NLPGroup8 at SemEval-2026 Task 2: Diverse Ensembles and Hierarchical Transformers for Emotional State Prediction

Troy Arthur[1], Aidan Kelley[1], and Sierra Reschke[2]
[1]Department of Computer Science and Linguistics
[2]Department of Computer Science
University of Colorado Boulder
{trar3243, aike6451, sire7023}@colorado.edu

Abstract

This paper presents the systems developed by NLPGroup8 for SemEval-2026 Shared Task 2. Our approach combines a diverse ensemble for Subtask 1 with a context-aware transformer aggregation architecture for temporal forecasting in Subtasks 2A and 2B. The diverse ensemble achieves state-of-the-art performance for the Subtask 1 Valence metric, ranking first in Valence prediction. Our Subtask 2B independent architecture ranks second overall among competitive submissions. We also report results for Subtask 2A, analyzing challenges our architecture faces with next-entry affect forecasting. These findings underscore the significance of our methodology for affective prediction, achieved without reliance on external affective datasets or large-scale computational resources.

trar3243/NLPGroup8AtSemeval-2026

1 Introduction

SemEval-2026 Shared Task 2: Predicting Variation in Emotional Valence and Arousal over Time from Ecological Essays (Soni et al., 2026) introduces subjective affective assessments to the NLP discourse by relying on users' self-reported affective state, along dimensions of Valence and Arousal. The goal of this shared task is to predict how people report they feel emotionally or how people will report they feel emotionally, based on their written text. Valence intends to capture how 'good/bad' an individual feels, and Arousal intends to capture how 'physiologically activated' an individual feels. By precluding annotators, this approach forces models to learn affective states as experienced, rather than as perceived by observers.

Shared Task 2 compiles a longitudinal list of English 'ecological essays' provided by U.S. service industry workers, including self-reported user affective states in terms of Valence/Arousal. Such data enables development of models that react to the affective 'rhythms' of individuals over time.

The goal of Subtask 1 is to predict affective state at entry t based on user written text at entry t. The goal of Subtask 2A is to forecast change in affective state between entry t and entry $t+1$ based on entries 1 to t per user. The goal of Subtask 2B is to forecast change in affective state between the average of entries $[1, .., t]$ and the average of entries $[t+1, ..., 2t]$ based on entries 1 to t per user.

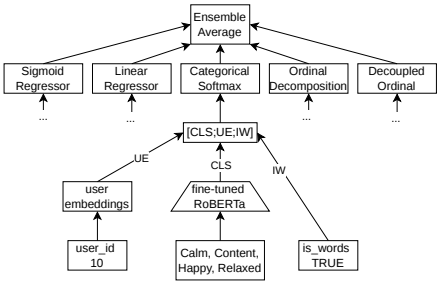

Figure 1: Subtask 1 Design

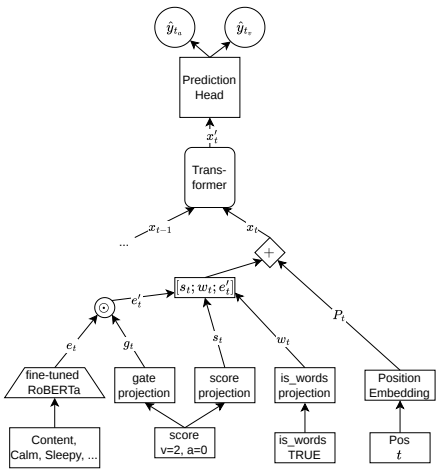

Figure 2: Subtask 2 Design

1.1 Primary System Strategies

For all Subtasks, we implement our architecture using the Hugging Face Transformers library (Wolf et al., 2020). We initialize our encoders with the pre-trained RoBERTa-base model (Liu et al., 2019) with 12 layers and a hidden size of 768.

Our Subtask 1 model relies on a diverse five-model ensemble. Models are all encoder-regressors, with various activation functions, loss functions, and model output shapes. These diverse prediction strategies aim to capture different nuances of the relationships between affect and lan-

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 379–386
July 3-4, 2026

guage. Development began with individual model prediction and moved to an ensemble approach to combine best results.

