IJWR

Transformer-Based Personality Trait Recognition Enhanced by Contextual Augmentation

Hossein Saberi, Reza Ravanmehr*
Department of Computer Engineering, Central Tehran Branch, Islamic Azad University, Tehran, Iran; hossein.saberi1998@gmail.com, r.ravanmehr@iau.ac.ir

ABSTRACT

Although personality detection from written text has clear applications, such as tailored interfaces and psychological research, it often suffers from label interference, vocabulary-driven overfitting, and limited labeled datasets. As a result, models are brittle: they can fail with small training samples and behave inconsistently across trait ranges. To address this, we employ a practical single-trait approach that uses five independent ELECTRA-based classifiers, each corresponding to one of the big five dimensions, and trained them as separate binary tasks to prevent cross-trait interference. To reduce lexical bias and double the Pennebaker and King essay corpus from 2,467 to 4,934 samples, the team applied careful synonym-replacement augmentation using WordNet and additionally incorporated contextual augmentation generated by the Gemma model. Models were adjusted methodically to ensure fair comparisons. With test AUCs above 0.75, the ensemble achieves an average test accuracy of 0.724 on the Pennebaker and King benchmark, with per-trait accuracies of 0.72, 0.71, 0.74, 0.73, and 0.72 for openness, conscientiousness, extraversion, agreeableness, and neuroticism (OCEAN), respectively. These results substantially reduce inter-trait interference while matching or surpassing LIWC baselines and other transformer approaches.

Keywords— ***Personality Recognition, Natural Language Processing, Transformer Models, ELECTRA, Big Five Personality Traits, Computational Psychology.***

1. Introduction

Personality recognition is a rapidly growing field in artificial intelligence (AI) and computational psychology, aiming to infer an individual's personality traits from various data sources, such as text [1], speech [2], or behavior [3]. Among the most widely used frameworks for personality assessment is the big five personality model, also known as the five-factor model (FFM) [4]. This model delineates personality across five broad dimensions: openness, reflecting creativity and intellectual curiosity; conscientiousness, denoting organization and dependability; extroversion, characterized by sociability and assertiveness; agreeableness encompassing compassion and cooperativeness; neuroticism, associated with emotional instability and anxiety. Together, these five dimensions form the OCEAN traits. The big five model is highly regarded for its robustness, empirical validation, and applicability across cultures, making it a cornerstone in personality recognition research [5]. However, alternative models, such as the Myers-Briggs Type Indicator (MBTI), which classifies individuals into 16 personality types based on dichotomies like introversion/extroversion and thinking/feeling [6], and the HEXACO model, which adds a sixth dimension (honesty-humility) to the big five, offer complementary perspectives. While the big five provides a continuous and quantitative assessment, MBTI and HEXACO cater to different research needs, such as categorical classification or enhanced trait granularity. Understanding these models and their interplay is crucial for advancing personality recognition systems and tailoring them to specific applications, such as personalized recommendations, mental health diagnostics, and human-computer interaction (HCI) [7].

Automating personality recognition under the big five framework requires models capable of decoding subtle linguistic patterns that correlate with trait dimensions [2]. For instance, openness may surface

http://dx.doi.org/10.22133/ijwr.2025.543527.1305
Citation H. Saberi, R. Ravanmehr, "Transformer-Based Personality Trait Recognition Enhanced by Contextual Augmentation", *International Journal of Web Research*, vol.9, no.1,pp.1-24, 2026, doi: http://dx.doi.org/10.22133/ijwr.2025.543527.1305.

*Coressponding Author
Article History: Received: 27 August 2025 ; Revised: 9 December 2025; Accepted: 20 December 2025.

through creative metaphor or lexical diversity, while neuroticism might be inferred from repetitive syntactic structures or negative emotional valence. While lexicon-based methods (e.g., LIWC) and statistical models have laid foundational groundwork, their reliance on handcrafted features limits their ability to generalize across diverse linguistic styles and contexts. Recent advances in transformer architectures, such as BERT, have demonstrated significant improvements in capturing contextual relationships [8], yet challenges persist in optimizing these models for specialized tasks like personality recognition. Fine-tuning large transformers on smaller, domain-specific datasets risks overfitting, and their computational complexity raises barriers to deployment in resource-constrained settings. These factors highlight the need for approaches that balance performance, efficiency, and adaptability to diverse linguistic inputs.

The following contributions are made in this paper:

- A trait-specific ELECTRA fine-tuning framework in which five independent ELECTRA-base classifiers (one per big five trait) are trained to generate more discriminative, trait-specific representations and reduce inter-trait interference.
- A strengthened data augmentation pipeline that combines synonym replacement and contextual generation, supported by semantic-similarity and large language model-based (LLM-based) contextual filtering to ensure linguistic quality and mitigate data scarcity.
- A comprehensive empirical evaluation on the Pennebaker and King essay corpus, including accuracy, precision, recall, F1-score, and ROC-AUC, with comparisons to LIWC-based and transformer baselines; all training protocols, hyperparameters, and preprocessing steps are fully documented, with limitations and ethical considerations discussed.

The remainder of this paper is organized as follows: Section 2 reviews related work in personality recognition within computational psychology and natural language processing (NLP). Section 3 outlines the methodology, including the dataset, augmentation strategy, model design, and training setup. Section 4 reports the experimental results for the five single-trait ELECTRA classifiers. Section 5 provides an overall discussion of the findings, addresses the study's limitations, and outlines directions for future research. Section 6 concludes the paper.

2. Related Work

Contemporary personality recognition from text has been transformed by transformer-based language models [9]. Pretrained transformers such as BERT, RoBERTa, ELECTRA, DeBERTa, GPT, and their variants have been applied to big five or MBTI classification tasks with strong results [10, 11]. These models leverage self-attention to capture long-range context and subtle linguistic cues (e.g., sarcasm, implicit meaning), which are critical for inferring personality [12]. For example, Shum et al. fine-tuned both BERT and RoBERTa on a large Reddit corpus, the PANDORA dataset, which contains 17 million comments collected from more than 10,000 users to predict continuous big five scores [13]. They found RoBERTa, especially the large version, consistently outperformed BERT across all traits, and that jointly modeling all five traits to capture their intercorrelations, further improved accuracy [13]. Similarly, Tsani and Suhartono applied an ensemble of BERT and RoBERTa to Twitter and YouTube posts for big five classification, reporting average F1-scores of 0.73 [14]. Naz et al. trained DistilBERT on an MBTI corpus to predict the openness dimension, achieving 0.92 accuracy, exceeding the 90% of traditional TF-IDF or word2vec-based methods [15]. In dialogue domains, Guo et al. used a fine-tuned BERT model on user–machine conversations and found that open-domain (chitchat) dialogues generally yielded higher personality-prediction accuracy than goal-oriented dialogues [16]. Across tasks, surveys note that transformer models have transformed personality trait detection by integrating contextual embeddings and can learn complex language patterns, though they require large labeled datasets and heavy computation [17].

Much of the recent work adapts standard fine-tuning setups: attaching a classification or regression head to the transformer encoder and updating on trait-labeled text [9]. For instance, Shum et al. trained RoBERTa to regress continuous big five scores [13], whereas Tsani used majority-vote ensembles of BERT/RoBERTa classifiers for each trait [14]. Jain et al. survey a range of transformer approaches, noting that models like BERT or GPT capture deep semantics but vary in performance and may be complemented by feature-based techniques [11]. In practice, configurations (batch size, learning rate, etc.) vary, but fine-tuning often uses similar recipes to other NLP tasks [18]. Table comparisons in recent studies show transformers often beat earlier methods: e.g., on an MBTI dataset, DistilBERT achieved 0.92 accuracy versus 0.86–0.88 for deep RNN/CNN models [11, 15, 19]. Ensembles of transformers, combining multiple models or output heads, have also been explored to boost robustness, as in Tsani and Suhartono's voting scheme [14].

2.1. Traditional Approaches

Before transformers, most personality-prediction systems relied on conventional deep networks or feature-based machine learning (ML). Recurrent neural networks (RNNs/LSTMs/GRUs) and convolutional nets have been widely used for textual personality tasks. These models capture sequential patterns in text: for example, Mohan et al. used word2vec embeddings fed into an LSTM for personality classification [20]. Such sequential models can effectively capture long-range dependencies, which are crucial for predicting personality-related text. In one review, sequential deep learning (DL) models like LSTM, BiLSTM, CNN, and GRU applied to MBTI text achieved accuracies in the mid-80s. CNNs have also been applied, treating text as local n-gram features, sometimes augmented with attention or hybrid LSTM layers. Early work, such as Bahri et al., analyzed tweets with CNNs; Mohan et al. integrated CNN and LSTM networks. Overall, deep networks typically outperformed bag-of-words baselines but were still limited by data size and feature design [21].

Classical ML approaches have also played a role, especially as baselines [22]. Studies often extract linguistic features such as LIWC categories, psycholinguistic or lexical cues, or use simple counts of words and phrases, then train SVMs, logistic regression, or random forests. For example, mental-health detection works note that previous work applied traditional methods such as SVM, Naive Bayes, and Random Forests. Similarly, personality-prediction studies historically used LIWC and n-gram features with SVMs or decision trees. Surveys report that ensemble and DL methods now tend to outperform shallow ML models, though classical models remain interpretable and useful in low-data regimes [23]. In summary, earlier DL architectures including CNNs, RNNs, LSTMs, hybrid models, and classical models like LIWC-based SVM or random forest laid the groundwork for automated personality inference. However, the emergence of transformers between 2018 and 2020 introduced a substantial improvement in both accuracy and flexibility [24].

2.2. Personality Taxonomies

Most work assumes a formal personality model such as the big five or MBTI. The big five or OCEAN framework is dominant due to its empirical validation. It represents personality as continuous scores on openness, conscientiousness, extraversion, agreeableness, and neuroticism [25]. Many text corpora and prediction tasks use big five annotations, including Pennebaker and King's essays dataset and various social media datasets labeled for the big five [26]. MBTI, a four-dichotomy categorical model, is also frequently used, particularly in social-media contexts like internet forums or Kaggle MBTI datasets. For example, Kerz et al. evaluated their models on both the big five essays dataset and the MBTI Kaggle dataset [27], and the Kaggle MBTI resource with its sixteen personality types has driven the development of many neural models. Researchers sometimes map MBTI categories onto big five proxies or treat each MBTI dichotomy (E/I, N/S, F/T, J/P) as an individual trait.

Less commonly, alternative models like HEXACO, which introduces honesty-humility and recognizes several traits, appear in the literature. A small number of studies in psychology or user modeling consider HEXACO traits, but within NLP, most benchmarks remain centered on the big five and MBTI. For completeness, we note that other taxonomies, including Enneagram, Cattell's 16PF, exist, although they are rarely used in recent computational research.

2.3. Transformer Model Architectures

Transformer-based systems for personality use the standard pretraining paradigms of NLP. Models like BERT and RoBERTa are pretrained with masked-language objectives on massive text of Wikipedia, BooksCorpus, and news [10]. ELECTRA employs replaced-token detection (RTD), while DeBERTa uses disentangled attention mechanisms; both approaches achieve improved language representations and have been applied to personality trait prediction. For instance, Elourajini and Aïmeur trained ELECTRA on MBTI data and reported an accuracy of approximately 0.75–0.78 [28]. Generative models like GPT or encoder–decoder LLMs could in principle be used via prompt-based classification or finetuning, though most published studies focus on encoder models [29]. During fine-tuning, practitioners typically append a simple feedforward output layer for classification or regression. For example, Shum et al. fine-tuned RoBERTa to regress continuous personality scores directly [13], while Tsani and Suhartono used pretrained BERT/RoBERTa as feature extractors with an added prediction head for each big five dimension [14]. Some studies treat all traits jointly or multi-output, others train separate models per trait. Batch sizes, learning rates, and regularization follow common NLP practice, though some analyses report that data augmentation like back-translation can modestly improve low-resource cases [14, 30].

Large models generally yield better results. Shum et al. report that RoBERTa-large outperforms RoBERTa-base and BERT-base when large data are available [13]. Naz et al. show that even the smaller DistilBERT can beat older methods on moderate-sized MBTI sets [15]. Guo et al. used standard BERT-base to capture 25 traits with high accuracy in dialogue text [16]. These findings echo wider NLP trends: transformer size, pretraining data, and fine-tuning strategy all significantly impact performance on personality tasks. Notably, some works have

combined transformers with additional network layers: for example, Kerz et al. used BERT plus a BiLSTM operating on sequences of psycholinguistic feature contours [27], demonstrating that transformer representations can be augmented with psychological features to boost accuracy [31].

2.4. Applications

Personality recognition has been proposed for many practical domains. A common motivation is mental health monitoring: by detecting personality patterns or shifts in text, systems could flag risk of depression or anxiety. Hasan et al. compared transformers like BERT, RoBERTa with LSTM models on Reddit posts about mental health, finding that a RoBERTa model achieved extremely high F1-score of approximately 0.995 on held-out Reddit data and outperformed all LSTM baselines [32]. This suggests transformer-based classifiers could underpin automated mental-health screening tools for online content [33].

Another area is personalized recommendations and marketing [34]. Personality profiles can improve recommender systems by matching users with suitable products or content [35]. For example, Shum et al. highlight that personality prediction could enhance recommendation systems for e-commerce or media. Similarly, Fernau et al. built a dialogue-based job recommender with MBTI profiles, finding higher satisfaction when the agent's personality matched the user's [36]. In hiring, personality analysis of social media or resume text can serve as a screening tool. Tsani and Suhartono note that their social-media personality model can be used as a parameter to screen candidate attitudes in the company recruitment process [14]. This points to potential application in human resources and candidate evaluation, although ethical concerns have been highlighted in the literature [37].

