Proceedings of the 3rd International Conference on Intelligent Cyber Physical Systems and Internet of Things (ICoICI-2025)
IEEE Xplore Part Number: CFP25VG0-ART; ISBN: 978-1-6654-5753-8

Efficient Personality Prediction using Multi-Output Regression with ALBERT and RoBERTa Models on Social Media Data

Nakka Sai Spoorthi
Dept of CSE (AI & ML)
Institute of Aeronautical Engineering, Dundigal, Hyderabad, India
22951A66A8@iare.ac.in

Dr. M Nagaraju
Dept of CSE (AI & ML)
Institute of Aeronautical Engineering, Dundigal, Hyderabad, India
m.nagaraju@iare.ac.in
0000-0001-8898-1090

Shaik Farahana
Dept of CSE (AI & ML)
Institute of Aeronautical Engineering, Dundigal, Hyderabad, India
22951A66C2@iare.ac.in

Shaik Imeen Roshni
Dept of CSE (AI & ML)
Institute of Aeronautical Engineering, Dundigal, Hyderabad, India
22951A66C3@iare.ac.in

2025 3rd International Conference on Intelligent Cyber Physical Systems and Internet of Things (ICoICI) | 978-1-6654-5753-8/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICoICI65217.2025.11252561

***Abstract*—Using social media content to predict personality traits has become relevant to the development of user modeling and personalization. A transformer-based Natural Language Processing (NLP) model developed to predict the Big Five personality traits will be presented, utilizing posts from the social media platform Reddit. The Personality and Demographics of Reddit Authors (PANDORA) dataset is leveraged to train and evaluate ALBERT and RoBERTa models in a multi-output regression design. To facilitate performance improvement and efficiency analysis, we employ several optimization strategies, including gradient clipping, weight decay, early stopping, hyperparameter configuration, and mixed-precision training. To further enable use in realistic contexts, personality scores are converted into narrative, descriptive summaries. In terms of the relative performance of RoBERTa and ALBERT, results show that ALBERT outperforms RoBERTa, based on all key evaluation metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R^2 score. Results also suggest that smaller transformer models perform satisfactorily in the context of personality traits. Directions for future extension of this work include better support for multilingual inputs and improved interpretability of outputs.**

***Keywords*—ALBERT, Interpretability, Mixed-Precision Training, Multi-Output Regression, PANDORA Dataset, Personality Prediction, RoBERTa, Transformer Models**

I. INTRODUCTION

The growth of user-generated content on sites like Reddit has opened up new pathways for deriving individual characteristics in a text-embedded format, with the use of the term "personality prediction" becoming increasingly more commonplace due to its application for use in adaptive user interfaces, modeling digital behavior, and intelligent recommendation systems. This research investigates the use of transformer-based language models to develop a multi-output regression framework to predict individual personality through the Big Five personality traits: Agreeableness, Openness, Conscientiousness, Extraversion, and Neuroticism in posts through Reddit. Two models, ALBERT and RoBERTa, are being used and evaluated through the PANDORA dataset to explore the models' ability to indistinctively generate accurate and compelling scores for the prediction of personality traits.

The motivation for this work originates from the limitations of traditional personality assessment methods that typically have a heavy reliance on structured questionnaires and interviews that are slow and not scalable. Social media content illustrates user expression naturally and in real-time via text. Content shared on social media employs informal styles with inconsistent grammar and varied vocabulary compared to the structured questionnaire context associated with limitations inherent to surveys. It represents a major effort to understand and predict personality. Therefore, advanced NLP models will be useful to recognize earlier patterns in real-world language from unique social contexts. The purpose of this research is to develop an accurate, easy-to-maintain, lightweight, and interpretable system predicting Big Five personality traits from social media text. The research compares ALBERT and RoBERTa, using a multi-output regression approach. The goal includes optimizing systems for performance, limiting training costs, and increasing applicability by converting numeric predictions from each step in the regression. After evaluating comparisons, the output is converted into a meaningful summary of personality that can be used in practice.

The contribution of this work is a comparative assessment of ALBERT and RoBERTa for predicting personality traits using Reddit text in a multi-output regression framework. The framework takes advantage of regularization approaches, mixed precision on the training stage, and early stopping on the training stages to enhance efficiency while gaining performance. To facilitate practical use, the interpretability layer converts raw model outputs into a readable personality description for the user. The main findings show the benefit of smaller models like ALBERT, where we see competitive accuracy while utilizing fewer computing resources. To verify the proposed approach, experiments were conducted on the PANDORA dataset and evaluated using standard regression metrics, including MAE, MSE, RMSE, and R^2 score. The results demonstrate the potential of transformer models regarding efficiency and interpretability for personality predictions.

II. LITERATURE SURVEY

Multiple literature items have addressed Automated Personality Prediction (APP) through Natural Language processing (NLP) methods applied to social media data. Specifically, reports presented findings of studies leveraging NLP and social media data. Scholars have presented the advantages of exploring machine learning and deep learning approaches to modeling linguistic cues defining personality. The literature outlines several modeling strategies (machine learning and deep learning), each with specific qualities, advantages, and drawbacks concerning not just predicting respondent personality traits from text, but also accurately predicting personality traits.

