prediction within the valence–arousal space. Mitsios et al. introduce an ordinal classification framework that arranges emotions according to valence and arousal levels, reducing the severity of misclassifications by explicitly modeling perceptual distances between emotions [3]. While their approach focuses on classification rather than regression, it highlights the value of structuring emotion prediction around affective dimensions. Additionally, Christ et al. investigate continuous emotion modeling at the story level, proposing a Transformer-based regression approach augmented with weakly supervised learning to capture emotional trajectories over long narratives [1]. Their results further confirm that valence is more reliably predicted than arousal and underscore the challenges of modeling emotional intensity from text alone. Together, these studies provide strong motivation for regression-based, dimension-aware Transformer architectures for valence–arousal prediction.

3. Methodology

This study focuses on Subtask 1 of SemEval-2026 Task 2, which aims to predict emotional valence and arousal from ecological essays. Since both emotions are represented as continuous values, the task is treated as a supervised regression problem. The dataset provided by the task organizers contains approximately 2,800 essays, each labeled with valence and arousal scores. Because the official test labels are not available, model evaluation is carried out using K-fold cross-validation to ensure reliable and fair performance estimation.

Before training the model, the essays are lightly preprocessed. All texts are converted to lowercase, and punctuation and numerical characters are removed. The essays are then tokenized using the tokenizer of the RoBERTa model. No heavy text normalization is applied, since pretrained Transformer models benefit from preserving the original structure and context of the text.

An exploratory analysis of the dataset is conducted to better understand its properties and potential challenges. Figure 1 shows the distribution of valence scores, which are mostly centered around neutral and slightly positive values, with fewer samples at extreme negative and positive levels. Figure 2 presents the arousal distribution, which is more uneven, with many essays labeled with low arousal and fewer samples at high arousal levels. This imbalance suggests that arousal prediction is more difficult than valence prediction. Figure 3 illustrates the distribution of essay lengths measured in number of words, showing that most essays are short, often containing fewer than 20 words, while a small number of essays are much longer. Figure 4 displays the distribution of the number of essays per user, indicating that most users contribute only a few essays, while a small number of users contribute many more. Figure 5 presents the correlation matrix between valence, arousal, essay length, and collection phase, showing very low correlation values and suggesting that essay length and collection phase do not strongly influence emotional labels.These dataset characteristics motivate the use of cross-validation and regularization techniques to improve generalization and reduce user-level bias.

The main model used in this study is based on the RoBERTa architecture, which serves as a pretrained text encoder. The output representations produced by RoBERTa are passed to a split-head regression structure, where one output head predicts valence and another predicts arousal. Both heads share the same text representation but learn separate mappings for each emotional dimension. This design allows the model to capture shared semantic information while also modeling differences between valence and arousal.

The model is trained using the Concordance Correlation Coefficient Loss (CCCLoss), which is commonly used in continuous emotion prediction tasks. This loss function directly measures the agreement between predicted and true values in terms of both correlation and scale. To further improve performance, Optuna is used for automatic hyperparameter optimization. During this process, different configurations such as learning rate, weight decay, dropout rate, loss weighting between valence and arousal, layer-wise learning rate decay, and parameter freezing strategies are explored.

Training is performed using a K-fold cross-validation strategy, where the dataset is divided into multiple folds and each fold is used once for validation. This approach reduces the effect of data imbalance and provides a more stable estimate of model performance. The system is evaluated using standard metrics for continuous emotion prediction, including the Concordance Correlation Coefficient, Pearson correlation, and composite scores combining both emotional dimensions. All experiments are implemented in Python using common deep learning and natural language processing libraries, and the entire pipeline is reproducible.

4. Results

The proposed system was evaluated on the task of predicting continuous emotional valence and arousal values from text. Quantitative results obtained via five-fold cross-validation are reported in Table 1. Overall, the model achieves strong performance across both dimensions, with consistently higher scores for valence than for arousal. This performance pattern is consistent with prior Transformer-based approaches for dimensional emotion prediction, where valence is generally predicted more accurately than arousal due to its more explicit lexical cues [2],[4]. Figures 6 and 7 present scatter plots of predicted versus ground-truth values for valence and arousal, respectively. The valence predictions show a close alignment with the identity line, indicating a strong linear relationship between predicted and true values.

