IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Revue d'Intelligence Artificielle
Vol. 37, No. 1, February, 2023, pp. 73-81
Journal homepage: http://iieta.org/journals/ria

Personality Classification Based on Textual Data using Indonesian Pre-Trained Language Model and Ensemble Majority Voting

Ghinaa Zain Nabiilah[1*], Derwin Suhartono[2]

[1] Computer Science Department, BINUS Graduate Program-Master of Computer Science, Bina Nusantara University, Jakarta 11480, Indonesia
[2] Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta 11480, Indonesia

Corresponding Author Email: ghinaa.nabiilah@binus.ac.id

https://doi.org/10.18280/ria.370110

Received: 21 December 2022
Accepted: 5 January 2023

Keywords:
ensemble learning majority voting, IndoBERT, Indonesian RoBERTa Base, Multilingual BERT, personality classification, pre-trained model

ABSTRACT

Personality is a collection of striking traits and behaviors of a person. The use of personality models can be applied in employee recruitment systems or to analyze characteristics and potential in more depth. Personality models are usually made using psychological test data or filling out questionnaires. However, this requires a long time. Building a personality classification model using NLP and deep learning is considered one of the best solutions. However, the performance of the classification model still needs to be improved, especially for Indonesian Language data. So, this research makes a personality classification model with Indonesian Language data using BERT-based architectures such as Multilingual BERT, IndoBERT, and Indonesian RoBERTa Base with an ensemble majority voting technique. Data limitations and imbalances were addressed using synonym replacement by incorporating words from a pre-trained model, MBERT. Information contained in social media often has ambiguous meanings because the words conveyed are not standardized, so this study tries to retain the information contained in the text by translating emoticons and slang words at the preprocessing stage to help keep the meaning of words in context. The proposed approach's research results can improve the classification model's results in classifying personality.

1. INTRODUCTION

Personality is a collection of people's characteristics, feelings, and behaviors in interacting with other individuals [1]. Personality information can be implemented to identify deeper characteristics and potential that can help develop oneself and minimize deficiencies. In addition, personality information can also be used in the employee recruitment system making it easier for companies to get employees who meet the criteria and needs [2]. Personality assessment is usually done using psychological tests. However, it takes a long time. So, it needs a model that can classify personality automatically.

The rapid development of social media allows someone to write, tell stories, or comment on many things continuously. This makes social media switch roles to become a place that saves someone's activity on social media in real-time without realizing it. Based on data compiled by Kemp [3], around 170 million Indonesians are active social media users. One of the social media platforms often used is Twitter which occupies the fifth position of social networking platforms in Indonesia. So, it can be concluded that the use of social media is commonplace and reasonable in Indonesia. The behavior of social media users has also been shown to have a close relationship with the user's personality [4]. This is because users use social media to express themselves to the world. Therefore, a strong relationship exists between user personality and user behavior on social media [5]. Based on this, the information stored on social media allows it to be used for classifying personality.

Research on personality classification using social media data requires analysis and a deep understanding of the context and textual meaning conveyed through writing. Natural Language Processing (NLP) is a well-known method of extracting data from text. The NLP method is necessary to understand the meaning of a language and sentence structure by applying syntactic and semantic analysis to classify personalities based on text data in tweets shared by users [6].

Adi et al. [7] previously conducted research related to personality classification using Indonesian Twitter data. Semi-supervised learning techniques are used to overcome data collection limitations. This study compares the machine learning algorithm Stochastics Gradient Descent (SGD) and Super Learner using the n-gram feature extraction method and the TF weighting scheme. The optimal result of this research is to use a super learner.

However, the machine learning model architecture has shortcomings in retaining information from the previous word, so the resulting context does not pay attention to the order of words in sentences. So, text classification research has started using deep learning model architecture, as in the research conducted by Ahmad et al. [8], using a combination of Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) models, which produce better accuracy.

The success of the bidirectional technique in LSTM and Recurrent Neural Network (RNN) inspired the creation of the Embedding from Language Model, which considers all the word vectors in a sentence and inserts the sentence through the two-way language model to generate contextual words [9, 10].

For the Language Model model to predict words correctly and allow them to be used in other tasks, it needs to be trained repeatedly with massive data. A language model trained using large data is called a Pre-Trained Language Model [11].

One of the well-known pre-trained models for handling various NLP tasks is the Bidirectional Encoder Representation from Transformers (BERT). BERT uses a transformer architecture. Generally, Transformer consists of two main parts: the encoder and the decoder [12]. However, BERT only uses the Encoder part of the Transformer. BERT is designed to use two-way (left-right and right-left) representation of text simultaneously and combine Mask Language Model (MLM) with Next Sentence Prediction (NSP) together. So far, BERT is considered the best method for understanding texts with complex contexts and preventing context ambiguity [13]. Research conducted by Santos and Paraboni [14] has also used BERT to classify personality. The BERT model is proven to have optimal, with an F1 score of 0.97 on Dutch Tweet data.

The development of the BERT model continues to be carried out so that it can cover various languages other than English. One of them is the Multilingual BERT (MBERT), which develops the BERT model for various languages such as Indonesian. Especially for Indonesia, the BERT model was developed with IndoBERT and IndoRoBERTa Base. Research related to personality prediction using Indonesian has also started using BERT or IndoBERT, such as research conducted by Kelvin et al. [15] using a combination of BERT and IndoBERT.