Personality recognition also enhances adaptive HCI. Dialogue systems or chatbots that infer user personality can adjust their style, including formality and humor, to improve engagement [38]. By predicting a user's personality, a dialogue system can adapt its behavior to personality of a user, leading to better task success and user satisfaction [16]. Liu et al. provide datasets for developing emotion-aware and personality-aware dialogue systems by integrating big five and emotional annotations [39]. Beyond chatbots, Bhin and Choi mention human–robot interaction (HRI) as a field that recognizes human personality via multimodal cues, making them more personable [3, 40]. Likewise, personalized education or gaming systems could use personality labels to adapt content or difficulty [41]. In healthcare, knowledge of patient personality from text or speech may guide therapy or improve communications [42].

Furthermore, personality computing has been explored in social science and psychology contexts. Researchers apply NLP-based personality prediction to analyze large-scale trends, such as the relationship between personality and political preference or demographic differences, and to validate psychological theories at scale [43]. The comprehensive surveys by Hashemi et al. and others provide context for these applications, reviewing how AI methods serve diverse use cases [44]. In summary, transformer-based personality models have been applied or proposed in domains ranging from mental health and marketing to personalized conversational agents, reflecting broad interest in automated personality inference. Table 1 summarizes prior personality recognition methods by category, outlining their main modeling approaches, data sources, key strengths, and principal limitations.

3. Methodology

3.1. Dataset

The essays dataset, often referred to as the Pennebaker and King dataset, is a valuable resource for personality recognition research. It consists of 2467 essays written by individuals who participated in a study of linguistic styles and personality traits. The essays are labeled with the OCEAN traits [45]. These traits are widely recognized in psychology as fundamental dimensions of human personality. The dataset provides a rich source of textual data that researchers can use to explore the relationship between language use and personality.

The essays were collected as part of a study conducted by J.W. Pennebaker and L.A. King. Participants were asked to write in a stream-of-consciousness style, allowing their thoughts to flow freely without much structure or editing [45]. This method was chosen to capture natural language use and provide insights into the participants' personalities. The essays cover a wide range of topics, reflecting the diverse experiences and perspectives of the participants. The dataset has been widely used in research to develop and evaluate models for personality trait recognition, providing valuable insights into how linguistic styles can reflect individual differences [9, 46].

Researchers have used various ML and DL models to analyze the essays dataset. Techniques such as SVM, BERT-based models, and attentive networks with contextual embeddings have been employed to classify the essays based on the big five personality traits. The dataset has been instrumental

Table 1. Summary of representative approaches for personality recognition from text

Category	Approach	Data Source	Key Strength	Key Limitation
Classical machine-learning	LIWC features, n-grams + SVM / Logistic Regression / Random Forest	Essays, social media posts, blogs	Interpretable features; performs reasonably well in low-data settings; computationally inexpensive	Limited ability to capture deep semantics; performance plateaus on complex language; requires manual feature engineering
Early deep-learning models	CNN, LSTM, BiLSTM, GRU, hybrid CNN–LSTM	MBTI datasets, Twitter, blog posts	Captures sequential patterns; handles long-range dependencies better than classical models; outperforms bag-of-words baselines	Still dependent on handcrafted embeddings (word2vec, GloVe); struggles with nuanced contextual cues; performance varies with dataset size
Transformer encoders (BERT, RoBERTa, ELECTRA, DeBERTa, DistilBERT)	Fine-tuning with classification or regression heads	Reddit corpora (e.g., PANDORA), Twitter/YouTube comments, MBTI datasets, dialogues	Strong contextual modeling through self-attention; SOTA accuracy on Big Five/MBTI tasks; effective across domains; can jointly model multiple traits	High computational cost and memory requirements; sensitivity to domain shift and annotation noise; limited interpretability of trait predictions
Ensemble transformer systems	Voting, multi-model ensembles of BERT/RoBERTa variants	Social media datasets, combined trait-labeled corpora	Increased robustness; typically improves F1/accuracy over single models; reduces variance	Higher computational cost; more complex deployment
Transformer + auxiliary features	BERT + BiLSTM for psycholinguistic feature contours; transformer + handcrafted linguistic features	Essays, MBTI datasets, hybrid corpora	Combines contextual embeddings with psychological cues; improved trait detection for subtle constructs	Requires additional feature engineering; added model complexity
Multimodal personality models	Text + audio + visual fusion networks; audiovisual transformer pipelines	Job-interview videos, HRI interactions, ChaLearn datasets	Leverages complementary non-textual cues; strong performance on traits sensitive to prosody or expression	Small datasets; domain-specific models; limited generalizability across text-only tasks

in advancing the field of personality recognition, enabling researchers to develop more accurate and robust models. By leveraging the rich textual data in the essays dataset, researchers can gain a deeper understanding of the complex relationship between language use and personality traits, contributing to the broader field of psychological and linguistic research.

3.2. Model Development

In this study, researchers selected the ELECTRA (Efficiently Learning an Encoder that Classifies Token Replacements Accurately) model for personality recognition from essays. ELECTRA represents a significant advancement in transformer-based architectures, offering a unique pre-training objective that sets it apart from other models like BERT, RoBERTa, and GPT. Unlike BERT and RoBERTa, which use masked language modeling (MLM) to predict randomly masked tokens, ELECTRA employs a RTD task [47]. In this approach, a generator network replaces some tokens in the input sequence, and a discriminator network is trained to distinguish between the original and replaced tokens. This allows ELECTRA to train on all tokens in the input sequence, rather than just the masked ones, making it significantly more computationally efficient. Additionally, ELECTRA achieves SOTA performance on many NLP benchmarks, often outperforming BERT and

RoBERTa while using fewer computational resources. These advantages make ELECTRA particularly well-suited for tasks like personality recognition, where capturing nuanced linguistic patterns is critical [48].

Several alternative transformer-based models were considered for this task, each with its own strengths and limitations. BERT, for instance, is a widely adopted model that uses bidirectional transformers to capture deep contextual representations [10]. However, its MLM objective is computationally expensive, as only 15% of tokens are used for training at each step. RoBERTa, an optimized version of BERT, improves performance by removing the next sentence prediction (NSP) objective and using dynamic masking, but it still relies on the less efficient MLM approach. GPT models, on the other hand, are based on unidirectional transformers and excel in generative tasks, but their lack of bidirectional context makes them less suitable for tasks requiring a comprehensive understanding of text, such as personality recognition. Other models like T5, DeBERTa, ALBERT, and DistilBERT were also considered [48]. T5's encoder-decoder architecture increases computational complexity, while DeBERTa's enhanced performance comes at the cost of higher computational requirements. ALBERT and DistilBERT offer efficiency gains but may not achieve the same level of performance as ELECTRA on tasks requiring high accuracy.

In addition to these models, recent advancements in NLP have introduced models like LLaMA, GPT-4, and PaLM. LLaMA is known for its impressive performance on various NLP tasks, but it requires substantial computational resources, making it less feasible for use on platforms like Google Colab. GPT-5, the latest iteration of the GPT series, offers remarkable generative capabilities but shares the same limitations as its predecessors in terms of bidirectional context. PaLM is another cutting-edge model that excels in understanding and generating human language, but its high computational demands make it challenging to implement in resource-constrained environments.

The decision to use ELECTRA was driven by its unique combination of efficiency, performance, and scalability. ELECTRA's RTD objective not only reduces training time but also enables the model to learn more effectively from the data, as it processes all tokens in the input sequence [47]. This is particularly important for personality recognition, where subtle linguistic cues spread across the entire essay can provide valuable insights into an individual's personality traits [20]. Furthermore, ELECTRA's bidirectional transformer architecture ensures that it captures rich contextual information, which is essential for understanding the complex relationship between language and personality. The model's scalability also makes it easier to fine-tune on domain-specific datasets, further enhancing its applicability to this task. Finally, ELECTRA's discriminator provides a degree of interpretability, as it highlights which tokens are most informative for the task, aiding in the analysis of linguistic patterns associated with specific personality traits [47]. For these reasons, ELECTRA was chosen as the core model for this study, combining strong predictive performance with computational efficiency that makes it well-suited for personality recognition from essay data.

3.3. ELECTRA

ELECTRA is a pre-training method for NLP models that addresses the inefficiencies of traditional MLM approaches, such as those used in BERT. Introduced by Clark et al., ELECTRA introduces a novel pre-training task called RTD, which significantly improves sample efficiency and downstream task performance [47]. Unlike MLM, which trains on only a small subset of masked tokens, ELECTRA trains on all tokens in the input sequence by distinguishing between real and replaced tokens. In this part, researchers provide a comprehensive theoretical foundation for ELECTRA, including its architecture, training objectives, and mathematical formulations.

The core innovation of ELECTRA lies in its RTD task. Instead of masking tokens and predicting their identities, ELECTRA employs a two-component system: a generator and a discriminator. The generator, typically a small neural network, is trained to predict masked tokens using a standard MLM objective. These predictions are then used to replace a subset of tokens in the input sequence, creating a corrupted version of the text. The discriminator, which is the main ELECTRA model, is trained to classify each token in the corrupted sequence as either real (original) or replaced (generated). This approach allows ELECTRA to leverage all tokens in the input sequence for training, rather than just the masked subset, leading to improved efficiency and effectiveness.

The ELECTRA model is defined by two primary components: the generator *G* and the discriminator *D*. Let $x=[x_1,x_2,\ldots,x_n]$ denotes an input sequence of tokens, where n is the sequence length. The generator *G* is trained to predict masked tokens using a MLM objective. Specifically, for a masked token x_i, the generator outputs a probability distribution over the vocabulary *V*, as defined in Equation (1):

$$P_G(x_i|x_{masked}) = softmax(W_G h_i^G + b_G) \quad (1)$$

where h_i^G is the hidden representation of the *i*-th token produced by the generator, and W_G and b_G are learnable parameters. The generator loss L_{MLM} is

computed as the negative log-likelihood of the correct token, as shown in Equation (2):

$$L_{MLM} = -\mathbb{E}[\sum_{i\in\mathcal{M}} logP_G(x_i|x_{masked})] \quad (2)$$

where M is the set of masked token positions.

The discriminator D is trained to classify each token x_i as real or replaced. Let $x_{replaced}$ denote the sequence with tokens replaced by the generator's predictions. The discriminator outputs a probability for each token, as formalized in Equation (3):

$$P_D\left(real\middle|x_i, x_{replaced}\right) = \sigma(W_D h_i^D + b_D) \quad (3)$$

where h_i^D is the hidden representation of the i-th token produced by the discriminator, σ is the sigmoid function, and W_D and b_D are learnable parameters. The discriminator loss L_{RTD} is computed as the binary cross-entropy loss, as defined in Equation (4):

$$L_{RTD} = -\mathbb{E}[\sum_{i=1}^{n} (y_i logP_D\left(real\middle|x_i, x_{replaced}\right) + (1-y_i)\log\left(1-P_D(real|x_i, x_{replaced})\right))] \quad (4)$$

where $y_i = 1$ if x_i is a real token and $y_i = 0$ if x_i is a replaced token.

The joint training objective of ELECTRA is formalized in Equation (5), where the discriminator loss L_{RTD} is combined with the generator's masked language modeling loss L_{MLM} through a weighted summation.

$$L = L_{RTD} + \lambda L_{MLM} \quad (5)$$

where λ is a hyperparameter that controls the relative importance of the generator loss. This joint objective ensures that the generator produces plausible replacements for the discriminator to classify, while the discriminator learns robust representations of the input text.

ELECTRA's RTD task provides two key advantages over traditional MLM-based models. First, it trains on all tokens in the input sequence, rather than just a small subset, leading to faster convergence and better sample efficiency. Second, the RTD task is more aligned with downstream tasks, where the model processes unmasked text and makes predictions based on the entire input sequence. As a result, ELECTRA achieves SOTA performance on a wide range of NLP benchmarks, including GLUE, SQuAD, and SuperGLUE, while requiring significantly less compute during pre-training.

The generator and discriminator in ELECTRA share the same input embeddings but have separate parameters for their respective tasks. The generator is typically much smaller than the discriminator to reduce computational overhead. For example, in the base ELECTRA model, the generator has 12 layers, while the discriminator has 12 layers and a larger hidden size. Both components are trained simultaneously on large text corpora, such as Wikipedia and BooksCorpus, using the joint training objective described.

After pre-training, the discriminator can be fine-tuned on specific downstream tasks, such as text classification, named entity recognition (NER), question answering, and sentiment analysis [9]. ELECTRA's robust representations and efficient training make it a powerful tool for a wide range of NLP applications [49]. The overall workflow of the methodology can be seen in Figure 1.

4. Results and Performance Evaluation

4.1. Augmented Data Characteristics and Label Distribution

As previously mentioned, the Pennebaker and King essays dataset, comprising 2467 data points, was employed to train multiple classification models for predicting different traits of the big five personality model using the ELECTRA framework. These classification processes allow for the independent prediction of each personality trait, which researchers and end-users can subsequently combine to generate a comprehensive personality profile for an individual. To enhance the model's predictive accuracy, researchers implemented data augmentation techniques, a common practice in text-based ML tasks to improve generalization and robustness.