In contrast, arousal predictions exhibit greater dispersion, particularly at higher intensity levels, reflecting the increased difficulty of modeling emotional intensity from textual input alone. Similar behavior has been reported in earlier studies, where arousal is shown to be more sensitive to contextual and implicit signals [2],[5]. When compared with existing work, the observed results are in line with findings reported in the literature. Mendes and Martins demonstrate that large Transformer models improve correlation-based metrics across languages while preserving the relative performance gap between valence and arousal [2]. Additionally, Xie et al. show that explicitly modeling relationships between emotional dimensions can improve regression performance, particularly for arousal [5]. During experimentation, multiple alternative configurations and loss functions were explored; however, these variants resulted in lower validation performance or less stable convergence and were therefore excluded from the final system.

5. Discussion

The selection of the ecological essay dataset has a notable impact on system performance. While the dataset provides rich and natural emotional expressions, it also introduces subjectivity and noise, particularly for arousal prediction, which is less explicitly encoded in text. The use of a Transformer-based regression model built on RoBERTa allows the system to capture contextual and semantic information effectively, leading to strong overall performance. In particular, the split-head design enables the model to share a common textual representation while learning separate output mappings for valence and arousal, which helps capture both shared and dimension-specific patterns. This design offers improved flexibility compared to a single-head regression setup, but the approach remains limited by its reliance on textual input alone, especially for emotionally intense or implicit arousal signals.

The observed performance trends are consistent with prior work on dimensional emotion prediction. Similar to findings reported by Mendes et al., valence is predicted more accurately than arousal, reflecting inherent differences in how these dimensions are expressed in language [2],[4]. Additionally, Xie et al. highlight the importance of modeling inter-dimensional relationships to improve regression performance, particularly for arousal [5]. Despite achieving competitive results, the proposed system has several limitations, including the absence of explicit temporal modeling and the lack of multimodal information. With additional time and resources, future improvements could include incorporating temporal sequence models, leveraging multimodal data such as audio or visual signals, or further refining the interaction between valence and arousal representations to enhance prediction robustness.

6. Conclusion

This project studied the task of predicting emotional valence and arousal from ecological essays as part of SemEval-2026 Task 2, Subtask 1, with the aim of modeling emotions on continuous scales rather than fixed emotion categories. To address this task, a RoBERTa-based regression model with a split-head structure was used, which allows valence and arousal to be predicted separately while sharing the same text representation. The experimental results show that the proposed approach performs well for both emotional dimensions, and the ensemble model trained with Concordance Correlation Coefficient Loss and optimized using Optuna achieved strong and stable performance across different cross-validation folds. As expected, valence prediction was more accurate than arousal prediction, since valence is usually expressed more clearly in text, whereas arousal is often more subtle and harder to capture using language alone. The analysis of numerical results and visual plots also indicates that the model successfully captures general emotional patterns in the data but becomes less reliable when predicting very intense emotional states. Overall, this study demonstrates that combining a pretrained Transformer model with an appropriate loss function and careful hyperparameter tuning is an effective method for continuous emotion prediction from text. However, predicting arousal remains challenging when only textual information is available, and future work could improve performance by incorporating additional contextual information, modeling emotional changes over time, or using data from other modalities such as audio or visual signals.

7. Individual Contributions

The project was carried out collaboratively by all team members. İlhan contributed to model training and designing the overall modeling approach, implementing the RoBERTa-based regression architecture. Atacan focused on data preprocessing, dataset exploration, and supporting the model training pipeline by preparing input representations and validation splits. Egehan was mainly responsible for hyperparameter optimization using Optuna, conducting the cross-validation experiments to evaluate model robustness, and coordinating the experimental setup. Ece worked on the evaluation process, including computing performance metrics, analyzing results, and preparing visualizations such as scatter plots to interpret model predictions. Oğuzhan contributed to the literature review, assisted with result interpretation, and supported the writing and organization of the report. All team members actively participated in group discussions, experiment planning, and report writing. The workload was shared evenly across the team, and each member contributed to multiple stages of the project, including implementation, evaluation, analysis, and documentation, ensuring a balanced and collaborative workflow throughout the project.

8. References

[1] Christ, L., Lauscher, A., Reiter, N., & Frank, A. (2024). *Modeling emotional trajectories in written stories utilizing transformers and weakly-supervised learning*. In Findings of the Association for Computational Linguistics: ACL 2024 (pp. 7144–7159).

[2] Mendes, G. A., & Martins, B. (2023). *Quantifying valence and arousal in text with multilingual pre-trained transformers*. arXiv preprint arXiv:2302.14021.