In improving model performance, some researchers sometimes combine several models. The Ensemble technique is used to find the best solution from several algorithms used by combining them. Research conducted by Kazameini et al. [16] uses Bagged BERT SVM by combining several BERT SVM classification models, and then the final prediction model is carried out using Ensemble Majority Voting. This approach is proven to get better results.

Based on this, this study conducts a personality classification based on the activities of social media users in Indonesia, namely Twitter, using several development models from BERT, such as IndoBert, MBERT, and Indonesian RoBERTa Base by applying the Ensemble Learning Majority Voting technique. Due to limitations in Indonesian language data collection, data augmentation techniques using Easy Data Augmentation by synonym replacement using contextual embedding MBERT. The data are added randomly according to the number of unbalanced labels. This aims to increase the variety of the dataset and overcome extreme data imbalances. Words or sentences generated from user Tweet data in Indonesia are often not standardized or not following the correct Indonesian writing rules or spellings. So, translate slang words and emoticons added at the preprocessing stage.

2. RELATED WORK

Text processing is one of the main methods of converting text data into information. Text processing applications can be implemented in several ways, such as text summarization, opinion mining, and text classification [17]. Text classification is the process of grouping text (for example, articles, product reviews, or tweets) into specific categories or groups. Generally, text classification is used for sentiment analysis and topic classification, including personality classification [18].

Various approaches are used to complement research related to personality classification, for example, Ren et al. [19] proposed a personality classification model combining emotional and semantic features with BERT and CNN. Based on the experiments conducted, this approach gives better results. The ensemble learning method has also been applied to research conducted by Saini and Sharan [20] which gives optimal results. Research using the application of ensemble learning is not only used for personality classification, as was done by Wang et al. [21] implementing BERT, Electra, and Resnet with ensemble learning methods to detect Chinese grammatical errors. The optimal result of this study is the f1-score of 0.8166.

Meanwhile in Indonesia, Christian et al. [22] Also uses an ensemble model for BERT and its developments, such as RoBERTa and XLNET, to classify personality. However, the Indonesian language data used is then translated into English data. Another study conducted by Adi et al. [23]. The optimal results of this study were obtained using the Super Learner and XGBoost algorithms which were performed with hyperparameter tuning, feature selection, and sampling to overcome the imbalance in the amount of data. The amount of data used in this study is 300 Twitter user data that has been adjusted and annotated by psychologists with the Big Five personality theory. Ong et al. [24] also used 250 Twitter user data in Indonesian to classify the Big Five personalities. The XGBoost method is used to classify user tweets. The optimal result of this research is the Openness class with ROC AUC 0.71. Research related to personality classification using Indonesian Twitter data is still quite limited. The use of the BERT model and its development in the Indonesia Language still needs further exploration so that it can help improve the accuracy of the personality classification model. Moreover, ensemble techniques in several studies have been shown can increase the accuracy of classification models.

3. RESEARCH METHODOLOGY

This section describes the method proposed in this study. The pre-training model used in this study was previously trained using Indonesian data, such as MBERT, IndoBERT, and IndoRoBERTa Base. In addition, experiments were also carried out with other pre-training models, such as ALBERT and RoBERTa. The three models with the best classification results were determined for the Embedding Majority Voting process. In this study, data augmentation was also carried out using synonym replacement with contextual MBERT embedding and additional emoticon and slang word translation processes at the preprocessing stage. At the evaluation stage, an experiment was carried out using data testing. This testing data does not go through the model training process but is directly tested with a model that has been previously trained using training data and validation data; as a differentiator in Figure 1, the stages through which the data testing goes are distinguished by a red line. Figure 1 contains the flow of the experimental process carried out.

3.1 Dataset

This study uses a dataset developed in a study conducted by Adi et al. [7]. In this study, there were 958 data on Twitter users in Indonesia, consisting of 508 labeled data and 450 unlabeled data. However, this study only used 508 labeled data.

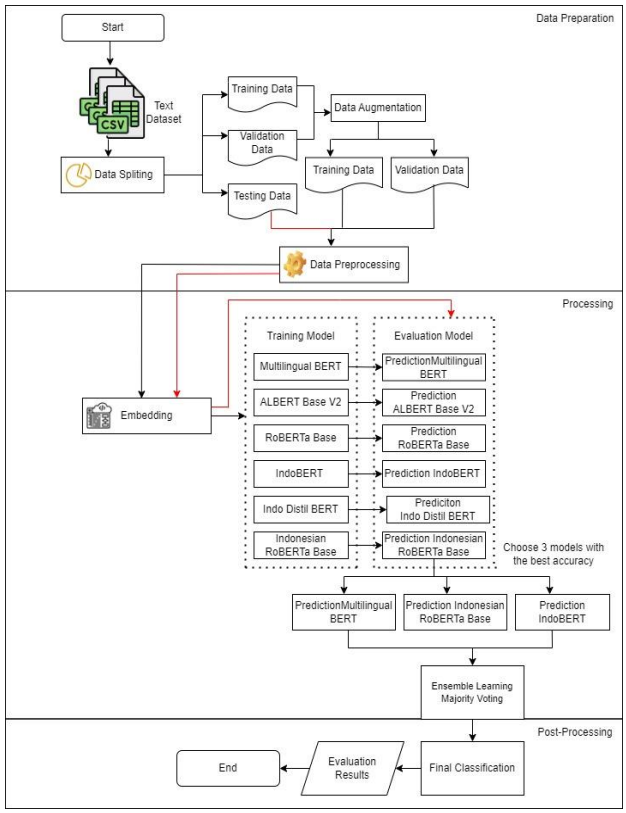

Figure 1. Proposed process flow

The data are labeled according to the Big Five personality types. Labels are given manually with the help of psychologists so that the data can be tested for relevance. Other information obtained in this dataset is the social features associated with each individual. This feature contains additional information related to user interactions on Twitter, such as the number of followers, the number of followings, the number of mentions, and other interactions. This is also considered in labeling and evaluation. This dataset is multi-label, where each user can be grouped into one or more than one personality label. Figure 2 contains the distribution of data for each personality type. In Figure 2, the data distribution for each personality label adds up to more than 508 because the data used is multi-label.

