Anual Proj DeepLearning

Maor Moshe, David Oriel, Idan Salomon,

September 2025

1 Abstract

We address prediction of affective dimensions from annotated text. The dataset exhibits substantial missingness and severe class imbalance, both known to bias models toward majority classes [1]. We mitigate these issues with a correlation-preserving imputation that maintains the emotion–emotion structure, and with a contrastive augmentation procedure based on Label–Indicative Coefficients (LIC) [5] to strengthen minority classes. We then employ Concept Bottleneck Models (CBMs) [2] that map sentence embeddings to interpretable concepts (valence, arousal) before predicting interest. On a clean test set (no augmented sentences), the sentence-embedding backbone attains the strongest macro/weighted F1 across targets (Table 2). Error and calibration analyses reveal adjacent-class confusions in the mid-range and mild over-confidence, while bottleneck interventions indicate that valence+arousal alone are insufficient to fully recover interest—motivating a richer yet interpretable concept set (e.g., VAD or basic emotions). We also show that adding augmented sentences to the test set causes semantic leakage that inflates reported performance.

2 Exploratory analysis

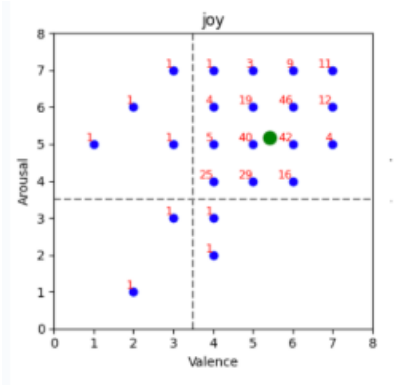
The exploratory phase was designed to uncover patterns in the dataset, quantify data quality issues, and identify which emotional dimensions are most promising for predictive modeling.

Category Distributions and Missing Data

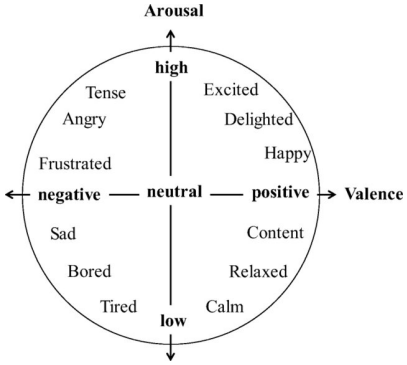
An initial inspection revealed substantial missing values and strong class imbalance across emotion categories. Some emotions appear frequently, while others are sparsely represented, with minority categories sometimes orders of magnitude smaller than majority ones. Such imbalance risks biasing models toward dominant classes and therefore requires mitigation. Missing data further complicates matters: in 24 categories, roughly half the observations are absent. Our preliminary imputation strategy (assigning 1 with probability 0.8 and 2 with 0.2) assumed missingness reflected irrelevance rather than neutrality. However, this simplistic approach distorted correlations between emotions, flattening the maps. For example, among 1,100 missing *Joy* cases, *Happiness* is not 1 in 270 instances and exceeds 2 in 100, showing that the naive 80/20 fill fails to capture the observed joint structure. More sophisticated imputation is required.

To address this issue, we adopted a two-step approach. First, we focused on emotions that were more *predictable*, defined by sufficient variance, relatively few missing values, and stable distributions. Second, we explored imputation methods designed to preserve correlation structures, such as k-Nearest Neighbors and correlation-preserving techniques. Ultimately, we selected imputing missing values while *preserving the correlation map* between emotions obtained from the unimputed data. Since imputed values were not always integers, we rounded them to the nearest integer using a 0.5 threshold to maintain ordinal consistency.

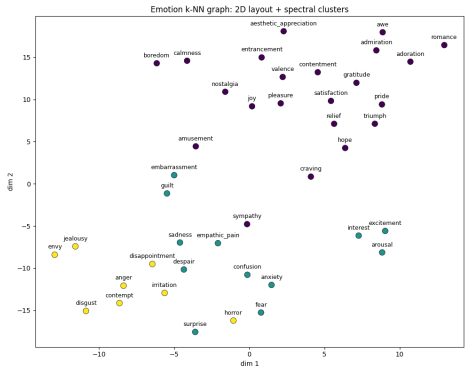
This strategy allowed us to preserve the original correlation maps while minimizing distortions. Moreover, by examining the variance and outlier prevalence of each emotion, we identified *interest* as the most predictable and stable target for downstream modeling. Crucially, we first verified that the empirical correlation structure is theory-consistent: the emotion–by–emotion correlation matrix exhibits the expected large-scale organization along valence and arousal. To substantiate this, we built a mutual k-nearest-neighbors (kNN) graph over emotions using correlation distance ($D = 1-r$), converted it to a Gaussian-weighted affinity, obtained groups via spectral clustering, and embedded the nodes in two dimensions with t-SNE for visualization. Because the construction is correlation-based, spatial proximity reflects co-occurrence patterns (e.g., *joy* with *pleasure*/*happiness*); the discovered groups are semantically coherent and align with the empirical and theoretical maps shown below.



