Accredited SINTA 2 Ranking
Decree of the Director General of Higher Education, Research, and Technology, No. 158/E/KPT/2021
Validity period from Volume 5 Number 2 of 2021 to Volume 10 Number 1 of 2026

Published online at: **http://jurnal.iaii.or.id**

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 9 No. 1 (2025) 20 - 30 e-ISSN: 2580-0760

Evaluating Transformer Models for Social Media Text-Based Personality Profiling

Anggit Dwi Hartanto[1]*, Ema Utami[2], Arief Setyanto[3], Kusrini[4]
[1]Faculty of Computer Science, Universitas Amikom Yogyakarta, Yogyakarta, Indonesia
[2,3,4]Doctor of Informatics, Universitas Amikom Yogyakarta, Yogyakarta, Indonesia
[1]anggit@amikom.ac.id, [2]ema.u@amikom.ac.id, [3]arief_s@amikom.ac.id, [4]kusrini@amikom.ac.id

Abstract

This research aims to evaluate the performance of various Transformer models in social media-based classification tasks, specifically focusing on applications in personality profiling. With the growing interest in leveraging social media as a data source for understanding individual personality traits, selecting an appropriate model becomes crucial for enhancing accuracy and efficiency in large-scale data processing. Accurate personality profiling can provide valuable insights for applications in psychology, marketing, and personalized recommendations. In this context, models such as BERT, RoBERTa, DistilBERT, TinyBERT, MobileBERT, and ALBERT are utilized in this study to understand their performance differences under varying configurations and dataset conditions, assessing their suitability for nuanced personality profiling tasks. The research methodology involves four experimental scenarios with a structured process that includes data acquisition, preprocessing, tokenization, model fine-tuning, and evaluation. In Scenarios 1 and 2, a full dataset of 9,920 data points was used with standard fine-tuning parameters for all models. In contrast, ALBERT in Scenario 2 was optimized using customized batch size, learning rate, and weight decay. Scenarios 3 and 4 used 30% of the total dataset, with additional adjustments for ALBERT to examine its performance under specific conditions. Each scenario is designed to test model robustness against variations in parameters and dataset size. The experimental results underscore the importance of tailoring fine-tuning parameters to optimize model performance, particularly for parameter-efficient models like ALBERT. ALBERT and MobileBERT demonstrated strong performance across conditions, excelling in scenarios requiring accuracy and efficiency. BERT proved to be a robust and reliable choice, maintaining high performance even with reduced data, while RoBERTa and DistilBERT may require further adjustments to adapt to data-limited conditions. Although efficient, TinyBERT may fall short on tasks demanding high accuracy due to its limited representational capacity. Selecting the right model requires balancing computational efficiency, task-specific requirements, and data complexity.

Keywords: *Profiling analysis; Transformer, BERT Variants;*

How to Cite: A. Hartanto, Ema Utami, Arief Setyanto, and Kusrini, "Evaluating Transformer Models for Social Media Text-Based Personality Profiling", *J. RESTI (Rekayasa Sist. Teknol. Inf.)*, vol. 9, no. 1, pp. 20 - 30, Jan. 2025
DOI: https://doi.org/10.29207/resti.v9i1.6157

1. Introduction

Personality profiling, a psychological process aimed at understanding individual characteristics and behavioral tendencies, has become increasingly important in various applications, from psychological assessment to targeted marketing and human resource management [1]. Personality traits were traditionally assessed using psychometric instruments, such as the Big Five Inventory, where individuals self-reported responses to structured questionnaires [2], [3], [4]. However, these conventional methods can be time-consuming and prone to bias, as respondents may provide socially desirable answers rather than authentic ones.

With the rise of social media, a new paradigm has emerged: personality profiling based on linguistic analysis of online user behavior. Platforms such as Facebook [5], Twitter (now X) [6], and Instagram [7] generate vast amounts of textual data that reflect individual opinions, emotions, and personality traits [8]. Analyzing this data has the potential to provide a more dynamic and real-time understanding of users' personalities, making personality profiling scalable and applicable in new contexts such as social media

Received: 31-10-2024 | Accepted: 23-12-2024 | Published Online: 21-01-2025

marketing, recommender systems, and even mental health diagnostics [9].

Personality prediction has seen significant advancements in recent years, particularly in integrating deep learning models and natural language processing techniques [10]. The use of social media data for personality profiling has become increasingly popular due to the vast amount of user-generated content available online [11]. Studies have explored various machine learning and deep learning approaches to predict personality traits, often utilizing the Big Five personality model as a framework [12]. For instance, research has demonstrated the effectiveness of using BERT and other transformer-based models for extracting semantic features from text data, which are crucial for accurate personality prediction [13].

Natural Language Processing (NLP) techniques have rapidly evolved, enabling the extraction of meaningful insights from large volumes of unstructured text data [14]. Transformer models revolutionized NLP by allowing models to handle dependencies between words across long sequences more effectively than earlier models like RNNs or CNNs [15]. With their self-attention mechanisms, transformers have become the backbone of state-of-the-art language models, enabling significant breakthroughs in various NLP tasks, including text classification, sentiment analysis, and question-answering [15].

Transformer architectures, such as BERT (Bidirectional Encoder Representations from Transformers), have been particularly influential in advancing the capabilities of personality prediction models [16]. These models leverage attention mechanisms to capture contextual information from text, allowing for a more nuanced understanding of language. Integrating pre-trained language models like BERT with additional layers for fine-tuning has shown promising results in various personality prediction tasks [17]. For example, studies have highlighted the success of ensemble models that combine BERT with other architectures, such as ULMFiT, to enhance prediction accuracy and F1 scores [18].

Among the most prominent Transformer-based models besides BERT are RoBERTa (Robustly Optimized BERT Pretraining Approach) [19] and ALBERT (A Lite BERT), all of which have demonstrated superior performance across many benchmarks. These models excel at understanding context and semantics by processing input bidirectionally, capturing richer representations of language [16]. However, the wide variety of Transformer variants available, each optimized for different choices in acceleration, size, and accuracy, presents challenges when selecting the most suitable version for particular assignments like personality profiling. While BERT and RoBERTa focus on maximizing accuracy, smaller models like DistilBERT, TinyBERT, and MobileBERT prioritize computation efficiency, rendering them appealing choices for implementation in resource-limited settings [20].

Moreover, the application of transformer models in personality profiling is not limited to English text. Research has extended these methodologies to other languages, such as Indonesian and Turkish, demonstrating the versatility and adaptability of transformer architectures across different linguistic contexts [1], [21]. The development of language-specific models, like IndoBERT [22], has further improved the accuracy of personality predictions by tailoring the pre-training process to the linguistic characteristics of the target language. This highlights the importance of considering language diversity in designing and implementing personality prediction models [23].

