Big Five Personality Prediction from Essays through Vocabulary Graphs and Contextual Transformer Representations

Yuliant Sibaroni[1)]*, Sri Suryani Prasetyowati[2)]

[1] School of Computing, Software Engineering, Telkom University, Bandung, Indonesia
[2]School of Computing, Informatics, Telkom University, Bandung, Indonesia
* Corresponding author: yuliant@telkomunivesity.ac.id

Abstract

This study develops an approach to predict personality traits based on the Big Five model through essay text by combining global and contextual information using a Vocabulary Graph Convolutional Network (VGCN) and Bidirectional Encoder Representations from Transformers (BERT). VGCN is used to capture global relationships between words through a vocabulary graph representation built based on NPMI weights, while BERT plays a role in producing richer and more contextual text representations. Experimental results show that the integration of structural and contextual features can provide good performance in multi-label classification tasks. Based on the grid search process and 5-fold cross-validation, the best model configuration obtained a macro F1 score of 0.6827 and a weighted F1 score of 0.6829 in out-of-fold prediction. The evaluation results for each feature show that the recall value is close to 1.0, while the precision value ranges from 0.50 to 0.56. This condition indicates a tendency for the model to produce more positive predictions. Furthermore, the results of the study highlight the importance of optimizing the prediction threshold value to obtain a more optimal compromise between precision and recall. Overall, the integration of graph-based semantic information with contextual representations from BERT shows that the VGCN-BERT approach has good potential for application in multi-label personality classification.

Keywords- Personality Prediction, Big Five, Text Analysis, Graph-Enhanced Embedding, Multi-Label Evaluation

1. INTRODUCTION

Personality has a significant influence on how a people thinks, feels, and behaves. It also plays a crucial role in decision-making, motivation, and social interactions [1]. Understanding personality can be utilized in various fields, such as job suitability assessment, mental health assessment, and adaptive intelligent system development [2]. In psychology, one of the most widely validated personality theories is the Big Five Personality Traits, which divides personality into five OCEAN dimensions [2], [3]. These five dimensions include Openness to Experience which relates to curiosity and openness to new experiences, Conscientiousness which reflects discipline and order, Extraversion which relates to openness in socializing and self-confidence, Agreeableness which reflects empathy and the ability to work together, and Neuroticism which describes a tendency towards emotional instability and vulnerability to stress [2], [3]. Traditionally, personality is assessed using questionnaires such as the NEO Personality Inventory (NEO-PI) and the Big Five Inventory (BFI). However, these methods are relatively costly and time-consuming, difficult to implement on a large scale, and have the potential to introduce response bias [3]. Alternatively, natural texts such as essays can be analyzed automatically by leveraging developments in Natural Language Processing (NLP) and Artificial Intelligence (AI) [4]. One such development is Bidirectional Encoder Representations from Transformers (BERT), which, through its bidirectional self-attention mechanism, is able to capture context and long-range semantic relationships better than static representations such as TF-IDF and Word2Vec [5], [6]. These capabilities make BERT widely used in various text classification tasks, including personality detection [7].

Several studies have explored automated personality detection using the Essays-Big 5 dataset. Kazemeini et al. [8] developed a Bagged-SVM model that integrates BERT-based features with psycholinguistic information and obtained better results than traditional machine learning methods. Kerz et al. [1] integrated BERT with the BiLSTM-Attention architecture and achieved an average accuracy of 63.5%. In addition to sequence modeling-oriented methods, graph-based approaches are also increasingly being applied due to their ability to represent more complex relationships between words. Ramezani et al. [9] developed KGrAt-Net by utilizing the knowledge graph structure, while Waqas et al. [10] proposed TraitBertGCN, which integrates BERT with a three-layer GCN and achieved an accuracy of 77.42%.

DOI : https://doi.org/10.21063/jtif.2026.V14.2.100-109
Creative Commons Attribution-Share Alike 4.0 International License (CC BY-SA 4.0)

This finding aligns with the research of Lin et al. [11], who applied the BertGCN framework. This research demonstrates that integrating contextual representations with a graph-based learning approach has the potential to improve performance in classification tasks.

In the realm of graph-based approaches, Lu, Du, and Nie [12] proposed a Vocabulary Graph Convolutional Network integrated with BERT (VGCN-BERT). This model combines contextual embeddings generated by BERT with information about global relationships between words through a vocabulary graph representation. The graph is formed based on word co-occurrence statistics, and the resulting relationships are weighted using Normalized Pointwise Mutual Information (NPMI) [12]. valuation results show that VGCN-BERT outperforms both BERT and TextGCN on several text classification benchmarks, including CoLA and SST-2 [12]. Condori [13] reported similar findings when applying VGCN-BERT to the task of detecting aggressive language. Despite its promising potential, the use of VGCN-BERT in multi-label Big Five personality classification is still relatively limited. Previous research has tended to emphasize the use of context within documents and has not specifically explored global semantic relationships at the corpus level [12]. Graph-based approaches have the potential to capture complex linguistic patterns relevant for multi-label personality classification. Unlike TraitBERTGCN, which uses document-level graphs, VGCN-BERT constructs graphs at the corpus level and has previously been applied primarily to single-label classification. Therefore, the novelty of this research lies in adapting VGCN-BERT for multi-label classification on the Big Five personality dimensions. Binary labels are used as classification results based on algorithmic thresholds, not as clinical diagnoses. This model is intended for research purposes only and should not be used as diagnostic tool or for screening recruitment candidates to minimize the risk of demographic bias.