(a) Correlation map for Joy - it can be seen that the mode(in green) is inline with the theoretical value between valence and arousal



(b) Theoretical correlation structure



(c) kNN graph (t-SNE; labels above points)

Figure 1: Empirical, theoretical, and graph-based views of the emotion structure.

We then quantitatively compared three imputation strategies under a train/test masking protocol (holding out observed entries as pseudo-missing in the test set): a fixed 80%→1/20%→2 baseline, a correlation-based iterative imputer, and a kNN imputer. Performance on the test set showed that the correlation-preserving imputer achieved the lowest MAE and RMSE and exhibited no systematic bias (paired tests not significant), outperforming both baselines (Figure 2).

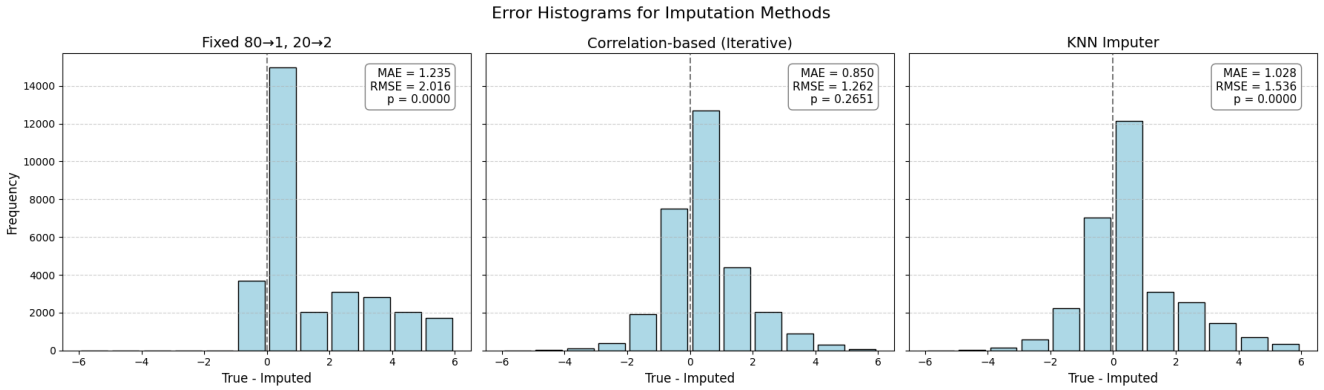


Figure 2: Error histograms on the test mask for three imputation methods (lower is better).

Feature Analysis: LLaMA Embeddings

The feature matrix X was provided precomputed: for each segment, we received a 4096-dimensional vector obtained by mean-pooling LLaMA layer-28 activations; stacking 10,361 segments yields $X \in \mathbb{R}^{10361\times4096}$. Unless stated otherwise, features were z-scored across segments prior to analysis.

Preprocessing and geometry. A correlation map of the 4096 units revealed strong blocks of positive correlations, indicating multicollinearity in X. Principal component analysis (PCA) quantified the effective dimensionality: the first 100 components explained about 60% of the variance, with an elbow around 60–120 components. In the remainder, we use PCA primarily as an orthogonalization and dimensionality-reduction step (and, where relevant, for pre-whitening), without ascribing semantic meaning to individual components.

Modeling implications. The strong multicollinearity and heavy singular-value decay in X motivate models that (i) regularize or sparsify coefficients and (ii) operate in a reduced subspace. Hence, as a baseline, we pursued ridge regression on either the raw features or on the top PCA components, alongside the ordinal/classification heads used later. This choice is supported by the PCA result (100/4096 ≈ 2.4% of directions explain 60% of the variance),

suggesting that good performance may be achievable using only a fraction of the features while improving stability and interpretability.

2.1 Evaluation Procedure

The evaluation of our models considered both the multiclass setting and the strong class imbalance present in the dataset. Relying solely on overall accuracy would bias results toward majority categories, as a model could achieve high accuracy by predicting only frequent classes. To address this, we prioritized metrics that balance sensitivity to minority categories with overall predictive performance. Specifically, we report **weighted F1-scores**, which average precision and recall while accounting for class frequencies; and **class-weighted metrics** more generally, which ensure that minority classes contribute proportionally to evaluation. In addition, the **confusion matrix** provides qualitative insight into systematic misclassifications, particularly between semantically adjacent classes.

Data Partitioning. We chose to split the data such that each participant contributes samples to both the training and the test sets. In principle, one might prefer a subject-level split, training on some participants and testing on others. However, with only twelve participants in total, the linguistic diversity within any subset is too limited to capture the semantic richness necessary for a model to generalize across individuals. In practice, the heterogeneity in style and expression of a small subgroup of twelve speakers does not provide a sufficient basis for learning generalizable patterns about human emotional expression. For this reason, we opted for a row-wise split across all participants, ensuring that every subject is represented in both sets.

