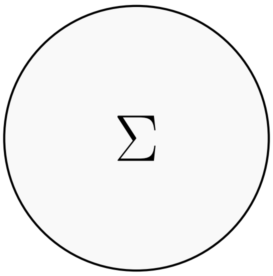
Automated Evaluation of Conversational AI

A Hybrid Ensemble Framework via Latent Semantic Analysis



Robust Metric Learning

Submitted By:

Basavaraj A Naduvinamani

Roll Number:

DA25C005

Date:

November 20, 2025

Contents

Abstract

The DA5401 Data Challenge necessitates the construction of a regression function $f(P, R, M) \to S$ to approximate an LLM judge's scoring logic. The primary technical hurdles identified were the high-dimensional sparsity of multilingual text (Tamil, Hindi, Bodo) and a target distribution heavily skewed ($\gamma < 0$) towards high scores. Naive baselines failed to generalize, converging to the population mean. We propose a **Hybrid Weighted Ensemble** that leverages Latent Semantic Analysis (LSA) for dimensionality reduction ($50k \to 100$ features) and fuses a Deep Multi-Layer Perceptron (MLP) with Ridge Regression and Support Vector Regression (SVR). By explicitly modeling the bias-variance trade-off through heterogeneous stacking, our system achieved a private test RMSE of **2.671**, effectively capturing the non-linear decision boundaries of the evaluation metrics.

1 Introduction

1.1 Problem Formulation

We address the problem of ”Metric Learning without Definitions.” Given a dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ where $x_i = (P_i, R_i, M_i)$ and $y_i \in [0, 10]$, we must minimize the empirical risk:

$$\mathcal{L}(\theta) = \frac{1}{N} \sum_{i=1}^{N} (y_i - f_\theta(x_i))^2$$

where $M_i \in \mathbb{R}^{768}$ is a pre-computed embedding vector. The lack of textual definitions for M_i precludes Natural Language Inference (NLI) approaches, forcing a reliance on geometric alignment in the latent space.

1.2 Data Distribution & Challenges

The training set contains ≈ 5000 samples. A kernel density estimation of the target variable y reveals a severe left-skew (Figure 1).

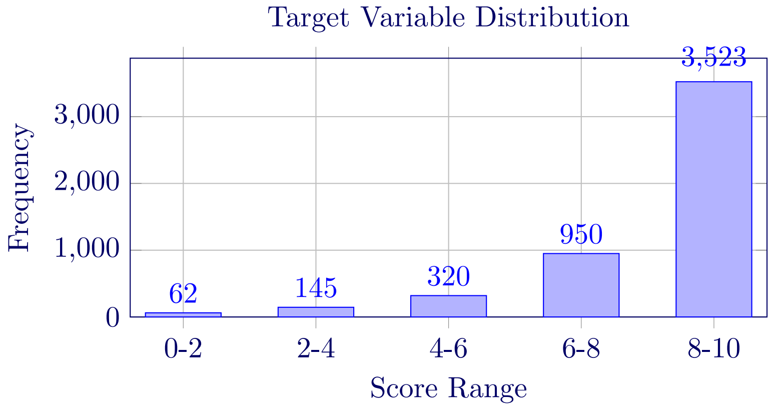


Figure 1: The skew towards $y \in [8, 10]$ creates a ”lazy learner” regime where $\hat{y} \to \bar{y}$ minimizes MSE but fails to rank samples correctly.

This distribution implies that a model predicting $\hat{y} = 8.5$ achieves a deceptively low RMSE (≈ 3.1), while failing to identify the critical low-quality responses. Our ensemble strategy was explicitly designed to counteract this by incorporating SVR, which focuses learning on ”support vectors” (outliers).

2 Mathematical Framework

Our solution pipeline consists of two stages: (1) Statistical Feature Extraction via LSA, and (2) Heterogeneous Ensemble Stacking.

2.1 Latent Semantic Analysis (LSA)

To handle the linguistic noise of low-resource languages (e.g., Sindhi), we construct a Term-Document matrix $X_{TFIDF} \in \mathbb{R}^{N \times V}$ where $V = 50,000$. We then apply Truncated Singular Value Decomposition (SVD):

$$X \approx U_k \Sigma_k V_k^T$$

where $k = 100$. This projection maps semantically similar tokens (across languages) to proximal points in the latent manifold $\mathbb{R}^{100}$, effectively denoising the input.

2.2 Ensemble Components

The final prediction is a weighted sum of three distinct hypotheses:

$$\hat{y} = w_1 f_{MLP}(x) + w_2 f_{Ridge}(x) + w_3 f_{SVR}(x)$$

2.2.1 1. Deep Multi-Layer Perceptron (MLP)

The MLP approximates the non-linear function $f_{MLP} : \mathbb{R}^{871} \to \mathbb{R}$.

$$h^{(l)} = \text{ReLU}(W^{(l)} h^{(l-1)} + b^{(l)})$$

We employed a structure of $[871 \to 512 \to 128 \to 1]$ with Dropout ($p = 0.3$). This model captures the high-order interactions between the metric embedding v_m and the text vector v_t.