In addition to language considerations, recent studies have addressed data imbalance and feature optimization challenges. Techniques such as synthetic minority oversampling (SMOTE) [24] and particle swarm optimization has enhanced model performance by balancing datasets and optimizing feature selection. These advancements underscore the potential of transformer architectures in providing robust and interpretable models for social media-based personality profiling, paving the way for future research to explore more sophisticated and context-aware approaches in this domain [25], [26].

Despite the promise of Transformer models, there is still a need for a comprehensive evaluation of their effectiveness in personality profiling tasks, especially when using real-world datasets like MyPersonality, which contains social media posts labelled with personality traits derived from the Big Five personality framework [27]. MyPersonality offers a unique dataset for this research, bridging the gap between traditional psychometrics and modern NLP techniques by allowing machine learning models to determine personality traits from social media behavior [5].

This paper aims to evaluate six Transformer-based models, namely BERT [16], RoBERTa [19], ALBERT [28], DistilBERT [20], TinyBERT [29], and MobileBERT [30] on the MyPersonality dataset. By applying a consistent fine-tuning strategy across these models, we systematically compare their performance in personality profiling, measuring key metrics such as classification accuracy, precision, recall, and F1 score. Furthermore, this study explores the trade-off between model accuracy and computational efficiency, considering the practical implications of deploying these models in real-world applications where resources may be limited.

Furthermore, we extend our analysis by modifying the fine-tuning strategy for ALBERT to investigate whether specialized fine-tuning can compensate for this model's smaller size and efficiency, thereby providing insights into how model architecture and training strategy influence performance in personality profiling.

This research contributes to developing more efficient and accurate Transformer models for social media-based personality profiling, offering guidance for researchers and practitioners on selecting and optimizing models based on specific task requirements.

2. Research Methods

This study evaluates the performance of six Transformer-based models, BERT, RoBERTa, ALBERT, DistilBERT, TinyBERT, and MobileBERT, on personality profiling tasks using the MyPersonality dataset. The dataset contains user-generated content from social media and labels based on the Big Five Personality Traits framework, making it suitable for text-based personality profiling.

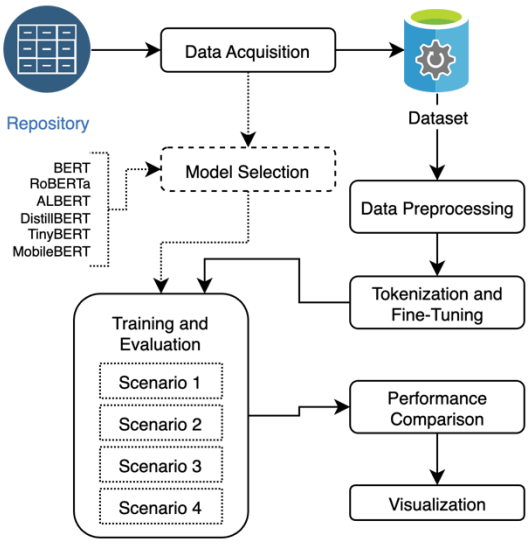

Figure 1. Research Flow

The goal is to analyze and compare these models under different fine-tuning strategies and dataset variations, as outlined in Figure 1.

2.1 Dataset

The MyPersonality dataset contains 9,920 raw data samples, including social media posts and personality labels. In this study, we use two different versions of the dataset. Scenario 1 and Scenario 2: The whole dataset of 9,920 raw samples. Scenario 3 and Scenario 4: A subset of the dataset consisting of 1,000 samples. Each sample in the dataset consists of user-generated content labelled with personality traits derived from the Big Five personality model. Nonrelevant columns such as sentiment scores and metadata are removed to process the data for classification tasks, and the text features are consolidated for input into the models.

The bar chart in Figure 2 illustrates the distribution of five categorical personality traits: Extraversion (cEXT), Neuroticism (cNEU), Agreeableness (cAGR), Conscientiousness (cCON), and Openness (cOPN). For each trait, the data is divided into two categories: "Yes" (indicating a high level of the trait) and "No" (indicating a low level). Openness (cOPN) has the highest "Yes" responses, with 7,370 instances, significantly higher than the "No" responses (2,547). In contrast, Neuroticism (cNEU) shows more "No" responses (6,200) compared to "Yes" (3,717). Extraversion (cEXT) and Conscientiousness (cCON) exhibit a relatively balanced distribution between "Yes" and "No" categories. Lastly, Agreeableness (cAGR) has slightly more "Yes" responses (5,268) than "No" (4,649). This chart provides a clear visual representation of how these traits are distributed across the dataset.

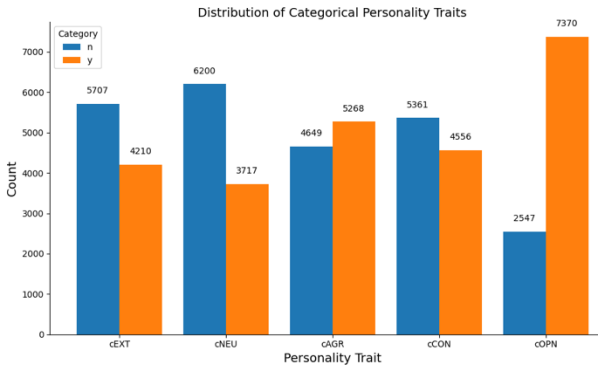

Figure 2. Dataset Distribution

2.2 Model Selection

The six Transformer models selected for comparison are BERT (bert-base-uncased), RoBERTa (roberta-base), ALBERT (albert-base-v2), DistilBERT (distilbert-base-uncased), TinyBERT (huawei-noah/TinyBERT_General_6L_768D), MobileBERT (google/mobilebert-uncased). These models represent a variety of architectures, ranging from standard models like BERT and RoBERTa to lightweight models such as TinyBERT and MobileBERT. While BERT, RoBERTa, and ALBERT are full-size models emphasizing accuracy, the smaller models (DistilBERT, TinyBERT, and MobileBERT) are optimized for speed and efficiency.

2.3 Data Preprocessing

The preprocessing pipeline consists of several key steps:

Removing Non-Relevant Features: Columns such as sentiment data and metadata (sentiment, #AUTHID, and DATE) were excluded from the dataset. The remaining columns were consolidated into a single text input field for each sample.

Label Encoding: The #AUTHID column, which represents user IDs, was used as the target variable (personality label). This column was encoded into numeric values using a LabelEncoder.

Splitting the Dataset: The dataset was divided between training and testing sets in an 80-20 ratio. This division was uniform across all cases to guarantee similar outcomes.

Dataset Conversion: The training and testing sets were converted into the Hugging Face Dataset format to facilitate tokenization and model training.

2.4 Tokenization and Fine-Tuning

Each model was tokenized using its corresponding tokenizer. Tokenization was done with padding and truncation to ensure uniform input length across samples (max sequence length: 128). The tokenizers used for each model are BERT (BertTokenizer), RoBERTa (RobertaTokenizer), ALBERT (AlbertTokenizer), DistilBERT (DistilBertTokenizer), TinyBERT (AutoTokenizer), MobileBERT (MobileBertTokenizer).