2. METHODS AND DATA

2.1 Research Stages

This study implemented systematic experimental stages to obtain reliable and valid personality classification results. The overall research flow is presented in Figure 1. The research stages began with processing the Essays-Big 5 dataset [14]. To avoid data leakage and maintain the objectivity of the evaluation process, the dataset was first partitioned into data train and data test before the preprocessing step was carried out. The preprocessing step includes cleaning the data and resampling. Next, vocabulary selection, word co-occurrence calculations, and NPMI statistical calculations for graph construction were performed using only the training data for each fold. After the dataset is split, the essay text is processed through two parallel pipelines. The first pipeline uses BERT to generate contextual representations at the document level [5], while the second pipeline builds an NPMI-weighted vocabulary graph using a graph-based approach [12]. Both representations are then combined in the Graph-Enriched Embedding (BERT + VGCN) stage to form the final document representation. The model is trained using grid search [15] and 5-fold multilabel stratified cross-validation to generate out-of-fold (OOF) predictions. The OOF predictions are then used for threshold tuning [16] to determine the optimal threshold for converting probabilities into binary labels. The tuning process uses only the OOF validation data, leaving the test set untouched. The optimal threshold obtained is then locked and applied to the test set to generate the final evaluation.

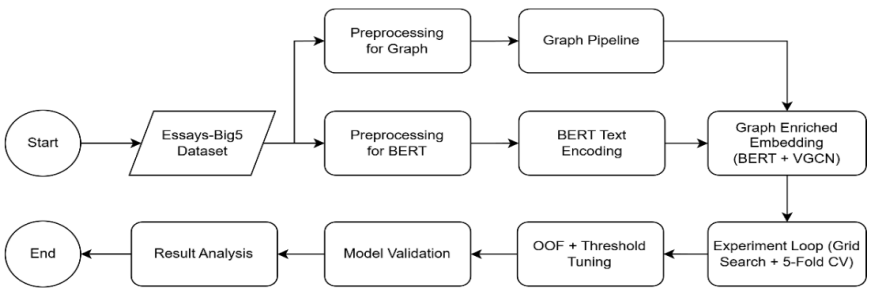

Figure 1. Experiment Flowchart

DOI : https://doi.org/10.21063/jtif.2026.V14.2.100-109
Creative Commons Attribution-Share Alike 4.0 International License (CC BY-SA 4.0)

2.2 Data Preprocessing

The dataset used is the Essays-Big 5 dataset from the Hugging Face Datasets, which consists of 2,467 English-language essays with binary labels for the five Big Five dimensions: Openness to Experience (O), Conscientiousness (C), Extraversion (E), Agreeableness (A), and Neuroticism (N). Label 0 indicates an unidentified trait, while label 1 indicates an identified trait. The distribution of labels across the five dimensions is relatively balanced, thus not showing significant class imbalance. Preprocessing was performed according to the requirements of the two pipelines. Both apply lowercase and remove URLs, non-ASCII characters, numbers, and punctuation. In the vocabulary graph pipeline, the text is then processed by removing English stopwords and performing lemmatization to reduce word form variation [17], before being tokenized to build a global vocabulary. Meanwhile, BERT input undergoes only minimal preprocessing, with no stopword removal or lemmatization to preserve linguistic context and word relationships [5]. Examples of preprocessing results from both pipelines are shown in Table 1.

Table 1. Dataset Preprocessing Results

Original Text	Graph Preprocessing	BERT Preprocessing
it is wednesday. I can't wait until friday because I am going home to see brandon. …	wednesday cant wait friday going home see brandon …	it is wednesday. I can't wait until friday because I am going home to see brandon. …

2.3 Vocabulary Graph Construction

The graph component in this study is constructed as a vocabulary graph (VGraph) to capture global semantic relationships across the entire essay corpus [12]. Unlike sequential models that focus on local context, VGraph models word-to-word relationships more broadly across the entire dataset. Co-occurrence relationships are obtained using a 20-token sliding window on the training data, with each word pair represented as an edge in the graph. Edge weights are calculated using Normalized Pointwise Mutual Information (NPMI) to measure the strength of semantic associations in a normalized manner, resulting in more stable relationships during training. As in Equation 1, the NPMI score uses the probabilities of each word, $p(i)$ and $p(j)$, as well as the co-occurance probability $p(i,j)$, estimated based on the proportion of word pairs in the sliding window.

$$NPM(i,j) = -\frac{1}{log\ p(i,j)} log \frac{p(i,j)}{p(i)p(j)} \quad (1)$$

NPMI values range from −1 to 1, with positive values indicating stronger semantic associations. To optimize computational efficiency and reduce matrix density, an NPMI threshold is applied. Word pairs with values below the threshold are removed, while those that meet the threshold are retained as edges in the graph.