[3] Mitsios, M., et al. (2024). *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) (pp. 808–813).

[4] Paz-Arbaizar, J., et al. (2025). *Emotion Forecasting: A Transformer-Based Approach*. IEEE Journal of Biomedical and Health Informatics / Conference Proceedings.

[5] Xie, H., Lin, W., Lin, S., Wang, J., Yu, L. (2021). *A Multi-dimensional Relation Model for Dimensional Sentiment Analysis.* Information Sciences, 579, 832–844. https://doi.org/10.1016/j.ins.2021.08.052

9. Appendix

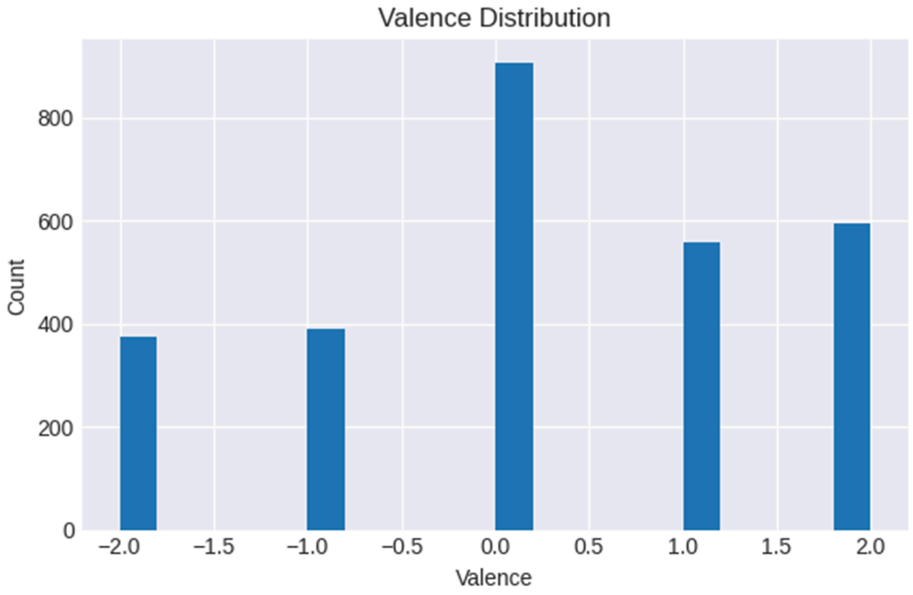


Figure 1. Valence Distribution Histogram

This figure shows how valence scores are distributed across the dataset, with clear peaks at 0, 1 and 2.

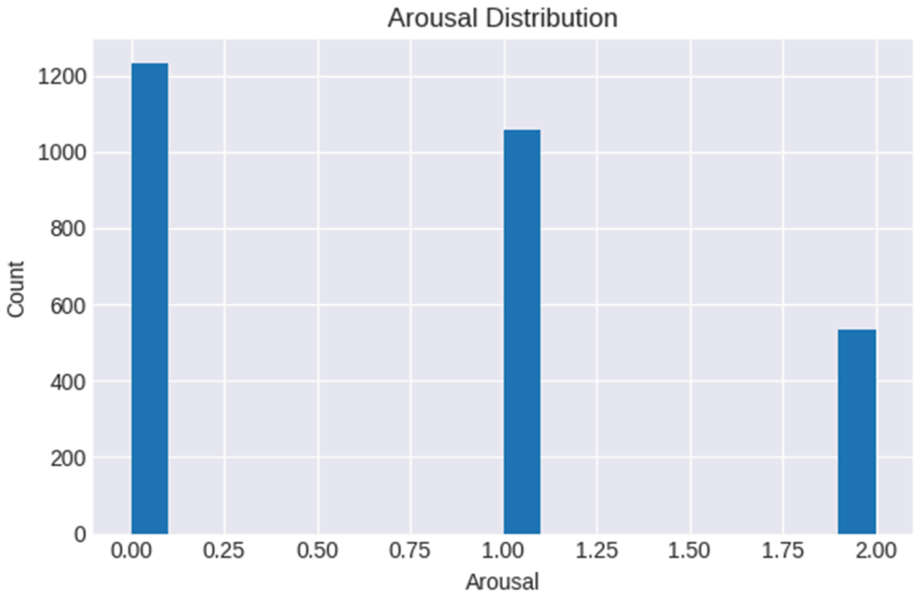


Figure 2. Arousal Distribution Histogram

This visualization displays the distribution of arousal scores, showing a strong right-skew with many texts labeled as 0 or 1.