To justify this design choice, we trained a simple MLP on the data of a single participant and then evaluated it both on the same participant and on the remaining participants. As expected, performance was much stronger in the within-subject setting, while cross-subject generalization was markedly weaker. This result indicates that models trained on such limited speaker diversity cannot reasonably be expected to generalize to new individuals. The comparison is shown in Table 1.

Table 1: Performance of a simple MLP trained on one subject and tested on the same vs. a different subject.

	Accuracy	Balanced Acc.	Weighted Acc.	Weighted F1
Same Subject	0.718	0.765	0.699	0.714
Other Subject	0.236	0.189	0.239	0.240

3 Baseline and Improved Prediction Models

3.1 Linear sparse baseline (Lasso / Ridge as ablation)

Motivation. We established a *linear* baseline for three reasons. (i) It provides a sanity check on the learnability of the task under simple hypotheses (additivity and monotonicity in the embedding coordinates). (ii) In our setting the feature matrix is high-dimensional (p=4096) and strongly correlated (Sec. §Exploratory analysis), while sample size is moderate; this is the classical regime where shrinkage estimators are preferred over unregularized least squares due to variance inflation under multicollinearity.

Model and objective. For each target $y \in \{\text{arousal}, \text{valence}, \text{interest}\}$ We apply Lasso regression with the following objective:

$$\min_{b,w} \frac{1}{n} \sum_{i=1}^{n} \left(y_i - b - x_i^\top w\right)^2 + \lambda \|w\|_1,$$

Let $X \in \mathbb{R}^{n \times p}$ denote the sentence-embedding feature matrix after dropping metadata columns, and let y be the corresponding $1-7$ ordinal labels for the target under study. We set λ=0.1 to impose stronger regularization. For each target, we retained only rows with non-missing labels. Stratified K=5-fold cross-validation was used on the discrete labels to preserve class proportions in each fold. Within each fold, the model was trained on the training partition and evaluated on the held-out split. To compare with ordinal classifiers, continuous predictions were mapped back to discrete classes by rounding and clipping to the valid range $[1, 7]$:

$$\widehat{\text{class}}(x) = \text{clip}\big(\text{round}\big(\hat{y}(x)\big),\, 1,\, 7\big)\,.$$

Evaluation metrics. Because the model is trained as a regressor, we report fold-level MAE and RMSE on the raw (continuous) outputs. For qualitative error structure we additionally compute confusion matrices and a classification report on the rounded predictions, which makes the linear baseline directly comparable to later models.

outputs The linear baseline was evaluated on the emotion dimensions of Valence, Arousal, and Interest and achieved very limited classification performance. Across the five folds, the macro-F1 score was consistently low($\approx$ 0.10 on average), and the weighted-F1 score was only slightly higher ($\approx$ 0.15–0.17). Accuracy across folds remained in the range of 15–19%. The qualitative error structure highlights the model's limitations: predictions collapsed toward the central categories (classes 4,5), while the minority classes (1, 2, and 7) were almost never identified. .

Limitations A purely linear model cannot represent interactions between linguistic cues; thus, it underfits whenever the mapping from text embedding space to affect is intrinsically non-linear (for example, combinations of cues that only jointly signal interest). Moreover, under strong collinearity Lasso's feature selection is not unique: different folds may select different—but statistically equivalent—coordinates. These limitations underscore why the linear baseline performs poorly and motivate the transition to non-linear models in the subsequent sections.

Takeaway. The sparse linear baseline establishes that a small, linearly separable subspace of the embeddings may carries predictive signal for Arousal, Valence, and (more weakly) Interest, with errors concentrated between adjacent classes in the mid-range. However, because the data remain tabular in structure but exhibit non-linear relationships, there is clear motivation to move beyond linear models and examine tree-based approaches, which are well suited to capturing such interactions. If these still prove insufficient in extracting signal, neural network models provide a natural next step, these directions establish a principled roadmap for building upon the linear baseline.

3.2 Improved Model: XGBoost

Since the linearity/additivity assumptions did not hold and lasso regression underperformed, we moved to a tree–based gradient boosting model (XGBoost), which is better suited to nonlinearity and complex interactions.

Motivation XGBoost is particularly well-suited for our setting because it **handles nonlinearity and interactions**, as decision trees naturally capture thresholds and feature relationships that linear models overlook. It also offers **built-in regularization**, applying L_1/L_2 penalties on leaf weights, shrinkage (learning rate), and row/column subsampling, which collectively help control overfitting. In addition, the method is **robust to multicollinearity and scaling**, since trees are less sensitive to correlated features and do not require feature standardization.

Training setup We used stratified k-fold cross-validation to respect the imbalanced and ordinal nature of the targets. Concretely, we binned the target into ordinal bins solely for the *splitter* to preserve the class/score distribution in each fold, trained an `XGBRegressor` on the continuous labels within each fold, and evaluated on the held-out fold. Alongside a rounded-class accuracy (by rounding predictions to the nearest integer), we report regression metrics (MAE/RMSE), which are more informative for an ordinal/continuous target.

