SemGroup8 at SemEval-2026 Task 2: A Diverse Ensemble for Emotional State Prediction

Troy Arthur
College of Engineering / University of Colorado, Boulder
trar3243@colorado.edu

Sierra Reschke
College of Engineering / University of Colorado, Boulder
sire7023@colorado.edu

Aidan Kelley
College of Engineering / University of Colorado, Boulder
aike6451@colorado.edu

Abstract

The second shared task of SemEval2026 aims to predict individuals' felt emotional states on the Affective Circumplex Model's emotional valence and arousal plane based on written text. While emotional valence and arousal are typically described as theoretically non-discrete states, categorical labels are often applied to emotional states for the purpose of linguistic communication. This discrete linguistic labeling may influence cognitive processing of emotion, which blurs the theoretically continuous nature of emotion. To resolve these issues, we propose an ensemble approach to emotional text classification such that several ensemble members rely on a categorical approach while others rely on a continuous approach. This diverse ensemble outperforms all individual models when predicting individuals' location on the affective circumplex plane.

GitHub Repository
Video Presentation

1 Introduction

1.1 Motivation

The determination of specific emotional states is relevant across a variety of applications, including mental health and consumer marketing. In these applications, conventional sentiment analysis focuses on a single dimension, valence. However, such approaches are inferior to dual-dimension analysis, as arousal provides key insights.

In the space of mental health, text-derived depression signals on social media have been used for targeted mental health outreach, with positive reception from those targeted individuals (Kelleher et al., 2018). The distinctions between valence and arousal for the purpose of mental health outreach are more useful than simple valence-based approaches because anxiety (low valence, high arousal) and depression (low valence, low arousal) differ in terms of effective treatments (Tiller, 2013).

In the space of marketing and customer feedback engagement, well-defined distinctions between the valence and arousal of individuals' mental states are particularly useful, as the degree of emotional arousal dictates an individual's willingness to take action while the degree of emotional valence dictates whether such action will be positively or negatively aligned with the goals of the product or service provider. While consumers may feel positively towards particular consumer categories, establishments which provide these products and engage in targeted marketing or consumer feedback engagement are better informed if the individual's arousal state is also determined. This is because an individual who feels positively towards a product category, yet has a low arousal state, is less likely to engage in consumer behavior with that product category than an individual who feels positively towards the product category and has a high arousal state (Han et al., 2007).

1.2 Conceptualizations of Affect

Russell (1980) developed the Circumplex Model of Affect, which conceptualizes affective states as some combination of a scalar arousal value (Y axis in fig. 1) and a scalar valence value (X axis in fig. 1). Consistent with this conceptualization, the training data provided included valence and arousal scores for each text entry. Russell's theoret-

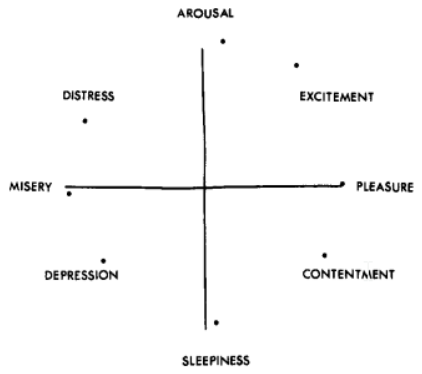

Figure 1: The Circumplex Model of Affect

ical framework would indicate that emotional states preexist as combinations of purely scalar valence and arousal values. However, other authors have suggested that linguistics play a role in affect such that if there does not exist a linguistic label for an emotional state then that state cannot exist. This work is ultimately derived from the Sapir-Whorf Hypothesis, which implies that emotional states are constrained to the categorical linguistic labels which we apply to our felt affect (Whorf, 2017). This conceptualization of affect has received some empirical support suggesting that emotional experiences are constituted by language-specific emotion words (Lindquist et al., 2015). This framework would support a categorical approach to the affect labeling task.

In lieu of an academic consensus, we opted for a diverse ensemble approach to sentiment analysis, where several different models are employed to predict valence and arousal scores such that several ensemble members are consistent with the categorical framework of affect while others are consistent with the continuous framework of affect.