Our Subtask 2 models rely on a hierarchical transformer architecture, with a fine-tuned RoBERTa model responsible for local feature extraction and a downstream transformer as a global context aggregator. For each user, encoded text embeddings for each entry are fed into a global transformer model. The model forecasts the next affective change relevant to the matching Subtask.

2 Background

2.1 Training Data

An example entry is shown in table 4 in the Appendix, incorporating fields from all subtasks. Subtask 1 training data consists of 2764 entries. Each entry consists of user id, text id, text, text timestamp, collection phase, a boolean indicating text type as essay-style or word list, and users' self-reported Valence and Arousal scores. Valence and Arousal scores are integers, with Valence scores ranging from -2 to 2 (5 values) and Arousal scores ranging from 0 to 2 (3 values).

Subtask 2A training data contain the same fields and entries as Subtask 1 with added fields, state change Valence and state change Arousal, to represent the Valence and Arousal of the current text subtracted from the following text. These values represent how differently the user is going to feel between their current entry and their next entry.

Subtask 2B training data contains the same fields and entries as Subtask 1 with added fields; group, disposition change valence, and disposition change arousal. Group 1 marks the first half of texts per user and Group 2 marks the second half. Each group is approximately equal in size. Disposition change valence is the first half of per-user texts' mean Valence deducted from the second half, and disposition change Arousal the same for mean Arousal. These values represent how different the user is going to feel in their second group of entries, compared to their first group of entries.

2.2 Related Work

Transformer-based models such as RoBERTa are widely used for sentiment and emotion tasks due to their strong contextual representations (Liu et al., 2019). Other work has explored ordinal regression for affect prediction, motivated by the ordered nature of Valence and Arousal labels (Shi et al., 2023). Ensemble methods are also well studied; prior research demonstrates that diverse ensembles can improve overall performance by reducing correlated errors across models (Durrant and Lim, 2020). Additionally, architectural heterogeneity in ensemble text classifiers has been demonstrated to to improve predictive performance. Kamateri and Salampasis (2025) show this approach outperforms base classifiers, and is generalizable across varied-length and multi-labeled text entries.

Our Subtask 1 system draws from this existing work. We combine continuous regression models, categorical classifiers, and ordinal-regression architectures into a single loss-diverse ensemble. This design reflects varied approaches to emotion representation, allowing the model to capture complementary signals.

Hierarchical attention networks were first formulated by Yang et al. (2016), with two levels of attention mechanisms applied at word and sentence levels. We extend this hierarchical intuition to the temporal domain for Subtask 2: our model uses a RoBERTa encoder to capture within-entry semantics and a following transformer to model between-entry dynamics. Similar approaches have been taken by Pappagari et al. (2019), who demonstrate that feeding windowed encoder embeddings into a secondary sequence model effectively bypasses encoders' fixed context windows.

3 System Overview

Software is available on GitHub and model weights are available on Zenodo (Arthur et al., 2026).

3.1 Subtask 1 Architecture

3.1.1 Ensemble Agreement

We have constructed an ensemble of five diverse models, as increased diversity improves performance (Durrant and Lim, 2020). The ensemble predicts a continuous score for both Valence and Arousal. Predictions are based on an average across all constituent model predictions. As seen in figure 1, each constituent model takes as input finetuned RoBERTa CLS embeddings from the text field concatenated with $4D$ learned user embeddings and a $1D$ feeling words/open-response boolean value. Each model corresponds to its own RoBERTa encoder, which during training is initially pulled from RoBERTa-base (Liu et al., 2019) and then fine-tuned. All five constituent models, with the exception of Decoupled Ordinal, which has separate hid-

den layers for Valence and Arousal, contain a single hidden layer of dimensionality $d_{input}/2$ followed by a GELU activation function and 0.1 dropout. All constituent models use an AdamW (Loshchilov and Hutter, 2019) optimizer. All model linear layers are initialized with the Xavier uniform initialization and all bias terms are initialized to zero.