Results Across folds, the mean rounded-class accuracy was approximately 0.19. Since this metric is coarse for an ordinal regression task, we emphasize MAE/RMSE and provide per-fold confusion matrices (after rounding) to visualize typical confusions (e.g., adjacent classes).

4 Feature importance analysis

We interpret the trained XGBoost model using SHAP (SHapley Additive exPlanations) [3, 4], which assigns each feature i a local contribution $\phi_i(f, x)$ to a specific prediction $f(x)$ while satisfying the Shapley axioms (efficiency, symmetry, dummy, and additivity).

SHAP on the XGBoost model. We computed TreeSHAP on each fold of the K-fold cross-validation used to train the model. Figure 3 compares the top features across three folds and shows that the identities of the highest-ranked features vary across resamples.

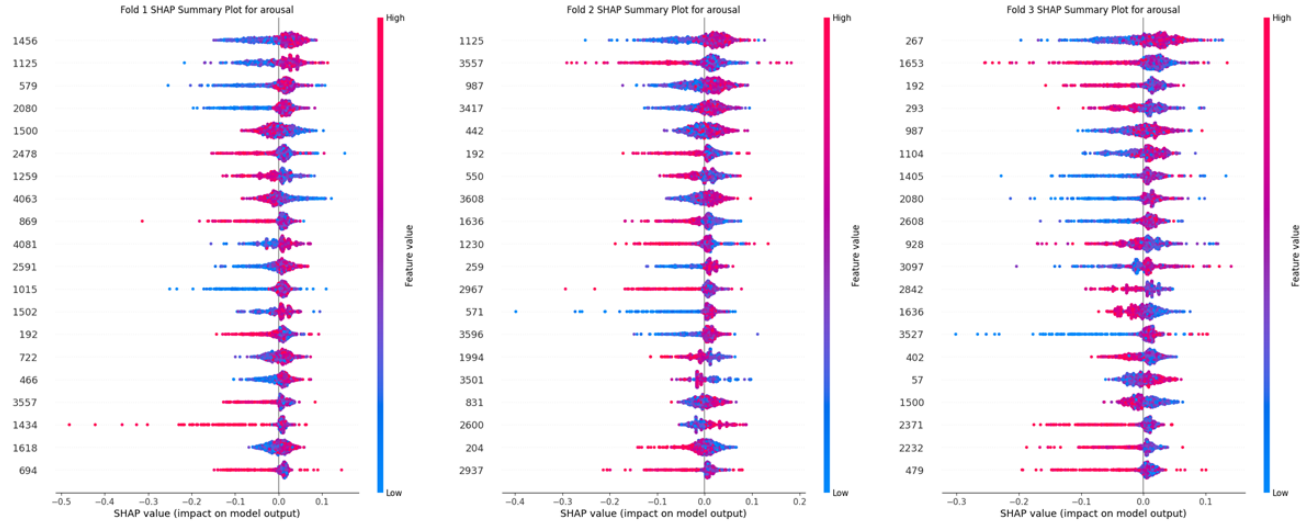


Figure 3: Top features across three folds (left to right). The set of top-10 features changes markedly between folds.

We also compare targets: Arousal, Interest, and Valence. The overlap of their top-50 SHAP features indicates partially shared drivers between Arousal and Interest, while Valence relies on different predictors.

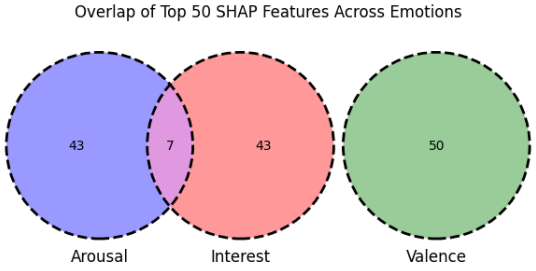


Figure 4: Overlap of top-50 SHAP features across emotion targets. Arousal–Interest share a small subset; Valence is largely disjoint.

From Fig. 3, high absolute SHAP values concentrate in a small number of features (sparse attribution), with mixed signs indicating both risk-increasing and risk-decreasing effects depending on the feature value (color). Fig. 3 also reveals pronounced *instability* of the top-ranked features across folds: the identities of the top-10 features change markedly even though their *magnitude* of contribution is comparable. Finally, the cross-target comparison in Fig. 4 shows a modest overlap between *Arousal* and *Interest* (seven shared features among their top-50), whereas *Valence* relies on a largely distinct set of predictors.

Taken together, these findings indicate that the signals are unstable across folds, with no single set of "leading" features emerging. This suggests that, despite tuning, the improved XGBoost model did not succeed in capturing robust and generalizable predictive structure,

5 Research Task

In this section, we describe two complementary research directions implemented in our project: (1) constructing contrastive samples to mitigate class imbalance in emotion-labeled text data, and (2) applying *Concept Bottleneck Models* (CBMs) to introduce an interpretable intermediate representation for downstream prediction.