2 Related work

Traditional sentiment analysis has focused nearly entirely on single-dimension polarity classification: deciding whether text is positive, negative, or occasionally neutral (Bordoloi and Biswas, 2023). While this works for broad judgments, it does not capture the level of physiological activation behind an emotional state, meaning an important part of affective experience is left out (Han et al., 2007). Modeling both dimensions provides a more complete picture of emotion, especially in cases where arousal meaningfully changes the interpretation of negative affect, such as differentiating anxiety from depression (Kelleher et al., 2018).

In modern NLP, transformer-based models such as RoBERTa are widely used for sentiment and emotion tasks because of 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, and prior research demonstrates that diverse ensembles can improve overall performance by reducing correlated errors across models (Durrant and Lim, 2020).

Our system draws from all of these directions. We combine continuous regression models, categorical classifiers, and ordinal-regression architectures into a single ensemble. This design reflects the lack of consensus in the literature about how emotion should be represented and allows the model to capture complementary signals that a single modeling approach might overlook.

3 Data and Task Description

3.1 Data

Subtask 1 Training data was released by task administrators as a .csv file. It contained text entries provided by service industry workers, prompted to describe their emotional state. The original training set consisted of 2764 entries across 137 distinct users. We split the original training set into a development set (10%: 277 entries) and a true training set (90%: 2487 entries). We will refer to these two sets derived from the original training set as the training set and the development set. The true test set for task 2 will be released in January 2026, and so we used the development set to mimic the test set. The following data was contained in each CSV column:

Data Entry Field	Content
user id	10
text id	408
text	Tired , Exhausted , Calm , Content , Happy
timestamp	2021-06-09 12:11:03
Collection Phase	1
is words	TRUE
valence	-1.0
arousal	0.0

Table 1: Sample Entry From Dataset

Each participant received a unique User ID; useable to track multiple text entries from the same

Arousal Value	Arousal Count	Arousal %
-1	1097	44.11
0	913	36.71
1	477	19.18

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

Figure 2: Analysis of Arousal and Valence values in training set.

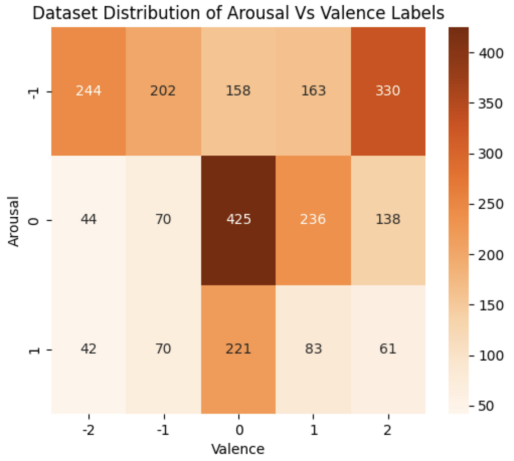

Figure 3: Heatmap of Arousal and Valence value distribution in training set.

participant. Each individual text entry received a unique Text ID. The text column contained the written responses from the participants; open-ended English responses of varying lengths. Punctuation in the text was broken out with spaces. The timestamp column marked when the text entry was recorded; used to chronologically order the responses from a single participant. The collection phase field marked when in data collection an entry was received. The "Is Words" field is a TRUE or FALSE value describing the two possible types of text entries. TRUE designates a word list self-describing participant emotions, broken by commas. FALSE designates an essay style prompt, self-describing participant emotion in open ended form. The valence field contained a value in the set [-2,-1,0,1,2] and the arousal field contained a value in the set [-1,0,1]. Both the valence and arousal fields were based on the user's subjective judgment of how "good" they were feeling (valence) and how "physiologically aroused" they were feeling (arousal).

3.2 Dataset Analysis

Light exploratory analysis was run on the training set. This was done to reveal points of under/over-representation in the data - which can aid in analysis of model performance later. To begin, a tally of datapoints was taken for each possible arousal and valence value. Because all text entries in the dataset received an arousal and valence score, datapoints count for a tally along both axes.

