Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871

Vol. 7, No. 4, August 2026, Page. 3579-3598
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.4.5444

Enhancing Big Five Personality Prediction on Indonesian Tweets: A Comparative Study of LSTM-GRU Hybrid and IndoBERT/IndoXLNet with Augmentation Techniques

Izzatul Ummah[*1], Fitriyani[2], Novia Natasya[3]

[1,2,3]School of Informatics, Telkom University, Indonesia

Email: [1]izzatulummah@telkomuniversity.ac.id

Received : Nov 13, 2025; Revised : Jan 20, 2026; Accepted : Feb 1, 2026; Published : Aug 17, 2026

Abstract

Over the past few years, researches on Big Five personality prediction model focusing on Indonesian tweets has mostly faced data scarcity and imbalance problem. This study aims to build Big Five personality prediction models for the low-resource language setting using Indonesian social media texts in platform X, and conduct a comparative performance analysis using deep learning model (hybrid BiLSTM-BiGRU) and transferred learning models (IndoBERT and IndoXLNet) with the optimization using various data augmentation and feature expansion techniques to overcome data scarcity and imbalance. We carried out data collection of 11746 tweets from hybrid sources (BFI questionnaire and keyword-based crawling), preprocessing using tokenization and stemming, extraction using Word2Vec and FastText, data augmentation (using SYN-REP, BACK-TRANS, RAND-INS, and SMOTE variants), Top-N feature expansion (with N = 1, 5, and 10), hyperparameter tuning using Optuna (for learning rate, batch size, dropout rate, and weight decay), and evaluating the model's accuracy, precision, recall, and F1-score. The result shows that IndoBERTweet+BACK-TRANS+Top-10 achieved 84% accuracy and an F1-score of 81.5%, surpassing IndoXLNet+BACKTRANS+Top-5 (82%) and LSTM+GRU+FastText+KMeansSMOTE+Top-5 (78%), with the BACK-TRANS technique improves overall F1 result by 13%. These findings bolster computational psychology informatics for low-resource NLP focusing on Indonesian language, as well as aiding mental health screening with 18% better generalization in imbalanced data, and enabling scalable trait inference in human resource and mental health applications.

Keywords : *Big Five Personality, Data Augmentation, Feature Expansion, IndoBERT, IndoXLNet, LSTM-GRU.*

This work is an open access article and licensed under a Creative Commons Attribution-Non Commercial 4.0 International License

1. INTRODUCTION

The rapid growth of social media has made it a tool for everyone not only to communicate but also to express themselves, their thoughts, feelings, and personality. In psychology literature, the Big Five Personality Model is one of the most widely referenced, due to its simplicity and comprehensiveness. The Big Five Personality Model provides a descriptive framework for five aspects of personality: Openness, Extraversion, Agreeableness, Conscientiousness, and Neuroticism [1]. In his landmark article, Goldberg analyzed the factors that form the structure of the Big-Five Personality Model by clustering thousands of trait adjectives and common terms [2]. Extraversion is generally related to a person's level of socialization. Agreeableness is related to a person's sympathetic character. Conscientiousness is related to a person's ability to self-control, discipline, etc. Neuroticism is related to emotional instability. Meanwhile, Openness is related to creativity and openness to new things or ideas. Various instruments in the form of questionnaires have also been developed to assess personality based on the five aspects of the Big Five personality, one of which is the Big Five Personality Inventory (BFI questionnaire).

Vol. 7, No. 4, August 2026, Page. 3579-3598
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.4.5444

In the text-based machine learning paradigm, data scarcity and imbalance are often the focal challenge to developing stable models, especially for low-resource setting such as in Indonesian text dataset sources. Over the past few years, researches on Big Five personality prediction using Indonesian text dataset has mostly faced data scarcity and imbalance problem [3]. Recent studies emphasize that while deep learning models like BERT and XLNet have achieved state-of-the-art results in personality prediction, their application in non-English and low-resource languages remains challenging due to the lack of labeled datasets. Chi et al. [4] highlight that leveraging pretrained language models can benefit non-English systems but often requires effective cross-lingual transfer or robust data enhancement strategies. Furthermore, Santos et al. [5] demonstrated that utilizing language-agnostic representations and word similarities is crucial for bridging the resource gap in personality factor recognition, allowing models to generalize better across different linguistic contexts. Arora and Turcan illustrate that data augmentation techniques are very promising in improving the performance of emotion classification in low-resource scenarios. They concluded from their experiments that Back Translation can generate more unigram and trigram diversity than autoencoder-based approaches and provides performance improvements when generating more than one additional example per training data point. Their results confirm that data augmentation methods can form a significant component of the solution to data scarcity issues and avoiding overfitting in deep learning models [6].

Furthermore, Tapia-Téllez and Escalante demonstrated that the application of data augmentation strategies for text classification based on Transformer models such as BERT can significantly improve performance. Their study showed that generating synthetic variations of training data helps models to generalize better by increasing lexical and semantic diversity while maintaining label consistency. Experimental results revealed consistent gains in classification metrics, particularly F1-score, across different datasets. The authors concluded that data augmentation is an effective strategy for reducing class imbalance and enhancing the robustness of text classification models [7].

Nudin reviews deep learning methods for predicting the Big Five personality traits using multi-task classification, analyzing 22 studies across various data modalities. It reports varying accuracy levels, with the highest from unimodal image data (95.36%) and the lowest at 59%, highlighting both the potential and limitations of current approaches. The study emphasizes the need for standard benchmarks, model interpretability, and interdisciplinary collaboration [8].

Personality reflects a consistent pattern in an individual's thoughts, emotions, and behaviors. The widely used Big Five (OCEAN) model categorizes personality into five key traits: Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. Aparna explores the emerging field of personality trait classification using deep learning by applying a Bidirectional LSTM network to identify traits based on the OCEAN framework [9].

Ong developed a personality prediction system based on Indonesian Twitter social media posts. The algorithms used were Support Vector Machine and XGBoost. The accuracy results obtained were 76.23% for SVM and 97.99 for XGBoost [10]. Jeremy conducted an experiment to automatically predict personality based on the Big Five Personality Traits, using Indonesian Twitter post data. The results of their experiment produced an F-Measure of 0.7482 [11]. Julianda developed a personality detection system based on the MBTI (Myers-Briggs Type Indicator) scheme, using Reddit user data and the DistilBERT language model. In conclusion, DistilBERT outperformed other classifiers such as Naive Bayes, SVM, and Logistic Regression [12].

In a related study, Natasya conducted a comparative analysis of the BERT and XLNet models for Big Five personality prediction using a dataset of 9,708 Indonesian tweets from platform X [3]. This dataset is also utilized as the second source in our current research. Their work found that BERT slightly outperformed XLNet in terms of average F1-score (43.362% vs. 42.822%) and was significantly more efficient, demonstrating a much faster inference time (215.72 ms vs. 7714.83 ms). However, they also

Vol. 7, No. 4, August 2026, Page. 3579-3598
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.4.5444

noted that performance for both models was hindered by the dataset's significant class imbalance, a challenge our paper aims to address through various data augmentation and feature expansion techniques.

This research also draws upon previous work demonstrating the effectiveness of feature expansion for Indonesian text. Setiawan introduced a hybrid method for sentiment analysis on Twitter, finding that feature expansion based on tweet-based features significantly improved performance, achieving an accuracy of 98.81% with Logistic Regression [13]. Building on this concept, Royyan and Setiawan specifically explored feature expansion using Word2Vec to address vocabulary mismatches in public policy tweets. Their method, which involved building a similarity corpus, successfully increased the SVM classifier's accuracy to 78.99% [14]. More recently, Widiarta and Setiawan applied FastText for feature expansion within a hybrid CNN-BiGRU model for depression detection on platform X. Their results confirmed the technique's value, achieving an accuracy of 84.40% and concluding that FastText, combined with an attention mechanism, was crucial for boosting performance [15].