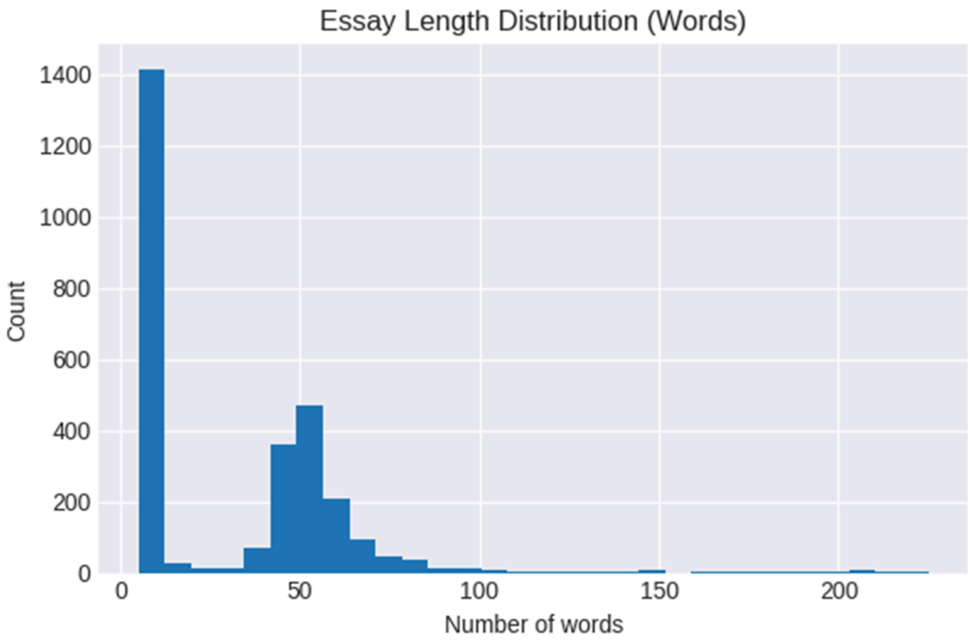


Figure 3. Essay Length Distribution (Words)

This histogram illustrates the skewed nature of essay lengths, with most texts being fewer than 20 words.

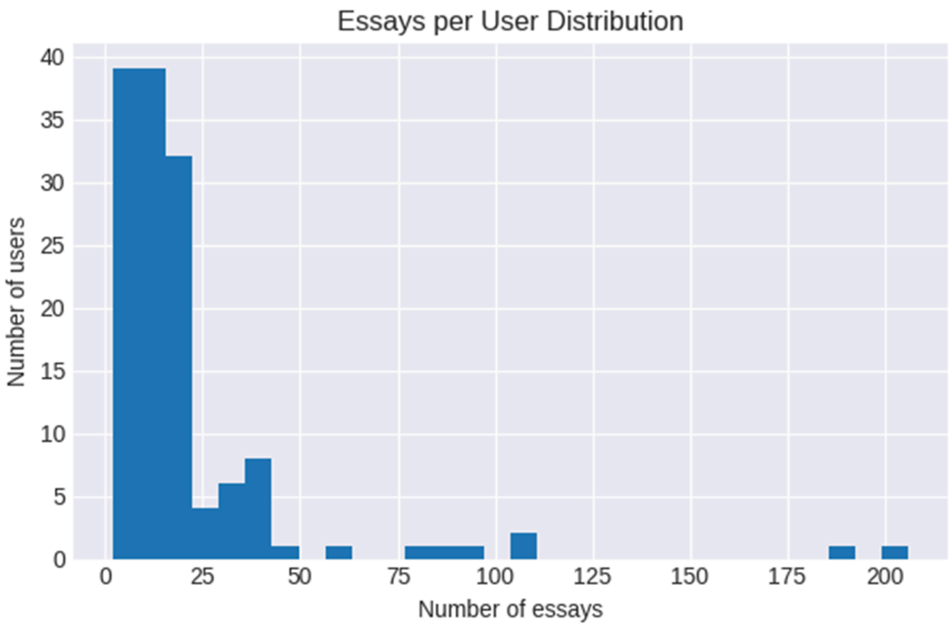


Figure 4. Essays per User Distribution

This figure shows the imbalance in user contributions, ranging from 2 to 206 essays per user