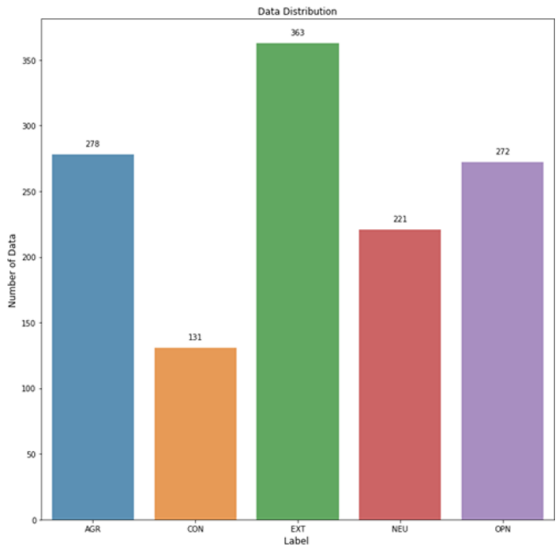

Figure 2. Data distribution

Information:
OPN: Abbreviation for the Openness personality type
CON: Abbreviation for Conscientiousness personality type
EXT: Abbreviation for the Extraversion personality type
AGR: Abbreviation for the Agreeableness personality type
NEU: Abbreviation for the Neuroticism personality type

The dataset used in this study is 508 Twitter user data which contains user tweets with more than 100 tweets per user. Datasets are divided into 3 types: training, validation, and test data. Because the data sharing ratio is adjusted to the amount of data and the model used, this study makes adjustments to determine the data distribution ratio. Namely 80% for training data, 11% for test data, and 9% for validation data.

This is because the model used to classify personality requires relatively high resources. To facilitate the training and validation process simultaneously, the validation data ratio is made smaller so that the model can more easily carry out the validation process with a small data set. So that the training process and model validation can be carried out simultaneously in a shorter time.

3.2 Personality theory

Personality can be defined as a pattern or characteristic of thought patterns, feelings, behaviors, and emotions that determine and influence an individual's style in interacting with his environment that is different from others and persists over time [25]. One personality assessment method is through social media interaction with various digital track records, such as audio, image, video, or text. Data sources derived from texts are considered to have textual factors that can reflect a person's personality traits. Many personalities assessment models, such as the Myres Briggs Type Indicator (MBTI), DISC Assessment, Strength Finder, and Big Five Personality.

This study uses the Big Five personality theory because it is considered a good personality structure, simple, and describes the general traits in personality [26]. The Big Five personality model is a personality taxonomy that is compiled based on a lexical approach by grouping words or language used in everyday life to determine individual characters. The Big Five Model groups personality into five personality dimensions described in Table 1. The Big Five Dimensions of Personality.

Table 1. Big five personality dimensions

Personality traits	High	Low
Openness	Imaginative, highly curiosity, love to learn new things, and creative	Prefer conventional ideas and less open to new things
Conscientiousness	Careful, organized, reliable, independent, disciplined, and detailed.	Relaxed, tend to be careless, less thorough, more disorganized, and irresponsible.
Extraversion	Friendly, chatty, sociable, and likes to socialize.	Tend to prefer to be alone, quiet, shy, and careful.
Agreeableness	Caring, empathetic, polite, and kind.	Competitive, short-tempered, irritable, suspicious.

3.3 Preprocessing

Preprocessing is the stage used to process data with much noise. In preprocessing, several stages are carried out to eliminate distracting words that are not needed in the classification process. The stages in preprocessing are as follows. Figure 3 contains the stages in preprocessing.

- Noise Removal, this stage will remove redundant spaces, URLs, punctuation marks, and numbers in text data.
- Case Folding is the stage to change the text into the same capitalization.
- Translating emoticon is a step to extract the emoticon contained in the text into meaning in the word.
- Tokenizing is the stage for breaking down each sentence into words.
- Translating slang word is changing non-standard words into standard words.
- Stemming is a process of removing affixes, prefixes, greetings, and suffixes in a word.
- Stopword Removal is a step that can eliminate words considered less influential with a high frequency of occurrence of words compared to other words.

At the stage of translating emoticons and slang words using datasets from Izzan [27]. However, the slang word dataset was modified by removing or adding non-standard words according to the dataset used in this study.