In [1], a navigating pathway to automate personality prediction is provided of ALBERT and RoBERTa models for multi-output regression in personality prediction. The authors used Reddit text data to generate continuous scores for the Big Five personality dimensions and showed that transformer models are effective in modeling subtle linguistic features. The authors' significant finding from their analysis is that ALBERT, a much smaller model than RoBERTa, can provide competitive performance and reduce computational costs. In a survey of methods [2] for predicting personality, it covered the transition from conventional psychometric testing and questionnaires to text analysis based on natural language processing (NLP) techniques on public social media postings. The authors noted particular difficulties regarding problems, including sparseness of data, the use of informal language, and the requirement for domain adaptation of the predictive models. Overall, this review showed that personality prediction is increasingly important for researchers in areas such as behavioral analytics and personalized marketing. The examination of personality is studied [3] in pre-trained language models and investigated how the models represent and encode personality traits in the text. They examined these models' capability of detecting subtle signals of personality, and assessed the function for fine-tuning a model to output personality-oriented behaviors. They also contribute to understanding future work regarding the adaptability of a model for detecting personality. In [4], predicting personality from multiple social media sources recommended the use of data across multiple platforms to increase robustness to the assessed personality. They applied pre-trained language models and models averaged to reach better performance, while emphasizing the potential advantages of different inputs of social media to generate a more rounded assessment of personality. In [5], PANDORA dataset and analysis, explored the intersection of personal characteristics and demographic variables within Reddit data. They showed that when using demographic variables, differences in language use on social media can differ from demographic group to demographic group, resulting in personalized expression and detection of personality traits. This study suggested an important resource and conceptual basis to build upon for any future research tackling the demographic context for predicting personality.

The study [6] used an integrated linguistic cue with behavioral indicators, such as posting behavior, interaction frequency, and network structures, to increase the accuracy of personality prediction models. The study [7] adopted RoBERTa, a variant of BERT enhanced with dynamic masking, larger training batches, and by removing next-sentence prediction, thereby improving the representation of context. The authors [8] explored personality expression in large language models using specific prompts, revealing consistent trait patterns indicative of each model size. The study [9] leveraged deep learning methods with convolutional and recurrent networks on social network text, achieving better results than those using traditional methods. A domain-specific transformer pre-trained to analyze sentiment in finance is created [10]. It illustrates how pre-trained models are enhanced by specialization in pre-training for specific applications. An hybrid statistical and deep learning method is provided for language identification which is a vital consideration in personality modeling for multilingual customers [11]. A machine learning approach is presented in [12] to MBTI prediction using MBTI data from an equivalent number of participants selected from a distinguished case study with corresponding value. It provides some methodological principles that could also be considered with approaches to prediction using the Big Five. The study [13] developed a multi-document transformer that accumulates multiple textual documents from the same user and then uses all the documents to produce a predictive model. This provides user-based stability and accuracy in predictions. Authors of [14] compared data across multiple platforms and combined models with model averaging to enhance robustness in measuring personalities. A bidirectional, long short-term memory (LSTM) network on Twitter data is utilized to leverage greater semantic context and recorded better predictions on Big Five traits than unidirectional models [15].

III. METHODOLOGY

A multi-stage process has been implemented for the development of an efficient and interpretable personality prediction system based on transformer models that have been fine-tuned on PANDORA social media data. The process incorporates several key phases, which include data preprocessing, feature extraction, model selection, and architecture, model training with optimization techniques, and interpretability improvement after the model is trained. Figure 1 illustrates the conceptual architecture of the envisioned system for predicting personality from text. It starts with scraping Reddit posts that were then cleaned and pre-processed.

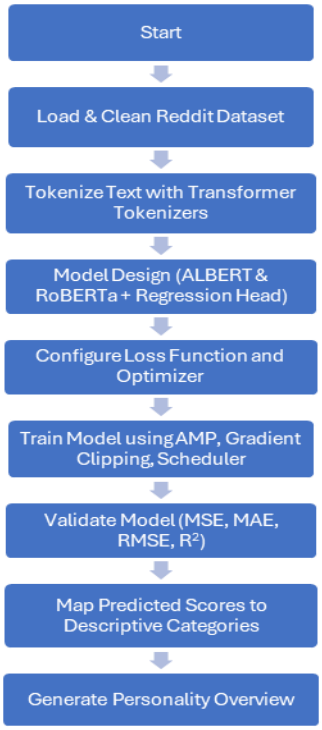

Figure 1. Conceptual Framework for Personality Prediction using transformer-based models

The dataset can be accessed here: https://github.com/Fatima0923/NLP (as cited by Habib et al., 2024 [1]).

After formatting, the cleaned data were tokenized by model-specific tokenizers compatible with ALBERT and RoBERTa. The encoded inputs were then passed into a transformer encoded file that was fine-tuned on the specific task for regression using a regression head that outputs five continuous values as Big Five personality traits. Finally, there is an interpretation layer that turns the raw trait scores into descriptions and gives the personality in a digestible way. The framework is designed to provide accurate predictions while allowing usability and efficiency for practical applications.