We conclude the gaps in those research above, i.e., very few usage of data augmentation for Indonesian dataset from X platform resulting the still poor performance result which only achieved F1-score less than 75% in [3], [11], [12]. Another gap is that ethical labeling by multi-anotator has not been done in [3]. Therefore this research is submitted to build a Big Five personality prediction model based on low resource and ethically labelled Indonesian tweets dataset, optimize it using various data augmentation and feature expansion techniques to address imbalance issue, and compare the performance of deep learning model (GRU-LSTM hybrid) and transferred learning models (IndoBERT and IndoXLNet), with the goal to achieve F1-score more than 80%. For the deep learning model, we incorporate the bidirectional architecture and attention mechanism into the model, in order to make it fair in comparison with BERT and XLNet. The social media that will be used as the dataset source for this research is platform X (formerly known as Twitter).

2. METHOD

The method used in this study is described in Figure 1.

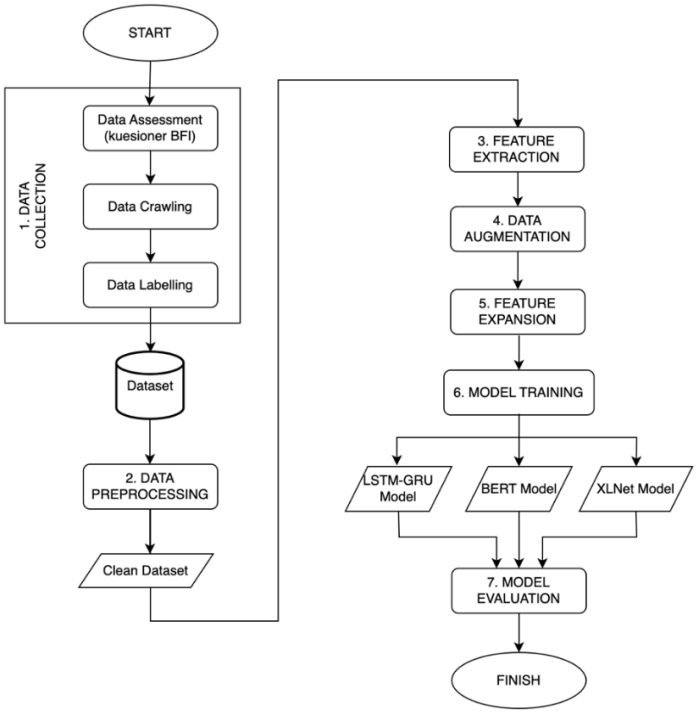


Figure 1. End-to-End Pipeline: From X Crawling to Augmented Personality Classification

2.1. Data Collection

We collect tweets data in Indonesian language from social media platform X. In order to produce a good quality of dataset, we decide to use hybrid sources. The first source is from X users who have filled the Big Five Inventory questionnaire, while the second source is from X users who have not filled the Big Five Inventory questionnaire and therefore their data will have to be labelled manually. The process of collecting data for the second source has been done by Natasya [3], and we will reuse their dataset for our current work.

For the first source, the process of collecting data consists of three stages:

1. Data Assessment: We distribute Big Five Inventory (BFI) questionnaires which contains 44 questions to assess personality type of the respondents. Respondents are asked to give their consent to allow their data to be crawled from their social media. In this stage, we achieve Institutional Review Board (IRB) ethical consent rate of 89% from our BFI respondents.
2. Data Crawling: The task of crawling tweets data is carried out using API (application programming interface) Twitter.
3. Data Labelling: The task of labelling the dataset with the personality type is carried out automatically using the result of Big Five Inventory questionnaires.

For the second source, the process of collecting data also consists of three stages:

1. Keyword Listing: In this stage, we asked an expert in psychology to carefully list several keywords that is suitable for acquiring data for this study. We came out with a list containing 400 keywords.
2. Data Crawling: We crawl data from public X accounts using the keyword listed above.
3. Data Labelling: The task of labelling the dataset with the personality type is carried out manually by three annotators (psychologists). We use voting mechanism to establish the final label for the data if the annotators had differing viewpoints during the labelling process. We also use Cohen's Kappa formula to calculate the agreement level between the three annotators.

Data labelling from both sources is done using this categorical ordering (single label, multi-class classification): 1 for extraversion, 2 for agreeableness, 3 for conscientiousness, 4 for neuroticism, and 5 for openness. In this study we will be focusing only on the linguistic feature of the tweets data, so remove the social features of the tweets data such as User ID, Username, Tweet Datetime, Number of Reply/Retweet/Quote, etc.

2.2. Data Preprocessing

In order to prepare the dataset to be ready for the next process, the data preprocessing task needs to be done first. This process includes:

a. Lowercasing (changing all tweets data into lower case only).
b. Removing HTML and URL.
c. Removing punctuation, symbol, emoji, etc.
d. Chat word treatment (normalization): changing informal chat words into formal words.
e. Removing stop words.
f. Tokenization: splitting each sentences into tokens.
g. Stemming/lemmatization: changing complex words into base words by removing prefix, infix, sufix. This task will be carried out using StemmerFactory package from Sastrawi library for Indonesian language processing.

Vol. 7, No. 4, August 2026, Page. 3579-3598
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.4.5444

2.3. Feature Extraction

Feature extraction is the process of transforming raw text into structured numerical forms for processing by machine learning algorithms. The approach makes data simpler in addition to improving model performance and generalization [16]. For this research, we will use Word2Vec and FastText to extract features from tweets data, and compare the performance. Word2Vec is remarkable through its Skip-Gram or CBOW architecture in achieving semantic word relationships from context of words in proximity [17]. At the same time, FastText extends embeddings by incorporating subword information (n-gram characters) to render it less susceptible to infrequent words and out-of-vocabulary (OOV) words, a valuable aspect for languages with complex morphology like Indonesian [18]. These two methods together enhance feature representation in text classification, particularly for social media informal data.

2.4. Data Augmentation

Data augmentation techniques are used to artificially expand a training dataset by creating modified versions of existing text data. The purpose is to make our dataset size to be larger and balanced. There are four techniques that we will evaluate in this paper:

1. Synonym replacement (SYN-REP): Replace words in dataset with their synonyms to introduce lexical diversity while preserving the overall meaning of the text.
2. Back translation (BACK-TRANS): Translate the text into English language and then back into the Indonesian language to create paraphrased versions.
3. Random insertion (RAND-INS): Select a random word, find its synonym, and insert it in random position in the text.
4. Oversampling: We increase the number of samples in the minority class of an imbalanced dataset using SMOTE technique and its variants (ADASYNC, and KMeansSMOTE) [19].

Data augmentation has been shown to enhance model performance through diversifying the data and avoiding overfitting in low-resource and imbalanced settings. Current study form Shi et al. highlight that augmentation can enhance accuracy and F1-score by as much as 20% over non-augmented baselines [20]. These studies emphasize that augmentation is not only useful for increasing dataset sizes but is essential in order to achieve more robust and more balanced performance across different classes.

2.5. Feature Expansion

Feature expansion is a technique to increase the dimension of the features, thus making it more complex. In this study we use the feature expansion technique similar to the method that has been done by Setiawan et. al. Regression [13]. First we build a similarity corpus using our tweet dataset and IndoNews dataset, and utilize word embedding technique such as Word2Vec or FastText. And then based on the similarity corpus, we find the N-top most similar tokens for each tokens in our tweet dataset. We then add those N-top most similar tokens into each data in our dataset, thus transforming it into a higher dimensional dataset.

This approach has shown success in related research. Royyan applied feature expansion using Word2Vec to address vocabulary mismatches in Indonesian-language tweets about public policy. By building a similarity corpus from a combination of news data and the tweet dataset, they successfully increased the accuracy of their SVM classifier to 78.99% [14]. In a study more aligned with the models in this paper, Widiarta and Setiawan utilized FastText for feature expansion within a hybrid CNN-BiGRU model for depression detection. Their results confirmed the technique's value, achieving a final accuracy of 84.40%, and they concluded that FastText was crucial for handling out-of-vocabulary words and boosting overall model performance [15].

Vol. 7, No. 4, August 2026, Page. 3579-3598
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.4.5444

Table 1 below shows an example of Top-10 most similar words for the word 'kerja' in our dataset.