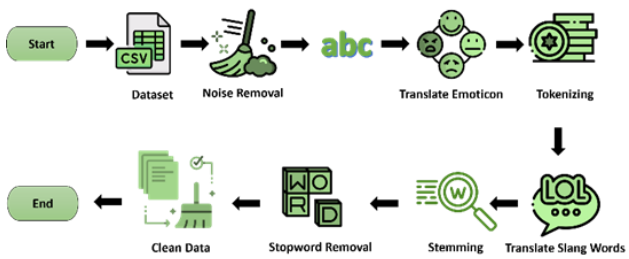

Figure 3. Preprocessing stages

3.4 Data augmentation

Data Augmentation is part of the preprocessing stage which performs the data manipulation process by generating additional synthetic data using the data owned. This stage aims to increase the data diversity in the training data by not collecting new data. One of the techniques related to data augmentation is Easy Data Augmentation (EDA). EDA consists of four simple operations: synonym replacement, random insertion, random swap, and random deletion. EDA can help improve model performance on small datasets [28]. In this study, the synonym replacement technique was used by randomly choosing a word from a sentence to be replaced with another word with a similar meaning. Sentences with these words are then added to new synthetic data. When changing the synonym of a word, it is possible to select the synonym of the word from the pre-trained model. This study uses Contextual Embedding MBERT as a pre-trained model in synonym replacement.

3.5 Pre-trained model

The Pre-Trained Model is a model that has been trained on massive data even in several different languages. This model aims to understand language and make machines read and understand the context like humans. In the end, a model with a training process with a large amount of data is expected to be reused on smaller datasets to refine and create a final model for the task across languages.

3.5.1 Multilingual BERT (MBERT)

BERT (Bidirectional Encoder Representations from Transformers) is a popular model because it applies bidirectional training of Transformer and Attention Mechanisms into language modeling. BERT modifies the development of previous language models that look at the left-right text order or a combination of right-left and left-right. By applying bidirectional training, the trained language model can have a deeper understanding of the context [12, 29]. The attention Mechanism of the Transformer applied in BERT can study the relationship and meaning between words as a whole.

The basic architecture of the Transformer has two separate mechanisms: the encoder reads the input, and the decoder generates predictions. However, because BERT is used as a language model, it only uses an encoder mechanism. The encoder can directly read the entire word order, so this is what causes it to be considered bidirectional. This capability also allows the model to learn and evaluate the overall meaning of words in sentences in either left or right word order. Figure 4 shows the BERT input representation. However, before entering the process stage, BERT requires additional inputs as follows:

- Token Embedding by adding the [CLS] token at the beginning and [SEP] at the end of the sentence.
- Segment Embedding is a marker that allows the encoder to distinguish between sentences.
- Positional Embedding to show its position in the sentence.

In the training process, BERT also applies a technique, MLM (Masked Language Model), which will randomly cover words in a sentence and try to predict the word order. In predicting these words, the BERT model looks at the entire sentence in bidirectional and uses the word's whole meaning to predict. So that in the process, the BERT model will take into account the token/word before and after it simultaneously.

It can overcome ambiguous in the same words but with different meanings. The BERT training process also uses NSP (Next Sentence Prediction), which is used to assist the model in understanding the relationship between two sentences, so that the model can take into account the relationship or correlation of meaning between sentences.

Meanwhile, MBERT is a development of the single language model of BERT that has been trained using monolingual corpora in 104 languages. MBERT was refined using specific training data from one language and evaluated in different languages, making it possible to be used across languages, and even MBERT was able to perform cross-language generalization tasks well [30]. The MBERT model has also been trained to use a dictionary in Indonesian so that it is possible to use it for assignments in Indonesian.

3.5.2 IndoBERT

IndoBERT is a pre-trained model based on Trans-former architecture in an Indonesian dataset called IN-DOLEM. The INDOLEM dataset consists of several Indonesian tasks, such as morpho-syntax, which leads to grammatical tasks or language rules (consisting of NER and POS Tagging), semantics, and discourse [31]. The IndoBERT model was trained and evaluated using the Dataset from INDOLEM and outperformed the experimental results with other algorithms, namely MBERT and Bi-LSTM-CRF.

3.5.3 IndoDistill-BERT

Distilbert Base Indonesian is a basic DistilBert model trained on 522 MB of Indonesian Wikipedia data and 1GB of Indonesian newspapers using a WordPiece with a size of 32,000 [32, 33]. The DistilBert model is a development of the BERT model with 40% fewer parameters than the BERT model. Although using fewer parameters, the DistilBert model can run 60% faster but maintain 95% of BERT performance [33].

3.5.4 ALBERT Base V2

ALBERT (A Lite BERT) is the development and modification of a simpler BERT model with fewer parameters without affecting model performance. In its modification, the ALBERT model introduces three main techniques, two of which are used to reduce parameters, namely [34]:

- Factorized Embedding Parameterization, by outlining the factorization of the embedding parameters into two matrices with smaller sizes.
- Cross-Layer Parameter Sharing, by sharing all parameters in each layer (reusing the same parameters in each layer).
- Sentence Order Prediction (SOP), to model coherence between sentences.

In this study, the ALBERT model used is ALBERT Base V2 which is available in hugging face. ALBERT Base V2 is the same as the ALBERT model and has been trained using book corpus data and Wikipedia across languages, including Indonesian.

3.5.5 RoBERTa Base

BERT is a language model previously trained for various NLP tasks, but the BERT model still has room for improvement. Therefore, a robust and optimized BERT development and modification model is proposed, namely RoBERTa (A Robustly Optimized BERT Pretraining Approach). Here are some things that are part of the modification of RoBERTa:

- RoBERTa is trained on more data sets than BERT, with 160GB of training data in text data.
- Changed the MLM (Mask Language Model) process on BERT, Static Masking, to Dynamic Masking to avoid masking the same word several times.
- Removed NSP (Next Sentences Prediction) on BERT and replaced it with Full Sentences without NSP.