5.1 Constructing Contrastive Samples

Method To rebalance the valence and arousal emotion classes, we follow the by the *Constructing Contrastive Samples* framework [5], we treat each discrete emotion rating as the target class to be predicted and synthesize minority–class variants of existing sentences. We prioritize edits that (i) are plausibly label–preserving and (ii) shift lexical content toward features most indicative of the target class. To that end, we compute a lexical importance score (LIC) for each word with respect to each class and generate *positive contrastive* samples for minority classes by replacing high–LIC, non–stopword tokens with semantically close alternatives retrieved from a Word2Vec model.

LIC Definition For each word w and class c, we define a *Label–Indicative Coefficient* (LIC) that compares how characteristic w is for c relative to all other classes:

$$\mathrm{LIC}(w, c) = \left(\frac{\mathrm{TF}_c(w) - \mu(\mathrm{TF}_{-c}(w))}{\sigma(\mathrm{TF}_{-c}(w)) + \varepsilon} \right) \cdot \mathrm{IDF}(w),$$

where $\mathrm{TF}_c(w)$ is the term frequency of w computed on the subset of documents with label c, $\mathrm{TF}_{-c}(w)$ is the vector of term frequencies of w over all *other* classes, $\mu(\cdot)$ and $\sigma(\cdot)$ are its mean and standard deviation, $\mathrm{IDF}(w)$ is inverse document frequency, and $\varepsilon > 0$ is a small constant (to avoid division by zero when $\sigma = 0$). Larger $\mathrm{LIC}(w, c)$ indicates w is more indicative of class c [5].

Implementation for each emotion - valance and arousal.

1. **Compute class counts and targets.** Let n_c be the number of samples in class c; let $N_{\max} = \max_c n_c$. For each class c, set the augmentation budget to $N_{\max} - n_c$.
2. **Score words per class.** For every word w and class c, compute its LIC score (higher means more indicative of class c).
3. **Generate positive contrastive samples.** For each minority class c:
   (a) Select a sentence labeled c and identify the highest–LIC token w that is not a stopword.
   (b) Using a pretrained Word2Vec model, retrieve a near–neighbor w' (semantically close but not identical to w) and replace $w \to w'$ to form a new sentence labeled c.
4. **Balance.** Repeat Step 3 until every class reaches its target size $N_c^{\text{target}} = \min\big(N_{\max}, 2\, n_c^{\text{orig}}\big)$. Stop when $n_c \geq N_c^{\text{target}}$ for all c. When we augmented the sentences for one emotion, the relative rates of the other emotions were copied as well, which introduced imbalance again. To address this, we implemented a recursive function that deletes augmented sentences until the desired balance is achieved, as illustrated in Figure 5.

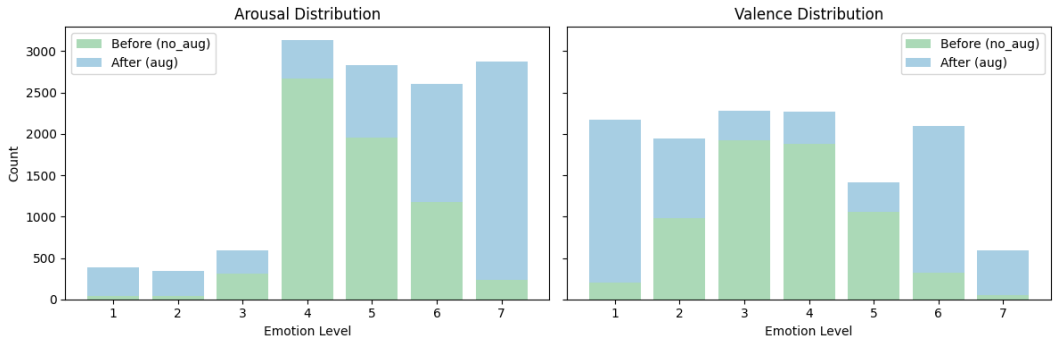


Figure 5: Class distribution histograms before and after applying the LIC-based augmentation. The **green bars (lighter)** represent the original (before augmentation) distribution, while the **blue bars** show the distribution after augmentation.

5.2 Concept Bottleneck Models

Background. *Concept Bottleneck Models* [2] introduce an intermediate, human-interpretable layer of concepts c between the raw input x and the final prediction y. The model decomposes into:

$$g : x \to c, \quad f : c \to y, \quad \hat{y} = f(g(x)),$$

where g predicts interpretable concepts and f uses these to predict the target label. Training can follow *Independent*, *Sequential*, or *Joint* paradigms [2], and the structure enables direct concept-level interventions during inference.