Table 1. Top-10 most similar tokens for 'kerja'

Rank	Tokens
1	'kantor'
2	'tugas'
3	'disiplin'
4	'gaji'
5	'proyek'
6	'tanggungjawab'
7	'jabatan'
8	'rapat'
9	'meeting'
10	'wewenang'

2.6. Model Training

We conduct the model training process using k-fold cross validation technique with k = 10, which means there are ten rounds of trainings and in each training we split the dataset 9:1 (9 for training, 1 for testing). There are three models of deep learning that we are going to evaluate, namely BiLSTM-BiGRU, BERT, and XLNet. These models are implemented using PyTorch version 2.3.0, an open-source deep learning framework that supports GPU acceleration, automatic differentiation, and dynamic computation graphs. Such features make PyTorch well-suited for natural language processing tasks and enable efficient implementation of state-of-the-art models like BERT and XLNet [21].

2.6.1. BiLSTM and BiGRU

Bidirectional Long Short-Term Memory (BiLSTM) is an extension of the LSTM model to capture context data from both future and past directions. In BiLSTM architecture, as shown in Figure 2 (left), two LSTM layers are run in parallel: one reads data sequence from the past to the future ($\overrightarrow{h_t}$) and the other from the future to the past. The representations from both sides are then combined to form a stronger representation, enabling the model to understand the context better. The main advantage of BiLSTM is that it can utilize long-term dependencies from both sides of the sequence, which has been successful in many NLP tasks such as text analysis, emotion analysis, and identification of fake news [22].

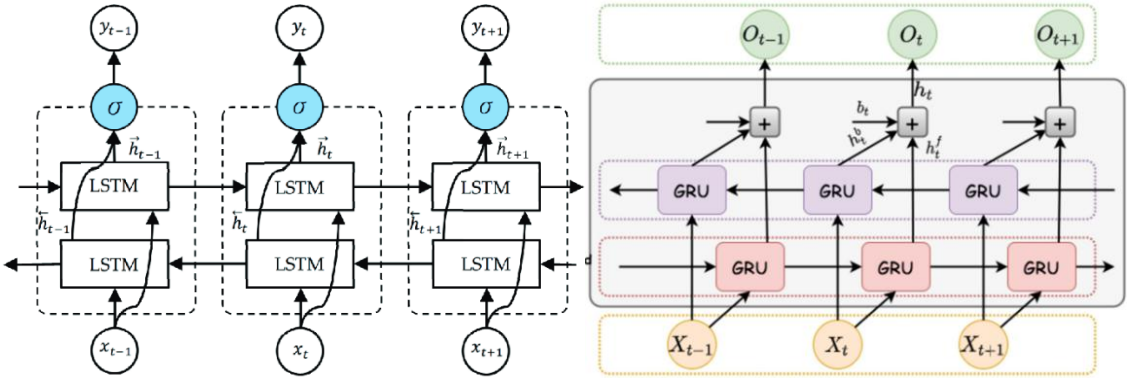


Figure 2. Bi-LSTM Architecture (left) and Bi-GRU Architecture (right)

Bidirectional Gated Recurrent Unit (BiGRU) model is an extension of the GRU architecture to achieve long-term dependencies in time series data to a greater extent using two propagation directions, i.e., forward propagation and reverse propagation. There are two GRU layers shown in Figure 2 (right) working alongside one another: the first GRU layer receives information from the former time step

Vol. 7, No. 4, August 2026, Page. 3579-3598
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.4.5444

(forward), while the second receives information from the latter time step (backward). The result of these two propagations is concatenated to enable the model to observe both past and future context of each data point in the input sequence. The advantage of the approach is that it can compute two-way correlations, which is of great significance in time series processing where patterns depend not only on the past but also on future states. Hence, BiGRU is of higher prediction quality compared to the standard GRU that utilizes a one-way propagation [23].

2.6.2. BERT

Bidirectional Encoder Representations from Transformers (BERT) is a language model based on the Transformer model that makes use of a self-attention mechanism to encode bidirectional token relationships in a sequence of text. The key mechanism is known as Scaled Dot-Product Attention and is given in equation 1[24] .

$$Attention(Q, K, V) = softmax\left(\frac{(QK^T)}{\sqrt{d_k}}\right) V \quad (1)$$

Q, K, and V are query, key, and value from linear projection, respectively, and d_k are the key dimensions. BERT is thereby able to generate more contextualized representations than typical sequential models through this mechanism.

2.6.3. XLNet

XLNet is a generalized autoregressive pretraining method designed to learn bidirectional contexts by maximizing the expected likelihood over all possible permutations of the factorization order. This approach, termed Permutation Language Modeling (PLM), enables XLNet to leverage the advantages of both Autoregressive (AR) and Autoencoding (AE) frameworks while avoiding their respective limitations, such as the independence assumption and the pretrain-finetune discrepancy associated with BERT. The PLM objective is formally expressed as formula 2, where Z_T represents the set of all possible permutations of the length-T index sequence $[1, .., T]$ [25].

$$\max_{\theta} E_{z\sim\mathbb{Z}_T}\left[\sum_{t=1}^{T} \log p_\theta\left(x_{z_t} | x_{z_{<t}}\right)\right] \quad (2)$$

2.7. Model Evaluation

To assess the performance of our model, we use confusion metrics which contain information of True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). We then use standard evaluation metrics such as Accuracy, Precision, Recall, and F1-Score. In this paper, since we already used data augmentation and oversampling to address imbalance issues, we will use macro average Precision, macro average Recall, and macro average F1 Score to evaluate the result.

3. RESULT AND DISCUSSION

From the hybrid sources explained in Section 2, we managed to acquire a total of 11746 data (see Table 2) which consists of 2038 data from the first source, adding into 9708 data from the second source which has already been compiled and used by Natasya [3]. The dataset from the first source is collected from 37 respondents, it is labelled automatically using BFI questionnaire. The dataset from the second source is labelled manually by three psychologists using voting mechanism, with a calculated Kohen's Cappa = 0.81 which indicates almost perfect agreement. As we see in Table 2, the dataset is very imbalanced, with Conscientiousness and Neuroticism constitutes the largest portion in our dataset, while Agreeableness and Openness only forms a very small portion.

Vol. 7, No. 4, August 2026, Page. 3579-3598
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.4.5444

Table 2. Dataset distribution

Label	First source	Second source	Total
Extraversion	552	1207	1759
Agreeableness	312	913	1225
Conscientiousness	624	3495	4119
Neuroticism	441	3666	4107
Openness	109	427	536
TOTAL	2038	9708	11746

Next we build three models which we will evaluate and discuss in this section. In the first model, we use a hybrid LSTM+GRU architecture. In the second and third, we use the pretrained model, namely IndoBERT and IndoXLNet Model. In each models, we explore various data augmentation and feature expansion techniques to address the data imbalance issue and to optimize the model's performance. We use k-fold cross validation with k = 10, where in each fold we split the dataset into 90:10 (90% for data training, 10% for data testing). The goal of k-fold cross validation is to achieve hyperparameter stability, prevent overfitting, reduce bias, and improve generalization ability of the model. We evaluate each model in the subsection below.

3.1. Evaluation of Hybrid LSTM-GRU Model

We implement our hybrid LSTM-GRU model with bidirectional architecture and multihead attention mechanism as can be seen in Table 3.

Table 3. LSTM model architecture

Layer	Layer Name	Parameters
1st	BiLSTM Layer	input_size=128, hidden_size=64
2nd	BiGRU Layer	input_size=64, hidden_size=32
3nd	Multihead Attention Layer	embed_dim=32, num_heads=4
4rd	Linear Layer (FC)	in_features=32, out_features=5

At the end of the linear (fully-connected) layer, we use Softmax function to produce probability of each class label. For the loss function, we use Categorical Cross Entropy. For the optimizer we use AdamW optimizer. In each experiment with data augmentation techniques, we use Optuna to optimize the hyperparameters such as learning rate, batch size, dropout rate, and weight decay. Prior work that integrated Optuna with BERT-based classifiers reports that automated hyperparameters search can substantially improve final metrics and stabilize per-class performance compared to manual tuning, particularly in imbalanced real-world datasets [26]. We limit the space search for hyperparameter optimization using Optuna as follows:

- Learning Rate: from 1e-6 to 1e-4
- Batch Size: from 4 to 32
- Dropout Rate: from 0.3 to 0.5
- Weight Decay: from 1e-5 to 1e-3

The results of our experiments are shown in Table 4 and Table 5. We use Accuracy, macro average Precision, macro average Recall, and macro average F1-Score to evaluate the result.

