VOL. 11. NO. 1 AUGUST 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480 /jitk.v11i1.6980

EVALUATING REGRESSION AND NEURAL NETWORKS FOR FIVE TRAIT TEXT-BASED PERSONALITY PREDICTION

Anggit Dwi Hartanto[1*]; Ema Utami[1]; Arief Setyanto[1]; Kusrini[1]

Faculty of Computer Science[1]
Universitas Amikom Yogyakarta, Yogyakarta, Indonesia[1]
http://amikom.ac.id[1]
anggit@amikom.ac.id*; ema.u@amikom.ac.id; arief_s@amikom.ac.id; kusrini@amikom.ac.id

(*) Corresponding Author
(Responsible for the Quality of Paper Content)

Abstract*—The aim of this study is to evaluate the effectiveness of several predictive modeling techniques in mapping the five major personality traits (extraversion, neuroticism, agreeableness, conscientiousness, and openness) from text-based data. The dataset consists of text-based features extracted from publicly available social media posts, providing a realistic basis for personality prediction. Performance was measured using mean absolute error (MAE), mean squared error (MSE), and R^2 score to evaluate prediction accuracy and generalization quality, along with training time for computational efficiency. The research compares linear regression, ridge regression, random forest, and neural networks implemented in PyTorch. Results indicate that ridge regression and random forest outperform linear regression and neural networks in error metrics and explained variance, with ridge regression offering a favorable balance between accuracy and training time. Random forest achieves slightly better accuracy but with significantly longer training duration, reducing its practicality for real-time use. Despite theoretical advantages in modeling non-linear relationships, neural networks showed suboptimal results, likely due to limited hyperparameter tuning and dataset size. These findings highlight trade-offs among model complexity, accuracy, and efficiency, suggesting ridge regression as a pragmatic choice for current personality prediction from text while encouraging future research on advanced neural architectures and enhanced datasets*

Keywords*: neural networks, personality trait prediction, random forest, ridge regression.*

Intisari— *Tujuan penelitian ini adalah untuk mengevaluasi efektivitas beberapa teknik pemodelan prediktif dalam memetakan lima ciri kepribadian utama (extraversion, neuroticism, agreeableness, conscientiousness, and openness) dari data berbasis teks. Dataset ini terdiri dari fitur-fitur berbasis teks yang diekstrak dari unggahan media sosial yang tersedia untuk umum, memberikan dasar yang realistis untuk prediksi kepribadian. Performa diukur menggunakan mean absolute error (MAE), mean squared error (MSE), dan skor R^2 untuk mengevaluasi akurasi prediksi dan kualitas generalisasi, serta waktu pelatihan untuk efisiensi komputasi. Penelitian ini membandingkan regresi linier, regresi ridge, hutan acak, dan jaringan saraf tiruan yang diimplementasikan dalam PyTorch. Hasil menunjukkan bahwa regresi ridge dan hutan acak mengungguli regresi linier dan jaringan saraf tiruan dalam metrik kesalahan dan varians terjelaskan, dengan regresi ridge menawarkan keseimbangan yang baik antara akurasi dan waktu pelatihan. Hutan acak mencapai akurasi yang sedikit lebih baik tetapi dengan durasi pelatihan yang jauh lebih lama, sehingga mengurangi kepraktisannya untuk penggunaan waktu nyata. Meskipun terdapat keunggulan teoretis dalam pemodelan hubungan non-linier, jaringan saraf tiruan menunjukkan hasil yang kurang optimal, kemungkinan karena keterbatasan penyetelan hiperparameter dan ukuran dataset. Temuan ini menyoroti trade-off antara kompleksitas model, akurasi, dan efisiensi, yang menunjukkan regresi ridge sebagai pilihan pragmatis untuk prediksi kepribadian saat ini dari teks sambil mendorong penelitian masa depan pada arsitektur saraf tingkat lanjut dan kumpulan data yang ditingkatkan.*

Kata Kunci*: jaringan saraf, prediksi ciri kepribadian, hutan acak, regresi ridge.*

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 11. NO. 1 AUGUST 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i1.6980

INTRODUCTION

Text-based social media has become a popular choice for interaction in today's digital era. User-generated text data is now essential in various disciplines, especially psychology and computer science [1]. This text data has excellent potential to unearth individual personality characteristics, which are usually measured through psychological questionnaires, but can now be predicted automatically and at scale using machine learning techniques.