Our Implementation. We structured the task as:

1. **Input representation:** three sentence embedding types (`bert_cls_embedding`, `bert_mean_embedding`, `sen_embedding_qwen0.6b`).
2. **Concept predictor g_θ:** a shared MLP trunk with two heads outputting logits for arousal and valence (each with $K = 7$ classes), trained using CrossEntropy loss.
3. **Label predictor f_ϕ:** a smaller MLP mapping concatenated concept logits (or probabilities) to the final interest label.
4. **Training:** sequential scheme—first train g_θ, then freeze and train f_ϕ.

Objective Function. The CBM can be expressed as:

$$\min_{\theta,\phi} \underbrace{\mathcal{L}_y\left(f_\phi(g_\theta(x)),\, y\right)}_{\text{label loss}} + \lambda \sum_{j\in\{\text{aro,val}\}} \underbrace{\mathcal{L}_{c_j}\left(g_{\theta,j}(x),\, c_j\right)}_{\text{concept loss}},$$

where λ controls the trade-off between label and concept supervision. In our sequential setup, the concept and label losses were optimized separately.

5.3 Advanced model- Balanced and Bottleneck model

Our pipeline by combining two complementary strategies: (1) **balancing** the dataset through LIC-based contrastive augmentation, which ensured sufficient representation of minority classes, and (2) **Concept Bottleneck Models (CBMs)**, which insert an interpretable intermediate layer of human concepts (valence, arousal) between text embeddings and the final label (interest). This dual approach both improves predictive performance and allows concept-level explanations of model behavior.

Pipeline. The overall architecture is illustrated in Figure 6. Given an input sentence x, we first compute its embedding using a pretrained encoder. These embeddings are passed to a *Concept Head*, an MLP trained to predict logits for arousal and valence (14-dimensional output). The resulting concept logits are then fed into a second MLP (target head), which predicts the final *Interest* label on a 1–7 scale.

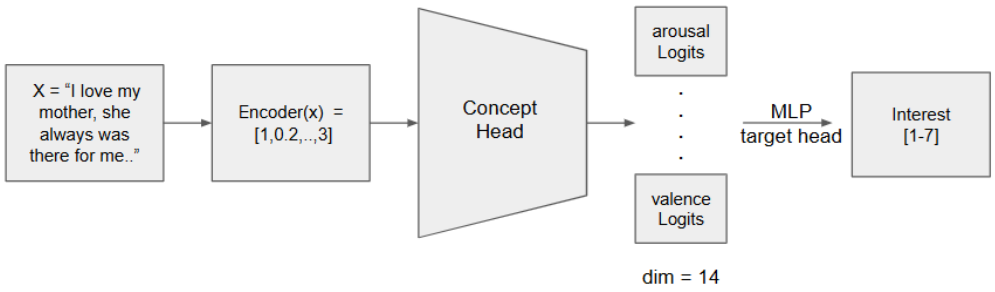


Figure 6: Pipeline of the advanced balanced + bottleneck model. Input text is embedded, mapped to concept logits (arousal, valence), and finally to the target label (*Interest*).

Table 2: Embedding comparison on the *clean* test set (F1 only; higher is better). The last row reports the same CBM evaluated on a test set that *includes* augmented sentences (see note).

	Arousal		Valence		Interest	
Embedding / Setting	$F1_{macro}$	$F1_{w}$	$F1_{macro}$	$F1_{w}$	$F1_{macro}$	$F1_{w}$
`bert_cls_embedding`	**0.515**	0.454	0.489	0.418	0.305	0.304
`bert_mean_embedding`	0.484	0.459	0.476	0.411	0.290	0.302
`sen_embedding_qwen0.6b`	**0.515**	**0.461**	**0.493**	**0.431**	**0.317**	**0.325**
`sen_embedding_qwen0.6b` (Aug. Test)	0.810	0.746	0.785	0.741	0.675	0.682

Note. The "Aug. Test" row evaluates the same CBM on a test set that includes high-LIC synonym augmentations. Because sentence embeddings map many such substitutions close to their originals, this induces *semantic leakage* (train–test near-duplicates) and inflates F1. Primary results are the first three rows on the clean test set.

Training Setup.

- **Stage 1 – Concept learning.** Using the balanced dataset, we trained the concept head g_θ with CrossEntropy loss for valence and arousal.
- **Stage 2 – Label prediction.** With g_θ frozen, the target head f_ϕ was trained to predict interest using CE loss (Multiclass classification to desrecte targets).
- **Balancing effect.** Because valence/arousal are themselves highly imbalanced, training on the augmented dataset improved the calibration and diversity of the concept logits, strengthening the downstream label predictor.

5.4 Quantitative comparison across embeddings

Table 2 summarizes results for three embedding backbones on the clean test set. The sentence–embedding model (`sen_embedding_qwen0.6b`) consistently achieves the strongest F1 scores across all tasks. Arousal is the easiest target, Valence is intermediate, and Interest is the hardest.