A. Data Collection and Preprocessing

The dataset used in the present work is the Personality and Demographics of Reddit Authors - PANDORA dataset, which features Reddit posts that are coded with continuous scores for Big Five personality traits: Agreeableness, Openness, Conscientiousness, Extraversion, and Neuroticism. It is a large and rich collection of text-based behavioral data, which is why the dataset was selected, as it is large, rich, and most directly relates to predicting personality. Since the dataset contains scores of traits rather than categorical labels, it is particularly well-suited for regression-based modeling. Although the original authors did not specifically discuss potential redundancy or inconsistencies within the dataset, the study completed additional checks to establish the reliability and appropriateness of the data for use with transformer architectures. A sample of the dataset is noted in Figure 2, illustrating how the inputs are delivered, as well as the personality traits associated with each input.

text	agreeableness	openness	conscientiousness	extraversion	neuroticism
his name was kim kimble originally wow thats some messed up p	9	61	13	4	72
theyre better than the normal posts on ryugioh id rather have th	50	85	50	85	50
it probably does ive learned a lot about myself by browsing this s	71	53	17	3	31
yea those are the same sound to me still	64	44	33	8	88
long term shifting is the cart titans gimmick though the fact that	50	85	50	85	50
texas is molly weasley i love it	79	84	86	53	1
yeah those are good points my experiences with recruiting is all v	85	95	15	50	15
public shame you put their reputations on the line and their view	15	85	15	85	15
thats how they get you man	0	38	43	77	12
any plan to do the same for the republican primaries im curious a	39	89	1	22	61
yeah the comicpost is a good one at least something i can relate	9	61	13	4	72
exploiting isnt cheating lel tell that to ppl who used external unde	46	27	31	10	84
professional quality almost all highlevel headphones are open pa	64	44	33	8	88

Figure 2. Sample entries from the PANDORA dataset.

- **Dataset Preparation**

To check on data quality, the post data from Reddit was cleaned by first converting to lowercase, followed by removing noise (i.e., URLs, special characters, excessive whitespace) via regular expressions. After cleaning, all duplicates were removed to maintain non-duplicate entries across the dataset. Instances where text was missing as part of the text data, or for situations where the personality trait scores were left incomplete, were removed as well. The systematically cleaned dataset solidified the quality of the dataset, in addition to improving the consistency of the model training downstream.

- **Data Cleaning and Transformation**

The cleaned text data was then passed through the tokenizers of ALBERT and RoBERTa. The tokenizers provided input IDs and an attention mask, both of which are requirements for transformer-based models. Padding and truncation were administered to ensure that input lengths remained the same across all datasets while ensuring that the research preserved the relevant contextual information needed for reliable predictions. This transformation step translated the raw text into structured numerical forms that could then be used and understood by the transformer models.

- **Dataset Preparation**

To properly evaluate the model, the data was allocated into two types of subsets: training and validation. To create efficient data pipelines, two of PyTorch's convenience classes, Dataset and DataLoader, were used to allow for efficient GPU training and seamless batching. Table 1 shows the important statistics after preprocessing. The average number of words, sentences, as well as the size of the data set allow the reader to have context for what the data looks like as it was used for model training and evaluation.

Table 1: Dataset Statistics Summary

Property	Count
Number of comments	16,407
Average word count	45.7
Maximum words	342
Minimum words	1
Average sentences	2.6
Maximum sentences	22
Minimum sentences	1

B. Feature Extraction

The most critical element of this analysis was the application of transformer-based tokenizers to evoke, retrieve, and derive rich semantic information from the Reddit datasets. Those tokenizers, ALBERT and RoBERTa in this case, converted the raw texts (described by the project) into structured tokenized sequences; successfully preserving the word-level meaning, while encapsulating both local and global contextual relationships. This required preprocessing was fundamental in allowing the models to learn subtle interactions and relationship patterns at the group level of textual data. The analysts were conceptualizing a new abstraction layer of the grouped data into language layers based on the richness of embeddings by the tokenizers. Using the pre-trained tokenizers in this analysis builds a strong basis for accurate personality predictions from social media posts. Plus, on a practical note, it explicitly allowed the models to learn on latent relations and connection dependencies.

C. Model Architecture

Transformers serve as the base for the proposed personality prediction framework, as these kinds of models are very good at modeling unstructured, contextualized text, which is also a main characteristic of social media data. Self-attention enables these models to model short and long-range dependencies within the text, making it possible to model the relationships among words and phrases comprehensively. This is beneficial when examining social media data, because language is often informal, creative, and users differ considerably in their linguistic choices. The inherent complexity of unique user styles allows transformers to pick up on smaller cues, which map to the concept of larger or more distinct personality traits. To predict continuous scores for the Big Five dimensions, the framework made use of the tuned versions of ALBERT and RoBERTa to transfer knowledge from the non-personality domain to personality. The text datasets were the posts made by Reddit users, and the three features per user for each

personality trait in the Big Five were measured as semantic feature sets that match each personality trait. Additional training of the full architecture used techniques such as gradient clipping, weight decay, and mixed-precision computations for accuracy and computational efficiency. The finished architecture not only showed promising predictive performance but is also capable of supporting downstream interpretability, which illustrates the system's potential in practical contexts.

- **Pre-trained Transformer Models**

In developing this framework, we have selected ALBERT and RoBERTa as the two primary pre-trained transformer models. Both models are well-known for their capability to deal with the intricacies of social media language and also facilitate rich semantic representations. The multi-headed self-attention allows these models to learn local dependencies and more global dependencies in text, providing an excellent means to very quickly extract the linguistic cues informing personality predictions.