Predicting personality traits from text data provides the basis for a wide range of applications, from personalized marketing, recommendation systems, the development of psychological interventions, to improving the user experience within digital platforms [2]. Previous studies have highlighted the importance of utilizing text-based feature representations such as TF-IDF, bag-of-words, and embeddings to capture relevant linguistic information in the prediction process [3], [4].

Predicting personality traits from textual data has become an increasingly important area of research, bridging the fields of psychology, computational linguistics, and artificial intelligence [2], [5], [6]. Personality assessment plays a crucial role in various applications, including mental health diagnostics, personalized marketing, job recruitment, and human-computer interaction systems [6], [7], [8].

Traditional methods for personality prediction have relied on statistical models such as linear regression, logistic regression, and support vector machines (SVMs), which analyze linguistic features—such as word frequency, sentiment, and syntactic patterns—to infer personality traits based on established psychological frameworks like the Big Five (Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism) [9], [10], [11].

In the methodological realm, various machine learning approaches have been applied to predict personality from text, including traditional regression models such as Linear Regression, Ridge Regression, and Random Forest, known for their ease of interpretation and computational efficiency [12]. Although these models offer interpretability and computational efficiency, the limitation of their predictive performance lies in their ability to capture non-linear patterns that often appear in complex and high-dimensional text data [13], [14], [15], [16].

Alternatively, Neural Networks have become a popular choice due to their ability to perform automatic feature extraction and more complex representation learning, allowing for significant improvements in the accuracy of personality trait prediction [17]. However, Neural Networks require substantial computational resources and have challenges regarding model interpretability, which remains a concern, especially in psychology applications that require transparency [18], [19].

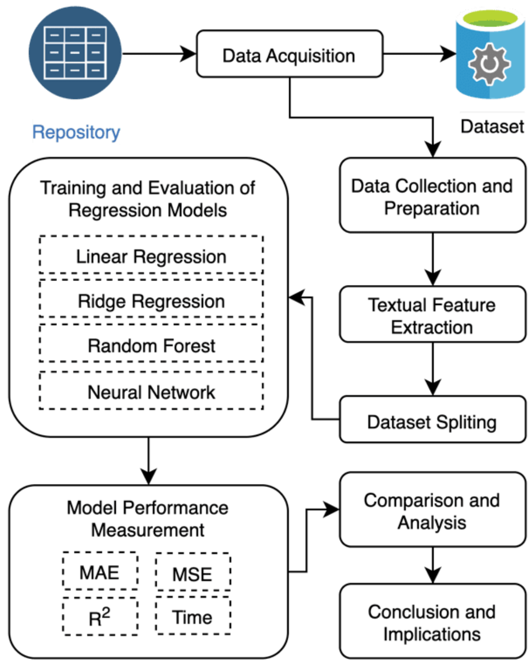
The current literature notes the lack of studies that systematically and comprehensively compare the effectiveness of traditional regression models and Neural Networks in predicting the five major personality dimensions of Extraversion, Neuroticism, Agreeableness, Conscientiousness, and Openness based on social media text data [20]. Some studies focus on a single approach or the latest model trends without providing an in-depth quantitative comparison.

Therefore, this research aims to fill this gap by evaluating the performance of both types of models using text datasets processed with the TF-IDF technique, measuring the prediction quality through standardized statistical metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and coefficient of determination (R2), and considering practical aspects such as computational time efficiency and ease of implementation in the context of real-time applications.

The main objective of this study is to provide a comprehensive overview of the advantages and limitations of each method in the context of text-based personality prediction, which can guide researchers and practitioners in choosing the most appropriate approach for their needs. In addition, this study is expected to provide insight into the trade-off between model accuracy, interpretability, and computational cost, thus supporting the development of more adaptive and responsive intelligence systems to user characteristics and preferences in various application fields, from psychology to digital marketing technology. Therefore, this study contributes to the development of science and technology at the intersection of psychology and data science and the practical application of artificial intelligence technology in improving personalization-based human-computer interaction.