Arousal predominantly consisted of negative datapoints, with the Arousal -1 comprising 44.11% of the overall training set. Positive values only comprised 19.18% of the arousal Data. Valence data was more evenly balanced between negative and positive values. Negative values (-1, -2 combined) comprised 27.02% of the valence data. Positive values (1, 2 combined) comprised 40.57%. Valence was also more evenly distributed among its component values - no single value occupied more than 33% of the data. For both arousal and valence, 0 (Neutral) was the most or next-to-most common value.

Using Seaborn, a heatmap was produced to visually highlight any combinations with high or low data representation. The combination 0 Arousal, 0 Valence received the highest representation. (28.8% more datapoints than the next leading combination). There are additional concentrations at the top corners of the chart. [-1 arousal, -2 valence] and [-1 arousal, -2 valence]. The least represented combination is [1 arousal, -2 valence]. This is in a generally cold region at the bottom-left corner of the map. The coldest spot is less of an outlier than the hottest, with only 4.7% fewer datapoints than the next leading combination

Numerically, the training data was unevenly distributed. There was a high preference for several combinations of arousal/valence, and both axes skewed towards one end (negative values for arousal, positive values for valence). This could potentially impact model learning. As will be explored further, the ensemble model scored lower for Pearson's R in arousal, which had more heavily skewed data.

The data distribution must also be considered in the real-world context of the task. Over a given period of time, human emotion is unlikely to be evenly distributed across all emotional states; extreme elation or sadness would likely present less

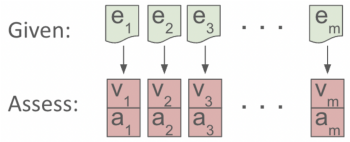

Figure 4: Subtask 1 required prediction.

often. The participant demographics must also be considered - data was only collected from United States service industry workers from 2021-2024. Participants from a single country and professional industry may trend toward emotional states. The data elicitation method may have produced bias towards particular emotional states (self-directed writing as compared to other possible methods such as interviews).

3.3 Subtask 1: Longitudinal Affect Assessment

Given a series of m texts in chronological order:

$$e_1, ..., e_m \tag{1}$$

A Valence and Arousal score must be assigned to each entry:

$$(v_1, a_1), ..., (v_m, a_m) \tag{2}$$

Texts may be of the essay or words list type.

For Subtask 1, there will be two marked groups in the test split of the data set to be released in January 2026. Unseen users will be new, with no previous entries in the training data. Seen users will have already appeared in the training data, but will appear in the test split at future timestamps.

4 Methodology

4.1 Ensemble Agreement

An ensemble approach to text classification can take the form of bagging, boosting, stacking, and voting. More generally, the ensemble approach uses a set (an ensemble) of different models which each contributes unique insights into the classification task. The models in the ensemble must be diverse, as redundant ensembles fail to outperform their constituent parts (Durrant and Lim, 2020). We constructed an ensemble of five models, with each ensemble member contributing unique advantages and limitations. The ensemble predicted four values; a continuous (float) score for valence, a continuous (float) score for arousal, a categorical (integer) bin for valence, and a categorical (integer) bin for arousal. Both the categorical and continuous predictions were essential for later errors analysis, although our primary evaluation metric, Pearson's R, relied on the continuous predictions. The categorical predictions were based on a majority vote of all constituent categorical predictions, while the continuous predictions were based on an average across all constituent continuous predictions.

4.2 Constituent models

Each of the five constituent models took as input RoBERTa-base CLS embeddings derived from the text concatenated with user-specific embeddings and a single value to indicate whether the entry is a set of feeling words or open-response. Each model corresponded to its own RoBERTa encoder, which during training was initially pulled from RoBERTa-base and then finetuned. The user-specific embeddings were also model-specific and learned during training. These user vectors supplied four dimensions to each model's feature vector. Therefore, all five constituent models have an input feature vector of size 773:

$d_{input} = d_{RoBERTa} + d_{user} + d_{is\_words}$
$= 768 + 4 + 1 = 773$

All five constituent models contained a single hidden layer of dimensionality $d_{input}/2$ followed by a GELU activation and 0.1 dropout:

$h = Dropout(GELU(W_1 x + b_1), 0.1)$

Of the five models, only model H had a separate hidden layer for valence and arousal; all other models shared the hidden layer between the valence and arousal heads.