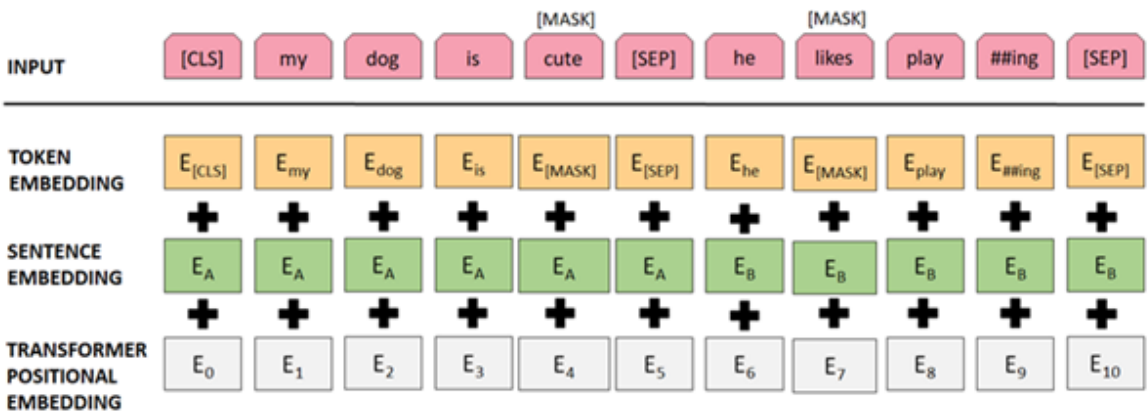

Figure 4. BERT input representation source [13], with modification

3.5.6 Indonesian RoBERTa Base

Indonesian RoBERTa Base is a language model of the RoBERTa model trained on the OSCAR (Open Super-Large Crawled ALManaCH corpus) dataset. OSCAR is a large multilingual corpus in 166 languages, one of which is Indonesian [35].

3.6 Ensemble learning

Ensemble learning is a machine learning algorithm that uses several learning models to achieve a better solution by combining them. Combining several classification models can significantly improve accuracy compared to using a single classification model [36]. Ensemble learning is generally used to improve classification performance or model prediction. Figure 5 shows Ensemble Learning Illustration.

3.6.1 Ensemble majority voting

Voting Ensemble is a heterogeneous learning ensemble. The way voting ensembles work is by combining predictions from several different models. The results of the classification between models are summed, and the final prediction is

chosen according to the most predictions among the overall models [37]. In the case of ensemble voting classification, there are two approaches, namely hard voting and soft voting. In hard voting, the summation process is based on the highest number of predictions from the entire model, while soft voting uses the most significant probability score in determining the final prediction. This study uses ensemble learning majority voting (ensemble voting with hard voting).

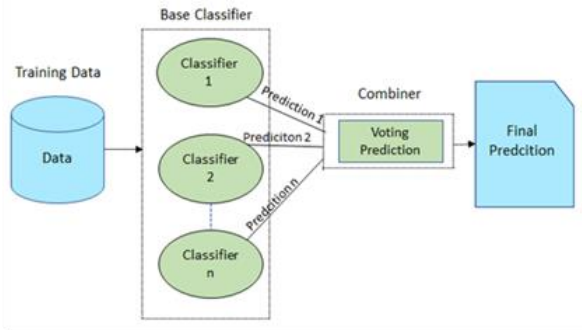

Figure 5. Ensemble learning illustration

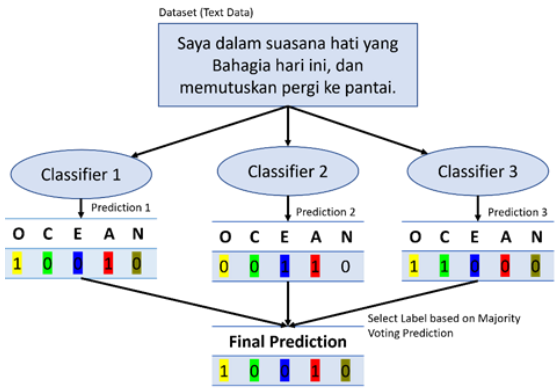

Figure 6. Ensemble majority voting illustration

The ensemble majority voting process is carried out on each class label in the dataset. So, it takes a process of classification and storage of prediction results from the model that the ensemble majority voting process will carry out. Figure 6 contains an example of the ensemble majority voting process in this study. This study has five labels based on the significant five personality types so that the resulting prediction has five numbers representing the personality label of the processed text data.

3.7 Evaluation model

Model evaluation is assessing the model's performance by calculating the confusion matrix. The confusion matrix calculation pays attention to the true positive (TP), true negative (TN), false positive (FP), and false negative (FN) assessments to obtain accuracy, precision, recall, and F1 scores. Accuracy describes the percentage of all data classified correctly in positive and negative classes. Eq. (1) is an equation for calculating accuracy.

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FN+FP)} \quad (1)$$

Precision is a measurement to determine how much the model correctly labels positive data from the total positive data labeled. The precision value can be calculated using the formula in Eq. (2).

$$\text{Precission} = \frac{TP}{TP+FP} \quad (2)$$

Recall is a measurement that determines how precise the model gives the correct label to positive data. The recall value can be calculated using the formula in Eq. (3).

$$\text{Recall} = \frac{TP}{TP+FN} \quad (3)$$

F1-Score is a combined value between precision and recall. The value of the F-1 Score can be calculated using the formula in Eq. (4).

$$\text{F1 Score} = \frac{2\times(\text{Precision}\times\text{recall})}{(\text{Precision}\times\text{recall})} \quad (4)$$

ROC (Receiver Operating Characteristics) describes the true and false positive rates. AUC (Area Under Curve) contains measurements of the area under the ROC curve. The equation used to plot the ROC value is found in Eq. (5) and Eq. (6).

$$\text{True Positive Rate} = \frac{TP}{TP+FN} \quad (5)$$

$$\text{False Positive Rate} = \frac{FP}{FP+TN} \quad (6)$$

4. RESULT AND DISCUSSION

The experiment in this study was conducted using the Google Colab Pro platform with a memory allocation of 25GB. Google Colab Pro is recommended for research using pre-trained models such as BERT and other models that are developments of BERT. This is because large parameters in the model require a large memory allocation to run and complete the model training process. For example, training the original BERT model also uses high resources, and it takes several days to complete the model training process.