MATERIALS AND METHODS

This study systematically evaluates the performance of traditional regression models and neural networks in text-based personality prediction, as illustrated in Figure 1.



Source : (Research Results, 2025)
Figure 1. Research Flow

A. Dataset

The dataset used in this study was taken from the Kaggle platform, precisely from the "MyPersonality" dataset available at https://www.kaggle.com/datasets/krishnanpalanisami/mypersonality. This dataset contains text data from social media status collected along with personality trait score labels based on the Big Five personality model. This data is particularly relevant for text-based personality prediction studies because it provides real-life examples of users' posts on social media that describe their self-expression. This data set has also passed the bare preprocessing stage, making it easier to further process the extraction of numerical features such as TF-IDF. This dataset provides a solid foundation for objectively evaluating various machine learning models on personality trait prediction tasks and allows for fair comparisons between methods.

B. Research Phase

Data Collection and Preparation

The data used was in the form of social media status text and personality trait scores that had been labeled based on the Big Five framework. The data is then processed to remove noise, normalize text by removing non-alphabetic characters, perform lowercasing, and prepare the data for further analysis.

Textual Feature Extraction

From the text of social media statuses, numerical features were extracted using the TF-IDF (Term Frequency - Inverse Document Frequency) method to represent words in numerical vectors that reflect the importance of words in the document and the corpus.

Dataset Splitting

The prepared data is then randomly divided into training data (training set) and testing data (test set), generally with a proportion of around 80% training and 20% testing, to evaluate model performance objectively.

Training and Evaluation of Regression Models

In this evaluation scenario, several machine learning models for regression are applied and compared to measure their predictive performance against the available dataset. The first model used is Linear Regression, which serves as a baseline to provide an initial benchmark against the performance of other models. Next, by applying regularization techniques, Ridge Regression is applied to overcome potential overfitting that may occur in linear models. As a non-linear approach, Random Forest Regression is used to explore the ability of ensemble models to capture complex patterns and interactions between features. Finally, a Neural Network with an artificial neural network architecture is applied to identify deeper and non-linear data representations, which are expected to improve prediction accuracy.

Model Performance Measurement

Each model is evaluated using error metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and coefficient of determination (R2 Score). In addition, the time required to train and perform predictions is recorded for computational efficiency analysis.

Comparison and Analysis

The results of all models are arranged in a table to compare the prediction performance and computation time. From here, which model provides the best balance between accuracy and efficiency that suits practical needs is analyzed.

Conclusion and Implications

Based on the evaluation, the study concludes the advantages and disadvantages of each model. It recommends using traditional regression models or neural networks according to the context of text-based personality trait prediction applications.

JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)
VOL. 11. NO. 1 AUGUST 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i1.6980

RESULTS AND DISCUSSION

This study evaluates four different prediction models—Linear Regression, Ridge Regression, Random Forest, and Neural Network—in the task of predicting five personality trait dimensions: Extraversion (sEXT), Neuroticism (sNEU), Agreeableness (sAGR), Conscientiousness (sCON), and Openness (sOPN) using a text-based dataset. The evaluation focuses on several key metrics, namely Mean Absolute Error (MAE), Mean Squared Error (MSE), coefficient of determination (R2 score), and computational time spent on training and testing the models. This analysis aims to explore the prediction accuracy, efficiency, and practical utility of each method.

