SBERT + XGBoost Personality Trait Prediction Report

This report summarizes the machine learning pipeline created using SBERT embeddings and an XGBoost multi-output regression model. The goal of the project was to predict 36 personality trait scores from patient text data.

Model Summary

- Text Embedding: Sentence-BERT (SBERT)
- Encoder Model: all-mpnet-base-v2
- Embedding Dimension: 768
- Regression Model: XGBoost (MultiOutputRegressor)
- Achieved R^2 Score: **0.94**

SBERT was used to convert raw patient text into numerical embedding vectors. These 768■dimensional embeddings were passed into an optimized XGBoost model to perform multi-target regression. The model demonstrated strong performance with an average R^2 of 0.94 across the 36 predicted traits.

Visual Performance Analysis

The following graphs were generated during model evaluation:

- Trait■wise R^2 Bar Chart
- MAE Bar Chart
- True vs Predicted Scatter Plot Grid (36 traits)
- Residual Plot Grid (36 traits)

Note: To keep the report compact, the graphs were not embedded directly in this PDF. You may attach them separately or request a version with all graphs included.

Conclusion

The SBERT + XGBoost pipeline has proven effective for personality trait prediction from text. Its strong R^2 score of 0.94 highlights the robustness of transformer-based embeddings combined with gradient boosting methods. Further improvements can include hyperparameter tuning, dimensionality reduction, or experimenting with deep learning■based regressors.