For each model, the tokenized datasets were fine-tuned using a consistent fine-tuning strategy across all scenarios, except for ALBERT in Scenario 2 and Scenario 4, where a customized fine-tuning approach was used to explore model-specific optimization.

2.5 Training and Evaluation

This experiment employs four distinct scenarios to fine-tune various transformer models for text classification. The models used include BERT, RoBERTa, DistilBERT, TinyBERT, MobileBERT, and ALBERT, with different fine-tuning strategies applied across the scenarios. Each scenario explores variations in hyperparameters such as training epochs, learning rates, batch sizes, and weight decay, with particular attention given to ALBERT, which is fine-tuned separately in certain cases. Additionally, the dataset size differs between the scenarios, with Scenarios 1 and 2 using a full dataset of 9,920 data points, while Scenarios 3 and 4 operate on a reduced dataset size (30% of the total). This variation is intentionally designed to simulate real-world conditions in personality profiling, where available datasets are often limited in size due to privacy concerns and the challenge of collecting extensive, high-quality data. By evaluating model performance on a smaller dataset, the study aims to understand the impact of fine-tuning strategies and assess each model's robustness in handling data constraints commonly encountered in practical personality profiling applications. The defined experimental setups for these scenarios are detailed in Tables 1 and 2.

In this research, we evaluate the performance of various Transformer models for personality profiling using metrics beyond simple accuracy. Given the multi-class nature of the task, metrics such as precision, recall, and F1-score provide a more complete picture of model effectiveness. Precision measures how accurate the positive predictions are, while recall captures the model's ability to identify all relevant positive cases. The F1-score balances precision and recall, offering a useful metric when both false positives and false negatives are important. These metrics help assess the trade-offs between model performance and efficiency in social media-based personality profiling. The metrics were computed using Formulas 1- 5.

Table 1. Scenario 1 and 2

Aspect	Scenario 1	Scenario 2
Number of models for fine-tuning	6 models (BERT, RoBERTa, DistilBERT, TinyBERT, MobileBERT, ALBERT)	5 models (BERT, RoBERTa, DistilBERT, TinyBERT, MobileBERT), ALBERT handled separately
ALBERT Fine-Tuning	Same fine-tuning process as other models	Specialized fine-tuning with different batch sizes, epochs, and weight decay
Training Arguments	Standard across all models	Standard for 5 models, custom training arguments for ALBERT
Training Epochs	10 epochs for all models	10 epochs for 5 models, 20 epochs for ALBERT
Learning Rate	Defaults to transformers library's default	Standard models: 2e-5, ALBERT: 2e-5
Batch Size	Standard (8 for training, 16 for eval)	Standard (8 for training, 16 for eval), ALBERT: 16 (training), 32 (eval)
Weight Decay	0.01 for all models	0.01 for standard models, ALBERT: 0.3
Dataset Size	9920 data points	9920 data points

Table 2. Scenario 3 and 4

Aspect	Scenario 3	Scenario 4
Number of models for fine-tuning	6 models (BERT, RoBERTa, DistilBERT, TinyBERT, MobileBERT, ALBERT)	5 models (BERT, RoBERTa, DistilBERT, TinyBERT, MobileBERT), ALBERT handled separately
ALBERT Fine-Tuning	Same fine-tuning process as other models	Specialized fine-tuning with different batch size, epochs, weight decay, and learning rate
Training Arguments	Standard across all models	Custom training arguments for ALBERT (batch size, learning rate, weight decay)
Training Epochs	3 epochs for all models	3 epochs for 5 models, 6 epochs for ALBERT
Learning Rate	Defaults to transformers library's default	Standard models: 2e-5, ALBERT: 2e-5 (custom)
Batch Size	Standard (8 for training, 16 for eval)	Standard (8 for training, 16 for eval), ALBERT: 16 (training), 32 (eval)
Weight Decay	0.01 for all models	0.01 for standard models, ALBERT: 0.03
Dataset Size	30% of total dataset	30% of total dataset

$$Loss = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{ic} \cdot \log(\widehat{y_{ic}}) \quad (1)$$

N represent the total number of samples in the dataset and C the number of classes. For a sample i, the true label for the class c is denoted as $\widehat{y_{ic}}$. Meanwhile, the model generates a predicted probability for the sample i belonging to the class c, represented as $\widehat{y_{ic}}$.

$$\text{Accuracy: } \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

TP (True Positives) refers to the count of correct predictions where the actual class is positive. TN (True Negatives) represents the count of accurate predictions where the actual class is negative. FP (False Positives) indicates the number of incorrect predictions where the actual class is negative, but the model predicted it as positive. FN (False Negatives) refers to the number of incorrect predictions where the actual class is positive, but the model predicted it as negative.

$$\text{Precision: } \frac{TP}{TP+FP} \quad (3)$$

TP (True Positives) represents the count of accurate predictions where the true class is positive. FP (False Positives) refers to the count of incorrect predictions where the actual class is negative, but the model classified it as positive.

$$\text{Recall: } \frac{TP}{TP+FN} \quad (4)$$

TP (True Positives) is the count of valid predictions where the real class is positive. FP (False Positives) is the number of cases where the model mistakenly predicted negative for instances that are positive.

$$\text{F1-score: } 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (5)$$

Precision refers to the ratio of correct positive predictions to the total number of positive predictions made by the model. The recall represents the ratio of correct positive predictions to the total number of positive instances.

2.6 Performance Comparison

We follow a consistent evaluation process to compare the performance of the six models: BERT, RoBERTa, ALBERT, DistilBERT, TinyBERT, and MobileBERT. First, all models are fine-tuned using the same training configuration on the MyPersonality dataset. After fine-tuning, the performance of each model is evaluated using essential metrics: accuracy, precision, recall, and F1-score. These metrics are calculated on a reserved test set to assess the generalization capability of each model.

Finally, the results are visualized and compared to identify trade-offs between performance and computational efficiency across the models. This systematic comparison helps determine the most effective model for personality profiling. Additionally, a comparison of computational efficiency, such as model size and inference time, was made to highlight the trade-offs between model performance and resource constraints.

2.7 Visualization

To visualize the performance of the six models, we first create a results table summarizing each model's key metrics: accuracy, precision, recall, and F1-score. This table provides a clear comparison of how each model performs on various aspects of classification. Next, we generate bar charts to visualize evaluation loss and accuracy, allowing for an easy comparison of model performance at a glance.

These charts help highlight the differences between models, making it easier to spot trade-offs between accuracy and loss and offering a more intuitive understanding of the results.

3. Results and Discussions

3.1 Result

In Scenario 1, we evaluated the performance of six Transformer-based models, BERT, RoBERTa, ALBERT, DistilBERT, TinyBERT, and MobileBERT, on the MyPersonality dataset, consisting of 9,920 raw samples. All models were fine-tuned using the same configuration across 10 epochs. Table 3 shows the key findings for Scenario 1, focusing on accuracy, loss, precision, recall, and F1-score.