Analisis Performa Prediksi Berdasarkan Metrik

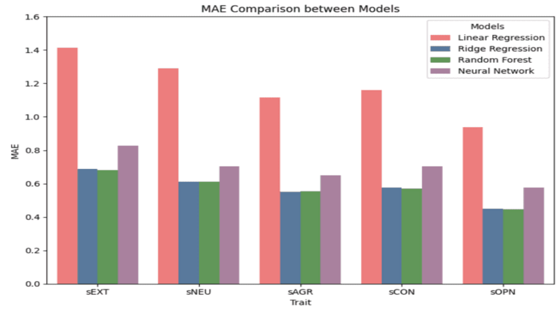
Based on the evaluation results in Table 1 and Graph 1, it can be seen that Random Forest and Ridge Regression consistently recorded the best performance in predicting text-based personality traits, with the lowest MAE and MSE values in almost all traits (sEXT, sNEU, sAGR, sCON, sOPN). Random Forest achieved the lowest MAE in sEXT (0.68) and sNEU (0.6101), while Ridge Regression excelled in sAGR (MAE: 0.5523) and sCON (MAE: 0.5777). Both models also recorded MSE below 0.75 for most traits, indicating high precision with minimal quadratic error.

Table 1. MAE & MSE Comparison Models

Model	Trait	MAE	MSE
Linear Regression	sEXT	1,4146	5,0042
Ridge Regression	sEXT	0,6859	0,724
Random Forest	sEXT	0,68	0,7017
Neural Network	sEXT	0,8262	1,0996
Linear Regression	sNEU	1,2915	4,1864
Ridge Regression	sNEU	0,6118	0,5791
Random Forest	sNEU	0,6101	0,5728
Neural Network	sNEU	0,7041	0,8005
Linear Regression	sAGR	1,1173	2,884
Ridge Regression	sAGR	0,5523	0,4493
Random Forest	sAGR	0,555	0,4605
Neural Network	sAGR	0,6488	0,6607
Linear Regression	sCON	1,1613	3,1949
Ridge Regression	sCON	0,5777	0,5314
Random Forest	sCON	0,5705	0,524
Neural Network	sCON	0,7043	0,8032
Linear Regression	sOPN	0,9372	2,2166
Ridge Regression	sOPN	0,4493	0,3486
Random Forest	sOPN	0,4463	0,3525
Neural Network	sOPN	0,5775	0,5719

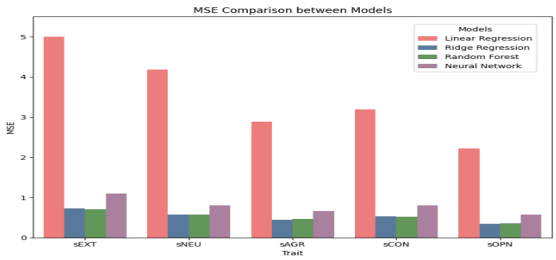
Source : (Research Results, 2025)

Table 1 shows that Linear Regression has the worst performance with MAE >1.1 and MSE >2.2 across all traits, indicating the unsuitability of linear models to capture the complexity of text patterns in personality prediction. Meanwhile, Neural Network ranks third with MAE ranging from 0.5775–0.8262 and MSE from 0.5719–1.0996. Although better than Linear Regression, its performance is still below that of Random Forest and Ridge, possibly due to limited training data or the need for further hyperparameter tuning.



Source : (Research Results, 2025)
Figure 2. MAE Comparison Models

Figure 1 and 2 provides an evaluation that This study's significant differences between traditional models, especially Random Forest/Ridge and Neural Network, reinforce the finding that ensemble and regularization-based models are more effective for text-based prediction with moderate dataset sizes. These results also highlight that model complexity does not always guarantee higher accuracy, as shown by the Neural Network's poor performance compared to simpler models. The implication is that model selection for personality trait prediction should consider the trade-off between accuracy and computational efficiency, especially considering the much longer training time of Random Forest, as shown in Table 3.



Source : (Research Results, 2025)
Figure 3. MSE Comparison Models

This is likely influenced by several factors, such as model complexity that has not been adjusted to the size and characteristics of the dataset, lack of optimal hyperparameter tuning, or perhaps also limited data volume that causes the neural network model to be susceptible to overfitting or underfitting.