Effect of including augmented sentences in test (semantic leakage). For completeness, we also evaluate the best backbone on a test set that *includes* augmented sentences (high-LIC synonym substitutions). The last row in Table 2 shows substantially higher F1 scores. This uplift is *not* evidence of better generalization; rather, synonym substitution followed by sentence embedding makes many test samples near-duplicates of training samples in embedding space, creating **semantic leakage** between train and test. We therefore treat these numbers only as a robustness reference and report primary results on the clean test set.

5.5 Best backbone: sentence embeddings (`qwen0.6b`), Error patterns:

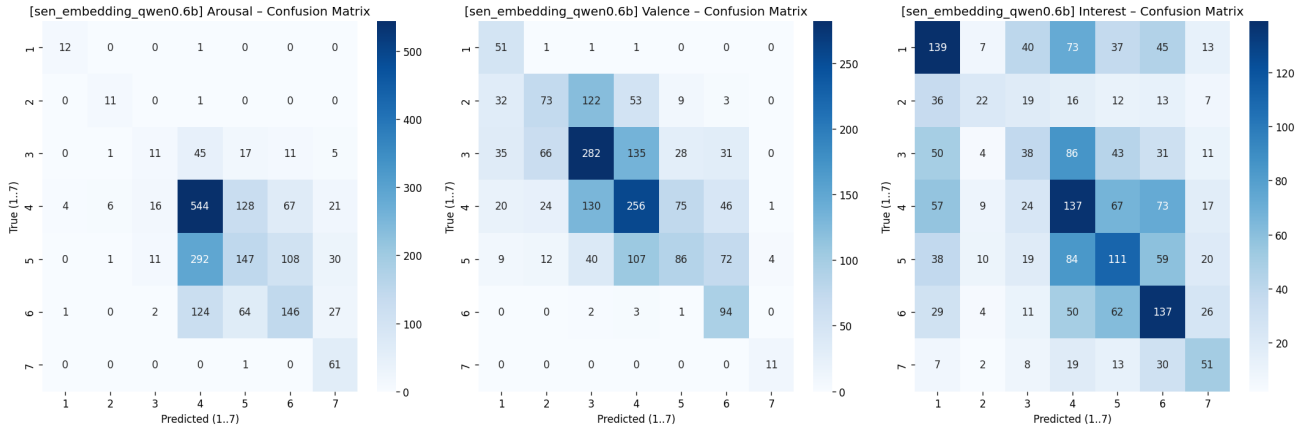


Figure 7: Confusion matrices (Arousal, Valence, Interest) for `sen_embedding_qwen0.6b` on the clean test set.

Figure 7 summarizes the error structure of the best backbone (`sen_embedding_qwen0.6b`) on the clean test set. Arousal and Valence show a strong diagonal and most mistakes occur between *adjacent* classes, reflecting the ordinal nature of the labels. Interest is clearly harder: errors spread across the middle band (classes 3–5), whereas the extremes (1 and 7) are better separated.

Despite balancing and the improvements over earlier baselines (see tab:embed-f1), the mid-range remains ambiguous. This is *expected*: human raters find it easy to detect very low or very high arousal/valence, but substantially harder to consistently distinguish, e.g., 3 vs. 4 or 4 vs. 5. As a result, the supervision signal that reaches the model is inherently noisy around the center of the scale, and the model's confusions concentrate exactly there. In other words, a non-trivial portion of the observed errors is label-ambiguity rather than purely model deficiency.

6 Bottleneck (CBM) Analysis

Aim. We analyze how the *Concept Bottleneck* mediates predictions from sentence embeddings to the target. Our goals are: (i) quantify concept quality and probability calibration; (ii) characterize the target's error structure; (iii) test how much the target depends on concepts vs. hidden features; and (iv) decide whether the current bottleneck (**Valence**+**Arousal**) is sufficiently expressive for predicting **Interest**.

6.1 Where do mistakes concentrate? (TCT - typical confusion trend)

The TCT in Fig. 7 shows 7-way confusion matrices for Arousal, Valence, and Interest using the best backbone (`sen_embedding_qwen0.6b`). Arousal/Valence exhibit strong diagonals with confusions concentrated in *adjacent* classes, which is exactly the pattern expected from ordinal structure. Interest is noticeably harder: errors spread across the middle bands (classes 4–6), indicating that **two-dimensional affect alone may not fully capture the variation relevant for perceived Interest**.

- **Concepts-only collapses.** The concepts-only CM concentrates on a few columns (e.g., class 7), showing that predicted A/V *alone* are insufficient to drive accurate Interest decisions.
- **Oracle concepts sharpen the diagonal.** Replacing A/V with their ground truth substantially improves Interest's confusion structure. This demonstrates that *better concept estimates* would directly lift the target—i.e., the bottleneck is *functionally relevant*.
- **Hidden features still matter.** The hidden-only variant retains non-trivial accuracy, indicating that information predictive of Interest exists outside the 2-D affect bottleneck.

TCT imply that, under our data and task, the **Valence+Arousal bottleneck is not sufficiently expressive** to explain or predict Interest on its own.