Table 2. Total data based on personality label

Personality label	Training data	Training data after augmentation data	Validation data	Validation data after augmentation data	Testing data
Agreeableness	224	699	27	115	27
Conscientiousness	106	564	11	93	14
Extraversion	293	674	34	111	36
Neuroticism	179	606	22	101	20
Openness	211	691	30	120	31

The programming language used in this research is python with the PyTorch library. The PyTorch library is used to train and develop neural network models. The description and amount of data distribution before and after the data augmentation process for each personality type are shown in Table 2.

The embedding and classification process was carried out using the six proposed models: MBERT, IndoBERT, Indonesian RoBERTa Base, IndoDistill-BERT, RoBERTa Base, and ALBERT Base V2. The experiment was carried out using the same number of epochs and batch sizes, namely 5 epochs and 8 batch sizes. The batch sizes are adjusted to the allocation and availability of resources on Google Colab Pro. Figure 7 shows the accuracy value based on training data and Figure 8 shows the accuracy value based on validation data.

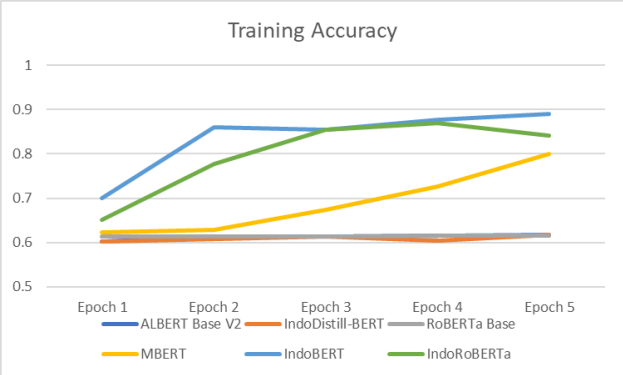

Figure 7. Training accuracy report

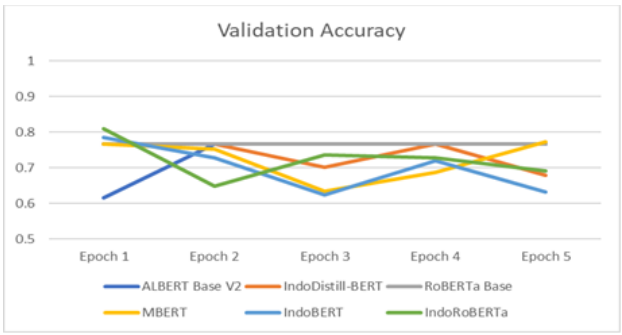

Figure 8. Validation accuracy report

Meanwhile, the evaluation results of the testing data on the entire model are shown in Table 3. The evaluation calculates the accuracy value, F1 Score, and ROC AUC to compare the results.

Three models have better testing accuracy from the experiments conducted than the other three models. The three models are MBERT, IndoBERT, and Indonesian Roberta Base. This result is dominated by a pre-trained model built specifically for the Indonesian Language. The dataset built when training the pre-trained model affects the model in handling specific NLP tasks for a certain language other than English. In addition, the model trained uses a cross-language dataset with many languages, so the model's knowledge of a particular language is not always good. However, it is different from MBERT, where the results of this experiment provide good accuracy; this is because the MBERT model has been trained with multilingual cases from various languages. Moreover, this model was created specifically to handle cross-language NLP tasks. So that when tested using Indonesian data, the accuracy obtained was 0.7372. So, the models chosen for the ensemble learning majority voting are MBERT, IndoBERT, and Indonesian Roberta Base.

Table 3. Result of all pre-trained models on testing data

Pre-trained model	Accuracy	F1 Score	ROC AUC
Multilingual BERT	**0.737254**	**0.717299**	**0.737543**
IndoBERT	0.737254	0.681818	0.726039
Indonesian RoBERTa BASE	0.705882	0.678111	0.706231
ALBERT Base V2	0.668407	0.50196	0.5
Indo Distil BERT	0.501960	0.650602	0.543922
RoBERTa Base	0.501960	0.668407	0.5

An ensemble majority voting process was then carried out using the three models with the best accuracy. At this stage, a comparison of the accuracy results is also carried out by not doing the entire preprocessing stage, namely using only noise removal and case folding and without using data augmentation. Table 4 contains a comparison of the results of model accuracy using all the stages proposed in this study (purposed model ensemble) without using augmented data and only using noise removal and case folding during preprocessing (non-purposed model ensemble).