The GELU activation function was chosen to maintain consistency with other authors' use of the RoBERTa encoder (Liu et al., 2019). All five constituent models used an AdamW optimizer for the same reason. All linear layers across all give models were initialized with the Xavier uniform initialization and all bias terms were initialized to zero.

4.2.1 Model A

Model A fits the continuous affect framework and contains a second linear layer which produces two logits, corresponding to valence and arousal. These logits are then passed through a sigmoid activation function to produce valence and arousal scores in the range [0,1]. These scores were then re-scaled to match the labels in the dataset during evaluation, while the loss function during training was based on the non-scaled values in comparison to down-scaled labels (such that the labels exist in the range

[0,1]).
$\hat{y}_{valence} = (\sigma(W_{valence}h + b_{valence})) \cdot 4 - 2$
$\hat{y}_{arousal} = (\sigma(W_{arousal}h + b_{arousal})) \cdot 2 - 1$
The categorical predictions for model A during evaluation were simply the continuous predictions, shown above, rounded to the nearest integer.
The rationale for using a sigmoid in this model was to avoid losses when the model produced high logits when the input feature vector contained values which map to "very high" or "very low" valence and arousal scores. The primary disadvantage is that the function space between two similar labels in the middle of the distribution-for example a valence score of 0 and 1-is very small.

4.2.2 Model B

Model B fits the continuous affect framework and contains a second linear layer which produces two scores, corresponding to valence and arousal scores. There was no activation function applied to these scores, and the loss function simply calculated the loss between the predicted values and the true labels.
$\hat{y}_{valence} = W_{valence}h + b_{valence}$
$\hat{y}_{arousal} = W_{arousal}h + b_{arousal}$
The categorical predictions for model B during evaluation were simply the continuous predictions, shown above, rounded to the nearest integer and clamped to exist in the range of the labels for both valence and arousal.
The rationale for the absence of an activation function on the two heads is that model B is meant to resolve model A's function density of middle-value predictions, while its primary disadvantage is that it creates losses when the feature vector and weights imply "very high" or "very low" scores, and the true label is simply the highest possible label for the category.

4.2.3 Model D

Model D fits the categorical affect framework and contains two heads; one for valence and one for arousal. The valence head contains five sets of weights, corresponding to the five different valence categorical labels supplied in the training data, while the arousal head contains three sets of weights, corresponding to the three different categorical arousal categorical labels. Each of the 8 sets of weights produced logits, and the predictions for valence and arousal were based on a softmax function applied to each logit within the relevant head.

$z_{valence} = W_{valence}h + b_{valence}$
$z_{arousal} = W_{valence}h + b_{valence}$
where z refers to class logits, and during evaluation the continuous predictions were based on the dot product between each label's predicted probability and its 0-indexed value, followed by rescaling from label indices to label values:
$\hat{y}_{valence} = SoftMax(z_{valence}) \cdot labels_{valence} - 2$
$\hat{y}_{arousal} = SoftMax(z_{arousal}) \cdot labels_{arousal} - 1$
While the categorical prediction was made simply by taking the argmax of each labels' logits.
During training, loss was calculated based on the difference between each class's softmax and the one-hot index label vector. The rationale for model D is that affect values, as previously described, may to some degree exist as independent categories with their own linguistic feature signals. The primary disadvantage of model D is that model D treats every score independently, without logic that scores which are numerically very close are more similar than scores that are numerically far apart.

4.2.4 Model G

Model G attempts to integrate both continuous and categorical approaches to the affect classifier task through the use of ordinal binary decomposition. This approach was based on work done by Shi et al. (2023). Model G Contains two heads, one for valence and the other for arousal. However, unlike model D, the valence head contains four sets of weights while the arousal head contains two sets of weights. The fundamental logic behind model G is that each head produces logits which represent the log odds of surpassing particular values. For example, if an individual has an arousal score of 1 (highest possible score), their model D class index label would be a one-hot $[0, 0, 1]$. Model G turns this one-hot vector into a vector of size two, with values $[1, 1]$. The first of two class values indicates that the first index was surpassed, and the second class value indicates that the second index was surpassed. During training, each class is trained based on its loss with the ordinal labels previously described after the class's logit is passed through a sigmoid. Put more formally,
$z_{valence} = W_{valence}h + b_{valence}$
$z_{arousal} = W_{arousal}h + b_{arousal}$
where z refers to class logits, and during evaluation the continuous prediction is based on a sum of sigmoid-converted values across all of the ordinal classes, followed by a rescale from 0-indexed labels to true values