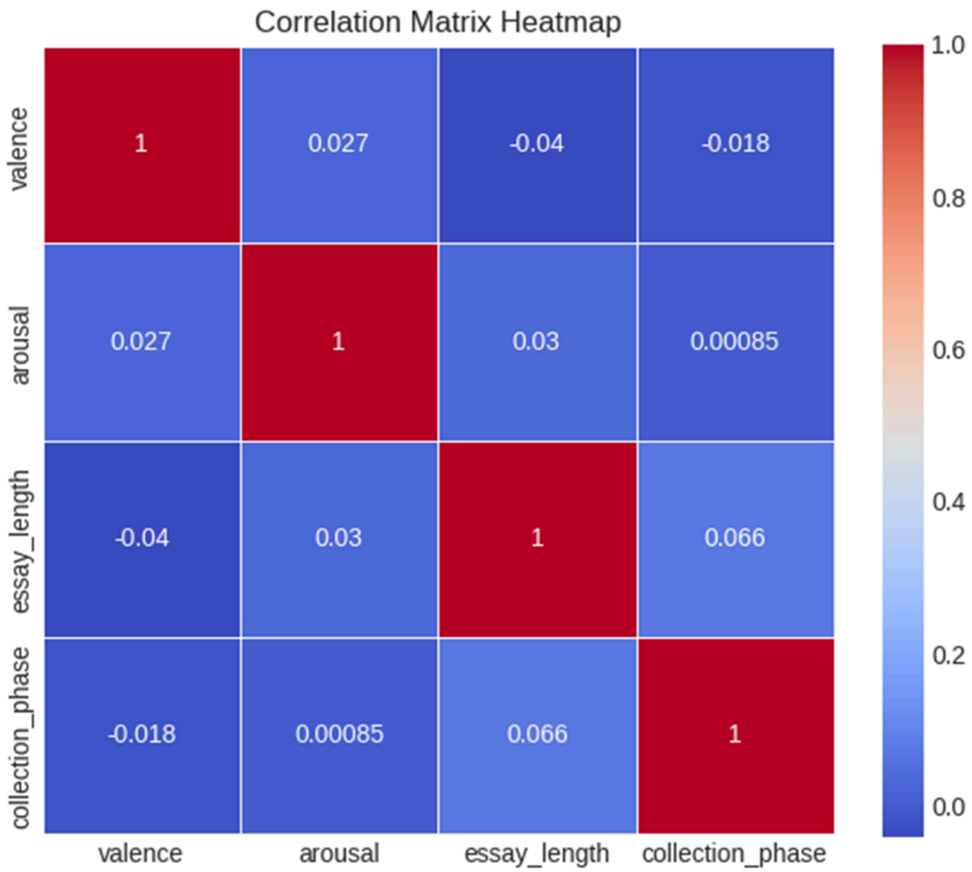
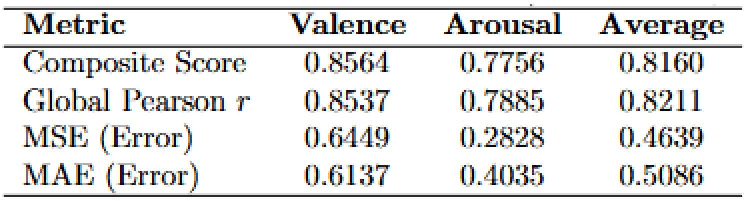


Figure 5. Correlation Matrix Heatmap

This heatmap presents the linear correlations among valence, arousal, essay length, and collection_phase, all of which are close to zero.



Metric	Valence	Arousal	Average
Composite Score	0.8564	0.7756	0.8160
Global Pearson r	0.8537	0.7885	0.8211
MSE (Error)	0.6449	0.2828	0.4639
MAE (Error)	0.6137	0.4035	0.5086

Table 1. Ensemble Performance Report (5-Fold Averaged)