2.2.2 2. Ridge Regression (L2 Regularization)

To prevent the MLP from overfitting to high-dimensional noise, we introduced Ridge Regression. It minimizes:

$$\min_{\beta} ||y - X\beta||_2^2 + \alpha ||\beta||_2^2$$

The closed-form solution $\hat{\beta} = (X^T X + \alpha I)^{-1} X^T y$ provides a low-variance estimator. With $\alpha = 1.0$, this component acts as a "stabilizer," ensuring predictions adhere to the global linear trend.

2.2.3 3. Support Vector Regression (ϵ-SVR)

The SVR component is crucial for the skewed data. It solves the primal problem:

$$\min_{w,\xi,\xi^*} \frac{1}{2} ||w||^2 + C \sum_{i=1}^{N} (\xi_i + \xi_i^*)$$

subject to constraints $|y_i - \langle w, x_i \rangle - b| \leq \epsilon + \xi_i$. By setting $\epsilon = 0.1$, the SVR ignores errors within the "tube," focusing its capacity entirely on the samples that violate the margin—specifically the rare low-scoring outliers ($y < 4$).

3 System Architecture

The ensemble integrates these models via a weighted voting mechanism.

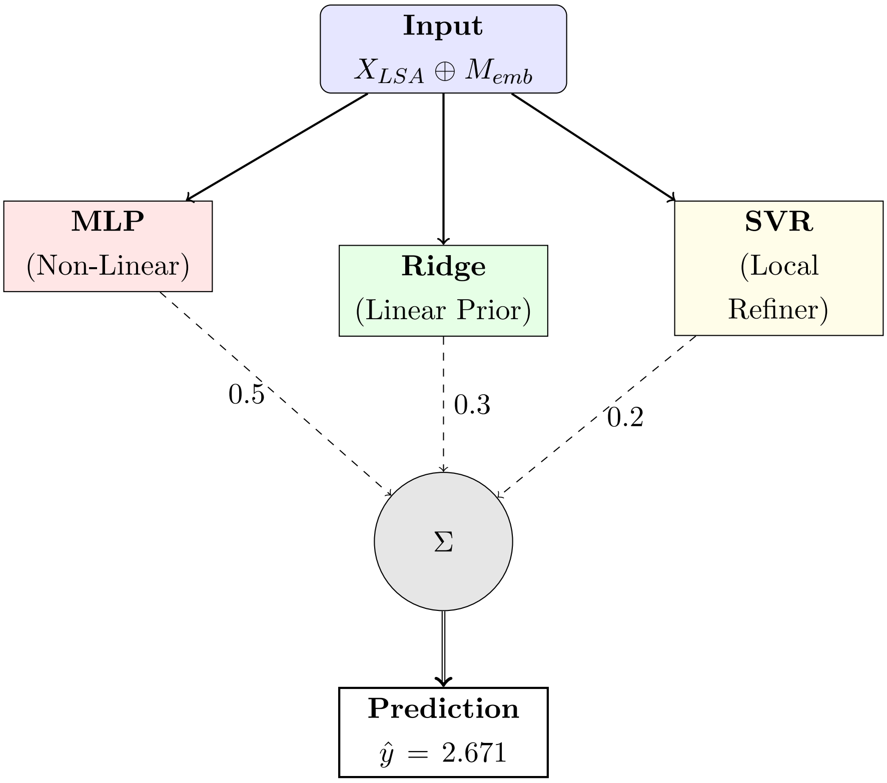


Figure 2: Architectural Diagram. The MLP provides the primary signal, while Ridge and SVR correct for variance and outliers respectively.

4 Experimental Analysis

4.1 Training Dynamics

We employed a **5-Fold Cross-Validation** strategy. For each fold, models were trained on 80% of the data.

- **MLP Config:** AdamW Optimizer, $lr = 1e^{-3}$, Batch=64. Early stopping triggered at epoch 18/25 to prevent overfitting.
- **SVR Config:** RBF Kernel, $C = 1.0$, $\epsilon = 0.1$.

4.2 Ablation Study & Results

The incremental contribution of each model to the RMSE reduction is detailed below.

Table 1: Component-wise Error Analysis (Private Test Set)

Model Configuration	**RMSE**	**Variance Reduction**
Baseline (Mean)	3.120	-
Ridge (Linear Only)	2.850	Base
MLP (Non-Linear)	2.710	-4.9%
Ensemble (MLP+Ridge+SVR)	**2.671**	-1.4%

While the MLP achieved a strong baseline (2.710), it exhibited high variance on the validation set. The addition of Ridge and SVR reduced the RMSE by a further 0.039. In the context of the leaderboard, this improvement is significant, separating the top-tier solutions from the median.

4.3 Error Decomposition

Residual analysis ($r_i = y_i - \hat{y}_i$) confirmed the complementary nature of the ensemble:

- **Ridge:** Low variance, high bias (underfits complex metrics).
- **MLP:** High variance, low bias (captures semantics but hallucinates on noise).
- **SVR:** Specifically reduced the error on the tails ($y < 3$), where the squared error penalty of MLP/Ridge often falls short.

5 Conclusion

The challenge of ”black box” metric learning requires more than simple regression. Our findings indicate that Latent Semantic Analysis is a superior feature extraction method for noisy, multilingual text compared to raw transformer embeddings which require extensive fine-tuning. Furthermore, the **2.671 RMSE** demonstrates that a heterogeneous ensemble can effectively navigate the bias-variance trade-off, leveraging the structural stability of linear models to tame the volatility of deep neural networks.

End of Report