Text data augmentation can be achieved through various approaches, including text-based, character-level, sentence-level, and rule-based methods. Text-based augmentation techniques involve transformations at the word level, such as synonym replacement, random insertion, random deletion, and random swapping. More advanced strategies include back translation, text paraphrasing, and word embedding perturbation using models like Word2Vec or FastText. Contextual word replacement, leveraging MLMs such as BERT, allows for the substitution of words based on their contextual meaning. Additionally, sentence shuffling and style transfer techniques can be employed to further diversify the dataset without altering the underlying semantics [50].

Character-level augmentation techniques introduce variations at the character level, such as random character insertion and deletion (e.g., transforming "hello" into "h3llo" or "helo"). Other methods include character swapping, where adjacent characters are interchanged (e.g., "hello" → "hlelo"), and keyboard typo simulation, which mimics common typing errors based on QWERTY keyboard

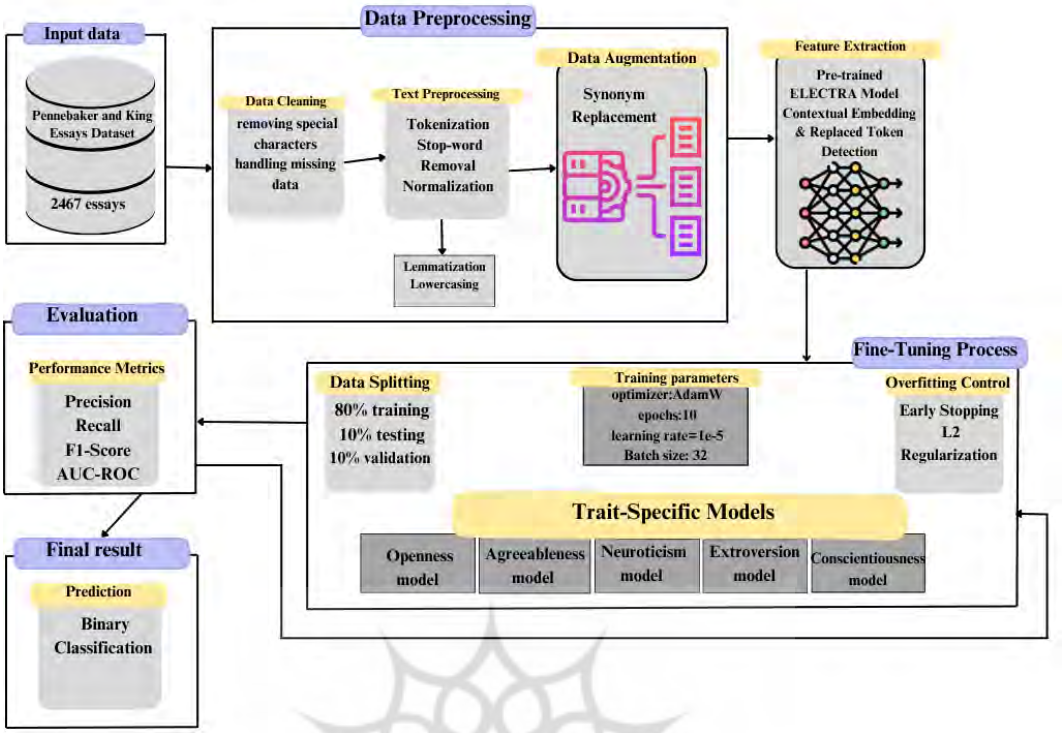

Figure. 1. Conceptual framework of the proposed approach

layout (e.g., "hello" → "hrllo"). Another approach is leetspeak conversion, where words are modified using numerical substitutions (e.g., "hello" → "h3ll0"). While these methods can introduce diversity, they may also generate noisy data, which can impact model performance if not carefully controlled.

Sentence-level augmentation focuses on modifying entire sentences rather than individual words or characters. This can be achieved through paraphrasing techniques using NLP models such as GPT-3, T5, or BART to generate alternative sentence structures while maintaining semantic meaning. Other methods include sentence splitting and merging, adversarial text perturbation, which introduces slight modifications while preserving intent, and the generation of new text using LLMs. Template-based augmentation is another effective strategy, where placeholders in sentences are replaced with varying values to create new training examples (e.g., "The [animal] is [adjective]" → "The cat is cute" or "The dog is friendly") [51].

Additionally, rule-based and external data augmentation techniques can enhance text diversity by incorporating knowledge from external sources. Thesaurus-based augmentation replaces words with their synonyms to introduce lexical variety, while knowledge-based augmentation injects domain-specific terminology to improve contextual relevance. Another method, mixup augmentation, involves merging elements from two different sentences to generate novel variations that retain semantic coherence [50].

For this study, synonym replacement was initially selected as a core data augmentation technique due to its ability to increase lexical diversity while preserving the semantic content of the original text. By substituting selected words with appropriate synonyms, the model becomes more robust to lexical variation and less sensitive to specific word choices. Synonym replacement is also computationally efficient, requiring no large external models, unlike resource-intensive methods such as paraphrasing or back translation. However, to address the limitations of synonym-based augmentation, like as potential unnatural phrasing, we extended the pipeline with contextual augmentation using the Gemma-27B-IT model, which generates fluent, contextually coherent paraphrases. This hybrid strategy increases linguistic variability while maintaining semantic fidelity.

Synonym replacement was implemented using WordNet, which groups words into synonym sets (synsets). For each candidate word, the algorithm retrieves valid synonyms and randomly substitutes up to two words per sentence to avoid semantic drift. Tokenization and candidate selection were performed with NLTK. This process expanded the dataset from 2,467 to approximately 4,934 samples. For example, the original sentence:

"Sitting here just writing stuff down on paper. Thinking about going out tonight. I'm pretty happy because the navy paid me some more money."

may be transformed into:

"Sitting here just writing material down on paper. Thinking about departure out tonight. I'm pretty happy because the navy paid Maine some more money."

To complement synonym replacement, contextual augmentation with Gemma-27B-IT was applied to produce semantically faithful paraphrases that more naturally vary sentence structure and lexical choices. This mitigates the rigidity and occasional unnaturalness of purely synonym-based augmentation.

Importantly, all augmented samples were used exclusively in the training set. The validation and test sets remained strictly unaugmented to ensure an unbiased evaluation of model performance and to prevent artificial inflation of results.

Figure 2 illustrates the class distribution (low/high) for each big five trait after augmentation. The counts remain approximately balanced (e.g., agreeableness: 2314 vs. 2620; conscientiousness: 2428 vs. 2506; extraversion: 2382 vs. 2552; neuroticism: 2468 vs. 2466; openness: 2392 vs. 2542), indicating that the dataset is well suited for model training and evaluation.

4.2. Experimental Setup and Training Protocol

The classification models are based on ELECTRA, a transformer-based architecture optimized for efficient language modeling. The framework adapts ELECTRA to classify five OCEAN traits, treating each trait as an independent binary classification task that determines whether an individual exhibits that trait. Central to this approach is transfer learning, where the model leverages ELECTRA's pre-trained knowledge of syntax, semantics, and discourse structure. Fine-tuning is achieved by replacing ELECTRA's pretraining head

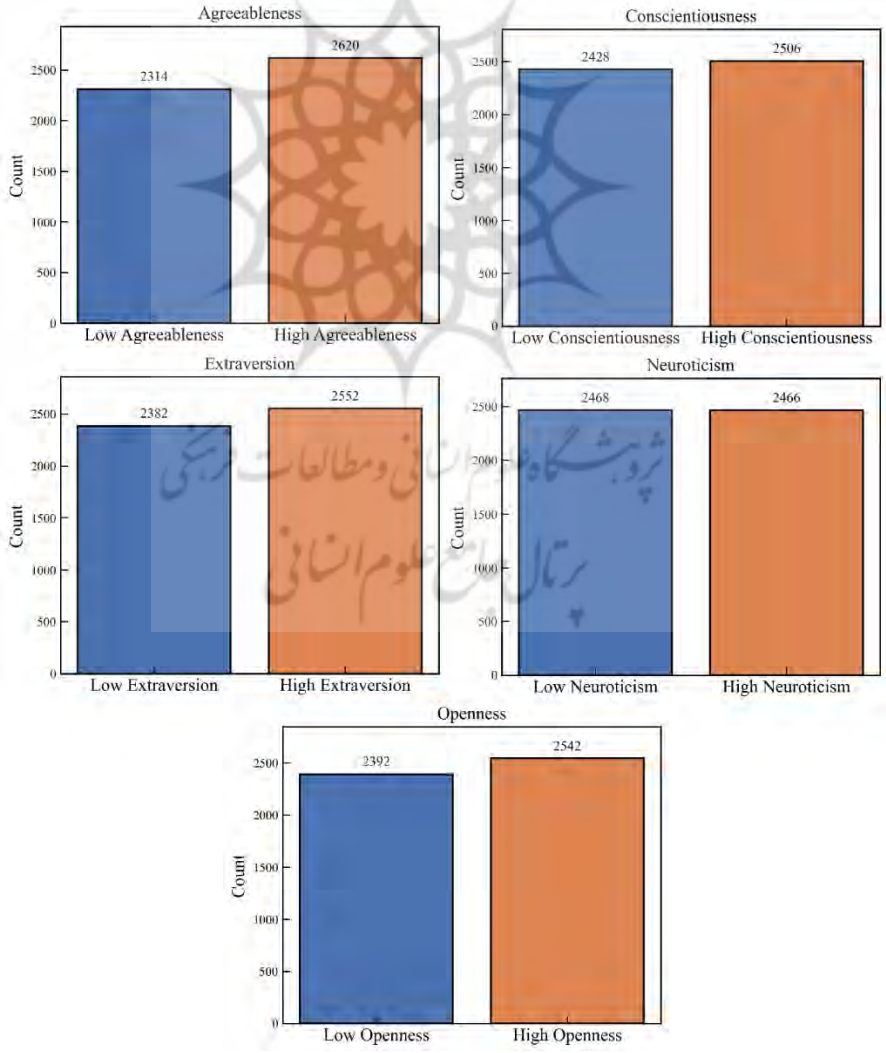

Figure. 2. Class distribution (low vs. high) across the five personality traits.

with a classification layer, mapping contextualized text embeddings to personality labels. This allows the model to detect linguistic and stylistic cues, such as cooperative language and politeness markers, which are indicative of certain personality traits, while still preserving its broader language understanding capabilities [52].

The input text undergoes regex-based normalization to remove non-alphabetic characters and standardize casing, ensuring that the model focuses on meaningful lexical content rather than noise. To enhance generalization, data augmentation is incorporated through synonym replacement, and contextual augmentation using Gemma-27b-IT generating paraphrased samples that maintain their original labels. This augmentation strategy helps the model learn invariant features across different phrasings, reducing overfitting to specific surface-level patterns. Given the nuanced nature of personality trait analysis, where small linguistic variations can carry significant meaning, this balance between lexical diversity and semantic consistency is crucial.

The dataset is split into 80 percent for training, 10 percent for validation, and 10 percent for testing, ensuring a reliable evaluation framework. Augmented samples are used exclusively in the training portion, while the validation and test sets remain strictly unaugmented to preserve unbiased performance assessment. Text is tokenized with a maximum sequence length of 256 tokens, and dynamic padding is applied at the batch level to optimize memory efficiency.

A 5-fold cross validation procedure with 10 epochs is employed to determine the optimal training configuration for each trait-specific classifier. Cross validation is carried out independently for all five traits to account for differences in linguistic patterns and label distributions. The search space includes learning rate, batch size, weight decay, and warmup ratio. For each trait, the hyperparameters that achieve the highest mean validation F1-score across the folds are selected for the final fine-tuning stage. Each personality trait is modeled using the ELECTRA-base discriminator (google/electra-base-discriminator). Fine-tuning is performed with Hugging Face's Trainer and optimized using the AdamW optimizer.

Table 2 summarizes the optimal hyperparameters identified for each classifier and presents the mean validation F1-scores across folds, illustrating the stability and generalization achieved through cross-validated tuning.

4.3. Evaluation Metrics

There are several statistical standards used to evaluate the accuracy of a classification models, and some of the commonly used ones include: confusion matrix, accuracy, recall, precision, Positive Predictive Value (PPV), Matthews Correlation Coefficient (MCC), G-Mean, Cohen's Kapp, specificity, and F1-score. In this study the confusion matrix, accuracy, recall, precision, F1-score were utilized to evaluate the classification performance of an ELECTRA model. Accuracy, recall, precision, and F1-score are formally defined in Equations (6)–(9), while the confusion matrix formulation is presented in Equation (10). Together, these metrics provide a comprehensive evaluation of how effectively the models assign essays to their corresponding personality categories.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (6)$$

$$Recall = \frac{TP}{TP + FN} \quad (7)$$

$$Precision = \frac{TP}{TP + FP} \quad (8)$$

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (9)$$

$$Confusion\ Matrix = \begin{matrix} \\ Class\ True \\ Class\ False \end{matrix} \begin{matrix} \begin{bmatrix} TP & FP \\ FN & TN \end{bmatrix} \end{matrix} \quad (10)$$

where, *TP* is true positive, *FP* represents false positive, *TN* is true negative, and *FN* represents false negative.

Accuracy measures the overall proportion of correctly classified essays and reflects how often the model correctly identifies whether a participant belongs to the high or low category of a personality trait.

Table 2. Cross validation hyperparameters and the final selected configuration.