This table presents the five-fold averaged results of the ensemble model

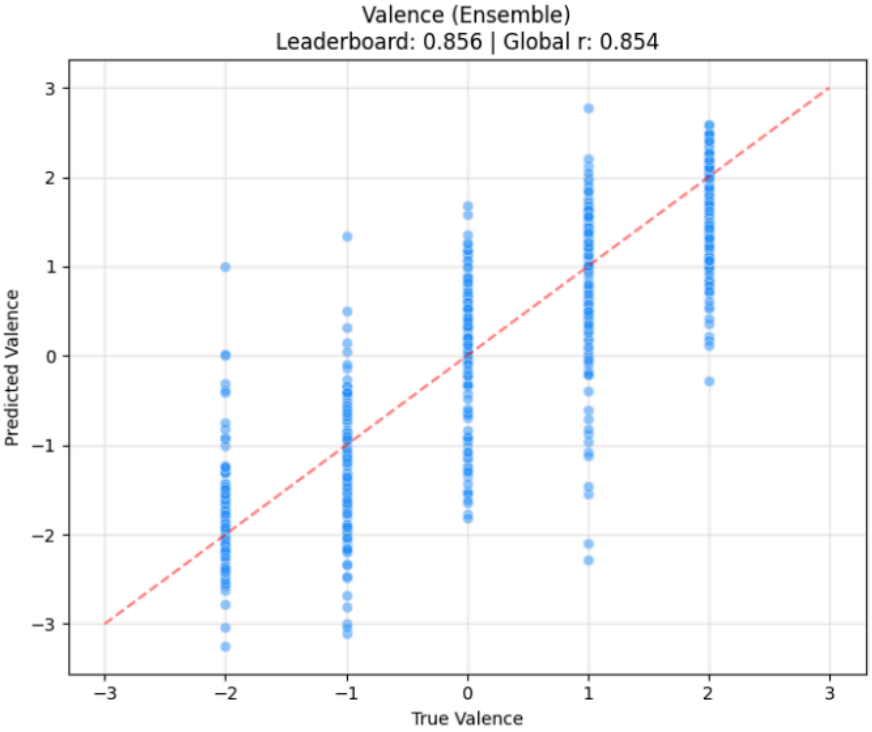


Figure 6. True vs. Predicted Valence Scores (Ensemble Model)

This figure shows a strong alignment between true and predicted valence values, indicating high prediction accuracy for the valence dimension.

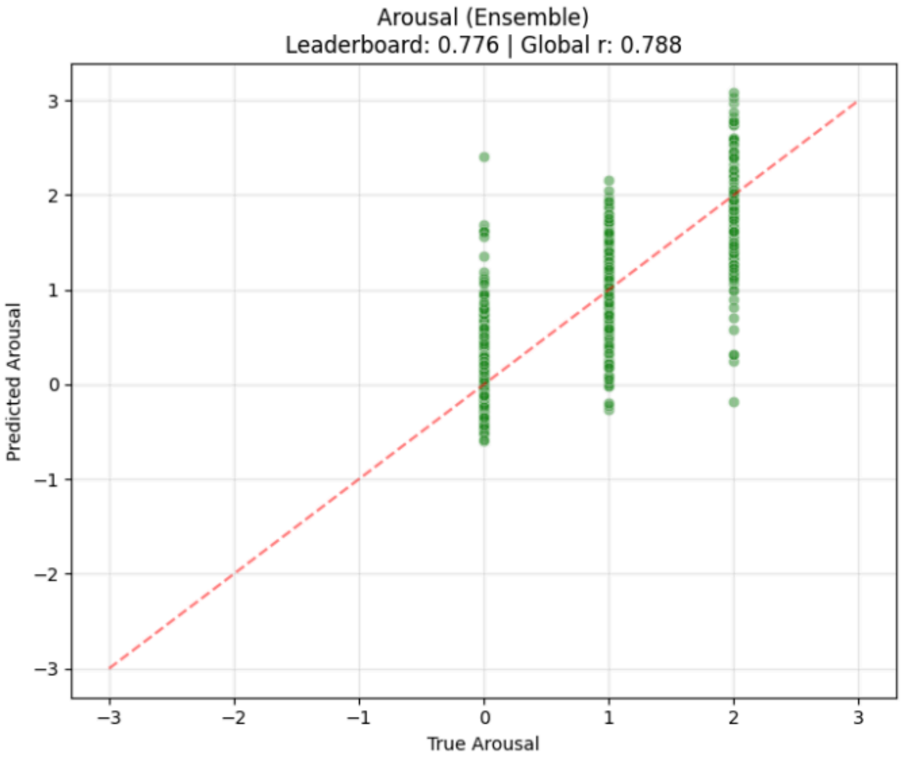


Figure 7. True vs. Predicted Arousal Scores (Ensemble Model)

This figure shows greater dispersion between true and predicted arousal values, reflecting the increased difficulty of predicting arousal compared to valence.