2.4 Transformer-Based Node Feature Extraction

Each unique word in the vocabulary graph is represented using a pretrained BERT embedding with bert-base-uncased checkpoints. The use of these checkpoints aims to maintain architectural consistency while obtaining richer semantic and contextual representations [5], [7]. Each word is tokenized using the WordPiece method and processed through the Transformer architecture to produce a hidden state representation [18]. Since a single word can consist of multiple subword tokens, a mean pooling strategy is used by calculating the average embedding of all its constituent subwords. The final representation is then arranged into a node feature matrix of size (5000 × 768), which represents 5,000 unique words with 768-dimensional vectors from the Transformer hidden layer.

DOI : https://doi.org/10.21063/jtif.2026.V14.2.100-109
Creative Commons Attribution-Share Alike 4.0 International License (CC BY-SA 4.0)

2.5 Vocabulary Graph Convolutional Network (VGCN)

Unlike traditional GCN architectures that generally build graphs at the document level, Vocabulary Graph Convolutional Network (VGCN) applies convolution directly to the relationships between words in the corpus-level vocabulary graph [12]. This approach allows the model to capture semantic relationships between words while simultaneously utilizing global structures in forming text representations [19]. For a document or sentence represented as a text vector x, the graph convolution operation is formulated as in Equation 2.

$$h = (\tilde{A}x^T)^T W = x\tilde{A}W \qquad (2)$$

In this equation, $\tilde{A}$ is the normalized vocabulary graph adjacency matrix, while W is a learnable weight matrix in the hidden layer. The multiplication of x and $\tilde{A}$ allows the model to dynamically select and assign more weight to parts of the global vocabulary graph that are relevant to the input document. For mini-batch processing, the operation on the first graph convolution layer is formulated as in Equation 3.

$$H = X\tilde{A}W \qquad (3)$$

To obtain the final graph-based representation, the VGCN architecture uses multiple graph convolution layers combined with a Rectified Linear Unit (ReLU) activation function. The structural feature extraction process is shown in Equation 4.

$$VGCN = ReLU(X_{mv}\tilde{A}_{vv}W_{vh})W_{hc} \qquad (4)$$

Where m denotes the mini-batch size, v is the vocabulary size, h is the hidden layer dimension, and c denotes the embedding dimension generated by the graph representation.

2.6 Hybrid VGCN-BERT Architecture Integration

The VGCN-BERT model combines contextual representations from Transformer with global semantic information obtained from VGCN [12]. The Transformer architecture is effective for personality detection because it is able to capture complex language patterns [20] as well as long-range semantic relationships in the entire input [21]. Both representations are processed through two parallel paths to preserve their individual characteristics before being combined. The graph representation from VGCN first goes through a linear transformation with the addition of bias to adjust its dimensions to the document representation from BERT. The projection process is formulated in Equation 5.

$$G_{embedding} = ReLU\big(X_{mv}\tilde{A}_{vv}W_{vh}\big)W_{hg} + b_g \qquad (5)$$

Where W_{hg} and b_g are the learnable projection weight and bias matrices, respectively. Both are used to adjust the embedding dimension of the VGCN to match the output dimension of the representation generated by the BERT encoder. The overall processing flow, from text tokenization to multi-label classification, is shown in Figure 2.

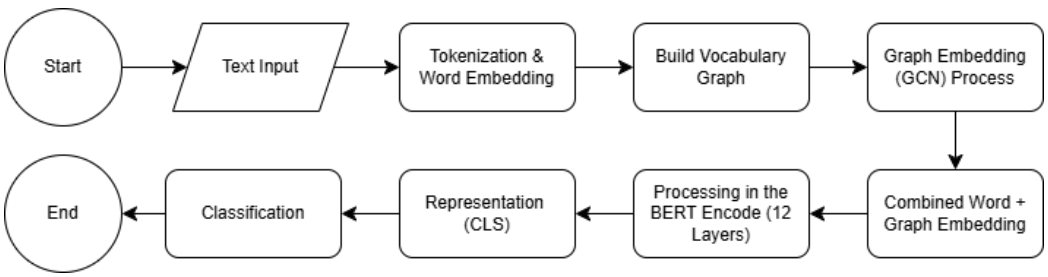

Figure 2. Flowchart of VGCN-BERT model

DOI : https://doi.org/10.21063/jtif.2026.V14.2.100-109
Creative Commons Attribution-Share Alike 4.0 International License (CC BY-SA 4.0)

The combined representation resulting from both paths is called a graph-enriched embedding. This representation combines local and global information to provide a more complete picture of the text's characteristics. The embedding enriched through graph information is passed to a fully connected classification layer to generate predictions in a multi-label scheme [22]. In the final stage, a sigmoid activation function is applied to obtain probability values independently on each Big Five (OCEAN) personality dimension.