Trait	Learning Rate	Batch Size	Weight Decay	Warm-up Ratio	Validation F1-Score
Agreeableness	5e-06	32	0.1	0.06	0.6970
Extroversion	2e-05	16	0.01	0.06	0.6850
Conscientiousness	1e-05	32	0.1	0.06	0.6784
Neuroticism	2e-05	16	0.01	0.06	0.6621
Openness	3e-06	16	0.0	0.06	0.6555

Additionally, recall or sensitivity indicates how effectively the model identifies all true high-trait cases, such as essays written by individuals with a high level of a trait; a higher recall means fewer high-trait individuals are overlooked.

Furthermore, precision measures the proportion of essays predicted as high-trait that are truly high-trait, which is essential in personality detection because low precision would imply many false alarms, such as incorrectly labeling low-trait individuals as high.

Moreover, the F1-score provides a balanced measure between precision and recall, making it valuable when both false positives and false negatives carry similar importance.

Together with these metrics, the confusion matrix summarizes the model's classification outcomes, showing where the model correctly or incorrectly predicts high-trait and low-trait classes. These evaluation metrics accuracy, recall, precision, F1-score, and the confusion matrix are widely used in ML due to their clarity, interpretability, and effectiveness in assessing classifier performance.

For all metrics, a binary averaging strategy is applied to account for the nature of the classification task. By combining transfer learning, data augmentation, and robust fine-tuning, this approach ensures an effective framework for predicting personality traits based on textual input.

4.4. Openness Model

This part evaluates the performance of ELECTRA-based model developed to classify individuals based on the personality trait "openness". The dataset used for training appears to be balanced, as observed in the distribution chart (Figure 2). The presence of nearly equal instances of both classes ensures that the model does not develop bias toward one particular class. A balanced dataset is crucial in classification tasks, as it allows the model to learn patterns effectively and make unbiased predictions.

One of the primary considerations in training DL models is selecting the optimal number of epochs to prevent overfitting [53]. The training, validation, and loss curves (Figure 3) provide insight into the model's learning behavior across ten epochs. The training loss decreases consistently, which indicates that the model is learning effectively from the data. However, the validation and testing loss curves exhibit fluctuations, especially around nine, where a noticeable spike is observed. This suggests potential overfitting, meaning the model is memorizing the training data rather than generalizing well to unseen data. The accuracy curves (Figure 3) further support this observation. While training accuracy continues to improve, validation and testing accuracy do not follow the same trend, showing instability in later epochs. The training and validation loss curves clearly indicate that after epoch five, the model begins to show signs of overfitting, with validation loss increasing while training loss continues to decrease. To mitigate overfitting, researchers selected the model from epoch five as our final version and evaluated its performance on the validation and test datasets.

The receiver operating characteristics (ROC) curves in Figure 4 demonstrate our model's ability to distinguish between the two classes. The area under the curve (AUC) values for training, validation, and test datasets are 0.96, 0.81, and 0.79, respectively. These results suggest that while the model performs exceptionally well on the training set, its performance on validation and test data is lower but still within an acceptable range. The fact that the validation and test AUC remain above 0.75 indicates that the model retains strong predictive power and generalizes reasonably well, despite the overfitting concerns observed in longer training periods.

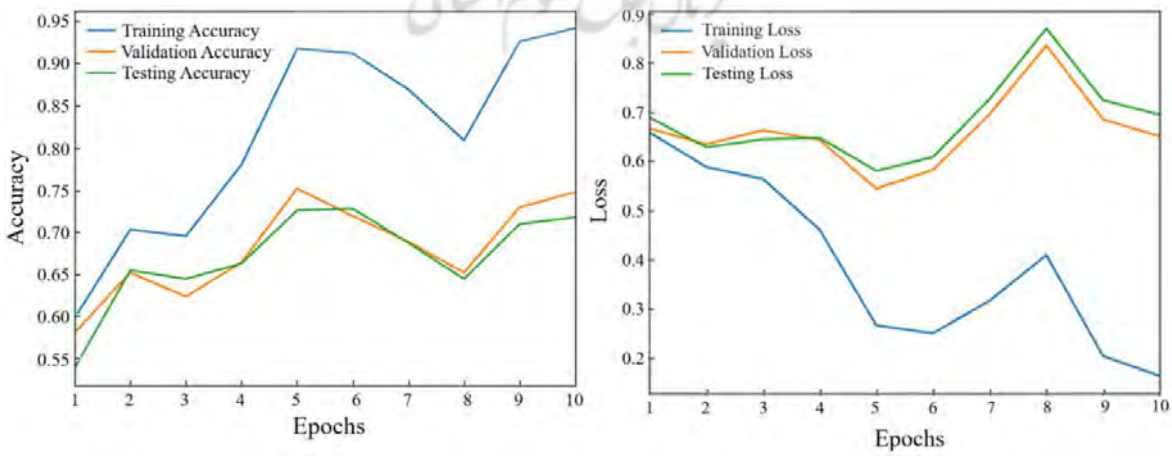

Figure. 3. Performance curves (accuracy and loss) of training, validation, and test for trait openness

A closer look at the classification reports in Table 3, and confusion matrices in Figure 5 reveals that in the training set, precision and recall values range between 0.88 and 0.95, leading to an F1-score of 0.92. However, in the test set, precision decreases to 0.71-0.75, and recall is slightly lower at 0.66-0.78, resulting in an F1-score of 0.72. A similar pattern is observed in the validation set, where the F1-score stabilizes at 0.75. These results confirm that selecting epoch five for evaluation was a reasonable choice, as training longer would have led to further performance degradation on unseen data.

The precision-recall (PR) curves further illustrate the trade-offs between precision and recall (Figure 4). The training set PR curve (AUC=0.96) is significantly higher than those for validation (AUC=0.81) and test (AUC=0.77) sets. While this confirms overfitting, it also demonstrates that our model maintains a good balance of precision and recall, ensuring that both false positives and false negatives are minimized.

The ELECTRA-based model performed well, achieving strong results on validation and test datasets when evaluated at epoch five. By selecting this epoch, researchers effectively mitigated overfitting while preserving high classification accuracy. The AUC values above 0.75 on unseen data confirm that the model generalizes well, making it a reliable predictor for the trait openness. The PR trade-offs and classification scores indicate that our model is well-optimized for this task.

Although this model performs well, some areas could still be improved. First, overfitting remains a concern, as shown by the gap between training and test performance. Regularization techniques such as dropout, weight decay, or adversarial training might help further mitigate this.

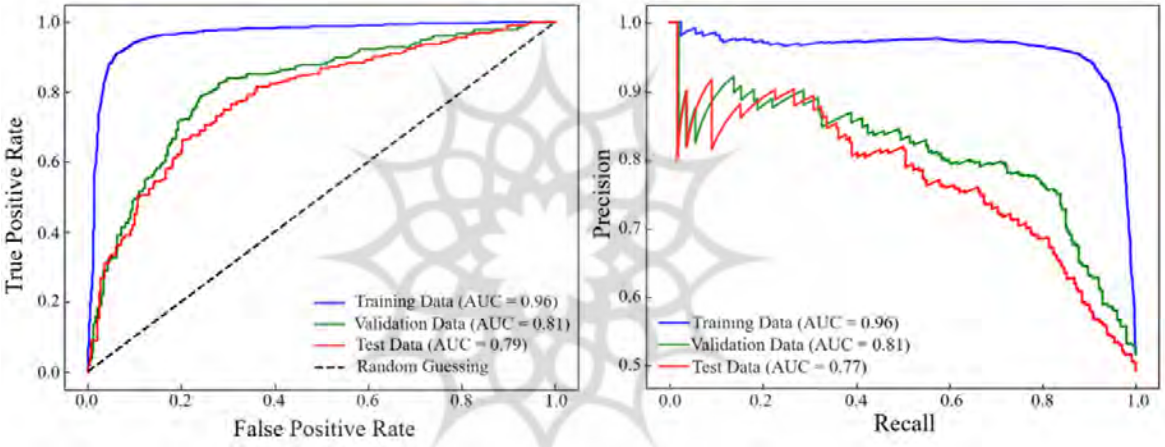

Figure. 4. ROC and PR curves for the openness classifier.

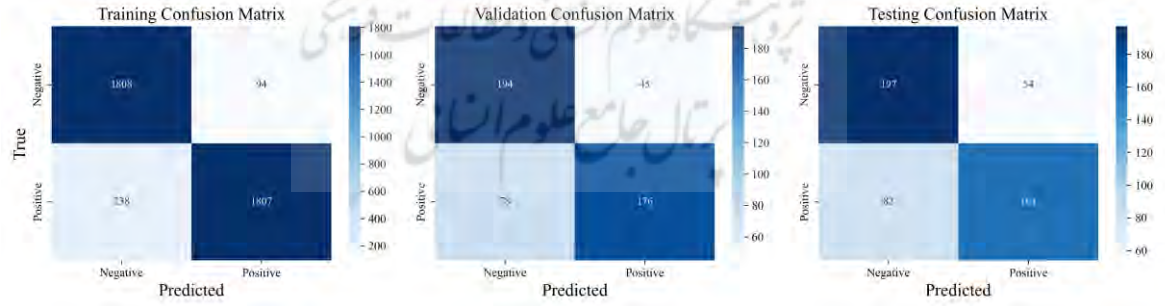

Figure. 5. Openness confusion matrices.

Table 3. Analysis of the trait openness.

Data Type	Class	*Accuracy*	*Recall*	*Precision*	*F1 Score*
Training	Positive	0.92	0.88	0.95	0.92
	Negative		0.95	0.88	0.92
Validation	Positive	0.75	0.69	0.80	0.74
	Negative		0.81	0.71	0.76
Testing	Positive	0.72	0.66	0.75	0.70
	Negative		0.78	0.71	0.74

Recall values are slightly lower in the test and validation sets, suggesting that some positive instances are being missed. Adjusting the classification threshold could improve recall without sacrificing precision too much. Moreover, Alternative architectures or fine-tuning strategies could be explored further to enhance generalization.

4.5. Conscientiousness Model

This part evaluates the ELECTRA-based model to classify individuals based on the conscientiousness personality trait. The class distribution analysis confirms a well-balanced dataset, which ensures that the model does not develop a bias toward either class. This balance is crucial in the personality classification task, as it allows the model to learn a fair representation of both positive and negative classes.

One of the critical aspects of DL model training is selecting the optimal number of epochs to prevent overfitting. The training, testing, and validation loss curves show (Figure 6) a clear overfitting after epoch 7, where the validation and test losses begin to rise while the training loss continues to decrease. Similarly, the accuracy curves shown in Figure 6 illustrate that the training accuracy continues to improve beyond epoch 7, but validation and test accuracies fluctuate, suggesting that the model is memorizing training patterns rather than generalizing to unseen data. To counteract this issue, researchers selected epoch seven as the final model checkpoint and evaluated its performance based on validation and test set results.

The ROC curves in Figure 7 demonstrate the model's ability to differentiate between classes. The AUC values for training, validation, and test datasets are 0.98, 0.78, and 0.79, respectively. The near-perfect training AUC indicates that the model has learned the training data exceptionally well, but the gap between training and validation/test AUC values suggests overfitting. However, with validation and test AUC values around 0.78-0.79, the model still exhibits strong generalization capability.

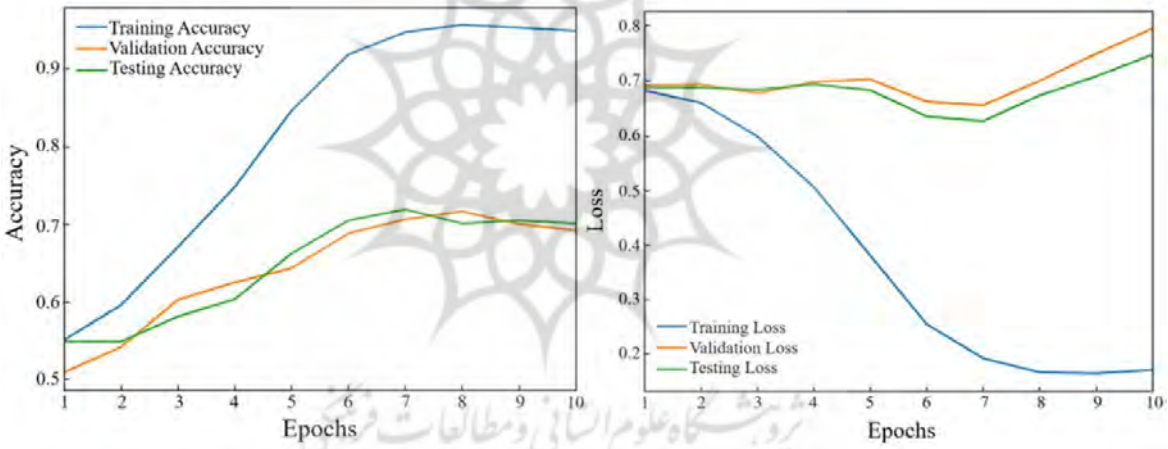

Figure. 6. Performance curves (accuracy and loss) of training, validation, and test for trait conscientiousness

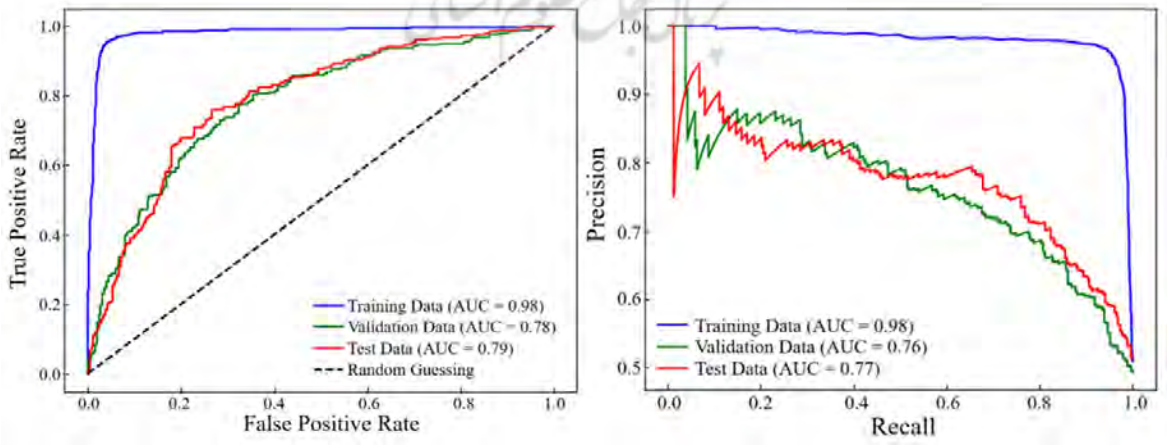

Figure. 7. ROC and PR curves for the conscientiousness classifier

The classification reports in Table 4, and the confusion matrices further break down the classification performance across datasets in the Figure 8. For the training set, the model achieves a high accuracy of 0.95, with only a small number of misclassifications. However, in the test set, accuracy drops to 0.72, and the validation accuracy stabilizes at 0.71. The classification reports indicate that in the training dataset, precision and recall values range between 0.92 and 0.97, leading to an F1-score of 0.95. However, for the test datasets, precision drops to 0.69-0.77, while recall ranges between 0.60-0.83, resulting in an F1-score of 0.70-0.71. The validation dataset follows a similar trend, with precision, recall, and F1-scores stabilizing around 0.70-0.72.