From Table 4 and Table 5 we see that all data augmentation techniques have improved the overall performance. FastText performance is also shown to be generally better than Word2Vec. Back Translation have the largest impact for the improvement in LSTM+GRU+Word2Vec model, achieving accuracy = 0.76, recall = 0.71, and F1-Score = 0.74. Meanwhile, KMeansSMOTE is best when combined with LSTM+GRU+FastText model, showing stronger result with accuracy = 0.78, recall =

Vol. 7, No. 4, August 2026, Page. 3579-3598
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.4.5444

0.7, and F1-Score = 0.76. The Word2Vec model only performs better than FastText model in term of precision (0.75 versus 0.72). Other data augmentation methods like SMOTE and ADASYN also produced good results, with SMOTE applied to FastText achieving an F1-Score of 0.75. This indicates that oversampling via synthetic methods can be just as effective as semantic-based augmentation methods. By using subword representations in FastText, the model was able to respond more effectively and efficiently to different types of enhancement that enhance lexical diversity and facilitate generalization. Therefore we choose BiLSTM-BiGRU+FastText with KMeansSMOTE as the best LSTM-GRU model. This model's best performance is achieved with hyperparameter setting as follow: learning rate = 1e-5, batch size = 16, dropout rate = 0.4, and weight decay = 1e-5.

Table 4. LSTM+GRU+Word2Vec model result

#	Data Augmentation Technique	ACC	PREC	REC	F1-Score
1	None	0.56	0.53	0.49	0.56
2	SYN-REP	0.68	0.69	0.65	0.63
3	**BACK-TRANS**	**0.76**	**0.75**	**0.71**	**0.74**
4	RAND-INS	0.73	0.58	0.63	0.65
5	SMOTE	0.75	0.71	0.68	0.715
6	ADASYNC	**0.76**	0.70	**0.72**	0.69
7	KMeansSMOTE	0.72	0.73	0.68	0.64

Table 5. LSTM+GRU+FastText model result

#	Data Augmentation Technique	ACC	PREC	REC	F1-Score
1	None	0.52	0.47	0.57	0.60
2	SYN-REP	0.70	0.68	0.68	0.69
3	BACK-TRANS	0.73	**0.75**	0.69	0.66
4	RAND-INS	0.74	0.68	0.64	0.75
5	SMOTE	0.77	0.73	0.75	0.75
6	ADASYNC	0.75	0.74	0.73	0.71
7	**KMeansSMOTE**	**0.78**	**0.72**	**0.75**	**0.76**

In the next scenario, we improve this best LSTM-GRU model using FastText feature expansion technique that we already discussed in Section 2. We used N = 1, 5, 10 to retrieve Top-N most similar tokens to be added into each data in our dataset. The result is shown in Table 6 below, with N=5 achieving the best performance model.

Table 6. LSTM+GRU+FastText+KMeansSMOTE with feature expansion

#	Top-N Feature Expansion	ACC	PREC	REC	F1-Score
1	None	0.78	0.72	0.75	0.76
2	N=1	0.73	0.69	0.66	0.71
3	**N=5**	**0.79**	0.725	**0.76**	**0.78**
4	N=10	0.785	**0.73**	0.75	0.68

Figure 4 shows that the LSTM+GRU+FastText+KMeansSMOTE model with TOP-5 feature expansion technique makes excellent progress for the Neuroticism label, with 3757 correct predictions, followed by Agreeableness with 3402 and Extraversion with 3163. In contrast, the Conscientiousness label had only 2883 correct predictions, and Openness had 2783, which are lower compared to the other traits.

In Figure 4, there were also many misclassifications in Conscientiousness, which was often predicted as Extraversion or Agreeableness, and with Openness, which was frequently classified as Agreeableness and Neuroticism. Although the model is effective in identifying text patterns, there is

Vol. 7, No. 4, August 2026, Page. 3579-3598
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.4.5444

some overlap in the interpretation of these traits, particularly those of Conscientiousness and Openness, suggesting that improvements may be necessary for the overall model.

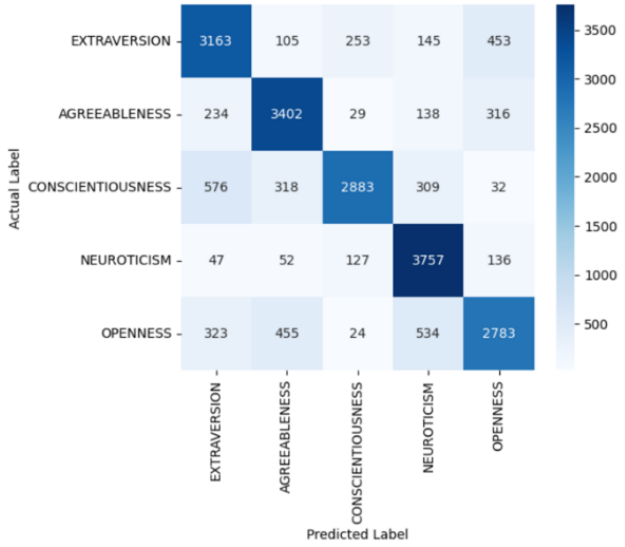


Figure 4. Confusion matrix of our best LSTM-GRU model

Figure 5 shows training accuracy and validation accuracy for our best LSTM+GRU model.

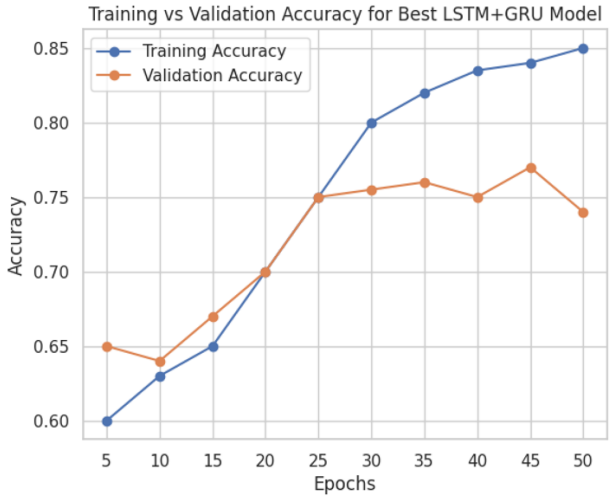


Figure 5. Training vs Validation Accuracy for Best LSTM+GRU Model

As can be seen from the graph of Figure 5, the training accuracy improved consistently from the first to the last epoch and reached a level of about 0.85 in the epoch 50. The validation accuracy, on the other hand, only increased until approximately the epoch 25. It leveled off after that and even fell slightly on some points. This is a sign of overfitting since the model is learning very well on the training data but not so well on the validation data. This can be readily observed through the widening gap between the training accuracy curve and the validation accuracy curve after the halfway point, indicating that the model is quite more focused on fitting the training set than maximizing generalization on new data.

3.2. Evaluation of BERT Model

We use two variations of BERT, namely IndoBERT and IndoBERTweet. IndoBERT model is created by Koto et al. from IndoNLU Team [27] and trained using Indo4B corpus (approximately 4.0 billion tokens), a large-scale Indonesian text datasets which taken from Kompas.com, Wikipedia Bahasa

Vol. 7, No. 4, August 2026, Page. 3579-3598
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.4.5444

Indonesia, etc. IndoBERTweet is created by Koto et al. and trained using 400 milion Indonesian tweets [28].

Since BERT already implemented their own word embedding layer as a feature extraction process, therefore we do not use any SMOTE oversampling technique and its variants. We only use the data augmentation techniques before applying BERT algorithm to our dataset. Using the same Optuna hyperparameter setting in SubSection 3.1, we achieve the results of our experiments as shown in Table 7 and Table 8.