- **Custom Regression Head**

To ensure the continuous prediction of the regression of the Big Five personality traits, a custom regression head was attached to both the ALBERT and RoBERTa models. The regression head encapsulates a dropout layer and a fully connected linear layer that allows the models to output five continuous values, the Big Five personality traits: Agreeableness, Openness, Conscientiousness, Extraversion, and Neuroticism. The dropout layer adds generalization properties to the models by preventing overfitting, while the linear layer maps the values produced by the final hidden states to the Big Five personality trait dimensions. The structure of the regression head incorporates the regression goal of the task while still allowing the transformer encoders to learn and transfer personality-related information from their contextualized input text representations. The regression operation can be expressed mathematically as follows:

$$\hat{Y} = W \cdot h + b \qquad (1)$$

where:

- $\hat{Y} \in R^5$, which means the model's output is a vector with 5 real numbers; each number corresponds to one of the Big Five personality traits.
- $h \in R^d$, is the hidden state output of the transformer model, either ALBERT or RoBERTa. It is a vector with d dimensions (depending on the model).
- $W \in R^{5 \cdot d}$, is the weight matrix that goes from the hidden state 'h' to trait scores. It has 5 rows for the 5 traits and d columns corresponding to the hidden size.
- $b \in R^5$, is the bias vector which was added after the linear transformation. It also has 5 values, which are traits as mentioned earlier.

Take the hidden state (h), apply a linear transformation using the weight matrix (W), add a tiny bias (b), and outcome the predicted trait scores (one for each of the five). The overall model process (input processing and score generation) is shown in Figure 3.

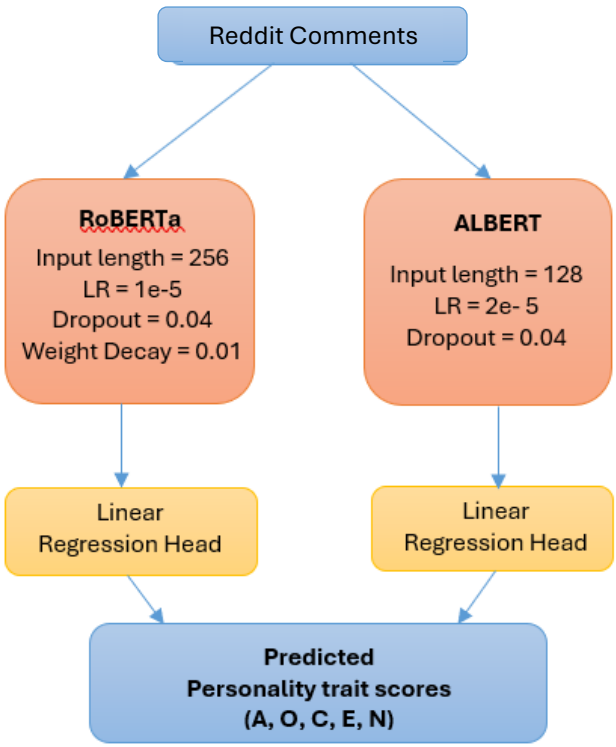

Figure 3. Architecture of ALBERT and RoBERTa-Based Regression Models

- **Hyperparameter Configuration**

To maximize performance and minimize instabilities in training, different hyperparameter configurations were selected for both ALBERT and RoBERTa due to architectural differences in how data was processed through the two models. These parameters were fine-tuned in attempts to optimize a balance between speed of convergence and generalization. ALBERT was set to a maximum input length of 128 tokens, with a learning rate of 2e-5, and dropout of 0.04, while RoBERTa was set to a maximum input length of 256 tokens, with a learning rate of 1e-5, and dropout of 0.04; and extra weight decay of 0.01 to enhance regularization. These values were found through iterative testing under the differing depth and representation capacity of the models. Figure 3 provides a visual overview of the architectures and the key hyperparameter configurations that informed the regression setup for both models.

D. Model Training and Optimization

A core quality of model optimization is being able to ensure the system generalizes past the training data well and can then appropriately deal with inputs it has not previously seen. This is done by choosing a proper loss function to evaluate the difference between predicted and actual scoring of personality traits. The optimizer is also critical in terms of how fast it converges and how long, in total, it takes to train the model. As well, it is often advantageous to bring in more sophisticated training methods such as mixed-precision training, gradient clipping, and weight decay to improve stability and performance. All of these functions work together to create a robust learning procedure with low risk of overfitting or unstable results among varying data distributions.

- **Loss Function and Optimizer**

To assess how accurately the model's predicted trait scores matched the actual values, we used the Mean Squared Error (MSE) loss function. MSE as a loss function is suited for this study as it emphasizes larger errors and assists the model's focus on correcting large errors. The AdamW optimizer was selected for the model due to AdamW's adaptive learning rate of 1e-5 and 2e-5 and a unique weight decay term of 0.01; this was a balance between accuracy and the model's ability to generalize. The learning rate and weight decay hyperparameters were tuned cautiously for convergence and the prevention of overfitting.