Table 3. Accuracy and Loss (Scenario 1)

Models	Accuracy	Loss
BERT	0.9975	0.0239
RoBERTa	0.9975	0.0202
DistilBERT	0.9975	0.0393
TinyBERT	0.9960	0.0383
MobileBERT	0.9975	0.0366
ALBERT	0.0166	5.0224

In the experimental results for Scenario 1, the performance of six Transformer models, BERT, RoBERTa, DistilBERT, TinyBERT, MobileBERT, and ALBERT, was assessed in terms of accuracy and loss, as shown in Table 3. The findings reveal that BERT, RoBERTa, and MobileBERT achieved the highest accuracy at 0.9975, indicating near-perfect classification performance. DistilBERT and TinyBERT followed closely, with accuracy scores of 0.9975 and 0.9960, respectively, demonstrating their effectiveness despite being smaller and more efficient versions of the original BERT model. ALBERT, however, displayed a significantly lower accuracy of 0.0166, accompanied by a high loss value of 5.0224, suggesting that it may not be suitable for this specific task under the conditions of Scenario 1. The variations in loss values further emphasize the stability of RoBERTa, which achieved the lowest loss (0.0202), indicating a minimal prediction error. In contrast, ALBERT's high loss suggests convergence or model compatibility issues with the dataset used in this scenario. Overall, these results highlight the strong performance of BERT, RoBERTa, and MobileBERT while underscoring potential limitations in ALBERT for this particular experimental setup.

Table 4. Precision, Recall, and F1-Score (Scenario 1)

Models	Precision	Recall	F1-Score
BERT	0.9964	0.9964	0.9964
RoBERTa	0.9959	0.9959	0.9959
DistilBERT	0.9961	0.9961	0.9961
TinyBERT	0.9928	0.9928	0.9928
MobileBERT	0.9957	0.9957	0.9957
ALBERT	0.0003	0.0003	0.0003

The precision, recall, and F1-score evaluation metrics provide insights into the performance consistency and reliability of various Transformer models, as shown in Table 4. BERT achieved the highest scores across all metrics (0.9964), demonstrating an exceptional ability to balance between precision and recall, resulting in high accuracy for both true positive and true negative classifications. This high performance from BERT suggests its robustness and reliability in handling the dataset effectively. DistilBERT and RoBERTa followed closely, with scores of 0.9961 and 0.9959, respectively, further underscoring their strong performance and making them suitable alternatives to the full BERT model, especially for scenarios where computational efficiency is a concern. TinyBERT and MobileBERT also showed commendable performance with F1-scores of 0.9928 and 0.9957, respectively, highlighting their effectiveness as compact and resource-efficient versions without significant loss in predictive quality. However, ALBERT presented an entirely different case, with extremely low scores across precision, recall, and F1-score (0.0003), which points to substantial challenges in classification accuracy within this specific scenario. This discrepancy in ALBERT's performance suggests that it may not be compatible with the dataset characteristics or requires further fine-tuning for this task. Overall, the results from Scenario 1 underscore the reliability of BERT, DistilBERT, and RoBERTa for high-precision tasks while revealing ALBERT's potential limitations in similar experimental setups, highlighting the importance of model selection based on task requirements.

In Scenario 2, the accuracy and loss metrics results for various Transformer models reveal their effectiveness in handling the experimental task, as illustrated in Table 5. RoBERTa and ALBERT achieved the highest accuracy, reaching 0.99748, indicating their strong predictive capability and reliable performance. Notably, ALBERT also recorded the lowest loss value at 0.01863, suggesting high efficiency and minimal error in its predictions for this scenario. RoBERTa followed closely with a loss of 0.030066, reinforcing its stability and effectiveness in handling the dataset.

Table 5. Accuracy and Loss (Scenario 2)

Models	Accuracy	Loss
BERT	0.996472	0.052605
RoBERTa	0.99748	0.030066
DistilBERT	0.996976	0.044164
TinyBERT	0.988911	0.178462
MobileBERT	0.996976	0.040961
ALBERT	0.99748	0.01863

Other models, including DistilBERT, MobileBERT, and BERT, also showed commendable accuracy, with scores of 0.996976, 0.996976, and 0.996472, respectively. This consistency across models highlights the robustness of the Transformer architecture in accurately classifying data. However, these models exhibited slightly higher loss values than ALBERT and RoBERTa, with BERT recording a loss of 0.052605 and DistilBERT at 0.044164. TinyBERT, while still achieving a respectable accuracy of 0.988911, showed the highest loss value (0.178462) among all models, which may indicate a greater tendency for error in this particular scenario, possibly due to its simplified architecture designed for efficiency.

Scenario 2 demonstrates the strong performance of RoBERTa and ALBERT in achieving high accuracy with minimal loss, while BERT, DistilBERT, and MobileBERT also deliver reliable results. TinyBERT, despite being efficient, might require additional tuning to improve its performance further. These findings underscore the importance of selecting a model that balances accuracy and computational efficiency based on specific task requirements.

The evaluation of precision, recall, and F1-score metrics for the various models, as shown in Table 6, highlights the overall effectiveness and nuanced differences in performance among the models. RoBERTa and ALBERT achieved the highest scores across all metrics, with both models attaining a precision of 0.995766, recall of 0.99748, and F1-scores of 0.996554 and 0.996474, respectively. This high level of performance indicates that RoBERTa and ALBERT are particularly well-suited to this task, as they maintain a strong balance between precision and recall, effectively capturing both true positives and minimizing false positives.

Table 6. Precision, Recall, and F1-Score (Scenario 2)

Models	Precision	Recall	F1-Score
BERT	0.994007	0.996472	0.99507
RoBERTa	0.995766	0.99748	0.996554
DistilBERT	0.994391	0.996976	0.995585
TinyBERT	0.980382	0.988911	0.98413
MobileBERT	0.99532	0.996976	0.99602
ALBERT	0.995766	0.99748	0.996474

BERT and DistilBERT also performed well, with BERT achieving a precision of 0.994007, recall of 0.996472, and F1-score of 0.99507, while DistilBERT recorded a precision of 0.994391, recall of 0.996976, and F1-score of 0.995585. These scores underscore their robustness and efficiency in classification, though they slightly trail RoBERTa and ALBERT in terms of balance between precision and recall. MobileBERT also demonstrated high precision and recall, with values of 0.99532 and 0.996976, respectively, resulting in an F1-score of 0.99602. This consistency highlights MobileBERT’s reliability in maintaining quality predictions. TinyBERT, while showing slightly lower metrics with a precision of 0.980382, recall of 0.988911, and F1-score of 0.98413, still achieved satisfactory performance, particularly for a smaller model designed for efficiency.

Overall, the metrics in Table 6 emphasize the robust performance of RoBERTa, ALBERT, and MobileBERT, with high precision, recall, and F1-scores, makes them suitable for tasks demanding high accuracy. Despite slightly lower scores, BERT and DistilBERT remain strong contenders, while TinyBERT offers a reasonable trade-off between

efficiency and predictive quality. These findings underscore the significance of choosing a model that aligns well with the specific requirements of the task in terms of both performance and computational efficiency.