Interpretasi Koefisien Determinasi (R2 Score)

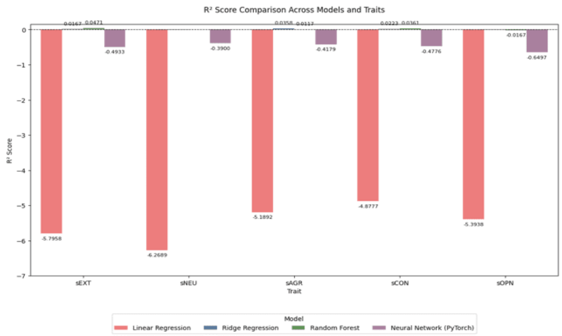
The evaluation results in Table 2 and Figure 4 show significant variations in model performance based on the coefficient of determination (R^2). Although Random Forest recorded the best performance with positive R^2 values for four of the five traits (sEXT: 0.0471, sNEU: 0.0055, sAGR: 0.0117, sCON: 0.0361), these values remain very low (<0.05), indicating that despite being the best among the models tested, the model's ability to explain the variance in the data is still very limited. This emphasizes a key limitation in the predictive power of the models used. Ridge Regression showed mixed results with R^2 values close to zero, ranging from -0.0056 to 0.0358, suggesting its predictive performance is only marginally better than a simple baseline model. The baseline here refers to a naive predictor that estimates the mean of the target variable for all instances, serving as a minimal reference point for performance comparison.

Table 2. R2 Score Comparison Models

Models	Trait	R2 Score
Linear Regression	sEXT	-5,7958
Ridge Regression	sEXT	0,0167
Random Forest	sEXT	0,0471
Neural Network	sEXT	-0,4933
Linear Regression	sNEU	-6,2689
Ridge Regression	sNEU	-0,0056
Random Forest	sNEU	0,0055
Neural Network	sNEU	-0,39
Linear Regression	sAGR	-5,1892
Ridge Regression	sAGR	0,0358
Random Forest	sAGR	0,0117
Neural Network	sAGR	-0,4179
Linear Regression	sCON	-4,8777
Ridge Regression	sCON	0,0223
Random Forest	sCON	0,0361
Neural Network	sCON	-0,4776
Linear Regression	sOPN	-5,3938
Ridge Regression	sOPN	-0,0056
Random Forest	sOPN	-0,0167
Neural Network	sOPN	-0,6497

Source : (Research Results, 2025)

Table 2 shows that Linear Regression and Neural Network recorded large negative R^2 values, namely Linear: -4.8777 to -6.2689; Neural Network: -0.39 to -0.6497. These negative values indicate that both models are worse than the baseline model, which only predicts the mean, especially in linear regression. This shows the inability of the default linear and neural network models to capture the relationship between text features and personality traits, possibly due to overfitting or the model architecture not matching the complexity of the data.



Source : (Research Results, 2025)

Figure 4. R2 Score Comparison Models

Figure 4 above provides an evaluation that these findings provide two critical implications. First, simple models with regularization, such as Random Forest/Ridge, are more stable for text-based prediction, although the achieved $R^2 \leq$ values are still low. This aligns with previous studies that highlight the limitations of personality prediction from text, even with state-of-the-art models. Second, there is a need for a better feature engineering approach or text representation, given the low overall $R^2 \leq$ values. The low $R^2 \leq$ may reflect noise in the data or nonlinearity in the relationship between linguistic features and personality traits.

Evaluation of Computational Time Efficiency

Based on the training time analysis in Table 3 and Figure 5, the models show a significant difference in computational efficiency. Neural Network recorded a consistent training time of 1.90–1.97 seconds per trait, while Ridge Regression showed a wider range of 1.74–3.28 seconds. Although the minimum training time for Ridge Regression is slightly faster than Neural Network, its average time tends to be higher, making Neural Network generally competitive in speed. Both models demonstrate efficiency suitable for text-based prediction tasks. In contrast, Random Forest requires substantially longer training times, ranging from 2,691.04 to 4,163.75 seconds (about 45–70 minutes per trait), approximately 1000 to 2000 times slower than Neural Network. Linear Regression occupies a middle position with training times of 48.19–70.36 seconds, still 25–40 times slower than Ridge Regression despite using a linear approach.