Table 4. Comparison result on testing data

Architecture	Accuracy	F1 Score	ROC AUC
Proposed Model Ensemble	**0.756862**	**0.725663**	**0.757320**
Non-Proposed Model Ensemble	0.670588	0.688888	0.670367

The optimal result of this research is the average ROC AUC value of 0.7573 using ensemble learning majority voting against three algorithms, namely MBERT, IndoBERT, and Indonesian RoBERTa Base. In addition, the use of data augmentation, the application of translated emoticons, and translating of slang words also affect accuracy. Where in the data used in this study, data augmentation can help improve accuracy. Models trained using additional synthetic data can learn more with more data variations to make determining patterns when classifying personality easier. The complete application of preprocessing, including the application of translating emoticons and slang words, also helps to keep the context of the words in the sentences intact, especially the text data used is Twitter user tweet data which contains much ambiguity in the meaning of the sentence. Translating each emoticon and slang word will help the model and make it easier to analyze the meaning and context of the sentence so that the correct classification pattern can be determined.

The application of the ensemble can also help improve the classification results, wherein in a single model, the optimal classification result is an average ROC AUC value of 0.7375. However, if the three models with the highest accuracy are combined, it can help increase the classification value to an average ROC AUC of 0.7573. The ensemble process makes learning models determine predictions based on the learning process from other models. So that the final prediction produced is already a combination and the selection of the best predictions from the entire model.

5. CONCLUSION

When using the ensemble learning model, several factors can affect the final result of the ensemble model. One of them is that selecting the right pre-training model to be used in model training that using a particular language greatly affects the classification results. Not all pre-workout models are good for new cases with certain languages, especially Indonesian. The IndoBERT model developed and modified from a version of BERT and has been specially trained using Indonesian language data, has proven to have good accuracy when used in this study. Another model with good accuracy in this study is the Multilingual BERT model, which is specially designed to handle NLP tasks in a particular language and has been shown to have good accuracy when used in the Indonesian dataset. In addition, another pre-trained model specifically for Indonesian has proven to have good accuracy, namely RoBERTa Base Indonesia. The RoBERTa Base Indonesia model is a new model published in 2021 and has good accuracy for Indonesian data in this study.

In addition, using augmented data on limited training data can also be an option; in this study, it was proven that augmented data could help improve model classification results. In addition, if using social media data, a complete preprocessing implementation that does not ignore any information from users also needs to be considered. For example, this study tries to save all information from user tweets by not deleting emoticons but translating them, and not directly deleting slang words but translating them into standard words, which are expected by doing this, the information contained in a sentence is well preserved. For further exploration in this study, future research can apply the use of other pre-trained models such as XLNet, Elmo, and ULMFiT, which are ensembled with other types of ensemble learning such as ensemble learning voting but with soft voting instead of majority voting, which may affect the results of model classification. In addition, future research can also use an approach that allows conducting a more in-depth analysis of the model to study and consider the meaning of each word.

REFERENCES

[1] Piechurska-Kuciel, E. Second Language Learning and Teaching The Big Five in SLA. [Online]. Available: http://www.springer.com/series/10129, accesed on Jan. 10 2023.

[2] Allal-Chérif, O., Aránega, A.Y., Sánchez, R.C. (2021). Intelligent recruitment: How to identify, select, and retain talents from around the world using artificial intelligence. Technological Forecasting and Social Change, 169: 120822. https://doi.org/10.1016/j.techfore.2021.120822

[3] Kemp, S. (2021). Indonesia Digital Report. https://wearesocial.com/uk/blog/2021/01/digital-2021-the-latest-insights-into-the-state-of-digital/, accessed on Jul. 26, 2022.

[4] Azucar, D., Marengo, D., Settanni, M. (2018). Predicting the Big 5 personality traits from digital footprints on social media: A meta-analysis. Personality and Individual Differences, 124: 150-159. https://doi.org/10.1016/j.paid.2017.12.018

[5] Huang, H.C., Cheng, T.C.E., Huang, W.F., Teng, C.I. (2018). Who are likely to build strong online social networks? The perspectives of relational cohesion theory and personality theory. Computers in Human Behavior, 82: 111-123. https://doi.org/10.1016/j.chb.2018.01.004

[6] Lauriola, I., Lavelli, A., Aiolli, F. (2022). An introduction to deep learning in natural language processing: Models, techniques, and tools. Neurocomputing, 470: 443-456. https://doi.org/10.1016/j.neucom.2021.05.103

[7] Adi, G.Y.N.N., Harley, M., Ong, V., Suhartono, D., Andangsari, E.W. (2019). Automatic personality recognition in bahasa indonesia: A semi-supervised approach. ICIC Express Lett, 13(9): 797-805. https://doi.org/10.24507/icicel.13.09.797

[8] Ahmad, H., Asghar, M.U., Asghar, M.Z., Khan, A., Mosavi, A.H. (2021). A hybrid deep learning technique for personality trait classification from text. IEEE Access, 9: 146214-146232. https://doi.org/10.1109/ACCESS.2021.3121791

[9] Jozefowicz, R., Vinyals, O., Schuster, M., Shazeer, N., Wu, Y. (2016). Exploring the limits of language modeling. arXiv preprint arXiv:1602.02410. https://doi.org/10.48550/arXiv.1602.02410

[10] Melamud, O., Goldberger, J., Dagan, I. (2016). context2vec: Learning generic context embedding with bidirectional lstm. In Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, pp. 51-61.

[11] Nguyen, D.Q., Vu, T., Nguyen, A.T. (2020). BERTweet: A pre-trained language model for English Tweets. arXiv preprint arXiv:2005.10200. https://doi.org/10.48550/arXiv.2005.10200

[12] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30.