$\hat{y}_{valence} = Sum(\sigma(z_{valence})) - 2$
$\hat{y}_{arousal} = Sum(\sigma(z_{arousal})) - 1$
while the categorical prediction is made by summing the number of class indices with values following the sigmoid greater than 0.5, followed by a rescale from 0-indexed labels to true values. Model G is considered to be the most theoretically aligned with the task, as it integrates both categorical and continuous affective frameworks.

4.2.5 Model H

Model H is a replica of model G, except the hidden layer is separated between the valence and arousal heads. Model H therefore has two separate hidden layers, each specialized in its own particular axis on the Circumplex Model of Affect. The rationale for the first four models all containing a single hidden layer is that valence and arousal interact with one another and correlate, while model H contributes diversity with a design that assumes a greater degree of independence between valence and arousal scores. The logic behind model H is that there are text features which uniquely predict valence, while other text features uniquely predict arousal.

4.3 Hyper Parameters

All hyper parameters are shown below. Each constituent model took its feature vector from RoBERTa CLS embeddings, and every constituent model had its own RoBERTa encoder which was fine-tuned with a seperate learning rate. All model hyperparamers were consistent across all constituent models.

Hyperparameter	Value
Optimizer	AdamW
Hidden Layer Dropout	0.1
Epochs	4
Batch Size	16
Model LR	.001
RoBERTa LR	.00005

Table 2: Hyper Parameters

5 Results

5.1 Evaluation Metrics

An evaluation function was implemented with a suite of metrics including Precision, Accuracy, Recall, F1, Mean Average Error (MAE), and Pearson's R. Because Arousal and Valence work on separate scales, evaluation metrics are computed separately for each. The following metrics refer to the ensemble model performance. With the exception of Pearson's R, metrics presume that all Arousal/Valence values exist as discrete, non-continuous values and use the categorical predictions described in section 4.

5.1.1 Precision, Accuracy, Recall, F1, MAE

Dev Precision: [Arousal: 0.6433, Valence: 0.5598]. Of the predictions that the model makes for a given class, precision describes how many were in that class in reality.

Dev Accuracy: [Arousal: 0.6273, Valence: 0.5407]. Of all the predictions made by the model, accuracy describes the portion that were correct.

Dev Recall: [Arousal: 0.6433, Valence: 0.5598]. Of the number of items that exist in a given class, recall describes how many the model identified.

Dev F1: [Arousal: 0.6334, Valence: 0.5433]. F1 measures how well the model makes predictions, averaged across 3 and 5 classes respectively for Arousal and Valence.

Individual F1 scores were also calculated for all component models. Model G performed the strongest for Arousal at 0.6145, and the strongest for Valence at 0.5049. This demonstrates that the ensemble schema outperformed any single component model. Training user embeddings also resulted in improvements for individual model performance; inducing improvements in average F1. Model A: [Before: 0.52715, After: 0.5479]. Model B: [Before: 0.5317, After: 0.5236]. Model D: [Before: 0.515, After: 0.55775].

Dev MAE: [Arousal: 0.3971, Valence: 0.5993]. MAE was implemented for its ability to assess the correctness of a model's guess. For both Valence and Arousal, there is a continuum of potential answers (-1 to 1, and -2 to 2, respectively.) MAE considers that a guess can be incorrect, but still better than others.

These metrics were also compared against results from a dummy baseline model which always predicted the average valence and arousal score across the training set. This comparison model received F1 [Arousal: 0.1883, Valence: 0.0964]. And MAE [Arousal: 1.0361, Valence: 0.6065], thus demonstrating significant ensemble performance improvements over the dummy model. Neither a baseline nor a starter pack was provided by task administrators.