Table 3. Training Time Comparison Models

Model	Trait	Time (s)
Linear Regression	sEXT	70,36
Ridge Regression	sEXT	3,28
Random Forest	sEXT	2818,39
Neural Network	sEXT	1,95
Linear Regression	sNEU	49,37

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)
VOL. 11. NO. 1 AUGUST 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i1.6980

Model	Trait	Time (s)
Ridge Regression	sNEU	1,76
Random Forest	sNEU	2691,04
Neural Network	sNEU	1,92
Linear Regression	sAGR	48,19
Ridge Regression	sAGR	1,78
Random Forest	sAGR	3206,91
Neural Network	sAGR	1,91
Linear Regression	sCON	48,32
Ridge Regression	sCON	1,82
Random Forest	sCON	3548,24
Neural Network	sCON	1,97
Linear Regression	sOPN	49,07
Ridge Regression	sOPN	1,74
Random Forest	sOPN	4163,75
Neural Network	sOPN	1,9

Source : (Research Results, 2025)

These extreme differences have several practical implications:

- Accuracy-Time Trade-off: While Random Forest provides the best MAE and MSE values, it requires significant computational resources, which is an important consideration in real-time applications or resource-constrained environments.
- Neural Network Efficiency: The neural network architecture with default parameters is faster than Linear Regression and outperforms it in terms of accuracy (MAE/MSE) and R^2, making it a more balanced choice for datasets of this size.
- Ridge Regression Advantage: With its short training time and stable predictive performance close to that of Random Forest, Ridge Regression serves as a strong baseline for studies like this.

Source : (Research Results, 2025)

Figure 5. Training Time Comparison Models

Figure 5 illustrates the comparative computational efficiency of the evaluated models, highlighting significant differences in training time. Neural Network and Ridge Regression demonstrate much faster training times relative to other models, with Neural Network maintaining a consistent range and Ridge Regression showing some variability but generally competitive speeds. In contrast, Random Forest requires substantially longer training times, often tens of minutes per trait, indicating potential scalability concerns for larger datasets. Linear Regression falls in the middle ground in terms of training time but is still markedly slower than the more efficient models. Based on these results, model selection should consider multiple factors: dataset scale, where Random Forest may struggle with large data volumes; computational resources, since Neural Networks have room for optimization through further tuning; and application objectives, as Ridge Regression and Neural Networks offer a practical balance of speed and accuracy suited for real-time scenarios, although with a slight trade-off in prediction performance.

Model Implications and Real-World Application

Based on a comprehensive analysis of three evaluation metrics—MAE/MSE, R^2, and computational time—this study reveals several critical implications for real-world applications of text-based personality prediction. Although Random Forest recorded the best predictive performance with MAE ranging from 0.4463 to 0.68, it incurs a prohibitive computational cost of up to 4163 seconds per trait. Furthermore, its R^2 values are low, with a maximum of 0.0471, limiting its suitability for real-time scenarios or environments requiring frequent model updates. Conversely, Neural Network offers excellent training time efficiency between 1.9 and 1.97 seconds with competitive accuracy (MAE 0.5775–0.8262), making it well-suited for applications such as automated personality analysis on social media or AI-driven recruitment. However, the negative R^2 values ranging from -0.39 to -0.6497 indicate that architectural improvements are needed to enhance prediction quality. Ridge Regression strikes an optimal balance, featuring fast training times of 1.74 to 3.28 seconds, stable MAE accuracy (0.4493–0.6859), and R^2 values close to zero. This makes it a practical choice for resource-constrained settings like edge computing or low-latency use cases.

Practical implications of these findings include:

a. Scalability priority: Neural Network or Ridge Regression are more feasible for enterprise systems with massive data than Random Forest.
b. Interpretability needs: Random Forest is a good choice if you want to know how important each feature or keyword is in the prediction. However, this method requires more time and computing resources.
c. Future optimization: Integration of pre-trained language models such as the BERT model that

can bridge the accuracy and efficiency gap, as proposed.