2.7 Training, Validation, and Evaluation Scenario

To systematically evaluate various hyperparameter combinations, model optimization was performed using a grid search strategy [15], [23]. The search space used included learning rate (1e-5 and 2e-5), dropout (0.2), maximum sequence length (200 and 256), NPMI threshold (0.2), GCN hidden layer dimension (128), number of GCN layers (1 and 2), and graph scale value (0.0, 0.5, and 1.0). The training process used the AdamW optimizer [24], Binary Cross-Entropy with Logits Loss, batch size 16, maximum 10 epochs, and random seed 42. To maintain the stability of the training process and reduce the risk of overfitting, a learning rate scheduler, gradient clipping, and early stopping with patience for three epochs were used based on the macro F1-score value on the validation data [25]. In addition to using the default threshold of 0.5, this study also applies empirical threshold tuning to out-of-fold (OOF) predictions to obtain a threshold value that can maximize the F1-score for each personality dimension [16]. The optimal threshold obtained is then locked and applied to the test set that is not used during the training or tuning process. The final classification performance is then evaluated using metrics obtained from the confusion matrix.

3. RESULTS AND DISCUSSION

This section presents a comprehensive empirical evaluation of the proposed VGCN-BERT model [12] for multi-label classification of the Big Five personality traits. The evaluation is conducted systematically by presenting the best hyperparameter configuration obtained through grid search with 5-fold cross-validation, followed by an analysis of the model's overall classification performance. One of the main focuses of this analysis is the empirical influence of decision threshold optimization on each personality trait [16]. By evaluating out-of-fold (OOF) predictions, this study demonstrates the importance of using an optimized threshold, rather than simply using the default value, to obtain a more optimal F1-score. In addition, a confusion matrix is used to provide a more in-depth overview of the trade-off between precision and recall. The evaluation results are also used to analyze the model's algorithmic tendency to provide positive predictions and demonstrate how threshold calibration can help balance high sensitivity with acceptable precision.

3.1 Experimental Results and Evaluation

A grid search procedure combined with 5-fold cross-validation [15], successfully determined the best hyperparameter configuration for the proposed VGCN-BERT model in multi-label classification of the Big Five personality traits. The best-performing model configuration was obtained using a learning rate of 1e-5, a dropout of 0.2, a maximum sequence length of 200 tokens, an NPMI threshold of 0.2, a GCN hidden layer dimension of 128, a single GCN layer, and a graph scale of 1.0. Based on this configuration, VGCN-BERT achieved optimal performance with a macro F1 score of 0.682719 and a weighted F1 score of 0.682902.

To obtain a more detailed overview of the performance for each personality trait, Table 2 displays the precision, recall, F1, and accuracy [26] values calculated based on the optimal out-of-fold (OOF) decision threshold. The evaluation results show that the model consistently produces near-perfect recall values (~1.0), while its precision values are at a moderate level, around 0.50–0.56. This condition results in a relatively consistent F1 score in the range of 0.66–0.70. This combination of metrics indicates a strong algorithmic tendency to predict the positive class. With high sensitivity, the model is able to identify most of the positive samples, but at the same time it also produces a large number of false positives.

DOI : https://doi.org/10.21063/jtif.2026.V14.2.100-109
Creative Commons Attribution-Share Alike 4.0 International License (CC BY-SA 4.0)

Table 2. Evaluation of Each Trait Based on Optimal Threshold

Trait	Precision	Recall	F1-Score	Accuracy
O	0.5585	0.9213	0.6955	0.5843
C	0.5109	0.9900	0.6740	0.5133
E	0.5191	0.9976	0.6829	0.5203
A	0.5311	1.0000	0.6937	0.5317
N	0.5010	1.0000	0.6675	0.5025
Macro Average	0.5241	0.9818	0.6827	0.5304
Weighted Average	0.5243	0.9818	0.6829	0.5304

3.2 Threshold Tuning Evaluation

To assess the effectivness of the threshold tuning strategy [16], a comparison was conducted between the default threhold of 0.5 and the optimal threshold determined based on the F1 score and its change (ΔF1), as shown in Table 3. The evaluation results showed that using the default threshold did not produce optimal performance. Conversely, using a lower decision threshold, in the range of 0.22–0.27, consistently produced the highest F1 scores across all personality traits. This finding suggests that threshold tuning is necessary to improve classification performance. Neuroticism experienced the most significant improvement, with an F1 score increase of 0.1963, from 0.4712 to 0.6675. Performance improvements were also found for other personality traits. Next, to determine the effect of threshold selection on other evaluation metrics, particularly subset accuracy, the model was re-evaluated using a predetermined threshold range of 0.50 to 1.00 with a 0.01 interval. The evaluation results are presented in more detail in Table 4.

Table 3. Comparison of F1 Scores at Default and Optimal Thresholds

Trait	Optimal Threshold	F1(0.5)	F1(Optimal)	ΔF1
O	0.22	0.5920	0.6955	0.1035
C	0.24	0.5236	0.6740	0.1504
E	0.27	0.5674	0.6829	0.1155
A	0.27	0.6194	0.6937	0.0743
N	0.23	0.4712	0.6675	0.1963

Using the default threshold of 0.5, the model achieved the highest macro-average accuracy of 0.547655 and a weighted average accuracy of 0.547948. However, this increase in accuracy was accompanied by a lower overall F1 score compared to using a lower optimal threshold. This finding suggests a trade-off in thresholding. A higher threshold tends to improve overall label classification accuracy, but simultaneously reduces the model's ability to detect positive samples. This condition then results in a decrease in the overall recall value and F1 score.