3.1.2 Component Models

Our ensemble models are motivated by a hybrid of categorical and continuous theoretical understandings of the interactions between language and affect. The ensemble consists of five distinct regression heads atop the RoBERTa encoder: (1) Sigmoid Regressor, which scales logits to the [0,1] range to dampen extremes, utilizing SmoothL1Loss; (2) Linear Regressor, a direct projection minimizing SmoothL1Loss; (3) Categorical Softmax, which treats affect as discrete classes (5 for Valence, 3 for Arousal) and computes expected values, using Cross Entropy Loss; (4) Ordinal Decomposition, following Shi et al. (2023) to model ordinal thresholds, using Binary Cross Entropy Loss; and (5) Decoupled Ordinal, which separates hidden layers for Valence and Arousal to enforce feature independence, using Binary Cross Entropy Loss.

3.2 Subtask 2 Architecture

To model the rhythm of users' affective states for future affective forecasting, we implement a Transformer-based aggregation layer which processes sequences of entry-level RoBERTa-derived text embeddings and other entry-level features (see figure 2). We use a weighted average of MSE (weighted 10%) loss and Pearson R loss (weighted 90%) for training. Pearson R loss is used when both the predicted and target labels have more than near zero variance among the batch; otherwise, MSE is used in isolation.

3.2.1 Input Construction, Text Modulation

Each entry's text data is passed through a fine-tuned RoBERTa-base model (Liu et al., 2019) to produce $768D$ encoder embeddings (e_t). For each entry, the Valence and Arousal scores are projected from a dimensionality of 2 to 64 to produce score embeddings s_t, and to a dimensionality of 768 to produce gate embeddings g_t. The is_words boolean is projected from a dimensionality 1 to 32 to produce w_t. These projections involve, sequentially, a learned linear layer and a LayerNorm, followed by a a GELU activation function for s_t and w_t or a sigmoid activation function for g_t. For feature fusion, the text embedding e_t is modulated by the gate projection (g_t) to produce the modulated text: $e'_t = e_t \odot g_t$. Modulated text embeddings are then concatenated with the is_words embedding and the score embeddings. To capture the sequential order of user entries, a learned positional embedding P_t, initialized from a normal distribution ($\mathcal{N}(0, 0.02)$), is added to the combined feature vector: $x_t = [s_t; w_t; e'_t] + P_t$.

3.2.2 Sequential Modeling

As Subtask 2 focuses on modeling emotional rhythms, we need to model specific stretches of time. We operationalize stretches of time with windows of user entries. These windows are comprised of timestamp-ordered lists of user entries, with one user for each window. Maximum window length is set to 25 entries. For windows mapped to fewer than 25 entries, all empty feature vectors are set to zero vectors. The window $X = [x_1, ..., x_n]$ is processed by a 4-layer Transformer Encoder with 8 attention heads and a feed-forward dimension of $4 \cdot d_x$ followed by a 0.1 dropout.

3.2.3 Prediction Head

The transformer output for the given context window's entry is sequentially linearly projected to a layer of dimensionality $d_x/2$, then passed through a GELU activation function and a 0.2 dropout. Resulting values are then passed through a final linear layer projecting to 2 outputs, being the final predicted Valence and Arousal scores. The window's final entry's contextualized output is used for prediction.

4 Experimental Setup

All experiments, including model training and hyperparameter tuning, were conducted locally on a single workstation equipped with one NVIDIA RTX Pro 2000 Blackwell GPU (16GB VRAM), an Intel Core i7-10700 processor, and 16GB of system memory.

We implement our models using Torch (v2.11.0.dev20260104+cu128) and the Hugging Face Transformers library (v4.57.3) (Wolf et al., 2020). The task baseline models include a linear (BERT) model and a random guess model for Subtask 1, and linear (BERT), linear (prev), linear (BERT; prev), and random models for Subtask 2.