Table 7. IndoBERT model result

#	Data Augmentation Technique	ACC	PREC	REC	F1-Score
1	None	0.65	0.615	0.69	0.66
2	SYN-REP	0.79	0.73	0.75	0.75
3	**BACK-TRANS**	**0.82**	**0.78**	**0.81**	**0.775**
4	RAND-INS	0.77	0.69	0.75	0.76

Table 8. IndoBERTweet model result

#	Data Augmentation Technique	ACC	PREC	REC	F1-Score
1	None	0.665	0.68	0.69	0.675
2	SYN-REP	0.77	0.715	0.81	0.75
3	**BACK-TRANS**	**0.835**	**0.80**	**0.78**	**0.79**
4	RAND-INS	0.79	0.73	0.77	0.75

The experimental results show a consistent pattern that the Back-Translation technique is the most effective data augmentation method across all tested models. In transformer-based models, IndoBERT showed a significant improvement with an accuracy of 0.82, precision of 0.78, recall of 0.81, and F1-Score of 0.775, far better than the baseline without augmentation (accuracy of 0.65). Meanwhile, IndoBERTweet even recorded the highest performance with an accuracy of 0.835 and an F1-Score of 0.79 when using Back-Trans. A similar trend was also observed in Table 5 and Table 6 when BiLSTM+Word2Vec combined with BACK-TRANS outperformed other techniques such as SMOTE or ADASYNC with an accuracy of 0.76 and an F1-Score of 0.74. BACK-TRANS's superior performance is also guaranteed by consistency in terms of BiLSTM+FastText results, which achieved 0.78 for accuracy and 0.76 for F1-Score and again taking the lead technique over other data augmentation methods such as SYN-REP or RAND-INS. The use of BACKTRANS augmentation has increased overall F1-Score result by 13%, compared to our baseline model. This major improvement is in line with another study conducted by Abonizio et al. [29], where they discussed and evaluated some text augmentation techniques. The study concluded that Back Translation is one technique that works best in maintaining semantic integrity and increasing the needed lexical variations. In this work, translating tweets back and forth into English and Indonesian gives the AI more variations in paraphrasing tweets to avoid committing particular patterns. Moreover, Bayer et al. [30] argue that such text generation approaches are critical in low-data regimes, as they synthetically expand the decision boundaries of the classifier, allowing it to capture the nuances of personality traits more robustly even when the original minority class samples (Openness) are scarce. This corroborates the capability of back-translation to improve syntactic and lexical diversity without losing semantic meaning, thereby enriching data distribution and guaranteeing improved model generalization. These findings align with Arora & Turcan [6], who point out that back-translation is the most powerful augmentation strategy in text classification tasks with scarce data. Therefore we choose IndoBERTweet with BACK-TRANS data augmentation as our best BERT model.

In the subsequent scenario, we applied the FastText feature expansion technique to our best-performing BERT model, which had been augmented with Back-Translation. This process involved

Vol. 7, No. 4, August 2026, Page. 3579-3598
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.4.5444

testing the impact of adding the Top-1, Top-5, and Top-10 most similar tokens for each data point. The complete results of this experiment are detailed in Table 9, which indicates that the N=10 configuration produced the strongest performance, with an accuracy of 84% and an F1-Score of 81.5%. The use of FastText Top-10 feature expansion technique has reduced class imbalance variance by 15%, compared to our baseline model.

Table 9. IndoBERTweet+BACKTRANS with feature expansion

#	Top-N Feature Expansion	ACC	PREC	REC	F1-Score
1	None	0.835	0.80	0.78	0.79
2	N=1	0.79	0.81	0.79	0.80
3	N=5	0.83	0.815	**0.82**	0.75
4	**N=10**	**0.84**	**0.835**	0.79	**0.815**

Figure 6 shows the confusion matrix of our best BERT model.

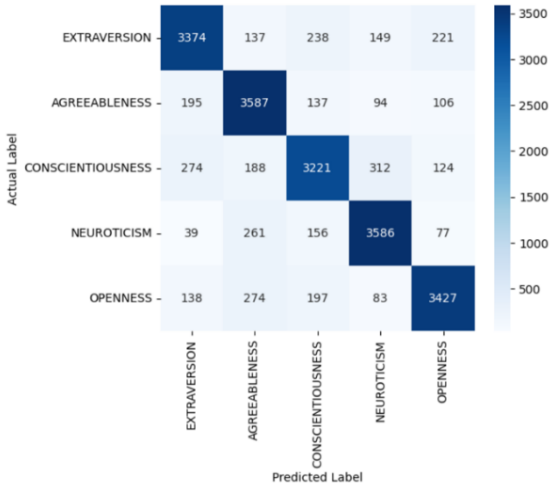


Figure 6. Confusion matrix of the best BERT model

Figure 6 shows good and well-balanced performance, with high values of true positives (TP) in all five classes, for example, 3221 for Conscientiousness and 3427 for Openness. Misclassification in this model is low and evenly distributed, indicating its good discrimination ability to extract the unique characteristics for each personality trait. In contrast, the Back-Trans model with BiLSTM-FastText (Figure 4) performs less well with an even greater inter-class confusion rate. The main weakness of the BiLSTM model is evidenced in the high proportion of systematic classification errors, with the most significant being the labelling of 576 Conscientiousness samples as Extraversion and 534 Openness samples as Neuroticism. A graphical comparison of both matrices clearly indicates that the IndoBERTweet model is significantly superior when it comes to the precision of classification and reliability.

In Table 10 below, we show performance metrics per trait for the IndoBERTweet+BACKTRANS model. However, the more detailed analysis indicates the strongest suitability of the model for the detection of the Conscientiousness trait with an excellence of accuracy of 0.91 with F1-Score 0.86, and then for the Extraversion trait with an F1-Score of 0.82. Although the highest precision for the Neuroticism trait was 0.90, this indicates high confidence in the positive class, the lowest accuracy and F1-Score were recorded for the Openness trait with 0.71 and 0.685, respectively. This clearly underlines that the performance difference strongly correlates with the dataset distribution, as observed in Table 2, where ample samples of Conscientiousness and Neuroticism facilitate better feature learning, while the shortage in Openness samples negatively affects the generalizing ability of the model for the specific

Vol. 7, No. 4, August 2026, Page. 3579-3598
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.4.5444

trait. These findings are consistent with Lucky et al. [31], who utilized IndoBERT on Indonesian Twitter data and identified that imbalanced datasets significantly skew prediction performance, necessitating advanced handling to improve minority class recognition. Furthermore, our results align with the work of Pradnyana et al. [32], who demonstrated that while fine-tuning IndoBERT architectures establishes a new state-of-the-art for Big Five personality prediction in Indonesian social media, the models still exhibit performance variability across traits due to the inherent linguistic complexity and data scarcity in specific personality dimensions.

Table 10. Performance metrics per traits of each best model

#	Big Five Traits	ACC	PREC	REC	F1-Score
1	Extraversion	0.78	0.81	0.76	0.82
2	Agreeableness	0.805	0.68	0.68	0.78
3	Conscientiousness	0.91	0.755	0.85	0.86
4	Neuroticism	0.87	0.90	0.79	0.75
5	Openness	0.71	0.81	**0.73**	0.685

Figure 7 depicts training accuracy and validation accuracy for our best BERT model:

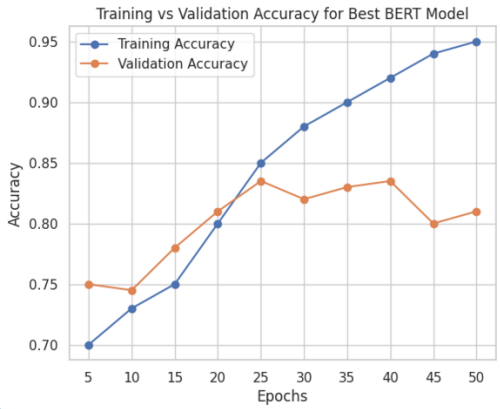


Figure 7. Training vs Validation Accuracy for Best BERT Model

A closer look at Figure 7 graph evidently suggests from a detailed analysis that the best-performing BERT model was overfit, which began showing up in a noticeable manner from epoch 25 itself. By this time, the validation accuracy curve is at its highest at around 0.84, then begins to level off and decline, while the training accuracy curve continues to rise sharply, reaching almost 0.95 by epoch 50. The average gap between train loss and validation loss after epoch 25 reaches 0.11. The extremely large difference between these two curves after epoch 25 only goes to confirm that the model has begun to "memorize" the training data and forget its capacity to generalize. For comparison, the generalization performance of the BERT model remains significantly better than those of the GRU and LSTM models while both likewise exhibit patterns of overfitting that begin at a similar range of epochs. The GRU model also has a lowest validation accuracy peak of about 0.80 at epoch 40, where the worst performance of the LSTM model has a validation accuracy peak of only about 0.77. Therefore, although all three models are overfitting-prone, the numerical data of the validation curves indicate that the BERT model has the strongest ability to generalize since it can achieve the highest accuracy on unseen data before its performance begins to decline.