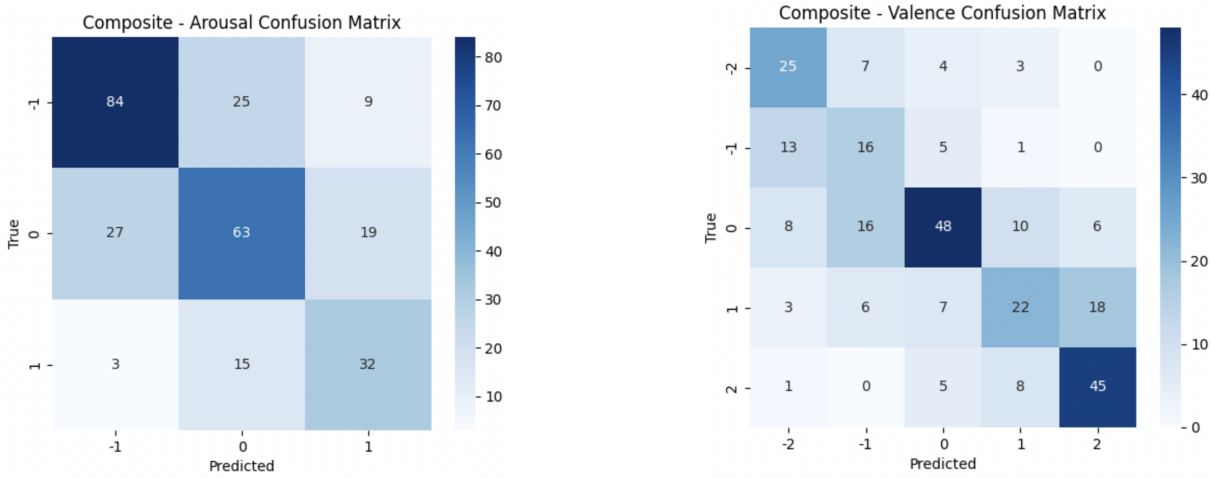

Figure 5: Confusion matrices demonstrating Ensemble prediction against true labels for Arousal and Valence.

5.1.2 Pearson's R

Although the task administrators did not clarify a particular evaluation metric, Pearson's R served as our primary evaluation metric for the models. This is because the administrators stated that predictions ought to be float values which implies a continuous evaluation metric. Pearson's R measures the linear relationship and strength of correlation between two datasets. A 1 or -1 value represents a perfect positive or inverse correlation, while values closer to 0 represent weak correlations. The Pearson's R values described are based on the continuous value predictions described in section 4. As with previous metrics, a separate Pearson's R value was calculated for Arousal and Valence. The final ensemble resulted in Pearson's R values [Arousal: 0.63, Valence: 0.77]. These are both considered strong positive correlations, indicating strong model performance over pure guessing.

Pearson's R (Arousal and Valence) was additionally calculated for all component models in the ensemble. The strongest individual Arousal performance was Model G with Pearson's R [Arousal: 0.61]. The strongest individual Valence performance was Model A with Pearson's R [Valence: 0.75]. This demonstrates that the ensemble model outperformed all component models in both Arousal and Valence.

Pearson's R for the dummy model resulted in [Arousal: 0, Valence: 0]. A weak correlation, demonstrating that the ensemble model could not have achieved its results through pure guessing.

5.1.3 Error Analysis and Confusion Matrices

To assist in error analysis, confusion matrices were generated from model class predictions. These compare model predictions (X-Axis) against true label values (Y-Axis). Within each matrix, the diagonal cells from top-left to bottom-right are correct predictions made by the model. Cells deviating from this diagonal show incorrect class predictions. Separate matrices were generated for arousal and valence due to the differing class amounts - a 3x3 matrix for arousal and a 5x5 matrix for valence. Cells display the number of model predictions, and a simple heatmap function was implemented for easy viewing of high/low activity areas.

The primary confusion matrices display the ensemble model performance. For ensemble class predictions, we retained the majority-votes compiled from all component models. These were compared against the true labels, which remained consistent regardless of model being used. Using the visual heatmap, the diagonal line of correct prediction/label combinations is visibly highlighted, indicating strong model performance. There are also no hotspots in the opposite corners of the matrices (such as Prediction 1, True Label -1), showing the model did not make a large amount of egregious mis-predictions.