The PR curves in Figure 7 provide further insights into the model's predictive performance. The training set PR AUC (0.98) is significantly higher than that of validation (0.76) and test sets (0.77), reinforcing the overfitting observation. Nevertheless, the model still maintains a reasonable balance of precision and recall, suggesting that while it does overfit the training data, it still retains moderate generalization capability on unseen data.

By selecting epoch 7, researchers managed to balance the trade-off between overfitting and generalization while achieving an accuracy of 0.72 on the test set and 0.71 on the validation set. The AUC values (0.78-0.79) confirm that the model is a strong predictor of the conscientiousness trait, despite some overfitting. The model's PR balance further suggests that it maintains reliable classification ability, effectively distinguishing between conscientious and non-conscientious individuals.

Despite the model's strong performance, some areas warrant further improvement. Overfitting remains a challenge, as evident from the large discrepancies between training and test/validation results. Applying techniques such as dropout regularization, data augmentation using backtranslation, or adding more context could help mitigate this.

Recall values fluctuate, especially for the negative class, indicating that some negative instances are being misclassified as positive. Adjusting the classification threshold might improve recall without significantly sacrificing precision. Alternative architectures or fine-tuning strategies could be explored to enhance generalization performance, especially for unseen data. By addressing these concerns, the model could become more robust in the future while maintaining its already promising classification performance for the conscientiousness trait.

4.6. Extroversion Model

For the prediction of extroversion, ELECTRA leverages its transformer-based architecture for classification. The class distribution in Figure 2 reveals a well-balanced dataset, ensuring that the model is not biased toward any specific class. A balanced dataset is essential for fair classification, particularly in personality assessment tasks where skewed distributions can lead to biased predictions.

The training, validation, and testing loss curves in Figure 9 demonstrate a common trend observed in DL models. As training progresses, the training loss decreases significantly, whereas validation and test losses initially decrease but start to fluctuate after

Table 4. Analysis of the trait conscientiousness.

Data Type	Class	*Accuracy*	*Recall*	*Precision*	*F1 Score*
Training	Positive	0.95	0.92	0.93	0.95
	Negative		0.97	0.97	0.94
Validation	Positive	0.71	0.81	0.67	0.73
	Negative		0.60	0.77	0.68
Testing	Positive	0.72	0.83	0.69	0.75
	Negative		0.60	0.77	0.67

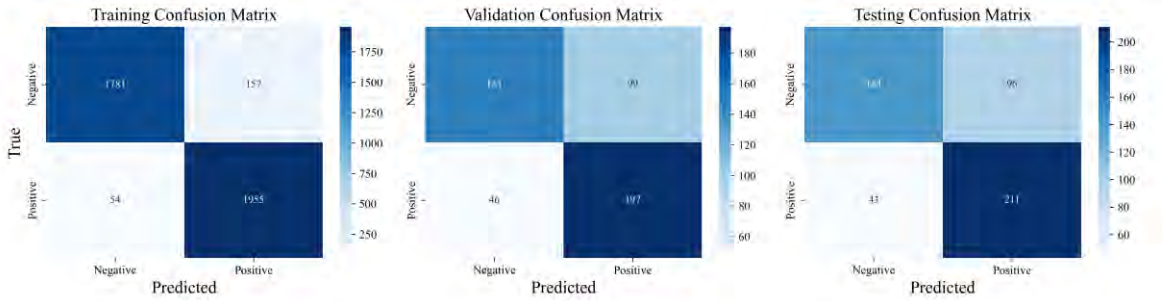

Figure. 8. Conscientiousness confusion matrices

epoch 8. This suggests that the model begins to overfit beyond this point, learning patterns specific to the training data rather than generalizing well to unseen data. Similarly, the accuracy curves in Figure 9 reveal that while training accuracy continues to increase steadily, validation and test accuracies reach a peak around epoch 8 before showing minor fluctuations. Based on these observations, epoch 8 was selected as the optimal checkpoint for extracting final evaluation results.

The ROC curves in Figure 10 provide insight into the model's classification capability across different datasets. The AUC values for training, validation, and test datasets are 0.98, 0.79, and 0.84, respectively. The exceptionally high AUC for the training set confirms that the model has learned training data features effectively, while the slightly lower validation and test AUC values indicate some degree of overfitting. However, with validation and test AUCs around 0.79-0.84, the model still demonstrates strong generalization ability for extroversion classification.

According to Table 5, the training set achieves an impressive 0.96 accuracy, with minimal misclassification. However, in the test set, the accuracy drops to 0.78, and the validation accuracy stabilizes at 0.74. The classification reports further reinforce these findings. The training dataset, precision, and recall values are consistently around 0.95-0.96, leading to an F1-score of 0.96. However, in the test dataset, precision values range from 0.74 to 0.82, with recall values between 0.73 and 0.83, leading to an F1-score of 0.77-0.78. The validation dataset follows a similar pattern, with F1-scores stabilizing around 0.74. The confusion matrices illustrate the classification performance in greater detail (Figure 11).

In Figure 10, the PR curves provide an additional perspective on the model's performance. The training PR AUC (0.98) is significantly higher than that of the validation (0.77) and test (0.82) sets, reinforcing the overfitting observed earlier. However, the model maintains a strong balance between precision and recall, demonstrating that it is effective in distinguishing between extroverted and non-extroverted individuals.

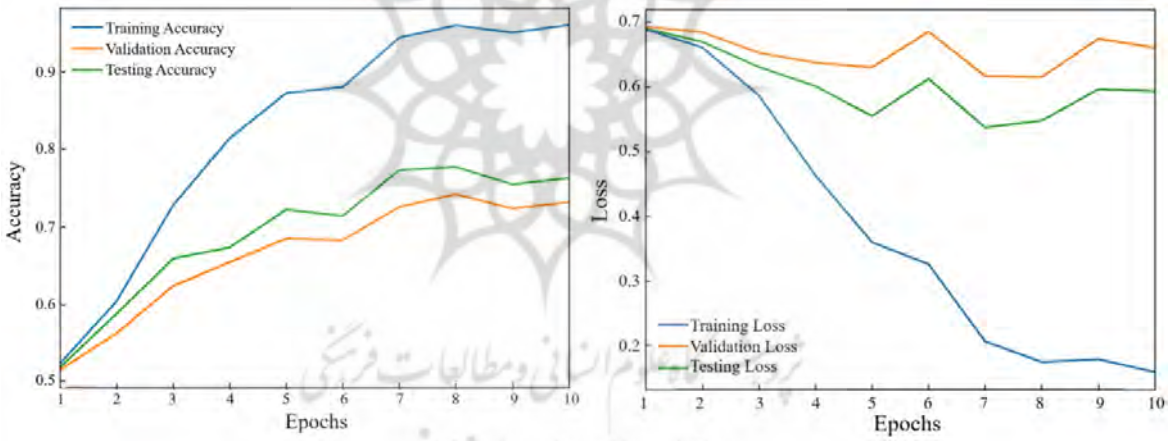

Figure. 9. Performance curves (accuracy and loss) of training, validation, and test for trait extroversion

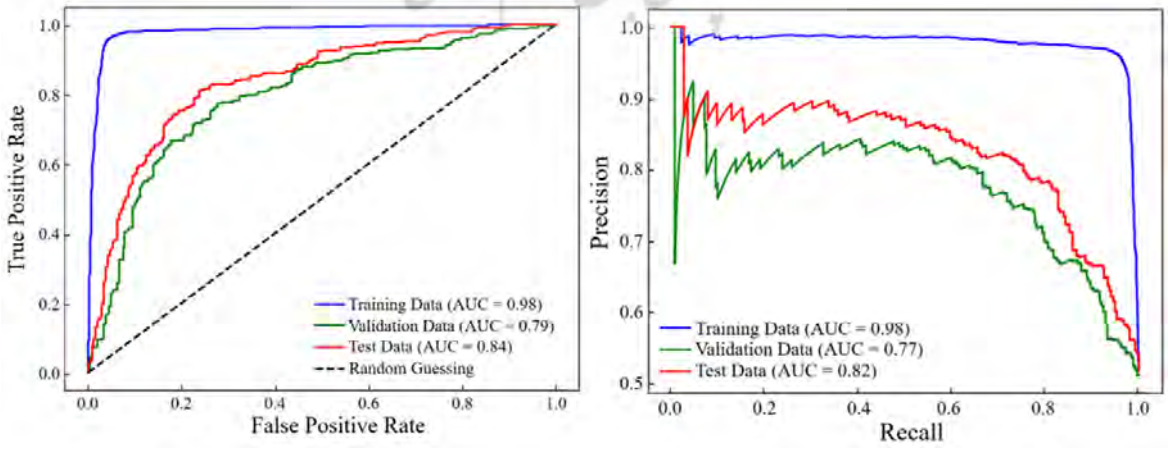

Figure. 10. ROC and PR curves for the extroversion classifier

Table 5. Analysis of the trait extroversion.

Data Type	Class	*Accuracy*	*Recall*	*Precision*	*F1 Score*
Training	Positive	0.96	0.95	0.96	0.96
	Negative		0.96	0.95	0.96
Validation	Positive	0.74	0.71	0.77	0.74
	Negative		0.78	0.72	0.75
Testing	Positive	0.78	0.73	0.82	0.77
	Negative		0.83	0.74	0.78

Figure. 11. Extroversion confusion matrices

By selecting epoch 8, we effectively balance the trade-off between overfitting and generalization, achieving a test accuracy of 0.78 and validation accuracy of 0.74. The AUC values (0.79-0.84) confirm that the model is a reliable predictor of the extroversion trait. The PR balance further suggests that the model maintains a strong classification performance across various datasets.

Despite the model's strong overall performance, several areas require further attention. Overfitting remains an issue, as seen in the discrepancy between training and validation/test performance. Implementing dropout layers, weight regularization, or additional data augmentation could mitigate this effect. The recall for the negative class is slightly lower than that of the positive class, indicating that some negative cases are being misclassified as positive. Adjusting the classification threshold could help address this imbalance.

4.7. Agreeableness Model

This section evaluates the performance of the ELECTRA-based model in classifying individuals based on the personality trait agreeableness. The class distribution in Figure 2 reveals a slight imbalance, with more instances of the positive class, but this does not appear to have significantly impacted the model's performance.

The training, validation, and test loss curves in Figure 12 show that while training loss steadily decreases, the validation and testing loss fluctuate but remain stable after epoch nine, making it the optimal checkpoint. The accuracy curves in Figure 12 indicate that while training accuracy continues to rise, the validation and test accuracies exhibit an upward trend, peaking around epoch nine. This suggests that the model has successfully learned features relevant to agreeableness while maintaining generalization to unseen data.

The ROC curves in Figure 13 for training, validation, and test datasets highlight the model's discriminative power. The AUC values for validation and test sets are 0.82 and 0.80, confirming that the model is effective in distinguishing between high and low agreeableness traits.

The classification reports in Table 6 provide deeper insight into performance. The training precision and recall are exceptionally high (0.96), indicating that the model has learned well from the training data. However, the test accuracy stabilizes at 0.73, with an F1-scores of 0.68-0.76. The validation accuracy stabilizes at 0.74, with F1-score of 0.69-0.78. These scores confirm that the model is not merely memorizing training data but has learned meaningful patterns applicable to new samples. The confusion matrices in Figure 14 further support this, showing that the model accurately predicts both classes with reasonable balance.

The PR curves in Figure 13 further validate the model's effectiveness. The test and validation PR AUC scores of 0.82 indicate a strong balance between precision and recall, suggesting that the model maintains high predictive capability while minimizing false classifications. This is particularly crucial in psychological assessment, where balanced classification ensures that no significant bias affects predictions.