In Scenario 3, the performance of different Transformer models in terms of accuracy and loss, as presented in Table 7, reveals substantial variations. ALBERT achieved the highest accuracy (0.996472) with a low loss value of 0.050416, indicating its strong capability to effectively classify data with minimal error in this scenario. BERT also demonstrated high accuracy (0.99244) but recorded a higher loss (0.230323), suggesting some trade-off between accuracy and error in its predictions.

Table 7. Accuracy and Loss (Scenario 3)

Models	Accuracy	Loss
BERT	0.99244	0.230323
RoBERTa	0.019657	5.018772
DistilBERT	0.019657	5.018816
TinyBERT	0.989919	0.079766
MobileBERT	0.534274	2.614055
ALBERT	0.996472	0.050416

TinyBERT performed well with an accuracy of 0.989919 and a relatively low loss of 0.079766, confirming its efficiency as a lightweight model with solid predictive capabilities. However, MobileBERT displayed moderate accuracy at 0.534274 and a higher loss value of 2.614055, which may reflect challenges in maintaining stability and precision for this particular task. In contrast, RoBERTa and DistilBERT experienced significant difficulties, recording a low accuracy of 0.019657 and high loss values exceeding 5.0. These results suggest that RoBERTa and DistilBERT may not be compatible with the data or conditions set in Scenario 3, as their high error rates indicate poor model fit or convergence issues. Table 7 highlights ALBERT as the most successful model in Scenario 3, closely followed by BERT and TinyBERT, which also performed reliably. MobileBERT's moderate results and the underperformance of RoBERTa and DistilBERT underscore the importance of model selection tailored to specific dataset characteristics and task requirements.

The precision, recall, and F1-score metrics for various models, as shown in Table 8, reveal significant performance differences. ALBERT achieved the highest scores across all three metrics, with a precision of 0.99376, a recall of 0.996472, and an F1-score of 0.995043, indicating its robust ability to classify data accurately and maintain a strong balance between precision and recall. BERT also demonstrated high performance with a precision of 0.987065, recall of 0.99244, and F1-score of 0.989373, reflecting its reliability in achieving accurate predictions in this scenario.

TinyBERT closely followed BERT, recording a precision of 0.982471, recall of 0.989919, and F1-score of 0.985764, highlighting its effectiveness as a compact model with commendable predictive power. However, MobileBERT exhibited a notably lower performance, with a precision of 0.427076, recall of 0.534274, and F1-score of 0.455623, indicating moderate success but a significant drop in classification quality compared to ALBERT and BERT. On the other hand, RoBERTa and DistilBERT struggled considerably, both achieving extremely low precision (0.000386), recall (0.019657), and F1-scores (0.000758). These poor results suggest that RoBERTa and DistilBERT faced substantial challenges in this scenario, likely due to misalignment with the dataset or task requirements, leading to inadequate model performance. In conclusion, Table 8 underscores ALBERT as the top-performing model in Scenario 3, with BERT and TinyBERT also showing strong results. MobileBERT's moderate performance and the significant underperformance of RoBERTa and DistilBERT highlight the importance of selecting a model compatible with the specific characteristics of the data and task.

Table 8. Precision, Recall, and F1-Score (Scenario 3)

Models	Precision	Recall	F1-Score
BERT	0.987065	0.99244	0.989373
RoBERTa	0.000386	0.019657	0.000758
DistilBERT	0.000386	0.019657	0.000758
TinyBERT	0.982471	0.989919	0.985764
MobileBERT	0.427076	0.534274	0.455623
ALBERT	0.99376	0.996472	0.995043

In Scenario 4, the performance evaluation in terms of accuracy and loss for different Transformer models, as shown in Table 9, highlights both the strengths and weaknesses of these models. ALBERT achieved the highest accuracy at 0.99496 with a low loss of 0.147728, indicating its capability to minimize error while achieving high classification performance effectively. MobileBERT also performed exceptionally well, with an accuracy of 0.990927 and an even lower loss of 0.084978, demonstrating its reliability and efficiency in this scenario.

Table 9. Accuracy and Loss (Scenario 4)

Models	Accuracy	Loss
BERT	0.972278	0.759298
RoBERTa	0.972278	0.504494
DistilBERT	0.970766	0.414988
TinyBERT	0.679435	2.435512
MobileBERT	0.990927	0.084978
ALBERT	0.99496	0.147728

BERT and RoBERTa showed identical accuracy scores of 0.972278, but RoBERTa outperformed BERT in loss, with values of 0.504494 for RoBERTa and 0.759298 for BERT. This indicates that while both models achieved similar accuracy, RoBERTa minimized the error more effectively, suggesting a slightly more stable performance. DistilBERT followed closely with an accuracy of 0.970766 and a loss of 0.414988, showcasing its ability to balance efficiency and predictive quality, although it slightly trails behind its larger counterparts. TinyBERT, however, displayed a significantly lower accuracy of 0.679435 and a high loss of 2.435512, indicating substantial limitations in

this scenario, likely due to its compact and simplified architecture, which might have struggled with the complexity of the dataset or task.

Table 9 illustrates that ALBERT and MobileBERT were the top performers in Scenario 4, combining high accuracy with low loss, making them suitable choices for tasks that require both efficiency and reliability. BERT, RoBERTa, and DistilBERT also delivered solid results, albeit with higher loss values. TinyBERT's lower performance emphasizes the trade-offs involved in using highly compact models, especially for tasks that may demand greater representational capacity.

The precision, recall, and F1-score results across various models, as shown in Table 10, provide a deeper insight into the classification performance. ALBERT achieved the highest scores across all metrics, with a precision of 0.991083, recall of 0.99496, and F1-score of 0.992801, underscoring its exceptional capability in identifying true positives and maintaining overall accuracy. MobileBERT closely followed, with a precision of 0.983963, recall of 0.990927, and an F1-score of 0.987027, demonstrating its effectiveness as a lightweight model still retaining high predictive quality.

Table 10. Precision, Recall, and F1-Score (Scenario 4)

Models	Precision	Recall	F1-Score
BERT	0.951065	0.972278	0.96038
RoBERTa	0.949396	0.972278	0.95981
DistilBERT	0.94679	0.970766	0.957813
TinyBERT	0.56119	0.679435	0.599095
MobileBERT	0.983963	0.990927	0.987027
ALBERT	0.991083	0.99496	0.992801

BERT, RoBERTa, and DistilBERT also performed well, with F1-scores of 0.96038, 0.95981, and 0.957813, respectively. While these models demonstrated strong recall (above 0.97), their slightly lower precision than ALBERT and MobileBERT suggests that they might generate more false positives, albeit minimally. This consistent performance across these models highlights their reliability in maintaining a balance between precision and recall, though with marginal differences. TinyBERT, however, showed a notable drop in performance, with a precision of 0.56119, recall of 0.679435, and an F1-score of 0.599095, indicating limitations in accurately capturing true positives and maintaining classification quality. This lower performance could be attributed to its compact structure, which may struggle to fully represent the complexity of the dataset or task requirements in this scenario. In conclusion, Table 10 highlights ALBERT and MobileBERT as the top performers in Scenario 4, excelling across precision, recall, and F1 scores, making them ideal for high accuracy and reliability tasks. BERT, RoBERTa, and DistilBERT also deliver strong, consistent results, while TinyBERT's reduced scores emphasize the trade-offs involved in model simplification, particularly when task complexity is high.