Future Works

Based on the findings obtained from the evaluation of the three metrics, this study recommends several development directions for future text-based personality prediction studies: Optimization of Neural Network with More Specialized Architecture. Although Neural Network shows high training time efficiency, the negative R^2 value indicates the need for architectural improvements. Transfer learning approaches with pre-trained models such as BERT or RoBERTa can be tested to improve the capacity of text feature modeling. In addition, experiments with attention mechanisms or transformer-based models can help capture more complex linguistic contexts.

Hybrid Model: Combination of Random Forest and Neural Network. Random Forest provides the best accuracy but with a high computational cost. One potential solution is to develop a hybrid model that leverages the strengths of Random Forest in important feature selection and Neural Network in deep text pattern extraction. Exploration of More Robust Feature Engineering. The low R^2 values in all models indicate that the text features used may not be representative enough. Future research could test the combination of linguistic features, such as LIWC, n-grams, or word embeddings. With more specific psycholinguistic features. Study [21] showed that integrating personality theory-based features, such as the Big Five lexicon, can improve the correlation of predictions.

Refinement of Evaluation with Relevant Metrics. In addition to MAE/MSE and R^2, future research could include metrics such as balanced accuracy or F1-score if a classification approach is used. Bias analysis and model fairness must be tested. Scalability and Real-Time Applications. For real-time needs, the development of lightweight neural network-based models, such as distilled versions of BERT or Random Forest, and inference time optimization through parallel computing could be a focus.

CONCLUSION

This study shows that Ridge Regression and Random Forest provide the best prediction performance. Low error values and positive or near-zero R^2 indicate the models' ability to capture data variation adequately. In contrast, Linear Regression yields the worst results with negative R^2, signaling its inability to model data complexity. Meanwhile, the neural network's performance is suboptimal, likely due to suboptimal hyperparameter settings and limited data, possibly resulting in underfitting or overfitting.

Regarding computational efficiency, Ridge Regression and Neural Network stand out with very short training times, suggesting their suitability for real-time or resource-constrained scenarios. Although Random Forest achieves the highest accuracy, its long training time limits its applicability in speed-critical contexts.

REFERENCE

[1] G. Rajalakshmi, A. Kansal, S. Thomas, M. P. Mali, D. Anandhasilambarasan, and D. Sharma, “A Comprehensive Study on the Real-Time Text Categorization of User Generated Content using Artificial Intelligence,” in *2024 Global Conference on Communications and Information Technologies, GCCIT 2024*, Institute of Electrical and Electronics Engineers Inc., 2024. doi: 10.1109/GCCIT63234.2024.10862668.

[2] 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.

[3] M. Liang and T. Niu, “Research on Text Classification Techniques Based on Improved TF-IDF Algorithm and LSTM Inputs,” in *Procedia Computer Science*, Elsevier B.V., 2022, pp. 460–470. doi: 10.1016/j.procs.2022.10.064.

[4] E. M. Dharma, F. Lumban Gaol, H. Leslie, H. S. Warnars, and B. Soewito, “THE ACCURACY COMPARISON AMONG WORD2VEC, GLOVE, AND FASTTEXT TOWARDS CONVOLUTION NEURAL NETWORK (CNN) TEXT CLASSIFICATION,” *J Theor Appl Inf Technol*, vol. 31, no. 2, 2022, [Online]. Available: www.jatit.org

[5] G. Angelini, “Big five model personality traits and job burnout: a systematic literature review,” *BMC Psychol*, vol. 11, no. 1, Dec. 2023, doi: 10.1186/s40359-023-01056-y.

[6] 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*

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 11. NO. 1 AUGUST 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i1.6980

Sciences, Oct. 2022, doi: 10.1016/j.jksuci.2019.10.008.

[7] A. A. Azmee, F. Nweke, M. Pederson, M. A. A. H. Khan, and Y. Pei, "Human AI Collaboration Framework for Detecting Mental Illness Causes from Social Media," Jun. 01, 2025, *Elsevier B.V.* doi: 10.1016/j.smhl.2025.100566.