Table 4. Performance Evaluation Based on Personality Traits

Trait	Threshold	Precision	Recall	F1-Score	Accuracy
O	0.5	0.6378	0.5523	0.5920	0.6077
C	0.5	0.5436	0.5050	0.5236	0.5330
E	0.5	0.5444	0.5924	0.5674	0.5323
A	0.5	0.5663	0.6834	0.6194	0.5545
N	0.5	0.5119	0.4365	0.4712	0.5108
Macro Average	-	0.5608	0.5539	0.5547	0.5477
Weighted Average	-	0.5612	0.5556	0.5557	0.5479

DOI : https://doi.org/10.21063/jtif.2026.V14.2.100-109
Creative Commons Attribution-Share Alike 4.0 International License (CC BY-SA 4.0)

3.3 Discussion

To identify the prediction patterns generated by the model, Figure 3 displays the confusion matrix of the best-performing configuration. Across all personality traits, the high number of True Positives (TP) and low number of False Negatives (FN) indicates near-perfect recall, although the precision obtained remains at a moderate level. This finding indicates a strong algorithmic bias in predicting the positive class. High sensitivity allows the model to recognize most potential indicators of personality traits in text, but also results in a relatively high number of False Positives. This pattern indicates that although graph-based representations are capable of capturing broader semantic information, the model still tends to over-predict positives. This condition is a common challenge in multi-label psychological prediction tasks. To further examine this pattern, Figure 4 shows that the F1 score peaks at a low threshold range of 0.22–0.27, then decreases as the threshold value increases. This finding aligns with the characteristic of the F1 score, which balances precision and recall [26], and also reinforces that using the default threshold of 0.5 does not produce optimal performance. Next, to examine the impact of applying stricter decision criteria, Figure 5 presents the confusion matrix obtained using a higher threshold, namely in the range 0.50–1.00.

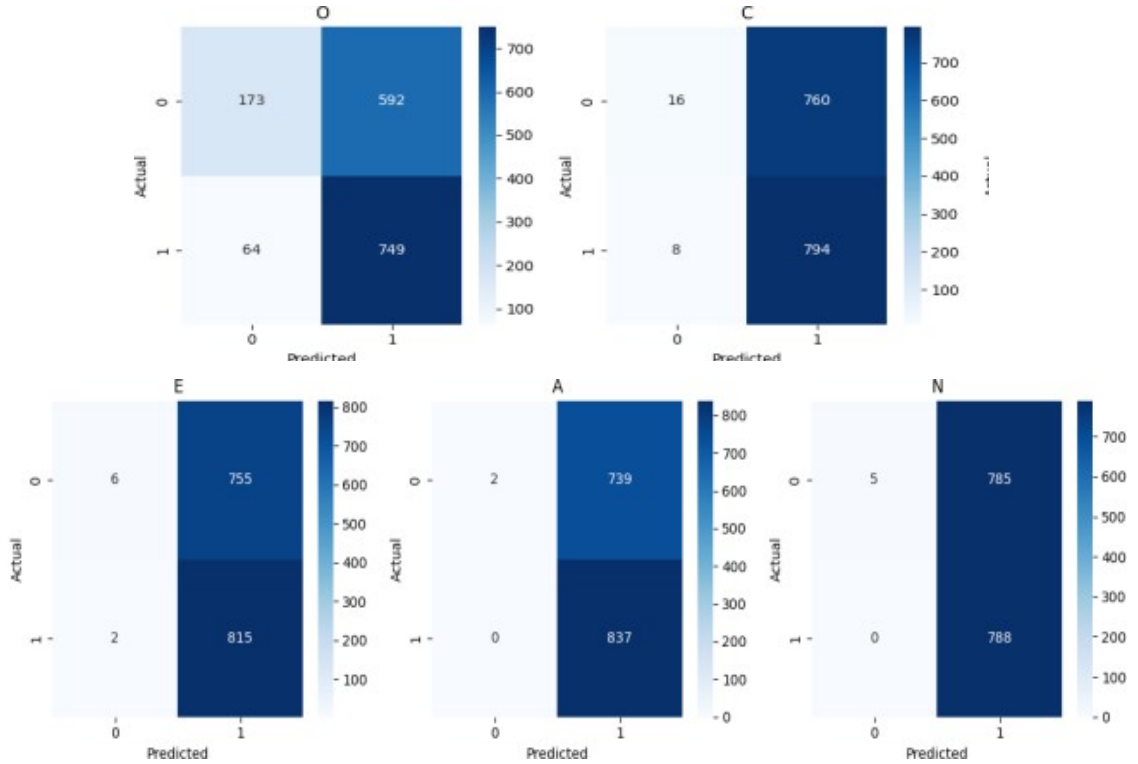

Figure 3. Confusion Matrix of the Best Model Configuration for the Five Personality Traits