- **Stability and Training Efficiency Measures**

Several advanced techniques were used to ensure stable and efficient training. First, gradient clipping was applied to limit the size of the updates applied during backpropagation to limit the potential for instability from large updates. Mixed precision training was applied to both speed up computation and reduce memory usage, utilizing the PyTorch Automatic Mixed Precision (AMP) module. Second, one of the linear learning rate schedules with a warm-up phase was chosen for the learning rate to help encourage smooth convergence towards minima. Finally, it utilized early stopping based on the training time since validation loss was not improving, to help reduce the potential for overfitting.

- **Learning Rate Scheduling and Early Stopping**

A linear learning rate scheduler as well as an initial warm-up were implemented to provide smooth, stable convergence during training. The linear reduction in the learning rate supported the ability to make gradual adjustments in learning toward an optimal solution. Early stopping was also implemented to increase generalization and avoid overfitting. Early stopping checked the validation loss and stopped training if the validation loss did not further decrease. These adjustments were useful to save computing power and also to trust the model.

E. Evaluation Metrics

In evaluating the predictive accuracy of our models, we modeled four conventional regression metrics: MAE, MSE, RMSE, and R^2. These metrics quantitatively measured the deviation between the actual and predicted values to express the precision of model predictions over all five traits. Let:

- y_i be the actual trait scores for the i^{th} sample from the dataset.
- $\hat{y}_i$ be the predicted trait score given by the model.
- $\bar{y}$ be the mean of all actual scores in the dataset.
- n be the total number of samples, i.e., the Reddit entries used for evaluation.

Mean Squared Error (MSE): Averages the squared differences between actual trait scores and predicted scores. It penalizes larger errors more than smaller errors.

$$\mathbf{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \qquad (2)$$

Mean Absolute Error (MAE): It measures the average magnitude of errors in a set of predictions, without considering their direction.

$$\mathbf{MAE} = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i| \qquad (3)$$

Root Mean Squared Error (RMSE): Represents the square root of the MSE model error; however, it is easily interpretable back into the original unit of target values and punishes larger errors relative to the smaller.

$$\mathbf{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \qquad (4)$$

R^2 Score (Coefficient of Determination): Shows how well the model can account for or explain the variance of actual scores. A value of 1 indicates perfect prediction and a value of 0 indicates the prediction is about as good as just taking the mean. If R^2 is negative, then the predictions were worse than taking a simple average of the mean.

$$\mathbf{R^2} = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y}_i)^2} \qquad (5)$$

We calculated each of the metrics presented for each of the five personality traits. All of the metrics provided a very complete sense of model performance across all the personality traits.

F. Interpretability and Personality Overview Generation

To assist in the interpretation and usability of continuous trait predictions, an interpretability component was added to the system. After the model has produced numerical scores for each of the Big Five personality traits: Agreeableness, Openness, Conscientiousness, Extraversion, and Neuroticism, the scores are passed through a logic-based post-processing logic. Instead of converting the scores into basic categories like low and high, the system uses hand-coded rules to map each score to an appropriate behavioral description. These descriptions have been carefully written to be representative of actual psychological tendencies witnessed at specific scores, while also tracing the accuracy of the model's outputs. When the created descriptions for the five traits are synthesized into a cohesive paragraph of the full personality summary, a one-line personality statement is also written as a quick and intuitive psychological snapshot. This process will not only guarantee that the regression model's outputs are predictive and valid, but also interpretable and usable in real-world contexts.

G. Implementation Details

All model building, including training and evaluation, was conducted using PyTorch in combination with the Hugging Face Transformers library. This endowed us with a great deal of flexibility in distinguishing some of the state-of-the-art transformer-based architectures, as well as in utilizing a considerable amount of data. Training occurred using Kaggle's GPU-based platform and allowed us to take advantage of that platform to meet the resource-heavy demand of training transformer models. During training, validation data were used to assess model performance, and the model with the lowest validation loss was maintained to ensure accuracy and replicability. The best weights from the training phase were saved in the .pt file type, which allows those same weights to be loaded in the future for other applications and integration into existing systems. In summary, this approach ensured reliable results and allowed for further improvements and real-world implementation.

IV. RESULTS AND DISCUSSIONS

A. Model Performance

The models ALBERT and RoBERTa both showed effective capacities in predicting the Big Five personality traits within a multi-output regression format. The ALBERT model reported MAE of 17.48, RMSE of 22.56, MSE of 508.60, and R2 of 0.52. A comparison indicated the RoBERTa model reported MAE of 22.36, RMSE of 26.38, MSE of 697.13, and R2 of 0.16. Based on these evaluation pieces of information, ALBERT maintained a consistent edge on lower errors and the best predictive potential. Rather than showing gains in specific areas, one of the key advantages of the ALBERT model is its consistent level of performance across the five personality characteristics, while, on the other hand, the RoBERTa model had more variability in its predictions across specific features, albeit still delivering strong overall results. ALBERT's persistence and variability for personality prediction problems were illustrated by the consistent and comparable predictive accuracies obtained for all characteristics. These consistent and comparable improvements are predictive of the potential of ALBERT for applied settings that would require accurate and reliable personality assessments from text.

B. Per-Trait Performance

A comprehensive evaluation was undertaken to evaluate the proposed ALBERT and RoBERTa models' predictive capability for each of the Big Five personality traits: Agreeableness, Openness, Conscientiousness, Extraversion, and Neuroticism. The models' predictive accuracy is reported in Table 2 for four main evaluation metrics: Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the coefficient of determination (R^2). These metrics summarize the models' predictive accuracy and can be compared directly across each trait.