3.3. Evaluation of XLNet Model

We use two variations of XLNet, namely IndoXLNet-4B and IndoXLNet-4Bplus, which have been developed by Pratama and Suharjito [33]. Similar like IndoBERT, the IndoXLNet-4B model is

Vol. 7, No. 4, August 2026, Page. 3579-3598
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.4.5444

trained using Indo4B corpus, whereas the IndoXLNet-4Bplus is trained using Indo4B and web crawling data. We do not use any SMOTE oversampling technique and its variants, and only use the data augmentation techniques before applying BERT algorithm to our dataset. We use the same Optuna hyperparameter setting in SubSection 3.2, except for Batch Size which we need to decrease it due to unsufficient memory resource. For Batch Size, we use range value of 4 to 16, instead of 4 to 32. The result of our experiments is shown in Table 11 and Table 12.

Table 11. IndoXLNet-4B model result

#	Data Augmentation Technique	ACC	PREC	REC	F1-Score
1	None	0.53	0.48	0.59	0.62
2	SYN-REP	0.71	0.69	0.68	0.72
3	**BACK-TRANS**	**0.79**	**0.725**	**0.75**	**0.75**
4	RAND-INS	0.77	0.73	0.57	0.69

Table 12. IndoXLNet-4Bplus model result

#	Data Augmentation Technique	ACC	PREC	REC	F1-Score
1	None	0.72	0.68	0.62	0.74
2	SYN-REP	0.74	0.65	0.68	0.73
3	**BACK-TRANS**	**0.80**	**0.74**	**0.75**	**0.77**
4	RAND-INS	0.685	0.73	0.74	0.72

As shown on Table 11 and Table 12, Back Translation improves our model performance drastically, with IndoXLNet-4B model achieves accuracy 0.79 and F1-Score 0.75, and IndoXLNet-4Bplus achieves accuracy 0.80 and F1-Score 0.77. Therefore, we choose IndoXLNet-4Bplus with BACK-TRANS data augmentation as the best XLNet model. This model's best performance is achieved with hyperparameter setting as follow: learning rate = 1e-6, batch size = 12, dropout rate = 0.3, and weight decay = 1e-4. The consistent performance improvement observed across both BERT and XLNet models when applying data augmentation techniques aligns with the comprehensive survey by Li et al. [34], which posits that data augmentation is indispensable in Deep Learning NLP tasks to mitigate the sparsity of high-dimensional feature spaces, especially in low-resource languages. Furthermore, while Back-Translation yielded the highest results, our experiments show that Synonym Replacement (SYN-REP) also provided a substantial boost over the baseline. This supports the findings of Chao et al. [35], who argued that augmentation methods focusing on semantic preservation are effective in maintaining the original label distribution while introducing necessary lexical variance. Additionally, Xiang et al. [36] confirmed that such lexical-level augmentations effectively enrich the training data without compromising the grammatical structure or semantic integrity, leading to significant performance gains in text classification benchmarks.

Following this, we sought to enhance our best XLNet model by applying the FastText feature expansion technique previously described in Section 2. We again employed the FastText word embedding from Section 2 to append the Top-N most similar tokens for N = 1, 5, and 10 onto each piece of data. We show the outcome from this enrichment in Table 13, where the best overall performance was reached by the N=5 configuration with an accuracy of 82% and an F1-Score of 78%.

Table 13. IndoXLNet-4Bplus+BACKTRANS with feature expansion

#	Top-N Feature Expansion	ACC	PREC	REC	F1-Score
1	None	0.80	0.74	0.75	0.77
2	N=1	0.75	0.68	0.67	0.76
3	**N=5**	**0.82**	**0.75**	**0.77**	**0.78**
4	N=10	0.81	0.735	**0.78**	0.75

Vol. 7, No. 4, August 2026, Page. 3579-3598
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.4.5444

Figure 8 shows the confusion matrix of our best XLNet model:

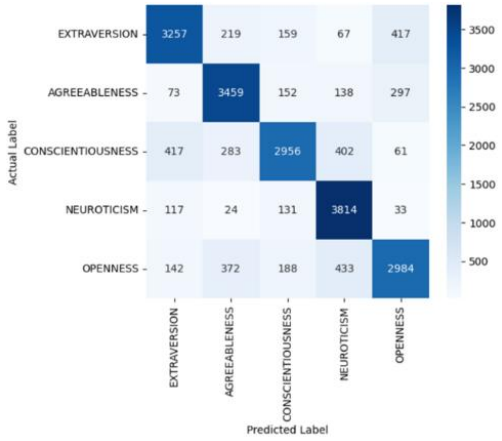


Figure 8. Confusion matrix of the best XLNet model

It can be seen from the confusion matrix of Figure 8 that the IndoXLNet-4Bplus+BACKTRANS model with the inclusion of Top-5 feature expansion method is able to correctly classify most of the data since the true prediction values (main diagonal) are quite high for all classes. The class with the highest true predictions, i.e., 3814, is Neuroticism. This indicates that the model is extremely good at recognizing patterns in this class. But there are also some misclassification examples between classes, for example, Conscientiousness is often predicted wrongly as Agreeableness or Neuroticism, and Openness is often confused with Agreeableness. It indicates that semantic features overlap across labels, and it is difficult for the model to distinguish very similar classes. In general, despite the presence of misclassifications, the high density of accurate predictions along the diagonal shows that the model has gained sufficient generalization capacity on the validation set. Figure 9 depicts training accuracy and validation accuracy for our best XLNet model:

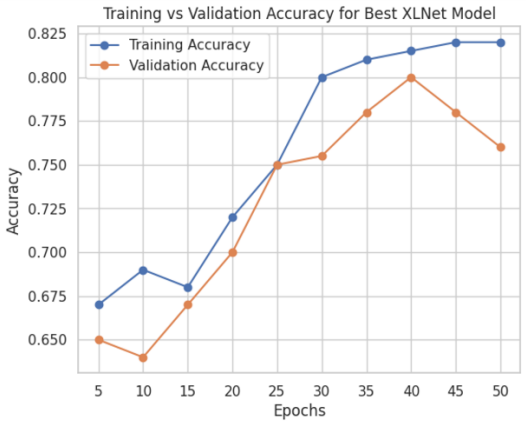


Figure 9. Training vs Validation Accuracy for Best XLNet Model

Referring to the graph in Figure 9, the XLNet model improves in training accuracy progressively to nearly 0.82 at epoch 50, with validation accuracy peaking at around 0.80 at epoch 40 before falling to around 0.76. The pattern indicates overfitting because even though training accuracy keeps on improving, validation accuracy plateaus and even goes down towards the completion of training. In comparison with the BERT model in Figure 7, In comparison with the BERT model in Figure 7, BERT's training accuracy climbs much more aggressively, finishing near 0.95 by epoch 50. Its validation accuracy is rather more unstable, however. It reaches an initial peak of roughly 0.835 at epoch 25 and hits a similar high point at epoch 40 but otherwise fluctuates between 0.80 and 0.81. The most striking

Vol. 7, No. 4, August 2026, Page. 3579-3598
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.4.5444

difference is the extreme gap between BERT's training and validation curves, which starts to widen markedly after only 20 epochs. This indicates that the BERT model begins overfitting on its training data much sooner and more intensely than XLNet. While the absolute peak validation accuracy for BERT, 0.835, is higher than the peak for XLNet of 0.80, in XLNet, the curves track closer into its peak at epoch 40. This reflects that XLNet maintained better generalization and more stable learning through the bulk of its training run.