Compared to Figure 3, applying a higher threshold yields a different pattern, characterized by an increase in True Negatives (TN) and a decrease in False Positives (FP), resulting in improved precision and accuracy. However, this is accompanied by an increase in False Negatives (FN) and a decrease in recall. This finding reinforces the trade-off between precision and recall in essay-based personality classification. A lower threshold tends to improve the F1 score through increased sensitivity, reflected in a high recall value, but results in more FPs. Conversely, a higher threshold prioritizes precision but potentially overlooks a number of true positive samples, resulting in decreased recall and increased FN. This trade-off is one of the main reasons why using the optimized threshold is able to produce better performance than the default threshold of 0.5.

Based on the evaluation at the personality trait level, the model exhibits near-perfect recall (~1.0), while precision remains at a moderate level, at 0.50–0.56. This pattern indicates a strong algorithmic bias in predicting the positive class. The high sensitivity makes the model relatively rare in missing potential personality trait indicators in the text, as indicated by the low number of False Negatives (FN). However,

DOI : https://doi.org/10.21063/jtif.2026.V14.2.100-109
Creative Commons Attribution-Share Alike 4.0 International License (CC BY-SA 4.0)

this condition also results in a significant number of False Positives (FP). Thus, although graph-based representations help the model capture broader semantic information, the model still has a tendency to over-label positives, which is one of the challenges in multi-label psychological classification.