4.1 Preprocessing

Input text is tokenized using the RoBERTa tokenizer (Liu et al., 2019). For initial hyperparameter tuning, we split the original training set into a development set (10%: 277 entries) and a training set (90%: 2487 entries). For final Subtask 2 model training, the entirety of the training set was used while Subtask 1 used the 90% training set. Hyperparameters are included in the appendix (5). All models across all Subtasks use a gradient clipping norm of 1.0. All hyperparameters were selected based on repeated experimentation on development sets, selecting the highest-performing hyperparameters on primary metric performance.

4.2 Subtask 2 Window Selection

During training, the Subtask 2A model generates $n_i - 1$ windows for each user i, where n_i refers to the user's number of associated entries. Each window for user i contains 1 to $\min(25, n_i - 1)$ entries. Then, the model predicts the change between the window's last entry and the following entry, and the loss function is used to compute the loss between the prediction and the training set's label. Importantly, each user's final entry is not included in any window. During inference, each user's last window's final entry's prediction is used for submission. Inference setup does not use each user's final entry, as no window is created to include it.

During training, the Subtask 2B model generates a single window for each user. This window includes the 25 latest entries within Group 1 for each user, zero-padding entries if the user's Group 1 is of size less than 25. For each user, the model predicts the Δ between the averages of Group 1 and Group 2. The loss function is used to compute the loss between the prediction and the training set's label. During inference, all training set user entries are considered a part of Group 1, and the model outputs the predicted Δ between the training group and the unseen group.

4.3 Evaluation Metrics

4.3.1 Subtask 1 Primary Metric

Following the task definition, a composite correlation is used as the primary metric for Subtask 1. This score is a composite of within-user and between-user correlation scores. The within-user correlation score reflects Pearson R correlation scores calculated within each user's set of scores, then averaged across users. The between-user correlation score reflects the Pearson R correlation scores of all participants' average scores. The composite correlation score is based on a Fisher's z-transformation of the within-user and between-user correlation scores.

4.3.2 Subtask 2 Primary Metric

Because the predictions for Subtask 2 include one prediction for each user, a simple Pearson's R correlation is computed for Subtasks 2A and 2B.

5 Results

Subtask	Track	Primary Metric	Ranking	Best Baseline
1	Valence	.688	1/26	.557
	Arousal	.416	21/26	.299
	Average	.552	12/26	.428
2A	Valence	.152	11/15	.615
	Arousal	.126	10/15	.67
	Average	.139	11/15	.643
2B	Valence	.354	2/12	.434
	Arousal	.388	4/12	.584
	Average	.371	2/12	.509

Table 1: Primary Metrics and Performance Ranking among Competitive Submissions across Subtasks

5.1 Subtask 1

5.1.1 Quantitative Findings

As seen in table 1, our model achieves state-of-the-art performance on Subtask 1 Valence, ranking first. Our model struggles with Subtask 1 Arousal, ranking 21st. This is primarily the result of poor performance on seen users. Although our model performs well on the primary metric for Subtask 1 Valence, our Valence MAE is the highest of all competing models ($\text{MAE}_{between} = 1.89, \text{MAE}_{within} = 1.925$). This is due to the fact that our model outputs 0-indexed Valence scores, rather than in the range $[-2, ..., 2]$.

5.1.2 Error Analysis, Post-Hoc Refinements

After submission and the release of gold labels, the 0-indexing issue resulting in large MAE scores was resolved and the resulting MAE scores significantly dropped ($\text{MAE}_{between} = .40, \text{MAE}_{within} = .777$). Detailed Subtask 1 metrics are in tables 6 and 7.

We conduct an ablation analysis for Subtask 1, omitting particular composite sub-models in each ablation run (see table 2). The results indicate that for the primary metric in Subtask 1, removing any model either decreases or does not affect the Valence performance of the Ensemble. Removal of Sigmoid Regressor results in the most significant

performance reduction; Valence composite correlation reduces to 0.683. Each ablation only reduces the Arousal composite correlation score, with the exception of Decoupled Ordinal, which marginally improves the Arousal composite correlation.