The experimental results in Table 11 and Figure 3 illustrate the performance of various transformer-based models BERT, RoBERTa, DistilBERT, TinyBERT, MobileBERT, and ALBERT across four scenarios. The metrics demonstrate the accuracy and robustness of each model in distinct experimental conditions. Notably, BERT and RoBERTa consistently exhibit high performance across scenarios, with RoBERTa showing a significant drop in Scenario 3. TinyBERT and MobileBERT provide competitive results that balance efficiency and performance, although TinyBERT's accuracy declines in Scenario 4. ALBERT, while performing well in Scenarios 2 to 4, shows a lower result in Scenario 1. These findings underscore the variability of model effectiveness depending on the scenario, emphasizing the need for careful model selection based on specific task requirements.

Table 11. Performance Comparison

Models	Sce-1	Sce-2	Sce-3	Sce-4
BERT	**0.99748**	0.996472	0.99244	0.972278
RoBERTa	**0.99748**	**0.99748**	0.019657	0.972278
DistilBERT	**0.99748**	0.996976	0.019657	0.970766
TinyBERT	0.995968	0.988911	0.989919	0.679435
MobileBERT	**0.99748**	0.996976	0.534274	0.990927
ALBERT	0.27873	**0.99748**	**0.996472**	**0.99496**

The nearly 100% accuracy results on several models indicate that Transformer models, especially BERT, RoBERTa, and MobileBERT, are very effective in learning patterns from the datasets used. This could be due to the high quality of data preprocessing, including appropriate tokenization and the use of default hyperparameters optimized for the classification task. However, these results could also indicate possible overfitting, especially since near-perfect accuracy is rare on real-world datasets. Further evaluation on more heterogeneous or independent datasets is needed to verify the generalization of the models.

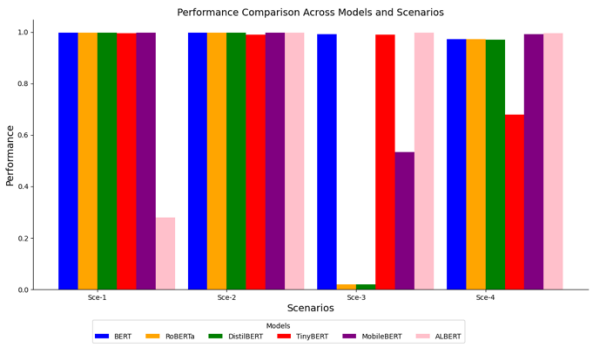

Figure 3. Model Comparison

Figure 3 shows a comparison of model performance in four scenarios. ALBERT excels in scenarios with limited data, while TinyBERT performs lower in scenarios with complex data. The bar chart shows BERT, RoBERTa, and MobileBERT consistently achieving high accuracy, while ALBERT improves significantly after Scenario 1. RoBERTa and DistilBERT drop sharply in Scenario 3, indicating sensitivity to smaller datasets.

Figure 4 shows a line chart highlighting stable performance for BERT, MobileBERT, and ALBERT,

while RoBERTa and DistilBERT struggle in Scenario 3. ALBERT's improvement across scenarios is clearly visible.

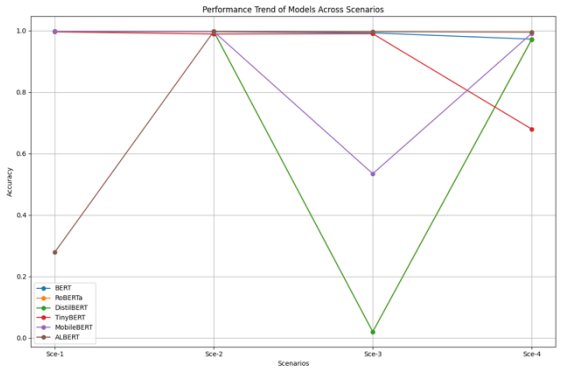

Figure 4. Performance Trend of Models Across Scenarios

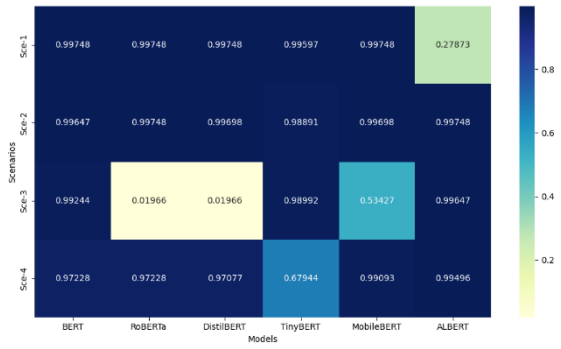

Figure 5. Model Accuracy Heatmap Across Scenario

Figure 5 shows that the heatmap visualizes accuracy intensity, with BERT and MobileBERT performing consistently. ALBERT improves significantly after Scenario 1, while RoBERTa and DistilBERT underperform in Scenario 3.

3.2 Discussion

The experimental results across the four scenarios provide valuable insights into the performance, efficiency, and limitations of Transformer-based models tested on the MyPersonality dataset. In Scenario 1, where all models were fine-tuned using the same configuration on the full dataset, most models displayed robust performance, with BERT, RoBERTa, DistilBERT, and MobileBERT achieving high accuracy scores close to 99%. This suggests their strong suitability for tasks that demand high precision and recall [16], [19]. However, ALBERT significantly underperformed in this scenario, indicating that the same fine-tuning strategy may not be appropriate for this smaller, parameter-efficient model [28]. The outcome in this scenario highlights the need for customized optimization for certain architectures, particularly those with fewer parameters.

Scenario 2 further reinforces this observation. When ALBERT was fine-tuned using a different, more tailored strategy, its performance improved drastically, achieving accuracy comparable to the other models. This improvement demonstrates that with a customized approach, ALBERT can reach similar performance levels, underscoring the importance of model-specific adjustments to optimize effectiveness [28]. While BERT, RoBERTa, DistilBERT, and MobileBERT performed well with the standard fine-tuning strategy, ALBERT's substantial performance boost with optimized fine-tuning reveals its potential when matched with the right optimization approach [31].