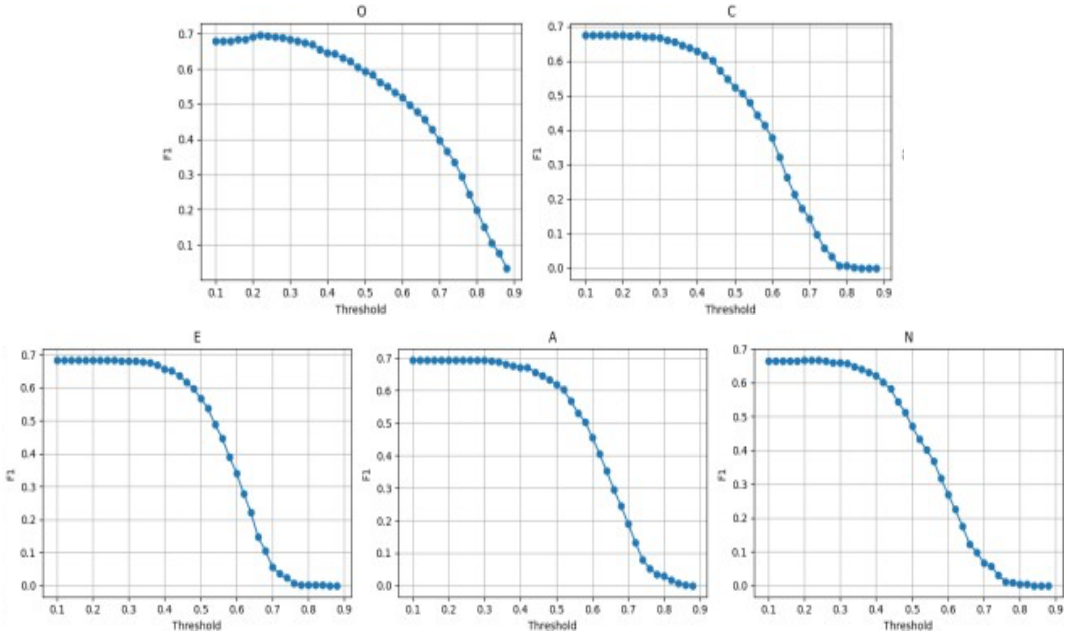

Figure 4. F1 Score Curves against Threshold for Five Personality Traits

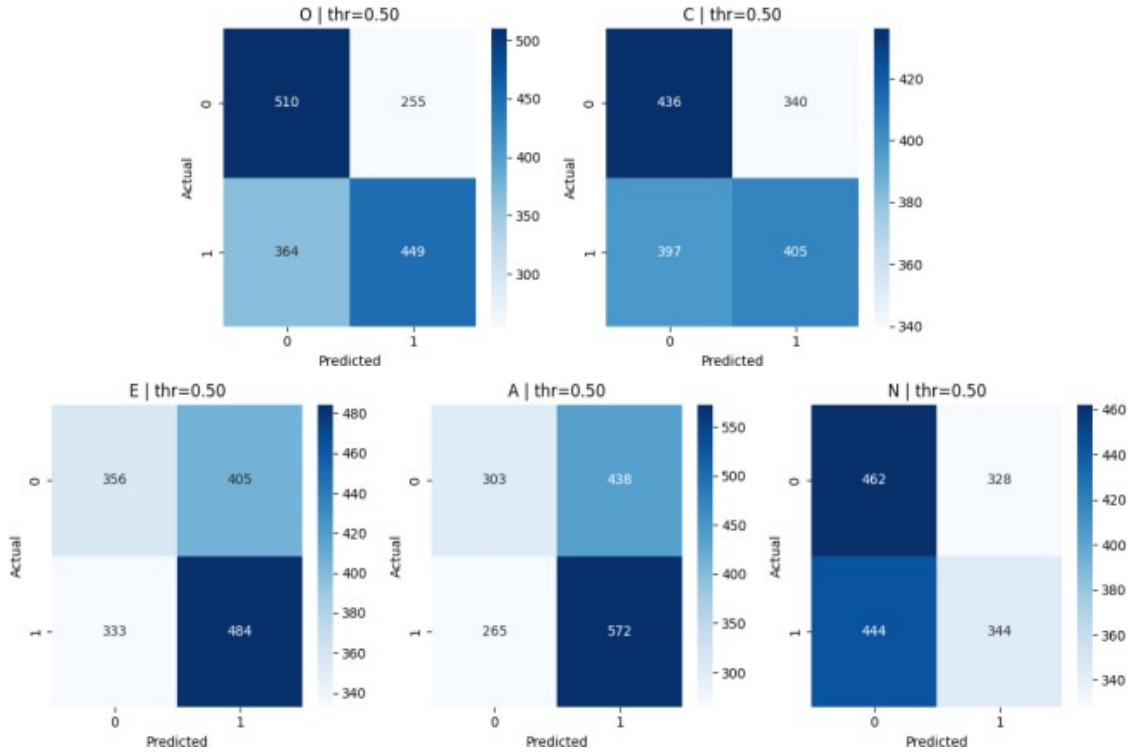

Figure 5. Confusion Matrix Based on Threshold Variation 0.50–1.00

4. CONCLUSION

This study demonstrates that the integration of local contextual representations from BERT with global semantic structures through a weighted vocabulary graph (VGCN-BERT) is capable of supporting essay-based Big Five personality trait classification. Through systematic hyperparameter and threshold tuning,

DOI : https://doi.org/10.21063/jtif.2026.V14.2.100-109
Creative Commons Attribution-Share Alike 4.0 International License (CC BY-SA 4.0)

the model achieved the highest macro F1 score of 0.6827 and a weighted F1 score of 0.6829. The evaluation results indicate that empirical threshold tuning plays a significant role in improving model performance, with the optimal threshold value consistently below the default value of 0.5. Error analysis also indicates an algorithmic tendency to over-predict positives, characterized by near-perfect recall but moderate precision. This condition indicates a trade-off between precision and recall when the model is optimized to achieve high sensitivity in multi-label psychological classification tasks.

Overall, the results provide empirical evidence that the VGCN-BERT adaptation has the potential to be used in personality prediction tasks. However, this study still has several limitations. The evaluation was conducted on only one English-language dataset, while the use of binary labels can simplify the complexity of personality characteristics. Furthermore, the reliance on specific threshold configurations potentially increases the model's susceptibility to demographic bias and domain shift. The integrated pipeline also limits the implementation of ablation studies with separate baselines. Therefore, future research should consider external validation using more diverse datasets, fairness evaluation, more robust baselines, and more comprehensive multi-label evaluation metrics, such as PR-AUC and precise subset accuracy. Automatic hyperparameter optimization can also be implemented to improve the model's robustness and generalizability.

REFERENCES

[1] E. Kerz, Y. Qiao, S. Zanwar, and D. Wiechmann, "Pushing on Personality Detection from Verbal Behavior: A Transformer Meets Text Contours of Psycholinguistic Features," *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pp. 182–194, Jan. 2022, doi: 10.18653/v1/2022.wassa-1.17.

[2] R. R. McCrae and O. P. John, "An Introduction to the Five-Factor Model and Its Applications," *J. Pers.*, vol. 60, no. 2, pp. 175–215, Jun. 1992, doi: 10.1111/j.1467-6494.1992.tb00970.x.

[3] Z. Ren, Q. Shen, X. Diao, and H. Xu, "A sentiment-aware deep learning approach for personality detection from text," *Inf. Process. Manag.*, vol. 58, no. 3, May 2021, doi: 10.1016/j.ipm.2021.102532.

[4] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, "Deep Learning-Based Text Classification," Apr. 30, 2022, *Association for Computing Machinery*. doi: 10.1145/3439726.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2019, pp. 4171–4186. doi: 10.18653/v1/N19-1423.

[6] M. Koroteev V., "BERT: A review of Applications in Natural Language Processing and Understanding," *arXiv (Cornell University)*, Mar. 2021, doi: 10.48550/arxiv.2103.11943.

[7] A. Subakti, H. Murfi, and N. Hariadi, "The performance of BERT as data representation of text clustering," *J. Big Data*, vol. 9, no. 1, Dec. 2022, doi: 10.1186/s40537-022-00564-9.

[8] A. Kazameini, S. Fatehi, Y. Mehta, S. Eetemadi, and E. Cambria, "Personality Trait Detection Using Bagged SVM over BERT Word Embedding Ensembles," *arXiv (Cornell University)*, Oct. 2020, doi: 10.48550/arxiv.2010.01309.