However, there were locations in the confusion matrices with more tendency for error. Negative arousal (at 71.19%) received more accurate prediction than neutral arousal (57.80%) or positive arousal (64%). This may have been impacted by the training data, which as can be seen in Figure 2, is more heavily skewed towards negative arousal, thus providing the models with more learning ma-

terial relative to the other classes.

For valence, the ensemble model struggled most with the intermediate values of 1, -1. With correct predictions rates of 39.29% and 45.72%. Valence value 0 received 54.55% accuracy. And the extreme ends of -2, 2 performed best with rates of 64.1% and 76.27% respectively. In this case, data distribution did not trend in the same direction as these error rates, so this was unlikely to be a major contributing factor. However, the component models may have had an easier time distinguishing the far ends of the Valence spectrum than in-between gradations of -1, 1. This can be seen in cells [Predicted: -2, True: -1] and [Predicted: 2, True: 1]. There are large clusters on both ends where the ensemble predicted an extreme -2 or 2 value that was actual true label -1 or 1.

Arousal and Valence confusion matrices were also constructed for all individual component models. These demonstrate different issue areas across the individual models, which the ensemble schema sought to smooth over and outperform. For example, Model B is seen to particularly struggle with predicting negative arousal values, at an over 50% error rate. Model G, meanwhile, prefers prediction of negative valence values, and this tendency is seen to be smoothed out in the ensemble confusion matrix.

6 Conclusion

This work presented SemGroup8's system for SemEval-2026 Task 2, focused on predicting valence and arousal from written text. Because the literature does not fully agree on whether emotions should be modeled as continuous states or as discrete categories, we designed an ensemble that incorporates both perspectives. Our system combines continuous regression models, categorical classifiers, and ordinal-regression architectures so that each model contributes different strengths to the overall prediction.

Across all evaluation metrics, the ensemble outperformed every individual model. Most importantly, it achieved strong Pearson's R values (0.63 for arousal and 0.77 for valence) showing a consistent linear relationship between predictions and true labels. The confusion matrix analysis supported this by illustrating that the ensemble corrected several systematic tendencies seen in individual models, especially at the edges of the affective spectrum.

Overall, the results suggest that emotional text analysis benefits from incorporating multiple theoretical viewpoints rather than relying on a single modeling assumption. As NLP systems continue to expand into mental-health applications, user experience, and longitudinal well-being tracking, approaches that bridge categorical and continuous representations may offer more stable and interpretable performance than methods that commit to only one interpretation of emotion.

7 Limitations and Future Work

Although the ensemble achieved strong results, several limitations point toward useful directions for future work.

First, because the training and development sets were created using a random split, almost every user appeared in both partitions. This prevented us from meaningfully evaluating performance on unseen users, even though the official test set will include both seen and unseen participants. As a result, we could not fully assess the generalization impact of the user embeddings. A more intentional data split, one that explicitly reserves certain users for development, would allow for a clearer analysis.

Second, the dataset is noticeably imbalanced, particularly for arousal, where negative values dominate. This skew likely influenced model behavior and may explain some of the error patterns visible in the confusion matrices. Future systems could address this with weighted losses, resampling, or representation-level techniques that help the model pay equal attention across the affective spectrum.

Third, the dataset consists of two very different types of text: short emotion word lists and longer free-form narrative responses. We included a binary feature to indicate format, but more explicit modeling of these differences such as domain adaptation or separate encoders could improve performance.

Finally, although our ensemble intentionally mixes diverse model types, we did not run extensive variations beyond the dummy baseline and the user-embedding comparison. Future work could examine how much each component contributes, explore alternate ensemble-weighting strategies, or evaluate shared versus independent encoders.

More broadly, extensions such as uncertainty estimation, temporally aware modeling, or personalized affect priors could further improve performance on real-world longitudinal emotion data.