The results were more varied in Scenario 3, where the dataset was reduced to 30% of the original size. BERT and TinyBERT maintained strong accuracy, suggesting their ability to generalize even with limited data [16], [29]. However, RoBERTa and DistilBERT experienced significant performance drops, possibly due to overfitting or a lack of adaptability to reduced data, which may indicate their reliance on larger datasets for optimal performance [19], [32]. With its specialized fine-tuning, ALBERT again outperformed several models, showcasing its resilience in data-constrained environments [28]. MobileBERT, on the other hand, exhibited a marked decline in performance, suggesting a sensitivity to dataset size that might require a larger dataset to sustain its robustness [30].

In Scenario 4, an optional scenario simulating conditions without data from the original dataset, MobileBERT and ALBERT emerged as the most stable and adaptable models, performing well even with minimal prior data exposure. This result indicates their robustness in handling tasks with little to no initial data [28], [30]. In contrast, TinyBERT showed a significant drop in accuracy, suggesting that while compact and efficient, lightweight models like TinyBERT may face challenges in complex or data-limited scenarios without further optimization [29], [33]. These findings emphasize that compact models offer computational efficiency but may require additional tuning or struggle with generalizing certain tasks.

The performance degradation of models like TinyBERT in Scenario 4 can be attributed to the lighter architecture, which while efficient, lacks the semantic representation capacity needed for complex tasks. In contrast, the high performance of ALBERT in the same scenario reflects the benefits of a tailored fine-tuning strategy, including batch size adjustment and weight decay. This finding confirms that optimization strategies tailored to a particular architecture can compensate for the inherent limitations of lightweight models.

These insights underscore the critical importance of aligning model choice with task-specific requirements, such as dataset size and the complexity of fine-tuning strategies. By carefully matching the model to the characteristics of the task, including the amount of data available and the model's architecture, it is possible to achieve an optimal balance between efficiency and predictive performance. This comprehensive evaluation of Transformer-based models highlights how nuanced fine-tuning and dataset considerations can unlock each

model's full potential, allowing maximum effectiveness across varying scenarios [34].

The findings of this study can be directly applied to various real-world scenarios. For instance, in psychology, the accurate profiling of personality traits can aid in developing personalized therapeutic interventions. Similarly, in marketing, these models can enhance customer segmentation and targeted advertising by predicting consumer preferences based on their social media activity. Furthermore, the models have potential applications in education and recruitment, where understanding individual personality traits can inform personalized learning paths and team-building strategies.

Despite the promising results, this study has certain limitations. First, the comprehensive dataset used, MyPersonality, may not fully represent the diversity of social media users worldwide. Second, the experiments focus primarily on English-language text, potentially limiting the applicability of the findings to other languages and cultural contexts. Finally, the high accuracy observed in some models may indicate potential overfitting, warranting further validation using more heterogeneous datasets. Future research should explore multilingual datasets and implement cross-validation strategies to address these limitations.

4. Conclusions

Experimental results show that ALBERT with custom fine-tuning achieves the highest accuracy of 99.7% in Scenario 4, while BERT consistently performs with an average accuracy of 99.6% across all scenarios. TinyBERT, although computationally efficient, has limitations in complex scenarios, achieving only 67.9% accuracy in Scenario 4. These results highlight the importance of selecting a model that fits the task requirements, considering the trade-off between efficiency and accuracy. This study shows that ALBERT, with a custom fine-tuning strategy, successfully achieves the highest average accuracy (99.6%) across all scenarios, especially under limited data conditions. On the other hand, TinyBERT shows lower accuracy (67.9% in Scenario 4), highlighting the importance of model selection based on task complexity and dataset. The results show that ALBERT and MobileBERT perform well in most conditions, especially in accuracy, precision, recall, and F1 score. ALBERT's custom fine-tuning settings are effective but underperform with default settings, highlighting the need for customized configurations. BERT, RoBERTa, and DistilBERT perform well on larger datasets with default fine-tuning, but RoBERTa and DistilBERT struggle with reduced data, while BERT remains stable. TinyBERT, while efficient, shows weaker performance with limited data, thus struggling with accuracy. In conclusion, fine-tuning is key for models like ALBERT, while BERT is reliable under a wide range of conditions, and TinyBERT may not be ideal for complex tasks. Model selection should balance efficiency and task requirements. This study's outcomes have practical implications for industries such as psychology, marketing, and education, where personality profiling can improve decision-making processes and provide deeper insights into individual behavior.

References

[1] E. Utami, A. D. Hartanto, S. Adi, I. Oyong, and S. Raharjo, "Profiling analysis of DISC personality traits based on Twitter posts in Bahasa Indonesia," *Journal of King Saud University - Computer and Information Sciences*, Oct. 2022, doi: 10.1016/j.jksuci.2019.10.008.

[2] M. A. Iqbal, F. A. Ammar, A. R. Aldaihani, T. K. U. Khan, and A. Shah, "Building Most Effective Requirements Engineering Teams by Evaluating Their Personality Traits Using Big-Five Assesment Model," 2019.

[3] P. Kavya and V. Kanchana, "Student Personality Analysis In Blended Mode Using Big Five," in *2023 International Conference on Network, Multimedia and Information Technology, NMITCON 2023*, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/NMITCON58196.2023.10276332.

[4] A. Koutsoumpis *et al.*, "Beyond traditional interviews: Psychometric analysis of asynchronous video interviews for personality and interview performance evaluation using machine learning," *Comput Human Behav*, vol. 154, May 2024, doi: 10.1016/j.chb.2023.108128.

[5] M. M. Tadesse, H. Lin, B. Xu, and L. Yang, "Personality Predictions Based on User Behavior on the Facebook Social Media Platform," *IEEE Access*, vol. 6, pp. 61959–61969, 2018, doi: 10.1109/ACCESS.2018.2876502.

[6] E. Utami, A. F. Iskandar, A. D. Hartanto, and S. Raharjo, "DISC Personality Classification using Twitter: Usability Testing," in *Proceedings - 2021 IEEE 5th International Conference on Information Technology, Information Systems and Electrical Engineering: Applying Data Science and Artificial Intelligence Technologies for Global Challenges During Pandemic Era, ICITISEE 2021*, Institute of Electrical and Electronics Engineers Inc., 2021, pp. 180–185. doi: 10.1109/ICITISEE53823.2021.9655937.

[7] C. Geary, E. March, and R. Grieve, "Insta-identity: Dark personality traits as predictors of authentic self-presentation on Instagram," *Telematics and Informatics*, vol. 63, Oct. 2021, doi: 10.1016/j.tele.2021.101669.

[8] A. D. Hartanto, E. Utami, Kusrini, and A. Setyanto, "A Survey of Semantic Approaches in Personality Traits Profiling Analysis," in *2024 International Conference on Smart Computing, IoT and Machine Learning, SIML 2024*, Institute of Electrical and Electronics Engineers Inc., 2024, pp. 35–42. doi: 10.1109/SIML61815.2024.10578100.

[9] S. Bazzaz Abkenar, M. Haghi Kashani, E. Mahdipour, and S. M. Jameii, "Big data analytics meets social media: A systematic review of techniques, open issues, and future directions," *Telematics and Informatics*, vol. 57, Mar. 2021, doi: 10.1016/j.tele.2020.101517.