Ablate	Dimension	r_{comp}	mae_{bw}	mae_{wi}
Baseline	Valence	0.688	1.89	1.925
	Arousal	0.416	0.25	0.523
Sigmoid Regressor	Valence	0.683	1.9	1.936
	Arousal	0.415	0.25	0.523
Linear Regressor	Valence	0.688	1.835	1.874
	Arousal	0.415	0.247	0.528
Categorical Softmax	Valence	0.688	1.923	1.958
	Arousal	0.414	0.253	0.524
Ordinal Decomposition	Valence	0.686	1.927	1.96
	Arousal	0.412	0.252	0.524
Decoupled Ordinal	Valence	0.687	1.864	1.901
	Arousal	0.417	0.253	0.523

Table 2: Composite Submodel Ablation Metrics for Subtask 1

Additionally, to evaluate the impact of the ensemble approach overall, we evaluate each submodel's predictions individually (see table 3). The results indicate that all individual sub-models under-perform the ensemble with both Valence and Arousal predictions. The sigmoid regressor model achieved the highest Valence composite correlation, with $r_{comp} = .684$, marginally less than the ensemble's $r_{comp} = .688$. The ordinal decomposition model achieved the highest Arousal composite correlation, with $r_{comp} = .402$, marginally less than the ensemble's $r_{comp} = .416$.

Model	Valence r_{comp}	Arousal r_{comp}
Sigmoid Regressor	0.684	0.388
Linear Regressor	0.655	0.387
Categorical Softmax	0.655	0.395
Ordinal Decomposition	0.67	0.402
Decoupled Ordinal	0.674	0.394

Table 3: Post-Hoc Subtask 1 Singular Model Test Set Composite Correlations

To evaluate the impact of ensemble model diversity on the ensemble's final performance, we retrain five instances of the highest-performing single model (sigmoid regressor) and re-run ensemble predictions on the test set using these post-hoc trained instances. For this new training setup, all hyperparameters are kept consistent with the original training setup. The non-diverse ensemble attains a composite Valence correlation of $r_{comp} = 0.663$ and a composite Arousal correlation of $r_{comp} = 0.467$. While the non-diverse ensemble achieves higher Arousal correlations than our diverse ensemble, our primary contribution in Subtask 1 is our state-of-the-art Valence prediction capability. For this objective, the architectural diversity of our submitted ensemble proves highly advantageous. The non-diverse ensemble may have outperformed the diverse ensemble on arousal predictions because its constituent models, all instances of the sigmoid regressor architecture, are designed to scale logits to the $[0, 1]$ range to dampen extremes. Therefore, because outputs on either extreme end of the arousal scale are difficult to fully predict using the sigmoid regressor, the non-diverse ensemble may be less vulnerable to overfitting.

While we have achieved competitive Valence performance, our system struggles with Arousal performance. As seen in table 7, this poor performance is overwhelmingly the result of a low correlation ($r_{comp} = .297$) among seen users, with unseen user predictions ($r_{comp} = .601$) far exceeding the composite score, ranking fifth among unseen user correlations. This is likely the result of overfitting to user-specific biases, to which small arousal ranges would be more sensitive. Our Subtask 1 system gains its competitive advantage on the Valence track from within-user correlation ($r_{within} = .572$, ranked first) and unseen users ($r_{comp} = .679$, ranked first). These suggest our system excels in contexts with decreased reliance on users' histories and patterns. When reviewing individual submissions (for example word-list entry "$Tired$, $Depressed$, Sad, $Conflicted$, $Trapped$" with gold label 2 and prediction -1.7), all significant model Valence mispredictions appear reasonable. The highest 30 Valence discrepancies between gold and predicted labels map to entries with average Valence of ≈ 1, in contrast to the test set average Valence score of $\approx .16$. This suggests that our model struggles most to predict high Valence entries. The 30 highest Arousal discrepancies map to entries with average Arousal score of ≈ 1.2, marginally larger than the test set average Arousal score of $\approx .76$.