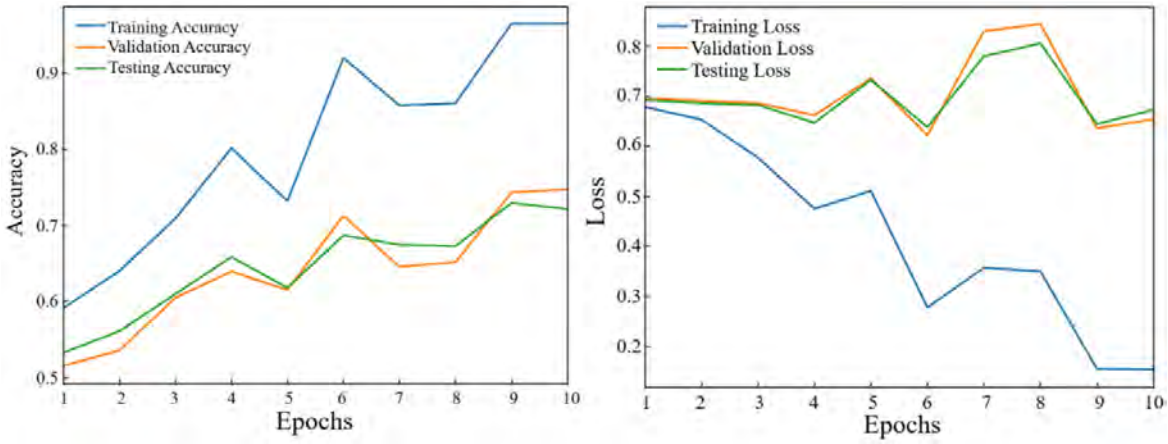

Figure. 12. Performance curves (accuracy and loss) of training, validation, and test for trait agreeableness

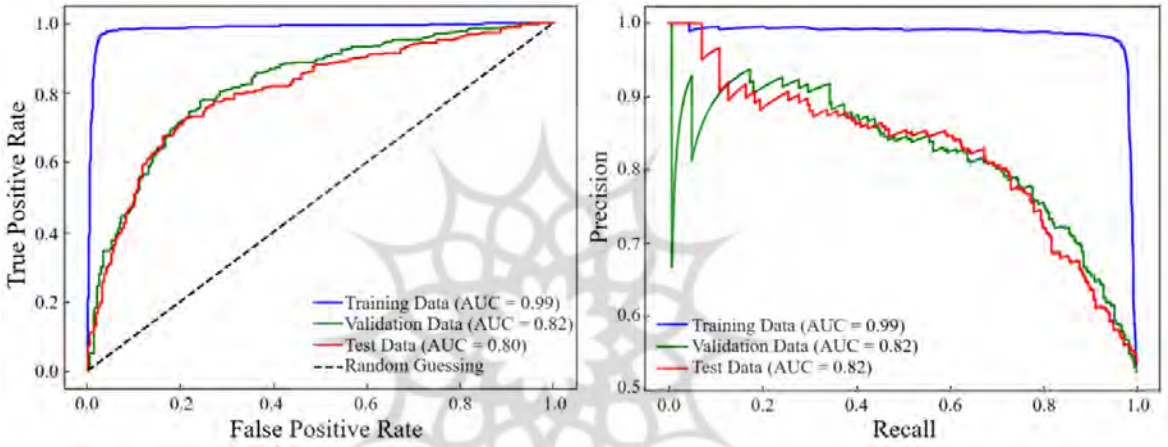

Figure. 13. ROC and PR curves for the agreeableness classifier

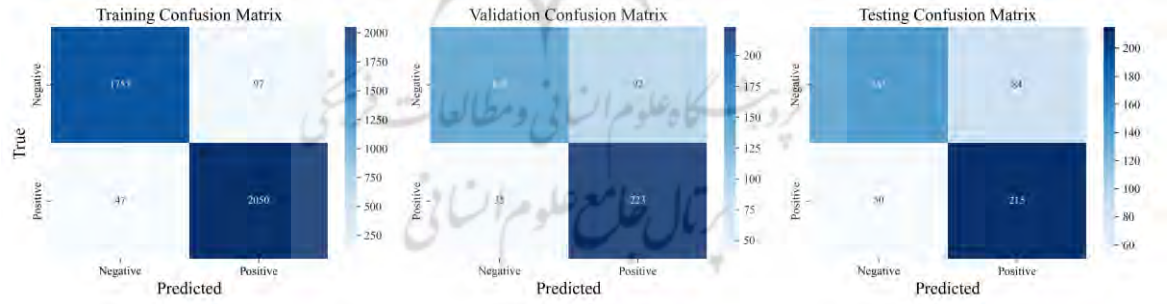

Figure. 14. Agreeableness confusion matrices

Table 6. Analysis of the trait agreeableness.

Data Type	Class	*Accuracy*	*Recall*	*Precision*	*F1 Score*
Training	Positive	0.96	0.98	0.95	0.97
	Negative		0.95	0.97	0.96
Validation	Positive	0.74	0.86	0.71	0.78
	Negative		0.61	0.80	0.69
Testing	Positive	0.73	0.81	0.72	0.76
	Negative		0.63	0.74	0.68

By selection epoch nine, researchers ensure that the model reaches its best generalization ability, striking a balance between learning and avoiding overfitting. With test accuracy of 0.73 and validation accuracy of 0.74, the model demonstrates a strong predictive capability for identifying agreeableness. The AUC values of 0.82 for both test and validation datasets in Figure 13 confirm that the model can reliably classify individuals based on this personality trait.

Despite a performance gap between training and test data, the model does not show excessive overfitting, as indicated by its ability to minimize generalization while still learning meaningful patterns. Further refinements, such as hyperparameter tuning and adding different data context, could further enhance its robustness while preserving its strong predictive performance.

4.8. Neuroticism Model

The performance of the ELECTRA-based model for classifying individuals based on the personality trait neuroticism is evaluated in this section. The dataset used for training appears to be balanced, as observed in the classification distribution in Figure 2. The dataset contains an approximately equal number of instances for both classes, ensuring that the model does not develop a bias toward any particular class. A balanced dataset is crucial in classification tasks, as it helps the model learn patterns effectively and make unbiased predictions.

The training, validation, and testing loss curves in Figure 15 show that the training loss consistently decreases, indicating that the model is learning effectively. However, both validation and test losses fluctuate, suggesting that the model's generalization ability is somewhat unstable. Notably, training accuracy continues to increase, while validation and test accuracy peak at the epoch, before slight fluctuations begin to emerge. Based on this, epoch seven is selected as the optimal model checkpoint for further evaluation.

The ROC curves in Figure 16 provide insight into the model's discrimination power. The AUC values for training, validation, and test datasets are 0.97, 0.79, and 0.79, respectively. The training AUC being significantly higher than the validation and test AUCs suggests that the model is likely overfitting, memorizing training patterns rather than generalizing to unseen data. Nevertheless, with test and validation AUCs close to 0.79, the model demonstrates moderate reliability in classifying neuroticism.

The classification reports in Table 7, and the confusion matrices in Figure 17 reveal further detail about model performance. The training set achieves a 0.95 accuracy, with very few misclassifications. However, in the test set, the accuracy drops to 0.72, and the validation accuracy is around 0.74. This discrepancy underscores the model's tendency to overfit. The confusion matrices for the test and validation sets show a balanced classification, with both positive and negative classes being identified with reasonable accuracy.

While training precision and recall remain high (0.94-0.96), the test performance stabilizes around 0.72 accuracy, with F1-scores of 0.72-0.74 for both test and validation datasets. These results demonstrate that despite a performance gap between training and unseen data, the model still maintains a strong predictive capability in classifying the neuroticism trait.

The PR curves in Figure 16 provide further confirmation of the model's effectiveness. The test PR AUC of 0.77 and validation PR AUC of 0.80 indicate that the model is capable of maintaining high precision while ensuring a balanced recall rate. This is particularly important in personality recognition tasks, where both false positives and false negatives can have implications for practical applications.

By selecting epoch seven, researchers ensure an optimal balance between learning and generalization. With a test accuracy of 0.72 and validation accuracy of 0.74, the model has demonstrated a strong ability to predict neuroticism in new data. The AUC values of 0.79 for validation and test datasets confirm that the model effectively distinguishes between individuals with high and low neuroticism scores.

The model shows a high predictive capability in identifying the neuroticism trait. The ROC and PR curves confirm its ability to generalize beyond the training data. Moreover, the model successfully maintains a stable recall and precision balance, ensuring reliable predictions. While some degree of overfitting is present, it does not significantly diminish the model's ability to make meaningful predictions. Further optimization, such as fine-tuning the learning rate, adding regularization techniques, or more data augmentation, could enhance its robustness while retaining its predictive power.

4.9. Overall Performance and Comparison

The proposed single trait ELECTRA framework achieves consistently strong performance across all five Big Five dimensions on the Pennebaker and King essays dataset. On the held-out test split, the five independently fine-tuned ELECTRA classifiers obtain accuracies of 0.72 (openness), 0.71 (conscientiousness), 0.74 (extraversion), 0.73 (agreeableness), and 0.72 (neuroticism), resulting in an average accuracy of 0.724. All traits achieve test ROC–AUC values above 0.75, with balanced PR profiles as reported in Table 3-7, indicating that the models maintain good discrimination rather than exploiting trivial decision thresholds.

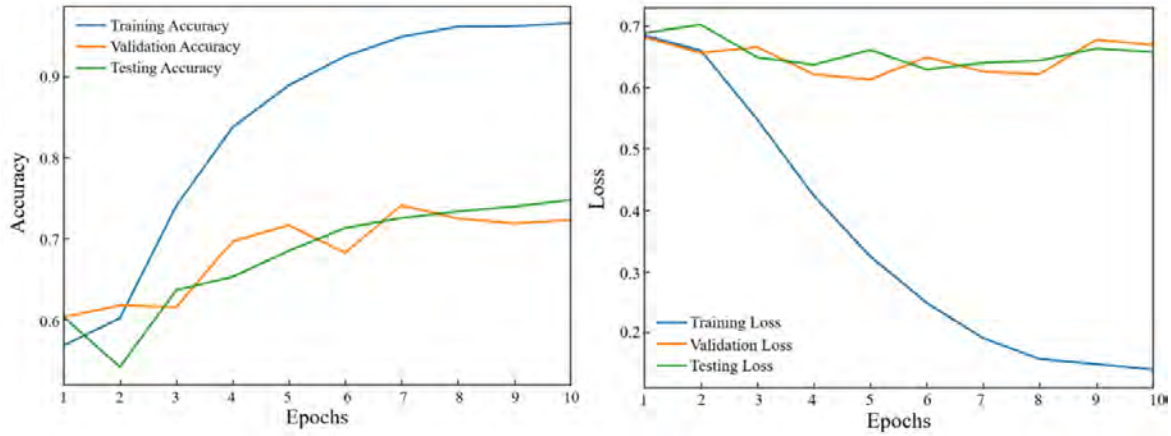

Figure. 15. Performance curves (accuracy and loss) of training, validation, and test for trait neuroticism

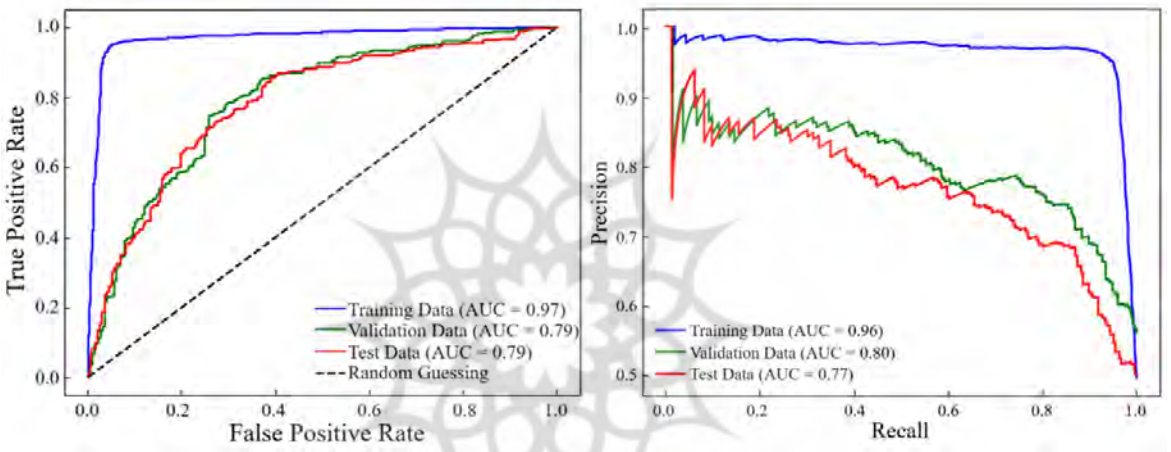

Figure. 16. ROC and PR curves for the neuroticism classifier

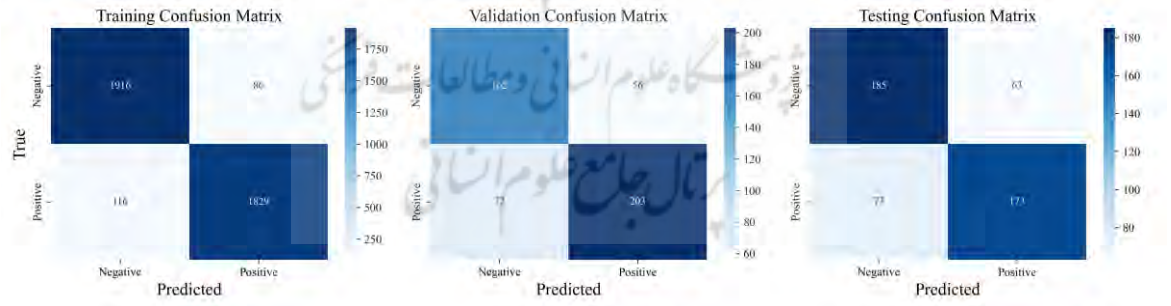

Figure. 17. neuroticism confusion matrices

Table 7. Analysis of the trait neuroticism.