As the results report indicates, the ALBERT model was consistently lower for all errors (MSE, RMSE, MAE) and consistently higher R^2 values observed for all traits compared to RoBERTa. Not only can we see that ALBERT will be more accurate in predicting personality, but it also better accounts for and explains the variance in trait scores. The overall equal performance across all traits, as discussed above, also speaks to ALBERT's reliability and versatility in obtaining consistent, accurate personality assessments from text. RoBERTa produced good results overall, but exhibited somewhat more variability across the traits. This variability again suggests that although RoBERTa performed well in a few instances, it does not generalize as well across all five dimensions of personality as ALBERT. This type of feedback is important for real-world situations, especially in cases where reliable, nuanced assessments of human personality are required and where stability and consistency are needed. A detailed comparison of the five different traits can be seen in Table 2 and shows the general advantages of ALBERT for reliable personality predictions, as well as demonstrating RoBERTa's inability to reliably produce the same results across traits.

Table 2: Comparison of Evaluation Metrics for Models

Trait	Model	MSE	RMSE	MAE	R^2
Agreeableness	ALBERT	339.49	18.43	16.76	0.59
	RoBERTa	714.53	26.73	22.16	0.16
Openness	ALBERT	599.48	24.45	16.76	0.85
	RoBERTa	786.06	28.04	25.30	0.01
Conscientious	ALBERT	557.62	23.62	15.35	0.65
	RoBERTa	619.87	24.90	19.07	0.20
Extraversion	ALBERT	529.77	23.69	18.79	0.11
	RoBERTa	720.84	26.85	23.45	0.24
Neuroticism	ALBERT	518.66	22.60	20.73	0.39
	RoBERTa	644.36	25.37	21.83	0.20

Insights: The data in Table 2 indicate that across all traits, ALBERT has consistently lower MSE, RMSE, and MAE, while producing higher values for R2 than RoBERTa. Therefore, ALBERT would be more suitable for any tasks that require consistent and reliable personality trait predictions, since we found that RoBERTa was less consistent across the various traits tested in this study. What we can say from these results is that ALBERT exhibited greater robustness and balanced predictive power in capturing the various nuanced patterns of personality-related text data.

C. Trait-wise Performance Comparison

To illustrate the model's performance more clearly across the Five-factor model, we created bar charts summarizing model predictions by comparing the mean absolute error (MAE), root mean square error (RMSE), and R^2 scores for both proposed ALBERT and RoBERTa models. This trait-wise comparison of the ALBERT and RoBERTa models is shown through three figures.

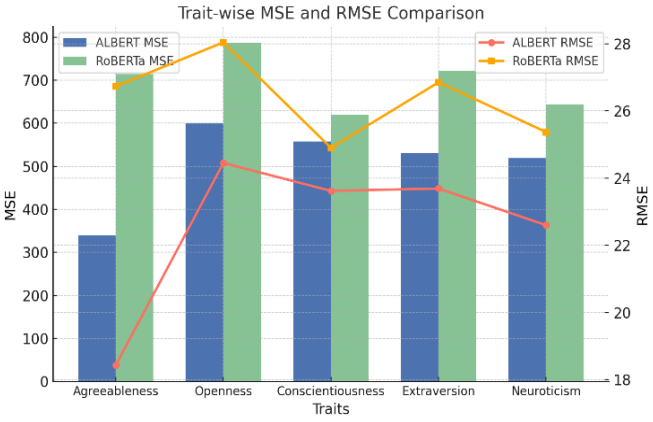

Figure 4. Trait-wise MSE and RMSE Comparison for ALBERT and RoBERTa Models

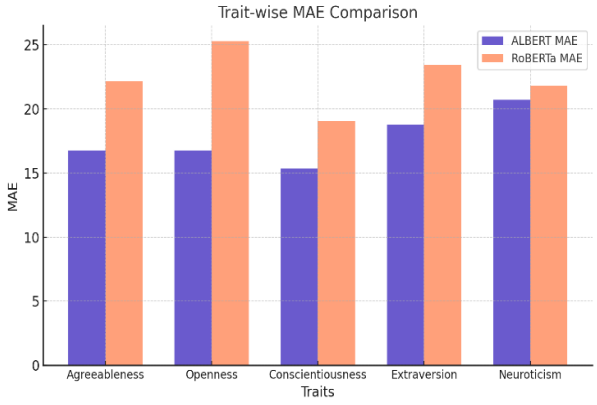

Figure 5. Trait-wise MAE Comparison for ALBERT and RoBERTa Models

Figure 4 is a dual-axis plot of the Mean Squared Error (MSE) and RMSE values. MSE measures the average of the squared differences between the predicted and actual trait scores and provides an indication of the overall prediction error value. RMSE, which is simply the square root of the MSE, simply affords a more interpretable error magnitude when returned to the original trait score measurement scale. It is easy to see from the image that the ALBERT model performs better overall, in comparison with the RoBERTa model, by making better predictions (lower MSE and RMSE) across all traits, especially Agreeableness and Conscientiousness. The results show that ALBERT can make more stable and more accurate predictions than RoBERTa. Figure 5 includes the Mean Absolute Error (MAE) of each trait given as a bar plot and adheres to the line plot for better visual readability. MAE is defined as the average absolute error between the predicted score and the actual score and provides an easy measure of how accurate the predictions are. Overall, ALBERT has a lower MAE than RoBERTa for each of the five traits. This shows us that ALBERT does not accumulate as much error and achieves strong performance at the trait level.