5.2 Subtask 2A

5.2.1 Quantitative Findings

Our model struggles with Subtask 2A Valence and Arousal, at competitive rankings of eleventh and twelfth, respectively. Inference stage outputs

predictions with a gold-label Valence correlation $r = .152$ and a gold-label Arousal correlation $r = .126$.

5.2.2 Error Analysis, Post-Hoc Refinements

It was determined that during the inference stage, where the goal is to predict the Δ between seen entry t and unseen entry $t + 1$, our model erroneously predicts the Δ between seen entry $t - 1$ and masked entry t. This off-by-one index error causes the model to predict the current masked difference rather than the future difference. In order to determine the degree to which this issue affected our final results, this index issue was resolved after submission and the resulting *post-hoc* predictions correlated with gold-label Valence at $r = .623$ and Arousal at $r = .631$.

5.3 Subtask 2B

5.3.1 Quantitative Findings

Our Subtask 2B model performs competitively across Valence and Arousal tracks, with correlations of 0.354 and 0.388, and respective rankings of 2nd and 4th among competitive submissions. The baseline model with the highest performance in Subtask 2B, the linear(prev) model, achieves the highest average correlation across all submissions, with a Valence correlation of 0.434 and an Arousal correlation of 0.584. On the Arousal metric, only a single competitive team surpasses the baseline model's performance.

5.3.2 Error Analysis

Although our Subtask 2B model performs well in comparison to competing teams, the highest-performing baseline model, which is a linear model based solely on prior scores, achieves the highest average Valence and Arousal correlation scores. This indicates that our team's modeling of the relationships between text and scores is inferior to simple mean-regression effects. Because our model caps the window size at 25, we evaluate users with 25 or fewer entries separately. Our analysis does not demonstrate that the window size cap significantly impacts our results at inference time. Analysis of large-discrepancy users demonstrates that our system is more likely to make mistakes when a user's average Valence or Arousal score in seen data is on one extreme end of the spectrum (for example, with an average Valence score of -2 and an average arousal score of 0). Our system forecasts small changes in these users' scores, while a model based purely on tendency to regress towards the mean would be more consistent with the gold labels.

To determine the impact of the fusion of scores and text embeddings, we post-hoc ablate the gate projection, resulting in the pre-transformer feature vector being a summation between positional embeddings and the concatenation of the score projection, the is_words projection, and the fine-tuned RoBERTa CLS token embeddings without a score fusion. All hyperparameters are kept consistent with the original model training setup, and a new training instance is run. The resulting ablation model's predictions on the test set perform worse than the original model's predictions, with a Valence correlation of $r = 0.323$ and an Arousal correlation of $r = 0.366$. These correspond to performance declines of $\approx 9\%$ and $\approx 6\%$, respectively. These results indicate support for our fusion approach to text and score inputs.

6 Conclusion

Without the use of external affective datasets or large-scale computational resources, our system achieves state of the art performance for predicting Valence affective scores from written text, and competitive performance for forecasting large-scale changes of affective state. Our architecture and results emphasize the potential for loss-diverse ensembles and hierarchical transformers in affective state prediction.

7 Limitations and Future Work

All entries were taken from English-speaking US service workers, which reduces generalizability to diverse users. Future work ought to combine our Subtask 1 and Subtask 2 approaches. Our loss-diverse ensemble proves advantageous, as does hierarchical transformer modeling. With better computational resources, these two approaches may be combined.

8 Ethics Statement

All data were used in accordance with shared task terms and no attempts were made to de-anonymize users. We acknowledge that predicting mental states has dual-use risk, as it can be used to survey users without their consent.

Acknowledgments

We acknowledge the use of Gemini (Google) for assistance in architectural strategizing and code refinement. All final implementations and results were independently verified by the authors.

References

Troy Arthur, Aidan Kelley, and Sierra Reschke. 2026. NLPGroup8 at SemEval-2026 Task 2: Model Weights.

Bob Durrant and Nick Lim. 2020. A Diversity-aware Model for Majority Vote Ensemble Accuracy. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, pages 4078–4087. PMLR.