Data Type	Class	*Accuracy*	*Recall*	*Precision*	*F1 Score*
Training	Positive	0.95	0.96	0.96	0.95
	Negative		0.94	0.94	0.95
Validation	Positive	0.74	0.74	0.78	0.76
	Negative		0.74	0.69	0.72
Testing	Positive	0.72	0.70	0.73	0.72
	Negative		0.75	0.72	0.73

Table 8 situates these results against several representative baselines spanning traditional feature-based methods, recurrent and graph based neural models, and recent hybrid or ensemble approaches. At the level of mean accuracy, our single trait ELECTRA ensemble and KGrAt Net form the top tier, with average scores of 0.724 and 0.722 respectively (KGrAt Net's original unrounded average is 0.7241). The difference between the two methods is below 0.3 percentage points and is unlikely to be statistically meaningful; the two systems should be regarded as essentially comparable in overall accuracy rather than one clearly dominating the other.

Per trait comparisons show a more nuanced picture. Our ELECTRA ensemble shares the best reported accuracies for openness and extraversion (0.72 and 0.74, tied with KGrAt Net), and attains the highest accuracies in agreeableness (0.73) and neuroticism (0.72) among the methods considered. In contrast, KGrAt Net achieves the strongest conscientiousness accuracy (0.73 vs. 0.71 for our model), indicating that its knowledge graph representation particularly benefits this trait. BiLSTM based classification over knowledge graph embeddings forms a competitive mid-tier baseline, with an average accuracy of 0.71 and per trait scores between 0.69–0.73, but generally trails both KGrAt Net and the proposed ELECTRA ensemble.

The advantage of transformer based and knowledge graph-based methods becomes clearer when compared with older graph and ensemble systems. Personality GCN reaches a mean accuracy of 0.606 on the essays dataset, despite being SOTA at the time and not relying on external embeddings. Ramezani et al.'s ensemble modeling method, which stacks several heterogeneous base learners, attains an average accuracy around 0.60, with the best single trait result of 0.64 for extraversion. SEPRNN and TF IDF+Bayes baselines perform similarly or worse, with averages of 0.592 and 0.53, respectively. Relative to these earlier systems, the single trait ELECTRA framework improves average accuracy by approximately 12–20 percentage points, highlighting the value of contextualized transformer embeddings and targeted fine tuning in capturing personality relevant linguistic cues.

Another desirable property of the proposed approach is its robustness across traits. Classical feature-based methods such as TF IDF+Bayes show substantial variation (e.g., dropping to 0.49 on agreeableness), whereas our system operates in a narrower band of 0.71–0.74 accuracy across all five traits. This stability is beneficial in downstream scenarios where all personality dimensions are equally important and large performance gaps across traits would be problematic.

Interpreting Table 8 requires several caveats that constrain the strength of any comparative claims. First, the evaluation protocols used across studies differ substantially. Our results are based on an 80/10/10 train–validation–test split with no augmentation applied to the validation or test sets, whereas methods such as Personality GCN and KGrAt Net often rely on k-fold cross-validation or alternative partitioning strategies. As a result, the reported accuracy values are not strictly comparable under identical experimental conditions.

A second limitation is that only accuracy is compared. Several baseline papers emphasize F-measure or other evaluation metrics, sometimes reporting stronger improvements in those measures than in accuracy. Because Table 8 focuses exclusively on accuracy, it may overstate or understate differences for models that prioritize precision over recall or vice versa.

Another source of variation arises from heterogeneous modeling assumptions. KGrAt Net, Personality GCN, and various ensemble approaches incorporate explicit knowledge graphs, multi-stage pipelines, or stacked classifiers that depend on external linguistic resources. In contrast, our method relies solely on transformer-based text encoders supported by WordNet-based and LLM-generated

Table 8. Performance comparison based on accuracy for essays dataset.

Reference	Model Architecture	Accuracy					
		Openness	Conscientiousness	Extraversion	Agreeableness	Neuroticism	AVG.
This Study	**Single-trait ELECTRA, 2025**	**0.72**	0.71	**0.74**	**0.73**	**0.72**	**0.724**
[54]	KGrAt-Net, 2022	**0.72**	**0.73**	**0.74**	0.71	0.71	0.722
[55]	BiLSTM, 2022	0.71	0.72	0.73	0.70	0.69	0.710
[56]	Personality GCN, 2020	0.64	0.59	0.60	0.57	0.63	0.606
[57]	Ensemble Modeling, 2022	0.56	0.59	0.64	0.60	0.61	0.600
[58]	SEPRNN, 2021	0.63	0.57	0.59	0.57	0.60	0.592
[58]	TF-IDF+Bayes, 2021	0.58	0.52	0.54	0.49	0.52	0.530

augmentation. Differences in resource usage, architectural complexity, and hyperparameter tuning introduce further variability and complicate efforts to attribute performance gains to any single design choice.

The use of augmentation and trait-specific models also influences comparability. Our model benefits from doubling the training set through synonym replacement and contextual paraphrasing, a strategy not employed by most earlier baselines. Additionally, training separate ELECTRA models for each trait increases model capacity relative to single multi-label architectures, which may partly account for the observed performance improvements and should be considered when comparing to more parameter-efficient baselines.

Finally, neither our study nor most of the related work reports confidence intervals or formal significance testing for differences in accuracy. Given that several performance gaps fall within a narrow range of 0.2–1.0 percentage points, the ranking of top-performing models, particularly our approach versus KGrAt Net, should be interpreted as indicative rather than definitive. A more detailed discussion of dataset, modeling, augmentation, and ethical limitations is provided in Section 5.

5. Discussion

The results show that a trait-specific ELECTRA framework can achieve stable performance across all five big five dimensions, with accuracies between 0.71 and 0.74 and ROC–AUC values consistently above 0.75. Training separate models reduces cross-trait interference and leads to more trait-focused representations, while the combined synonym-based and contextual augmentation strategy improves robustness under limited-data conditions. These outcomes indicate that contextualized embeddings, when paired with controlled augmentation and cross-validated fine-tuning, can extract relevant linguistic cues even from a relatively small corpus such as the Pennebaker and King essays. However, several constraints shape the generalizability of these findings. The study relies on a single, demographically narrow dataset written in an academic stream-of-consciousness format, which limits transfer to other domains such as social media, dialogue, or multilingual contexts. Although augmentation increases lexical diversity, WordNet substitutions and LLM-generated paraphrases may introduce stylistic drift or subtle semantic changes, and no systematic audit was performed to quantify label preservation. The use of independent ELECTRA models strengthens trait discrimination but increases computational cost and does not exploit known correlations among big five dimensions, which may be better captured by multitask or parameter-efficient alternatives.

A further limitation concerns interpretability, the current work focuses on predictive performance and does not provide an analysis of the linguistic features driving the models' decisions. Without attention-based inspection or gradient attribution, it remains unclear whether the classifiers rely on psychologically meaningful cues or on dataset-specific artifacts. This affects both theoretical value and responsible deployment, as personality inference from text is inherently indirect and grounded in self-reported labels rather than behavioral observations. The absence of statistical significance testing and confidence intervals also restricts the strength of cross-model comparisons, especially given that the differences between top-performing systems are small. Future work should incorporate interpretable modeling tools, evaluate performance on more diverse corpora, and use controlled augmentation validation to assess semantic fidelity. Multitask architectures, improved regularization, and broader evaluation protocols would further strengthen the robustness and applicability of trait-specific transformer models in computational psychology.

6. Conclusion

This study presented a transformer-based framework for personality trait recognition, employing five independently fine-tuned ELECTRA models to classify each big five dimension from free-form essays. By decoupling the traits into separate binary tasks, the framework reduces inter-trait interference and enables more specialized linguistic representations. Leveraging both WordNet-based synonym replacement and contextual augmentation generated by the Gemma-27B-IT model, the training corpus was substantially expanded while preserving semantic integrity, thereby improving model robustness under limited data conditions.

Across OCEAN traits, the proposed approach achieved strong and consistent performance, with test accuracies ranging from 0.71–0.74 and ROC-AUC values above 0.75. These results surpass or match several classical, deep-learning, and transformer-based baselines reported in the literature, demonstrating the effectiveness of contextualized embeddings combined with trait-specific modeling. Although signs of overfitting were observed, reflected in the gap between training and validation/test results, the adoption of careful checkpoint selection and cross-validated hyperparameter tuning helped mitigate these effects and ensured more reliable generalization.

Future research should explore larger and more diverse corpora, improved augmentation validation, and multitask or parameter-efficient architectures that jointly model personality traits while preserving trait-specific nuance. Incorporating multimodal signals such as speech, behavior, or interaction patterns may

enhance predictive reliability in real-world settings. Finally, integrating interpretable AI techniques, such as feature attribution or attention-based linguistic analysis, will be essential for linking model behavior to established psychological theory and ensuring responsible use in applied contexts.

Overall, the findings highlight the promise of transformer-based architectures, particularly ELECTRA, for personality recognition and provide a foundation for more robust, interpretable, and ethically grounded systems in computational psychology. All source code developed for this study is publicly accessible at the GitHub repository: https://github.com/hsisaberi/single-trait-electra.

Declarations

Funding

This research did not receive any grant from funding agencies in the public, commercial, or non-profit sectors.

Authors' contributions

HS: Conceptualization, methodology, validation, formal analysis, investigation, resources, data curation, writing original draft preparation, writing review and editing, visualization.

RR: Conceptualization, validation, writing review and editing, supervision, project administration.

Conflict of interest

The authors declare no conflicts of interest.

References

[1] N. M. Aljuhani, A. A.-M. Al-Ghamdi, H. S. Alghamdi, and F. Saleem, "Convolutional Bi-LSTM for Automatic Personality Recognition from Social Media Texts", IEEE Access, 2025, vol. 13, pp. 65582-65603. https://doi.org/10.1109/ACCESS.2025.3558714.

[2] M. Lukac, "Speech-based personality prediction using deep learning with acoustic and linguistic embeddings", Scientific Reports, 2024, vol. 14, p. 30149. https://doi.org/10.1038/s41598-024-81047-0.

[3] H. Bhin and J. Choi, "Multimodal Personality Recognition Using Self-Attention-Based Fusion of Audio, Visual, and Text Features", Electronics, 2025, vol. 14, p. 2837. https://doi.org/10.3390/electronics14142837.

[4] A. Feher and P. A. Vernon, "Looking beyond the Big Five: A selective review of alternatives to the Big Five model of personality", Personality and Individual Differences, 2021, vol. 169, p. 110002. https://doi.org/10.1016/j.paid.2020.110002.

[5] M. J. Shayegan and M. Valizadeh, "A method for identifying personality traits in telegram", in 8th International Conference on Web Research (ICWR), 2022: IEEE, pp. 88-93. https://doi.org/10.1109/ICWR54782.2022.9786253.

[6] M. Yang, J. Kim, M. Kim, and J. Han, ""What is your MBTI?": Predicting the Personality Types using Hierarchical Attention and Graph Learning", Expert Systems with Applications, 2025, vol. 297, p. 129295. https://doi.org/10.1016/j.eswa.2025.129295.

[7] H.-Y. Suen, K.-E. Hung, and C.-L. Lin, "TensorFlow-based automatic personality recognition used in asynchronous video interviews", IEEE Access, 2019, vol. 7, pp. 61018-61023. https://doi.org/10.1109/ACCESS.2019.2902863.

[8] A. Rasouli, E. Sadraiye, O. Ghahroodi, H. Rabiee, and E. Asgari, "AIMA at SemEval-2025 Task 1: Bridging text and image for idiomatic knowledge extraction via mixture of experts", in Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025), 2025, pp. 2270-2275. https://aclanthology.org/2025.semeval-1.296.

[9] H. Saberi, S. Ghofrani, and R. Ravanmehr, "Personality Recognition Using Transformer Model: A Study on the Big Five Traits", in 11th International Conference on Web Research (ICWR), 2025: IEEE, pp. 228-234. https://doi.org/10.1109/ICWR65219.2025.11006181.

[10] D. Jain, R. Beniwal, and A. Kumar, "Advancements in personality detection: unleashing the power of transformer-based models and deep learning with static embeddings on English personality quotes", International Journal of All Research Education & Scientific Methods, 2024, vol. 12, pp. 2235-2251. https://doi.org/10.56025/IJARESM.2023.1201242235.

[11] A. Naz, H. U. Khan, A. Bukhari, B. Alshemaimri, A. Daud, and M. Ramzan, "Machine and deep learning for personality traits detection: a comprehensive survey and open research challenges", Artificial Intelligence Review, 2025, vol. 58, p. 239. https://doi.org/10.1007/s10462-025-11245-3.