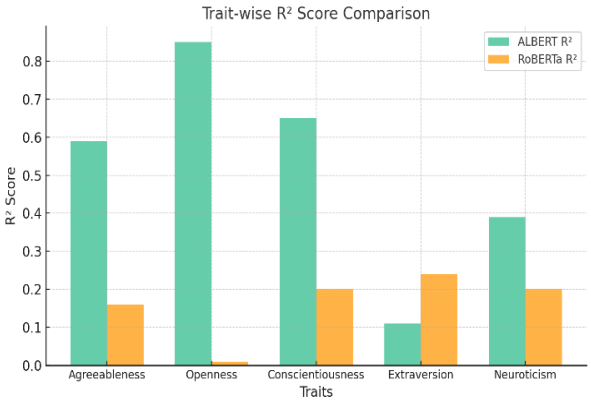

Figure 6. Trait-wise R² Score Comparison for ALBERT and RoBERTa Models

Figure 6 shows the R²scores for each trait, which represent how well the predictions align with the actual values. The closer the R² values approach 1, the more accurate the predictions. In this analysis, ALBERT consistently yields higher R² scores across most traits, indicating it explained more of the variance in the trait Scores compared to the other models; thus, ALBERT performed better by also detecting the subtle characteristics of the personality-related language data in this model

These visual analyses also show that not only did ALBERT reduce prediction errors (as seen in the lower MSE, RMSE, and MAE) but also maintained better and more reliable trait-level predictions (as highlighted in the higher R² values) compared to RoBERTa. This revealed ALBERT's overall performance and stability, leading to being a better predictor of personality traits using social media text data.

D. Interpretability and Personality Overview

In addition to producing ongoing trait predictions, the model's interpretability module was important for transforming raw numerical outputs into descriptive summaries of a person's behavioural tendencies. This assisted with making the connection from model predictions to understanding. After scoring each of the Big Five traits, the interpretability layer provided trait-wise behavioural explanations based on score strength. These natural-language descriptions showed likely psychological patterns from the predicted values.

```
Input Text: I prefer spending quiet evenings at home, reading a good book or watching a movie.
I enjoy deep conversations but need time alone to recharge.

Personality Trait Evaluation:

Openness (30): practical and conventional
Conscientiousness (49): moderately goal-focused
Extraversion (32): quiet and introspective
Agreeableness (36): analytical and assertive
Neuroticism (32): emotionally stable

Personality Overview:
This person is practical and conventional, moderately goal-focused, quiet and introspective, a
nalytical and assertive, and emotionally stable.

Overall, this person shows a balance between introversion and extroversion, indicating an ambi
vert personality.
```

Figure 7. Example of Model-Generated Personality Trait Evaluation and Overview

An example of the complete output is provided in Figure 7, which shows the cleaned input text, predicted scores, descriptions or interpretations, and a summary of personality. The summary paragraph represents a coherent picture of the user's total personality, while a brief one-line behaviour sense of a person is provided as an immediate, useful information snapshot of a user's behaviour tendencies. By providing this framework for interpretability, the user then has even more trust in the outputs of the model, and it is suitable for launching in the real world, including, but not limited to, social media analysis or digital profiling and recommendations systems.

E. Discussion of Results

The findings from this study provide insight regarding the predictive power and usefulness of transformer-based models for personality trait detection. The ALBERT model tended to outperform the other transformer-based models, particularly given that it had low error metrics and higher R² scores across all five traits. This indicates the reasonable utility of using smaller, more efficient models for personality prediction tasks. While ALBERT was a more compact model than RoBERTa, not only did it achieve high prediction scores, but it also evidenced greater generalization across all of the different aspects of personality than were evidenced with the other transformer-based models.

From this perspective, model design and parameter efficiency are very useful in developing predictions, especially when there are many text data types that have the potential for noise or language biases. Furthermore, the consistency of the patterns across each case illustrates the robustness of the suggested ALBERT-based framework. The practical relevance of the model is further increased when interpretability components are included, such as taking numerical predictions and interpreting them in terms of personality description. For potential applications such as social media screening, talent recruitment screening, and monitoring of mental health, the interpretability assures that the outcomes exhibit robustness and reliability and are easy to assimilate. In conclusion, the observed patterns of the study provide substantial evidence to support the utility of including interpretability into model outputs, as well as validate the proposed ALBERT-based model as a more robust alternative to the baseline and RoBERTa models. These patterns serve to position the ALBERT-based framework as a potential

instrument for further automated tools of personality assessment in the future.