Eleni Kamateri and Mike Salampasis. 2025. An ensemble framework for text classification. *Information*, 16:85.

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*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *ICLR*.

Raghavendra Pappagari, Piotr Żelasko, Jesús Villalba, Yishay Carmiel, and Najim Dehak. 2019. Hierarchical transformers for long document classification. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 838–844.

Xintong Shi, Wenzhi Cao, and Sebastian Raschka. 2023. Deep Neural Networks for Rank-Consistent Ordinal Regression Based On Conditional Probabilities. *Pattern Analysis and Applications*, 26(3):941–955.

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.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489, San Diego, California. Association for Computational Linguistics.

A Appendix

A.1 Tables

Metric	text_id 407	text_id 408	text_id 409	text_id 410
user_id	10	10	10	10
text	Tired, Exhausted, Calm, Content, Happy	Content, Calm, Sleepy, Lazy, Relaxed	Content, Calm, Relaxed, Happy	Calm, Content, Happy, Relaxed
timestamp	6/9/2021 12:11	6/10/2021 17:08	6/11/2021 12:03	6/11/2021 17:18
collection	1	1	1	1
is_words	TRUE	TRUE	TRUE	TRUE
valence	-1	2	2	2
arousal	0	0	0	1
state_change_valence (2A)	3	0	0	–
state_change_arousal (2A)	0	0	1	–
group (2B)	1	1	2	2
disposition_change_valence (2B)	1.5	1.5	1.5	1.5
disposition_change_arousal (2B)	0.5	0.5	0.5	0.5

Table 4: Sample Entry From Training Set

Task	Optim.	Epochs	Batch	Model LR	RoBERTa LR
1	AdamW	4	16	1.00E-03	2.00E-05
2A	AdamW	5	4	1.00E-04	2.00E-05
2B	AdamW	4	4	1.00E-04	2.00E-05

Table 5: Hyperparameters used for training.

Metric	Composite	Seen User	Unseen User	Words Only	Essay Only
r_composite	0.688	0.698	0.679	0.69	0.666
r_between	0.777*	.812*	.741*	.79*	.736*
r_within	0.572*	0.533*	.607*	.555*	.583*
mae_composite	NAN	1.796	1.986	1.953	1.868
mae_between	1.89	1.827	2.026	1.98	1.907
mae_within	1.925	NAN	NAN	NAN	NAN

Table 6: Performance metrics for Valence across different conditions. * Indicates $p < .01$

A.2 Dataset Analysis

Exploratory analysis was run on the training set. The original training set consisted of 2764 entries across 137 distinct users. Arousal had majority zero datapoints, with Arousal 0 at 44.11% of the training set. Valence data was more evenly balanced between negative and positive values. Negative values (-1, -2) comprised 27.02% of Valence data. Positive values (1, 2) comprised 40.57%. No single Valence value comprised more than 33% of the data.

Metric	Composite	Seen User	Unseen User	Words Only	Essay Only
r_composite	0.416	0.297	0.601	0.574	0.335
r_between	0.455*	0.306	.717*	.629*	.343*
r_within	0.376*	.287*	.452*	.514*	.326*
mae_composite	0.395	0.436	0.354	0.361	0.45
mae_between	0.25	0.314	0.185	0.256	0.313
mae_within	0.523	0.543	0.503	0.457	0.569

Table 7: Performance metrics for Arousal across different conditions. * Indicates $p < .01$

	Subtask 2a		Subtask 2B	
metric	**Valence**	**Arousal**	**Valence**	**Arousal**
R	0.152	0.126	0.354	0.388*
MAE	1.42	0.838	0.425	0.393

Table 8: Performance metrics (R and MAE) for Valence and Arousal across Subtask 2a and Subtask 2B. * Indicates $p < .01$

Value	Count	%
0	1097	44.11
1	913	36.71
2	477	19.18

Table 9: Arousal Distribution

Value	Count	%
-2	330	13.27
-1	342	13.75
0	804	32.33
1	482	19.30
2	529	21.27

Table 10: Valence Distribution