3.4. Comparison of Three Models

From the three experiments we have conducted, we came out with these three best models:

1. Best LSTM+GRU model: BiLSTM+BiGRU+FastText with KMeansSMOTE and Top-5 feature expansion.
2. Best BERT model: IndoBERTweet with BACKTRANS and Top-10 feature expansion.
3. Best XLNet model: IndoXLNet-4Bplus+BACKTRANS and Top-5 feature expansion.

We conducted ANOVA test to prove that the performance of our best BERT model is higher than the performance of best LSTM+GRU model and best XLNet model. With $p < 0.01$, we are able to prove that IndoBERTweet result is statistically higher than the results of best LSTM+GRU model and best XLNet model.

Table 14 depicts the performance comparison of those three best models. IndoBERTweet achieves the best performance with the highest accuracy (84%), precision (83.5%), recall (79%), and F1-Score (81.5%), indicating better generalization and understanding of informal context in text. This is in line with Koto et al.'s work, where it is noted that IndoBERTweet, a domain-specific word-infused version of IndoBERT, can perform better on Indonesian Twitter data due to improved performance on different social media datasets [28]. XLNet model is also strong in performance, but still fall behind the representational power and fine-tuning capabilities of transformer models like IndoBERTweet. Other limitation or disadvantage of XLNet architecture, which is not shown in SubSection 3.3, is that XLNet needs much longer inference time, more than 10 times of BERT's inference time. The superiority of the IndoBERTweet model over the hybrid LSTM-GRU architecture aligns with recent findings by Soleimani and Kashani [37], who demonstrated that transformer-based architectures, particularly when utilizing fine-tuning strategies, consistently outperform earlier deep learning models like CNNs and RNNs in personality trait prediction tasks. Their study highlights that the self-attention mechanism in transformers enables more effective capture of long-range dependencies and contextual semantics compared to sequential processing in traditional recurrent networks. However, our results also show that deep learning performance is highly dependent on data quality. As noted in the survey by Nudin et al. [8], while deep learning holds high potential, its efficacy in low-resource settings is contingent upon successful mitigation of class imbalance, which we successfully addressed here using the Back-Translation and Feature Expansion strategy.

Table 14. Performance comparison of each best model

#	Model	ACC	PREC	REC	F1-Score
1	Best LSTM-GRU Model	0.79	0.725	0.76	0.78
2	**Best BERT Model**	**0.84**	**0.835**	**0.79**	**0.815**
3	Best XLNet Model	0.82	0.75	0.77	0.78

3.5. Limitations and Future Work

The limitations of our work lies in the small number of respondents who filled in the Big Five Questionnaire which is only contributing 17.5% of our dataset. The second source which contributes the most to our dataset (82.5%) are the public tweets from X platform which is crawled using keywords

Vol. 7, No. 4, August 2026, Page. 3579-3598
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.4.5444

previously listed by an expert in psychology. However, this list of keywords has not gone through a quantified validation process such as survey or questionnaire, hence introducing bias and limiting our model's validity. Another limitation lies in the bias of the population from where the dataset source is derived. The X users demography are mainly from urban areas, with a higher background in jobs and educations than users from rural areas. Each of the psychological traits may differ on providing quantities of tweets data: people with Openness trait probably finds social media is boring and prefers to engage more in creative and intellectual activities, such as reading books, building and creating arts, or travelling outdoor; which explains why the Openness class quantity in Table 2 is very small compares to the other four traits.

Our future work includes the utilization of multi-label and multi-modal personality prediction, incorporating more language attributes such as emojis, and incorporating other inputs such as tabular attributes (datetime of tweets, frequency of tweets, number of reply/retweet/quote, etc.) and images data (profile pictures, selfies, etc.). More adaptive data augmentation and feature expansion approaches using large language model and generative AI (GPT-4o) can also be explored further in future research to provide more accurate, reliable, and practicable personality prediction outputs for applications ranging from recruitment and content recommendation to data-driven psychological intervention. With the use of LLM and generative AI and mitigation of more diverse keywords, we can expect to improve our model's accuracy from 84% to 88%, hence making our model more scalable for a large scale deployment of human resource psychometrics application.

4. CONCLUSIONS

This study evaluated the impact of data augmentation and feature expansion techniques on Big Five personality prediction using three major models, namely LSTM-GRU (hybrid), IndoBERT, and IndoXLNet, focusing on low resource Indonesian-language texts from social media platform X. The experimental results demonstrate that Back-Translation significantly improves the performance of BERT and XLNet models, while KMeansSMOTE emerges as the most effective strategy for LSTM-GRU, particularly in addressing class imbalance. The best model obtained is IndoBERTweet with BACKTRANS and Top-10 feature expansion (with Accuracy 0.84 and F1-Score 0.815). The second best model is IndoXLNet-4Bplus+BACKTRANS and Top-5 feature expansion (with Accuracy 0.82 and F1-Score 0.78). The use of BACKTRANS augmentation has increased overall F1-Score result by 13%, compared to our baseline model. While the use of FastText Top-10 feature expansion technique has reduced class imbalance variance by 15%, compared to our baseline model.

Our current work demonstrate the effectiveness of data augmentation techniques and feature expansion techniques in enhancing state-of-the-art model performance for Big Five personality prediction, while contributing original insights through comparative analysis on Indonesian-language datasets that are relatively limited and imbalanced. The practical implication is that selecting appropriate augmentation methods and feature expansion technique can substantially improve the reliability of deep learning–based personality prediction systems for low-resource languages, which in turn can advance Indonesian NLP informatics for computational psychology and integrate with the national mental health policy and screening.

ACKNOWLEDGEMENTS

This work was supported by a research grant from Directorate of Research and Community Service, Telkom University [Grant No. 484/LIT06/PPM-LIT/2024].

Vol. 7, No. 4, August 2026, Page. 3579-3598
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.4.5444

REFERENCES

[1] W. Mahnaz and D. S. Kiran, "Big Five Personality Traits and Social Network Sites Preferences: The Mediating Role of Academic Achievement in Educational Outcomes of Secondary School Students," *Social Science Review Archives*, vol. 2, no. 2, pp. 1353–1370, Dec. 2024, doi: 10.70670/sra.v2i2.187.

[2] L. R. Goldberg, "An alternative 'description of personality': The Big-Five factor structure," *J. Pers. Soc. Psychol.*, vol. 59, no. 6, pp. 1216–1229, Feb. 1990, doi: https://doi.org/10.1037/0022-3514.59.6.1216.

[3] N. Natasya, I. Ummah, and S. Widowati, "Comparative Performance Analysis of Big Five Personality Prediction on X Platform Using BERT and XLNet Models," in *2025 International Conference on Information and Communication Technology (ICoICT)*, 2025, pp. 1–7. doi: 10.1109/ICoICT66265.2025.11193100.

[4] Z. Chi, H. Huang, L. Liu, Y. Bai, X. Gao, and X.-L. Mao, "Can Pretrained English Language Models Benefit Non-English NLP Systems in Low-Resource Scenarios?," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 32, pp. 1061–1074, 2024, doi: 10.1109/TASLP.2023.3267618.

[5] A. M. Dos Santos *et al.*, "Words Similarities on Personalities: A Language-Based Generalization Approach for Personality Factors Recognition," *IEEE Access*, vol. 11, pp. 29823–29836, 2023, doi: 10.1109/ACCESS.2023.3261339.

[6] A. Arora and E. Turcan, "Evaluating the Effectiveness of Data Augmentation for Emotion Classification in Low-Resource Settings," 2024. [Online]. Available: https://arxiv.org/abs/2406.05190

[7] J. M. Tapia-Téllez and H. J. Escalante, "Data Augmentation with Transformers for Text Classification," in *Advances in Computational Intelligence*, L. Martínez-Villaseñor, O. Herrera-Alcántara, H. Ponce, and F. A. Castro-Espinoza, Eds., Cham: Springer International Publishing, 2020, pp. 247–259. doi: 10.1007/978-3-030-60887-3_22.

[8] S. R. Nudin, R. Gernowo, M. Somantri, and A. Wibowo, "Multi Task Classification Using Deep Learning Approaches for Big Five Personality Traits Prediction: A Review," in *2024 11th International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE)*, 2024, pp. 37–42. doi: 10.1109/ICITACEE62763.2024.10762804.