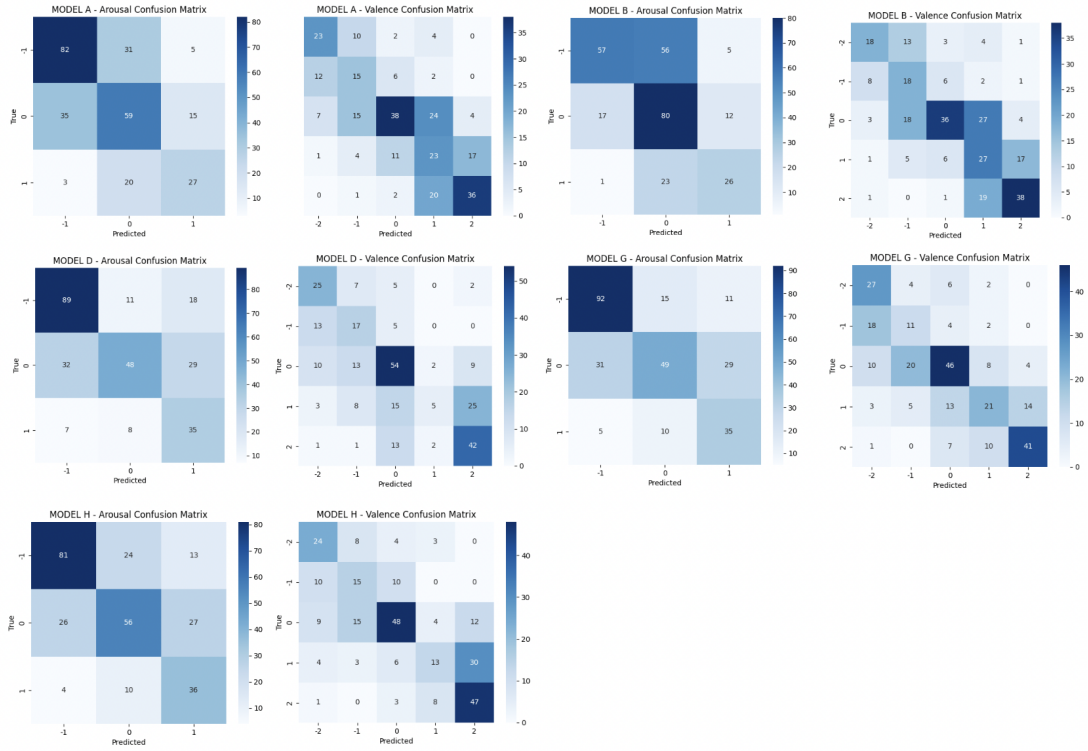

Figure 6: Arousal, Valence Confusion Matrices for all component models.

References

Monali Bordoloi and Saroj Kumar Biswas. 2023. Sentiment analysis: A survey on design framework, applications and future scopes. *Artificial Intelligence Review: An International Science and Engineering Journal*, 56(11):12505–12560. Place: Dordrecht Publisher: Springer Netherlands.

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. ISSN: 2640-3498.

Seunghee Han, Jennifer S. Lerner, and Dacher Keltner. 2007. Feelings and Consumer Decision Making: The Appraisal-Tendency Framework. *Journal of Consumer Psychology*, 17(3):158–168.

Erin Kelleher, Megan Moreno, and Megan Pumper Wilt. 2018. Recruitment of Participants and Delivery of Online Mental Health Resources for Depressed Individuals Using Tumblr: Pilot Randomized Control Trial. *JMIR Research Protocols*, 7(4):e9421. Company: JMIR Research Protocols Distributor: JMIR Research Protocols Institution: JMIR Research Protocols Label: JMIR Research Protocols Publisher: JMIR Publications Inc., Toronto, Canada.

Kristen A. Lindquist, Ajay B. Satpute, and Maria Gendron. 2015. Does language do more than communicate emotion? *Current directions in psychological science*, 24(2):99–108. Num Pages: 99-108.

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 [cs].

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

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. ArXiv:2111.08851 [cs].

John W. G. Tiller. 2013. Depression and anxiety. *Medical Journal of Australia*, 199(6).

Benjamin Lee Whorf. 2017. The Relation of Habitual Thought and Behavior to Language. *et Cetera*, 74(1/2):35–58. Num Pages: 35-58 Place: Concord, United States Publisher: Institute of General Semantics, Inc.