V. CONCLUSION

Transformer-based models have demonstrated a great deal of promise in the use of natural language processing techniques to predict personality traits from text on social media. Using a multi-output regression technique on the PANDORA dataset, ALBERT and RoBERTa were refined to estimate continuous scores for the Big Five personality characteristics. All assessment criteria showed that ALBERT consistently produced better results than the others, proving that smaller models may be computationally efficient and still provide good accuracy. The addition of an interpretability layer converts unprocessed numerical predictions into descriptive insights, categorized trait levels, and a succinct summary of personality. The practical usefulness of the model is increased in domains like user profiling, behavioural analytics, and digital well-being evaluation, adding to the human-centered layer. Trait-level analysis revealed significant differences in the way each model represented distinct aspects of personality, demonstrating the importance of detailed assessment as opposed to depending only on broad measurements. Notably, ALBERT not only performed consistently across all features but also outperformed the larger RoBERTa model in each category, supporting the notion that more accurate and dependable predictions can be made using smaller, optimized models in real-world scenarios. Although the framework has many advantages, it has limitations as well. Primarily, it inherits all English-language data, and because it is based on text-only input, it may not be expandable across different platforms, cultures, and data to the same degree. The model also does not have real-time flexibility, and it has only been tested on the Reddit platform, which limits its generalizability. In future studies, researchers could attempt to overcome some of these challenges by working with multilingual cohorts, using different modalities such as images, metadata, or interaction types, and using more complex interpretability techniques such as SHAP or LIME to increase trust and transparency in the process. These techniques would improve the adaptability of this framework as it could be used in other contexts where individuals are interacting with information in the real world, such as adaptive learning systems, digital engagement monitoring, and personalized learning. Further, the predicted personality trait scores have not yet been verified with formal psychological assessments, and this could be investigated in future work to further increase accuracy and usefulness in real-world applications.

REFERENCES

[1] F. Habib, Z. Ali, A. Azam, K. Kamran, and F. M. Pasha, “Navigating pathways to automated personality prediction: A comparative study of small and medium language models,” *Frontiers in Big Data*, vol. 7, Sep. 2024, Art. no. 1387325, doi: 10.3389/fdata.. 2024.1387325.

[2] Hetal Vora, Mamta Bhamare, and K. Ashok Kumar, “Personality Prediction from Social Media Text: An Overview,” *International Journal of Engineering Research and Technology*, vol. 9, no. 5, pp. 352–357, May 2020. Available online at: http://www.ijert.org

[3] Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu, “Evaluating and Inducing Personality in Pre-trained Language Models,” in *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS)*, 2023. Available online at: https://sites.google.com/view/machinepersonality

[4] Hans Christian, Derwin Suhartono, Andry Chowanda, and Kamal Z. Zamli, "Text-based personality prediction from multiple social media data sources using pre-trained language model and model averaging," *Journal of Big Data*, vol. 8, no. 1, p. 68, 2021, doi: 10.1186/s40537-021-00459-1.

[5] M. Gjurković, M. Karan, I. Vukojević, M. Bošnjak, and J. Šnajder, "PANDORA Talks: Personality and Demographics on Reddit," in *Proc. of the 9th Int. Workshop on Natural Language Processing for Social Media*, 2021, pp. 138–152.

[6] Kanchana, T. S., and B. Smitha Evelin Zoraida. "A framework for automated personality prediction from social media tweets." In 2022 IEEE World Conference on Applied Intelligence and Computing (AIC), pp. 698-701. IEEE, 2022.

[7] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., et al. *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. arXiv preprint arXiv:1907.11692, 2019.

[8] Hilliard, A., Munoz, C., Wu, Z., and Koshiyama, A. S. *Eliciting Personality Traits in Large Language Models*. In Proceedings of the ACM Conference on Fairness, Accountability, and Transparency, 2024, pp. 1–28.

[9] Xue, Di, Lifa Wu, Hong Zheng, Shize Guo, Liang Gao, Zhiyong Wu, Xiaofeng Zhong, and Jianshan Sun. "Deep learning-based personality recognition from text posts of online social networks." Applied intelligence 48, no. 11 (2018): 4232-4246.

[10] Araci, D.T.(2019).FinBERT: Financial sentiment analysis with pre-trained language models. arXiv [Preprint]. arXiv:1908.10063. doi: 10.48550/arXiv.1908.10063.

[11] Rajanak, R. S., Bhuvana, D., and Lakshmi, S. S., “Language Detection Using Natural Language Processing,” in Proceedings of the 2023 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 2023, pp. 1–5. doi: 10.1109/ICCCI56949.2023.10092353. Available online at: https://ieeexplore.ieee.org/document/10092353

[12] Brandon Cui and Calvin Qi, “Survey Analysis of Machine Learning Methods for Natural Language Processing for MBTI Personality Type Prediction,” 2023.

[13] Feifan Yang, Mingyang Zhou, Zhendong Mao, and Lei Zheng, “Multi-Document Transformer for Personality Detection,” in Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2021, pp. 4474–4484. doi: 10.18653/v1/2021.emnlp-main.364.

[14] Hans Christian, Derwin Suhartono, Andry Chowanda, and Kamal Z. Zamli, “Text-based Personality Prediction from Multiple Social Media Data Sources Using Pre-trained Language Model and Model Averaging,” Journal of Big Data, vol. 8, no. 1, p. 68, 2021. doi: 10.1186/s40537-021-00459-1.

[15] Chen, Yucheng. "Exploring the Intercorrelations of Big Five Personality Traits: Comparing Questionnaire-Based Methods and Automated Personality Assessment using BERT and RNN Models." Master's thesis, 2023.