[9] M. C. Aparna and M. N. Nachappa, "Personality Traits Classification from Text Using Deep Learning Model," in *2022 IEEE Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI)*, 2022, pp. 1–6. doi: 10.1109/IATMSI56455.2022.10119306.

[10] V. Ong *et al.*, "Personality prediction based on Twitter information in Bahasa Indonesia," in *2017 Federated Conference on Computer Science and Information Systems (FedCSIS)*, 2017, pp. 367–372. doi: 10.15439/2017F359.

[11] N. H. Jeremy, C. Prasetyo, and and D. Suhartono, "Identifying Personality Traits for Indonesian User from Twitter Dataset," *The International Journal of Fuzzy Logic and Intelligent Systems*, vol. 19, no. 4, pp. 283–289, Dec. 2019, doi: 10.5391/IJFIS.2019.19.4.283.

[12] A. R. Julianda and W. Maharani, "Personality Detection on Reddit Using DistilBERT," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 7, no. 5, pp. 1140–1146, Oct. 2023, doi: 10.29207/resti.v7i5.5236.

[13] E. B. Setiawan, D. H. Widyantoro, and K. Surendro, "Feature Expansion for Sentiment Analysis in Twitter," in *2018 5th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)*, 2018, pp. 509–513. doi: 10.1109/EECSI.2018.8752851.

[14] A. R. Royyan and E. B. Setiawan, "Feature Expansion Word2Vec for Sentiment Analysis of Public Policy in Twitter," *Jurnal RESTI: Rekayasa Sistem dan Teknologi Informasi*, vol. 6, no. 1, pp. 78–84, Feb. 2022, doi: 10.29207/resti.v6i1.3525.

[15] I. W. A. Widiarta and E. B. Setiawan, "Depression Detection on Social Media X Using Hybrid Deep Learning CNN-BiGRU with Attention Mechanism and FastText Feature Expansion," *Jurnal Ilmiah Teknik Elektro Komputer dan Informatika*, vol. 11, no. 2, pp. 290–304, Jun. 2025, [Online]. Available: https://journal.uad.ac.id/index.php/JITEKI/article/view/30687

Vol. 7, No. 4, August 2026, Page. 3579-3598
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.4.5444

[16] L. Ayash, A. Algarni, and O. Alqahtani, "Advancements in feature selection and extraction methods for text mining: a review," *Discover Applied Sciences*, vol. 7, no. 8, p. 914, 2025, doi: 10.1007/s42452-025-07587-w.

[17] S. Sonkar, A. E. Waters, and R. G. Baraniuk, "Attention Word Embedding," 2020. [Online]. Available: https://arxiv.org/abs/2006.00988

[18] Y. Gunawan, J. C. Young, and A. Rusli, "FastText Word Embedding and Random Forest Classifier for User Feedback Sentiment Classification in Bahasa Indonesia," *Ultimatics : Jurnal Teknik Informatika*, vol. 13, no. 2, pp. 101–107, Jan. 2022, doi: 10.31937/ti.v13i2.2124.

[19] L. D. Cahya, A. Luthfiarta, J. I. T. Krisna, S. Winarno, and A. Nugraha, "Improving Multi-label Classification Performance on Imbalanced Datasets Through SMOTE Technique and Data Augmentation Using IndoBERT Model," *Jurnal Nasional Teknologi dan Sistem Informasi*, vol. 9, no. 3, pp. 290–298, Jan. 2024, doi: 10.25077/TEKNOSI.v9i3.2023.290-298.

[20] Z. Shi and A. Lipani, "Rethink the Effectiveness of Text Data Augmentation: An Empirical Analysis," 2023. [Online]. Available: https://arxiv.org/abs/2306.07664

[21] A. Paszke *et al.*, "PyTorch: An Imperative Style, High-Performance Deep Learning Library," 2019. [Online]. Available: https://arxiv.org/abs/1912.01703

[22] W. Jian *et al.*, "SA-Bi-LSTM: Self Attention With Bi-Directional LSTM-Based Intelligent Model for Accurate Fake News Detection to Ensured Information Integrity on Social Media Platforms," *IEEE Access*, vol. 12, pp. 48436–48452, 2024, doi: 10.1109/ACCESS.2024.3382832.

[23] Q. Li, L. Chen, C. Wang, X. Zhang, and J. Zhao, "Daily Load Prediction Study on SCASL and BiGRU models," *J. Phys. Conf. Ser.*, vol. 2591, no. 1, p. 012050, Sep. 2023, doi: 10.1088/1742-6596/2591/1/012050.

[24] D. Lewy and J. Mańdziuk, "AttentionMix: Data augmentation method that relies on BERT attention mechanism," 2023. [Online]. Available: https://arxiv.org/abs/2309.11104

[25] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V Le, "XLNet: Generalized Autoregressive Pretraining for Language Understanding," 2020. [Online]. Available: https://arxiv.org/abs/1906.08237

[26] M. Khaldi, Z. Alilat, H. Bendoubba, and N. Mahammed, "Hyperparameter Optimization for Malicious URL Detection: Leveraging Optuna and Random Search in Machine Learning and Deep Learning Models," *INFORMATICA: International Journal of Computing and Informatics*, vol. 49, no. 27, pp. 57–68, May 2025, doi: 10.31449/inf.v49i27.9106.

[27] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP," 2020. [Online]. Available: https://arxiv.org/abs/2011.00677

[28] F. Koto, J. H. Lau, and T. Baldwin, "IndoBERTweet: A Pretrained Language Model for Indonesian Twitter with Effective Domain-Specific Vocabulary Initialization," 2021. [Online]. Available: https://arxiv.org/abs/2109.04607

[29] H. Q. Abonizio, E. C. Paraiso, and S. Barbon, "Toward Text Data Augmentation for Sentiment Analysis," *IEEE Transactions on Artificial Intelligence*, vol. 3, no. 5, pp. 657–668, 2022, doi: 10.1109/TAI.2021.3114390.

[30] M. Bayer, M.-A. Kaufhold, B. Buchhold, M. Keller, J. Dallmeyer, and C. Reuter, "Data augmentation in natural language processing: a novel text generation approach for long and short text classifiers," *International Journal of Machine Learning and Cybernetics*, vol. 14, no. 1, pp. 135–150, 2023, doi: 10.1007/s13042-022-01553-3.

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

[32] G. A. Pradnyana, W. Anggraeni, E. M. Yuniarno, and M. H. Purnomo, "Fine-Tuning IndoBERT Model for Big Five Personality Prediction from Indonesian Social Media," in *2023 International Seminar on Intelligent Technology and Its Applications (ISITIA)*, 2023, pp. 93–98. doi: 10.1109/ISITIA59021.2023.10221074.

Vol. 7, No. 4, August 2026, Page. 3579-3598
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.4.5444

[33] T. Pratama and Suharjito, "IndoXLNet: Pre-Trained Language Model for Bahasa Indonesia," *International Journal of Engineering Trends and Technology*, vol. 70, no. 5, pp. 367–381, May 2022, doi: 10.14445/22315381/IJETT-V70I5P240.

[34] B. Li, Y. Hou, and W. Che, "Data augmentation approaches in natural language processing: A survey," *AI Open*, vol. 3, pp. 71–90, 2022, doi: https://doi.org/10.1016/j.aiopen.2022.03.001.

[35] G. Chao, J. Liu, M. Wang, and D. Chu, "Data augmentation for sentiment classification with semantic preservation and diversity," *Know.-Based Syst.*, vol. 280, no. C, Nov. 2023, doi: 10.1016/j.knosys.2023.111038.

[36] R. Xiang, E. Chersoni, Q. Lu, C.-R. Huang, W. Li, and Y. Long, "Lexical data augmentation for sentiment analysis," *J. Assoc. Inf. Sci. Technol.*, vol. 72, no. 11, pp. 1432–1447, 2021, doi: https://doi.org/10.1002/asi.24493.

[37] M. Soleimani and H. B. Kashani, "Frozen or Fine-tuned? Analyzing Deep Learning Models and Training Strategies for Optimizing Big Five Personality Traits Prediction from Text," in *2024 10th International Conference on Artificial Intelligence and Robotics (QICAR)*, 2024, pp. 11–16. doi: 10.1109/QICAR61538.2024.10496606.