[9] M. Reza, F. Derakhshi, M. A. Balafar, M. Ramezani, M.-R. Feizi-Derakhshi, and M.-A. Balafar, "Text-Based Automatic Personality Prediction Using KGrAt-Net; A Knowledge Graph Attention Network Classifier", doi: 10.48550/arXiv.2205.13780.

[10] M. Waqas *et al.*, "TraitBertGCN: Personality Trait Prediction Using BertGCN with Data Fusion Technique," *International Journal of Computational Intelligence Systems*, vol. 18, no. 1, Dec. 2025, doi: 10.1007/s44196-025-00792-w.

[11] Y. Lin et al., "BERTGCN: Transductive text classification by combining GCN and BERT," *arXiv (Cornell University)*, May 2021, doi: 10.48550/arxiv.2105.05727.

[12] Z. Lu, P. Du, and J.-Y. Nie, "VGCN-BERT: Augmenting BERT with Graph Embedding for Text Classification," in *Advances in Information Retrieval* (ECIR 2020), Lecture Notes in Computer Science, vol. 12035, J. M. Jose et al., Eds. Cham: Springer International Publishing, 2020, pp. 369–382, doi: 10.1007/978-3-030-45439-5_25.

DOI : https://doi.org/10.21063/jtif.2026.V14.2.100-109
Creative Commons Attribution-Share Alike 4.0 International License (CC BY-SA 4.0)

[13] E. Mamani-Condori and J. Ochoa-Luna, "Aggressive language detection using VGCN-BERT for Spanish texts," *Lecture Notes in Computer Science*, pp. 359–373, Jan. 2021, doi: 10.1007/978-3-030-91699-2_25.

[14] J. W. Pennebaker and L. A. King, "Linguistic styles: Language use as an individual difference.," *J. Pers. Soc. Psychol.*, vol. 77, no. 6, pp. 1296–1312, 1999, doi: 10.1037/0022-3514.77.6.1296.

[15] A. S. Aribowo, N. H. Cahyana, and Y. Fauziah, "Enhancing Semi-Supervised Sentiment Analysis Through Hyperparameter Tuning Within Iterations: A Comparative Study Using Grid Search and Random Search," 2024, pp. 248–260. doi: 10.2991/978-94-6463-366-5_23.

[16] Z. C. Lipton, C. Elkan, and B. Narayanaswamy, "Thresholding classifiers to maximize F1 score," *arXiv (Cornell University)*, Feb. 2014, doi: 10.48550/arxiv.1402.1892.

[17] M. Apidianaki, "From Word Types to Tokens and Back: A Survey of Approaches to Word Meaning Representation and Interpretation," 2023, doi: 10.1162/coli_a_00474.

[18] T. Wolf et al., "Transformers: State-of-the-Art Natural Language Processing," *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, Jan. 2020, doi: 10.18653/v1/2020.emnlp-demos.6.

[19] B. Xue, C. Zhu, X. Wang, and W. Zhu, "The Study on the Text Classification Based on Graph Convolutional Network and BiLSTM †," *Applied Sciences (Switzerland)*, vol. 12, no. 16, Aug. 2022, doi: 10.3390/app12168273.

[20] 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.

[21] T. Agrawal, D. Agarwal, M. Balazia, N. Sinha, and F. Bremond, "Multimodal Personality Recognition using Cross-Attention Transformer and Behaviour Encoding," Jan. 2023, doi: 10.5220/0010841400003124.

[22] O. Galal, A. H. Abdel-Gawad, and M. Farouk, "Rethinking of BERT sentence embedding for text classification," *Neural Comput. Appl.*, vol. 36, no. 32, pp. 20245–20258, Nov. 2024, doi: 10.1007/s00521-024-10212-3.

[23] X. Liu and C. Wang, "An Empirical Study on Hyperparameter Optimization for Fine-Tuning Pre-trained Language Models," *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pp. 2286–2300, Jan. 2021, doi: 10.18653/v1/2021.acl-long.178.

[24] N. Gkouti, P. Malakasiotis, S. Toumpis, and I. Androutsopoulos, "Should I try multiple optimizers when fine-tuning pre-trained Transformers for NLP tasks? Should I tune their hyperparameters?," Feb. 2024, [Online]. Available: http://arxiv.org/abs/2402.06948

[25] J. Dodge, G. Ilharco, R. Schwartz, A. Farhadi, H. Hajishirzi, and N. Smith, "Fine-Tuning Pretrained Language Models: Weight Initializations, Data Orders, and Early Stopping," Feb. 2020, [Online]. Available: http://arxiv.org/abs/2002.06305

[26] M. C. Hinojosa Lee, J. Braet, and J. Springael, "Performance Metrics for Multilabel Emotion Classification: Comparing Micro, Macro, and Weighted F1-Scores," *Applied Sciences (Switzerland)*, vol. 14, no. 21, Nov. 2024, doi: 10.3390/app14219863.

DOI : https://doi.org/10.21063/jtif.2026.V14.2.100-109
Creative Commons Attribution-Share Alike 4.0 International License (CC BY-SA 4.0)