[13] Devlin, J., Chang, M.W., Lee, K., Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805. https://doi.org/10.48550/arXiv.1810.04805

[14] Santos, V.G.D., Paraboni, I. (2022). Myers-Briggs personality classification from social media text using pre-trained language models. arXiv preprint arXiv:2207.04476. https://doi.org/10.48550/arXiv.2207.04476

[15] Kelvin, Edbert, I.S., Suhartono, D. (2023). Utilizing indobert in predicting personality from twitter posts using bahasa indonesia. ICIC Express Letters, 17(1): 123-130. https://doi.org/10.24507/icicel.17.01.123

[16] Kazameini, A., Fatehi, S., Mehta, Y., Eetemadi, S., Cambria, E. (2020). Personality trait detection using bagged svm over bert word embedding ensembles. arXiv preprint arXiv:2010.01309. https://doi.org/10.48550/arXiv.2010.01309

[17] Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E.D., Gutierrez, J.B., Kochut, K. (2017). A brief survey of text mining: Classification, clustering and extraction techniques. arXiv preprint arXiv:1707.02919. https://doi.org/10.48550/arXiv.1707.02919

[18] Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., Gao, J. (2021). Deep learning--based text classification: A comprehensive review. ACM Computing Surveys (CSUR), 54(3): 1-40. https://doi.org/10.1145/3439726

[19] Ren, Z., Shen, Q., Diao, X., Xu, H. (2021). A sentiment-aware deep learning approach for personality detection from text. Information Processing & Management, 58(3): 102532. https://doi.org/10.1016/j.ipm.2021.102532
[20] Saini, M., Sharan, A. (2017). Ensemble learning to find deceptive reviews using personality traits and reviews specific features. Journal of Digital Information Management, 15(2): 84-94.
[21] Wang, S., Wang, B., Gong, J., Wang, Z., Hu, X., Duan, X., Shen, Z., Yue, G., Fu, R., Wu, D., Che, W., Wang, S., Hu, G., Liu, T. (2020). Combining ResNet and transformer for Chinese grammatical error diagnosis. In Proceedings of the 6th Workshop on Natural Language Processing Techniques for Educational Applications, pp. 36-43.
[22] Christian, H., Suhartono, D., Chowanda, A., Zamli, K.Z. (2021). Text based personality prediction from multiple social media data sources using pre-trained language model and model averaging. Journal of Big Data, 8(1): 1-20. https://doi.org/10.1186/s40537-021-00459-1
[23] Adi, G.Y.N., Tandio, M.H., Ong, V., Suhartono, D. (2018). Optimization for automatic personality recognition on Twitter in Bahasa Indonesia. Procedia Computer Science, 135: 473-480. https://doi.org/10.1016/j.procs.2018.08.199
[24] Ong, V., Rahmanto, A.D., Williem, W., Jeremy, N.H., Suhartono, D., Andangsari, E.W. (2021). Personality modelling of Indonesian Twitter users with XGBoost based on the five factor model. International Journal of Intelligent Engineering and Systems, 14(2): 248-261. https://doi.org/10.22266/ijies2021.0430.22
[25] Hogan, R., Sherman, R.A. (2020). Personality theory and the nature of human nature. Personality and Individual Differences, 152: 109561. https://doi.org/10.1016/j.paid.2019.109561
[26] McCrae, R.R., John, O.P. (1992). An introduction to the five-factor model and its applications. Journal of personality, 60(2): 175-215. https://doi.org/10.1111/j.1467-6494.1992.tb00970.x
[27] Izzan A. (2018). C.W.I.F.P. Indonesian Social Media Text Toxicity Dataset. https://github.com/ahmadizzan/netifier/tree/master/data/external, accessed on Jul. 21, 2022.
[28] Wei, J., Zou, K. (2019). Eda: Easy data augmentation techniques for boosting performance on text classification tasks. arXiv preprint arXiv:1901.11196.
[29] González-Carvajal, S., & Garrido-Merchán, E. C. (2020). Comparing BERT against traditional machine learning text classification. arXiv preprint arXiv:2005.13012.
[30] Pires, T., Schlinger, E., Garrette, D. (2019). How multilingual is multilingual BERT?. arXiv preprint arXiv:1906.01502.
[31] Koto, F., Rahimi, A., Lau, J.H., Baldwin, T. (2020). IndoLEM and IndoBERT: A benchmark dataset and pre-trained language model for Indonesian NLP. arXiv preprint arXiv:2011.00677.
[32] Wirawan, C. (2021). Distilbert-Base-Indonesian, huggingface.co, https://huggingface.co/cahya/distilbert-base-indonesian, accessed on Sept. 11, 2022.
[33] Sanh, V., Debut, L., Chaumond, J., Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.
[34] Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., Soricut, R. (2019). Albert: A lite bert for self-supervised learning of language representations. arXiv preprint arXiv:1909.11942.
[35] Flax Community, Indonesian Roberta Base, Hugging Face, (2021). https://huggingface.co/flax-community/indonesian-roberta-base, accessed on Oct. 22, 2022.
[36] Rincy, T.N., Gupta, R. (2020). Ensemble learning techniques and its efficiency in machine learning: A survey. In 2nd International Conference on Data, Engineering and Applications (IDEA), pp. 1-6. https://doi.org/10.1109/IDEA49133.2020.9170675
[37] Raza, K. (2019). Improving the prediction accuracy of heart disease with ensemble learning and majority voting rule. In U-Healthcare Monitoring Systems, Academic Press, pp. 179-196. https://doi.org/10.1016/B978-0-12-815370-3.00008-6