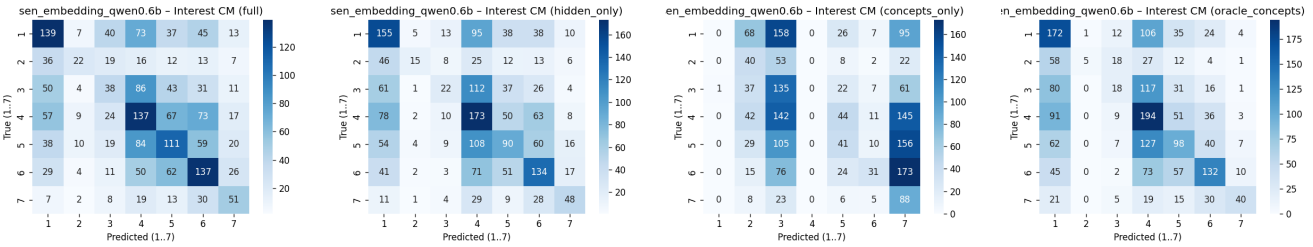


Figure 8: Interest confusion under full, hidden-only, concepts-only, and oracle-concepts settings.

6.2 Implication: widen the bottleneck

Given the evidence, we recommend replacing the 2-D affect bottleneck with a slightly richer—but still interpretable—set of concepts: **Basic Emotions (six-way) as logits.** Use discrete emotion primitives as the bottleneck (e.g., joy, sadness, anger, fear, disgust, surprise). This might provide a more expressive semantic basis while preserving interpretability.

7 Limitations and Future Work

Data limitations. A key limitation of our study is the small number of subjects: only twelve participants contributed data. While this setup suffices for within-subject modeling, it poses clear challenges for generalization. Human emotional expression is highly individual and stylistic, and with such a narrow pool of speakers $\{n = 12\}$ it is difficult to claim that the learned models generalize across humankind's styles of speaking. Future work should expand the dataset to include a larger and more diverse sample of individuals, which would allow the models to learn more universal patterns of emotional expression.

Modeling limitations. On the modeling side, our experiments revealed limitations in the choice of objective functions. When training with CrossEntropy loss, all misclassifications are penalized equally, regardless of how far apart the classes are on the ordinal scale. Conversely, Mean Squared Error (MSE) treats the task as purely continuous regression, which ignores the categorical structure of the labels. An appealing compromise is to employ **ordinal regression neural networks**, which explicitly account for the ordered nature of the labels. Such models can "walk between the drops" by penalizing large errors more than small, adjacent ones, thereby aligning better with the semantics of emotional ratings. Future research should explore ordinal regression architectures and loss functions as a way to improve predictive performance and interpretability simultaneously.

8 Conclusions

We studied emotion prediction from text under two practical difficulties: substantial missingness and severe class imbalance. Our pipeline combined (i) a correlation-preserving imputation that maintained the empirical emotion–emotion structure, (ii) LIC-based contrastive augmentation to bolster minority classes, and (iii) a Concept Bottleneck Model (CBM) that routes sentence embeddings through human-interpretable concepts (Valence, Arousal). On a *clean* test set without augmented sentences, the sentence-embeddings backbone (`sen_embedding_qwen0.6b`) achieved the strongest F1 scores across targets (Table 2). Error analysis (Fig. 7) showed that most confusions occur between adjacent classes and concentrate in the mid-range (3–5), which is consistent with annotator ambiguity and explains why accuracy alone is misleading for this task. Intervention analyses on the bottleneck (Fig. 8) revealed that predicted Valence+Arousal alone are not sufficiently expressive for recovering *Interest*; however, replacing them with oracle concepts markedly improves the target, demonstrating that better concepts would directly lift performance. We also showed that including augmented sentences in the test set creates semantic leakage and inflates scores, hence we report primary results on the clean split. Overall, CBMs provide useful structure and interpretability, but a two-dimensional affect bottleneck is too narrow for *Interest*. We therefore recommend widening the bottleneck to the six basic emotions, moreover, adopting ordinal-aware objectives and soft-labeling to reduce mid-range ambiguity, and expanding subject diversity to improve external validity.

References

[1] James M. Johnson and Taghi M. Khoshgoftaar. A survey of methods for addressing class imbalance in deep learning. *Journal of Big Data*, 6(27), 2019.

[2] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Been Kim, Ying Xiong, Yair Carmon, and Percy Liang. Concept bottleneck models. In *Proceedings of the 37th International Conference on Machine Learning*, pages 5338–5348. PMLR, 2020.

[3] Scott M. Lundberg, Gabriel Erion, Hsiang Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronen Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2(1):56–67, 2020.

[4] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems (NeurIPS 2017)*, volume 30, 2017.

[5] Tianyu Zhang, Jiashun Huang, Chang Feng, Xiaoqiang Jiang, Fei Wu, Wei Gao, and Zongyang Ma. Solving data imbalance in text classification with constructing contrastive samples. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1500–1511. Association for Computational Linguistics, 2022.