[8] Idris, E. Utami, A. D. Hartanto, and S. Raharjo, "Systematic Literature Review of Profiling Analysis Personality from Social Media," in *Journal of Physics: Conference Series*, IOP Publishing Ltd, Mar. 2021. doi: 10.1088/1742-6596/1823/1/012115.

[9] G. D. Salsabila and E. B. Setiawan, "Semantic Approach for Big Five Personality Prediction on Twitter," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 5, no. 4, pp. 680–687, Aug. 2021, doi: 10.29207/resti.v5i4.3197.

[10] M. A. Parameswara, F. S. P. Mahendra, M. R. C. Mardika, I. Puspasari, and N. P. Utama, "Big Five Personality Prediction Based on Indonesian Tweets and Personality Test," in *2023 International Conference on Electrical Engineering and Informatics (ICEEI)*, IEEE, Oct. 2023, pp. 1–6. doi: 10.1109/ICEEI59426.2023.10346812.

[11] 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.

[12] A. Perera and P. Fernando, "Cyberbullying Detection System on Social Media Using Supervised Machine Learning," *Procedia Comput Sci*, vol. 239, pp. 506–516, Jan. 2024, doi: 10.1016/J.PROCS.2024.06.200.

[13] Winarno, K. A. Azzahra, and S. W. Sihwi, "Dark Triad Personality Traits Prediction through Twitter Analysis using Support Vector Machine Method," in *APICS 2022 - 2022 1st International Conference on Smart Technology, Applied Informatics, and Engineering, Proceedings*, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 102–107. doi: 10.1109/APICS56469.2022.9918721.

[14] N. A. Utami, W. Maharani, and I. Atastina, "Personality Classification of Facebook Users According to Big Five Personality Using SVM (Support Vector Machine) Method," in *Procedia Computer Science*, Elsevier B.V., 2021, pp. 348–355. doi: 10.1016/j.procs.2021.12.023.

[15] R. P. Pratama and W. Maharani, "Predicting Big Five Personality Traits Based on Twitter User U sing Random Forest Method*," in *2021 International Conference on Data Science and Its Applications, ICoDSA 2021*, Institute of Electrical and Electronics Engineers Inc., 2021, pp. 110–117. doi: 10.1109/ICoDSA53588.2021.9617501.

[16] J.-T. Sohns, M. Schmitt, F. Jirasek, H. Hasse, and H. Leitte, "Attribute-based Explanation of Non-Linear Embeddings of High-Dimensional Data," *IEEE Trans Vis Comput Graph*, vol. 28, no. 1, pp. 540–550, 2022, doi: 10.1109/TVCG.2021.3114870.

[17] M. A. Kosan, H. Karacan, and B. A. Urgen, "Predicting personality traits with semantic structures and LSTM-based neural networks," *Alexandria Engineering Journal*, vol. 61, no. 10, pp. 8007–8025, Oct. 2022, doi: 10.1016/j.aej.2022.01.050.

[18] Z. Wang, T. Peng, and Z. Lu, "Comparative Research on Forest Fire Image Segmentation Algorithms Based on Fully Convolutional Neural Networks," *Forests*, vol. 13, no. 7, Jul. 2022, doi: 10.3390/f13071133.

[19] R. P. Putra and E. B. Setiawan, "RoBERTa as Semantic Approach for Big Five Personality Prediction using Artificial Neural Network on Twitter," in *ICACNIS 2022 - 2022 International Conference on Advanced Creative Networks and Intelligent Systems: Blockchain Technology, Intelligent Systems, and the Applications for Human Life, Proceeding*, Institute of Electrical and Electronics Engineers Inc., 2022. doi: 10.1109/ICACNIS57039.2022.10054870.

[20] X. Chen, J. He, E. Swanson, Z. Cai, and X. Fan, "Big Five Personality Traits and Second Language Learning: a Meta-analysis of 40 Years' Research," *Educ Psychol Rev*, vol. 34, no. 2, pp. 851–887, Jun. 2022, doi: 10.1007/s10648-021-09641-6.

[21] 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," *Artif Intell Rev*, vol. 58, no. 8, Aug. 2025, doi: 10.1007/s10462-025-11245-3.