[10] Z. Khan, "A Deep Learning Approach for Predicting Personality Traits," in *2023 14th International Conference on Computing Communication and Networking Technologies, ICCCNT 2023*, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/ICCCNT56998.2023.10307763.

[11] A. Salem, Jāmiʻat ʻAyn Shams, Egyptian Knowledge Bank, Institute of Electrical and Electronics Engineers. Egypt Section, and Institute of Electrical and Electronics Engineers, *Predicting Personality Traits from Social Media using Text Semantics*. 2018.

[12] W. Maharani and V. Effendy, "Big five personality prediction based in Indonesian tweets using machine learning methods," *International Journal of Electrical and Computer Engineering*, vol. 12, no. 2, pp. 1973–1981, Apr. 2022, doi: 10.11591/ijece.v12i2.pp1973-1981.

[13] M. Jayaratne and B. Jayatilleke, "Predicting Personality Using Answers to Open-Ended Interview Questions," *IEEE Access*, vol. 8, pp. 115345–115355, 2020, doi: 10.1109/ACCESS.2020.3004002.

[14] H. Chen, X. Zhang, and X. Wu, "Research on Improving Personalized Recommendation Accuracy Based on NLP Semantic Analysis," in *2023 IEEE International Conference on Control, Electronics and Computer Technology, ICCECT 2023*, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 1113–1118. doi: 10.1109/ICCECT57938.2023.10140740.

[15] W. Khan, A. Daud, K. Khan, S. Muhammad, and R. Haq, "Exploring the frontiers of deep learning and natural language processing: A comprehensive overview of key challenges and emerging trends," *Natural Language Processing Journal*, vol. 4, p. 100026, Sep. 2023, doi: 10.1016/j.nlp.2023.100026.

[16] J. Devlin, M.-W. Chang, K. Lee, K. T. Google, and A. I. Language, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," 2019. [Online]. Available: https://github.com/tensorflow/tensor2tensor

[17] B. Bsir, N. Khoufi, and M. Zrigui, "Prediction of Author's Profile Basing on Fine-Tuning BERT Model," *Informatica (Slovenia)*, vol. 48, no. 1, pp. 69–78, Mar. 2024, doi: 10.31449/inf.v48i1.4839.

[18] N. Halimawan, D. Suhartono, A. P. Gema, and R. Yunanda, "BERT and ULMFiT Ensemble for Personality Prediction from Indonesian Social Media Text," in *Proceeding - 2022 International Symposium on Information Technology and Digital Innovation: Technology Innovation During Pandemic, ISITDI 2022*, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 156–161. doi: 10.1109/ISITDI55734.2022.9944476.

[19] Y. Liu *et al.*, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," 2019. [Online]. Available: https://github.com/pytorch/fairseq

[20] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," Oct. 2019, [Online]. Available: http://arxiv.org/abs/1910.01108

[21] M. A. Kosan, H. Karacan, and B. A. Urgen, "Personality traits prediction model from Turkish contents with semantic structures," *Neural Comput Appl*, vol. 35, no. 23, pp. 17147–17165, Aug. 2023, doi: 10.1007/s00521-023-08603-z.

[22] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP," Online. [Online]. Available: https://huggingface.co/

[23] H. Lucky, Roslynlia, and D. Suhartono, "Towards Classification of Personality Prediction Model: A Combination of BERT Word Embedding and MLSMOTE," in *Proceedings of 2021 1st International Conference on Computer Science and Artificial Intelligence, ICCSAI 2021*, Institute of Electrical and Electronics Engineers Inc., 2021, pp. 346–350. doi: 10.1109/ICCSAI53272.2021.9609750.

[24] N. Mauliza, A. Shakila Iedwan, Y. Pristyanto, A. D. Hartanto, and A. N. Rohman, "Resampling Techniques on Model Performance Classification of Maternal Health Risks," *J. RESTI (Rekayasa Sist. Teknol. Inf.)*, vol. 10, no. 4, pp. 496–505, 2024, doi: 10.29207/resti.v8i4.5934.

[25] Sichuan Institute of Electronics and Institute of Electrical and Electronics Engineers, *Feature Analysis and Optimisation for Computational Personality Recognition*.

[26] S. Ouni, F. Fkih, and M. N. Omri, "Novel semantic and statistic features-based author profiling approach," *J Ambient Intell Humaniz Comput*, Sep. 2022, doi: 10.1007/s12652-022-04198-w.

[27] M. Hassanein, S. Rady, W. Hussein, and T. F. Gharib, "Predicting the Big Five for social network users using their personality characteristics," in *Proceedings - 2021 IEEE 10th International Conference on Intelligent Computing and Information Systems, ICICIS 2021*, Institute of Electrical and Electronics Engineers Inc., 2021, pp. 160–164. doi: 10.1109/ICICIS52592.2021.9694160.

[28] Z. Lan *et al.*, "ALBERT: A LITE BERT FOR SELF-SUPERVISED LEARNING OF LANGUAGE REPRESENTATIONS." [Online]. Available: https://github.com/google-research/ALBERT.

[29] X. Jiao *et al.*, "TinyBERT: Distilling BERT for Natural Language Understanding," Sep. 2019, [Online]. Available: http://arxiv.org/abs/1909.10351

[30] Z. Sun, H. Yu, X. Song, R. Liu, Y. Yang, and D. Zhou, "MobileBERT: a Compact Task-Agnostic BERT for Resource-Limited Devices," Association for Computational Linguistics. [Online]. Available: https://github.com/google-research/

[31] L. Sun and K. A. Aksyonov, "Fine-Tuning Bert on the Atis Dataset: Data Enhancement to Improve Intent Classification Accuracy," in *2024 9th International Symposium on Computer and Information Processing Technology, ISCIPT 2024*, Institute of Electrical and Electronics Engineers Inc., 2024, pp. 274–278. doi: 10.1109/ISCIPT61983.2024.10672674.

[32] Y. G. Xu, X. P. Qiu, L. G. Zhou, and X. J. Huang, "Improving BERT Fine-Tuning via Self-Ensemble and Self-Distillation," *J Comput Sci Technol*, vol. 38, no. 4, pp. 853–866, Jul. 2023, doi: 10.1007/s11390-021-1119-0.

[33] I. Panopoulos, S. Nikolaidis, S. I. Venieris, and I. S. Venieris, "Exploring the Performance and Efficiency of Transformer Models for NLP on Mobile Devices," in *Proceedings - IEEE Symposium on Computers and Communications*, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/ISCC58397.2023.10217850.

[34] H. Sobhanam and J. Prakash, "Analysis of fine tuning the hyper parameters in RoBERTa model using genetic algorithm for text classification," *International Journal of Information Technology (Singapore)*, vol. 15, no. 7, pp. 3669–3677, Oct. 2023, doi: 10.1007/s41870-023-01395-4.