[12] Y. O. Sharrab, H. Attar, M. A. H. Eljinini, Y. Al-Omary, and W. a. Al-Momani, "Advancements in Speech Recognition: A Systematic Review of Deep Learning Transformer Models, Trends, Innovations, and Future Directions", IEEE Access, 2025, vol. 13, pp. 46925-46940. https://doi.org/10.1109/ACCESS.2025.3550855.

[13] K.-M. Shum, M. Ptaszynski, and F. Masui, "Big Five Personality Trait Prediction Based on User Comments", Information, 2025, vol. 16, p. 418. https://doi.org/10.3390/info16050418.

[14] E. F. Tsani and D. Suhartono, "Personality identification from social media using ensemble BERT and RoBERTa", Informatica, 2023, vol. 47, pp. 537-544. https://doi.org/10.31449/inf.v47i4.4771.

[15] A. Naz, H. U. Khan, T. Alsahfi, M. Alhajlah, B. Alshemaimri, and A. Daud, "Using transformers and Bi-LSTM with sentence embeddings for prediction of openness human personality trait", PeerJ Computer Science, 2025, vol. 11, p. 38. https://doi.org/10.7717/peerj-cs.2781.

[16] A. Guo, R. Hirai, A. Ohashi, Y. Chiba, Y. Tsunomori, and R. Higashinaka, "Personality prediction from task-oriented and open-domain human–machine dialogues", Scientific Reports, 2024, vol. 14, p. 3868. https://doi.org/10.1038/s41598-024-53989-y.

[17] M. A. Akber, T. Ferdousi, R. Ahmed, R. Asfara, R. Rab, and U. Zakia, "Personality and emotion—A comprehensive analysis using contextual text embeddings", Natural Language Processing Journal, 2024, vol. 9, p. 100105. https://doi.org/10.1016/j.nlp.2024.100105.

[18] H. Bousselham and A. Mourhir, "Fine-tuning GPT on biomedical NLP tasks: an empirical evaluation", in 2024 International Conference on Computer, Electrical & Communication Engineering (ICCECE), 2024: IEEE, pp. 1-6. https://doi.org/10.1109/ICCECE58645.2024.10497313.

[19] H. Xu et al., "Temporal Shift for Personality Recognition with Pre-Trained Representations", in 2024 IEEE 14th International Symposium on Chinese Spoken Language Processing (ISCSLP), 2024: IEEE, pp. 446-450. https://doi.org/10.1109/ISCSLP63861.2024.10799950.

[20] T. Agrawal, D. Agarwal, M. Balazia, N. Sinha, and F. Bremond, "Multimodal personality recognition using cross-attention transformer and behaviour encoding", arXiv

preprint arXiv:2112.12180, 2021. https://doi.org/10.48550/arXiv.2112.12180.

[21] G. B. Mohan, R. P. Kumar, R. E, and S. Gorantla, "Enhancing Personality Classification through Textual Analysis: A Deep Learning Approach Utilizing MBTI and Social Media Data", in 2023 International Conference on Network, Multimedia and Information Technology (NMITCON), 2023: IEEE, pp. 01-06. https://doi.org/10.1109/NMITCON58196.2023.10276193.

[22] C. Yuan, J. Wu, H. Li, and L. Wang, "Personality recognition based on user generated content", in 15th International Conference on Service Systems and Service Management (ICSSSM), 2018: IEEE, pp. 1-6. https://doi.org/10.1109/ICSSSM.2018.8465006.

[23] C. Molnar, G. Casalicchio, and B. Bischl, "Interpretable machine learning–a brief history, state-of-the-art and challenges", in Joint European conference on machine learning and knowledge discovery in databases, 2020: Springer, pp. 417-431. https://doi.org/10.1007/978-3-030-65965-3_28.

[24] A. R. Sajun, I. Zualkernan, and D. Sankalpa, "A historical survey of advances in transformer architectures", Applied Sciences, 2024, vol. 14, p. 4316. https://doi.org/10.3390/app14104316.

[25] P. T. Costa Jr and R. R. McCrae, "The five-factor model of personality and its relevance to personality disorders", Journal of personality disorders, 1992, vol. 6, pp. 343-359. https://doi.org/10.1521/pedi.1992.6.4.343.

[26] M. Fatahian and R. Ravanmehr, "Personality Recognition in Social Media using Sentence Embeddings Based on Transformer Networks", SN Computer Science, 2025, vol. 6, pp. 1-22. https://doi.org/10.1007/s42979-025-04326-1.

[27] E. Kerz, Y. Qiao, S. Zanwar, and D. Wiechmann, "Pushing on personality detection from verbal behavior: A transformer meets text contours of psycholinguistic features", arXiv preprint arXiv:2204.04629, 2022. https://doi.org/10.48550/arXiv.2204.04629.

[28] F. Elourajini and E. Aïmeur, "AWS-EP: a multi-task prediction approach for MBTI/Big5 Personality Tests", in 2022 IEEE International Conference on Data Mining Workshops (ICDMW), 2022: IEEE, pp. 1-8. https://doi.org/10.1109/ICDMW58026.2022.00049.

[29] Y. Ji, W. Wu, H. Zheng, Y. Hu, X. Chen, and L. He, "Is chatgpt a good personality recognizer? a preliminary study", arXiv preprint arXiv:2307.03952, 2023. https://doi.org/10.48550/arXiv.2307.03952.

[30] M. Sobhi and A. Mazochi, "A Comparative Study of BERT-X for Sentiment Analysis and Stance Detection in Persian Social Media", International Journal of Information & Communication Technology Research, 2024, vol. 16, pp. 9-18. https://doi.org/10.61186/itrc.16.3.9.

[31] S. Leonardi, D. Monti, G. Rizzo, and M. Morisio, "Multilingual transformer-based personality traits estimation" Information 2020, vol. 11, p. 179. https://doi.org/10.3390/info11040179.

[32] Hasan, K., Saquer, J. and Ghosh, M., "Advancing mental disorder detection: A comparative evaluation of transformer and lstm architectures on social media", in 2025 IEEE 49th Annual Computers, Software, and Applications Conference (COMPSAC), 2025: IEEE, pp. 193-202. https://doi.org/10.1109/COMPSAC65507.2025.00033.

[33] C. M. Greco, A. Simeri, A. Tagarelli, and E. Zumpano, "Transformer-based language models for mental health issues: a survey", Pattern Recognition Letters, 2023, vol. 167, pp. 204-211. https://doi.org/10.1016/j.patrec.2023.02.016.

[34] N. Gholinejad and M. H. Chehreghani, "Heterophily-aware fair recommendation using graph convolutional networks", arXiv preprint arXiv:2402.03365, 2024. https://doi.org/10.48550/arXiv.2402.03365.

[35] S. Dhelim, N. Aung, M. A. Bouras, H. Ning, and E. Cambria, "A survey on personality-aware recommendation systems", Artificial Intelligence Review, 2022, vol. 55, pp. 2409-2454. https://doi.org/10.1007/s10462-021-10063-7.

[36] D. Fernau, S. Hillmann, N. Feldhus, T. Polzehl, and S. Möller, "Towards personality-aware chatbots", in Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue, 2022, pp. 135-145. https://doi.org/10.18653/v1/2022.sigdial-1.15.

[37] S. Garg, S. Sinha, A. K. Kar, and M. Mani, "A review of machine learning applications in human resource management", International Journal of Productivity and Performance Management, 2022, vol. 71, pp. 1590-1610. https://doi.org/10.1108/IJPPM-08-2020-0427.

[38] D. K. Kothari and O. N. N. Fernando, "Enhancing human-computer interaction through ai: A study on chatgpt in educational environments", in 2024 IEEE Conference on Artificial Intelligence (CAI), 2024: IEEE, pp. 500-503. https://doi.org/10.1109/CAI59869.2024.00100.

[39] Z. Liu et al., "Bilingual Dialogue Dataset with Personality and Emotion Annotations for Personality Recognition in Education", Scientific Data, 2025, vol. 12, p.514. https://doi.org/10.1038/s41597-025-04836-w.

[40] D. Han, T. McInroe, A. Jelley, S. V. Albrecht, P. Bell, and A. Storkey, "Llm-personalize: Aligning llm planners with human preferences via reinforced self-training for housekeeping robots", arXiv preprint arXiv:2404.14285, 2024. https://doi.org/10.48550/arXiv.2404.14285.

[41] H. K. Jach, L. Bardach, and K. Murayama, "How personality matters for education research", Educational Psychology Review, 2023, vol. 35, p. 94. https://doi.org/10.1007/s10648-023-09807-4.

[42] J. Hui, C. W. Espinola, T. Rodak, and D. M. Blumberger, "Electroconvulsive therapy in patients with trauma and personality disorders: what is the evidence?", Expert Review of Neurotherapeutics, 2025, pp. 1-33. https://doi.org/10.1080/14737175.2025.2542759.

[43] L. V. Phan and J. F. Rauthmann, "Personality computing: New frontiers in personality assessment", Social and personality psychology compass, 2021, vol. 15, p. e12624, 2021. https://doi.org/10.1111/spc3.12624.

[44] M. Hashemi, A. Darejeh, and F. Cruz, "Understanding User Preferences in Explainable Artificial Intelligence: A Mapping Function Proposal", ACM Transactions on Intelligent Systems and Technology, 2025, vol. 16, pp. 1-37. https://doi.org/10.1145/3733837.

[45] J. W. Pennebaker and L. A. King, "Linguistic styles: language use as an individual difference", Journal of personality and social psychology, 1999, vol. 77, p. 1296. https://doi.org/10.1037//0022-3514.77.6.1296.

[46] J. Killian Jr and R. Sun, "Detecting big-5 personality dimensions from text based on large language models", in International Conference on Deep Learning Theory and Applications, 2024: Springer, pp. 264-278. https://doi.org/10.1007/978-3-031-66705-3_18.

[47] K. Clark, "Electra: Pre-training text encoders as discriminators rather than generators", arXiv preprint arXiv:2003.10555, 2020. https://doi.org/10.48550/arXiv.2003.10555.

[48] H. Perera and L. Costa, "Personality Classification of text through Machine learning and Deep learning: A Review (2023)", Authorea Preprints, 2023. https://doi.org/10.36227/techrxiv.22337746.v1.

[49] H. Bashiri and H. Naderi, "Comprehensive review and comparative analysis of transformer models in sentiment analysis", Knowledge and Information Systems, 2024, vol.

66, pp. 7305-7361. https://doi.org/10.1007/s10115-024-02214-3.

[50] A. Taheri, A. Zamanifar, and A. Farhadi, "Enhancing aspect-based sentiment analysis using data augmentation based on back-translation", International Journal of Data Science and Analytics, 2025, vol. 19, pp. 491-516. https://doi.org/10.1007/s41060-024-00622-w.

[51] H. Dai et al., "Auggpt: Leveraging chatgpt for text data augmentation", IEEE Transactions on Big Data, 2025, vol. 11, pp. 907-918. https://doi.org/10.1109/TBDATA.2025.3536934.

[52] M. A. Khan, M. S. Khan, I. Khan, S. Ahmad, and S. Huda, "Non functional requirements identification and classification using transfer learning model", IEEE Access, 2023, vol. 11, pp. 74997-75005. https://doi.org/10.1109/ACCESS.2023.3295238.

[53] R. Ravanmehr and R. Mohamadrezaei, "Deep learning overview", in Session-Based Recommender Systems Using Deep Learning, 2023: Springer, pp. 27-72. https://doi.org/10.1007/978-3-031-42559-2_2.

[54] M. Ramezani, M.-R. Feizi-Derakhshi, and M.-A. Balafar, "Text-based automatic personality prediction using KGrAt-Net: a knowledge graph attention network classifier", Scientific reports, 2022, vol. 12, p. 21453. https://doi.org/10.1038/s41598-022-25955-z.

[55] M. Ramezani, M.-R. Feizi-Derakhshi, and M.-A. Balafar, "Knowledge Graph-Enabled Text-Based Automatic Personality Prediction", Computational intelligence and neuroscience, 2022, vol. 2022, p. 3732351. https://doi.org/10.1155/2022/3732351.

[56] Z. Wang, C.-H. Wu, Q.-B. Li, B. Yan, and K.-F. Zheng, "Encoding text information with graph convolutional networks for personality recognition", Applied sciences, 2020, vol. 10, p. 4081. https://doi.org/10.3390/app10124081.

[57] M. Ramezani et al., "Automatic personality prediction: an enhanced method using ensemble modeling", Neural Computing and Applications, 2022, vol. 34, pp. 18369-18389. https://doi.org/10.1007/s00521-022-07444-6.

[58] X. Xue, J. Feng, and X. Sun, "Semantic-enhanced sequential modeling for personality trait recognition from texts", Applied Intelligence, 2021, vol. 51, pp. 7705-7717. https://doi.org/10.1007/s10489-021-02277-7.

Hossein Saberi is an Artificial Intelligence (AI) Researcher at Shatel Company. He has a B.Sc. in Petroleum Engineering from Hakim Sabzevari University and is pursuing an M.Sc. in AI. With four years of experience in AI applications spanning petroleum engineering, medical diagnostics, and computer vision, he has presented at national conferences and published articles on predictive analytics, focusing on solutions for industry challenges.

Reza Ravanmehr graduated in computer engineering from Shahid Beheshti University, Tehran, in 1996. After that, he gained his M.Sc. and Ph.D. degrees, both in computer engineering, from IAU, Science and Research Branch, Tehran, in 1999 and 2004, respectively. His main research interests are social network analysis, distributed/parallel systems, and large-scale data management systems. He is currently an Associate Professor in the Department of Computer Engineering at the Central Tehran Branch, IAU.