International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

A Three-Order Ensemble Model for User-level Big Five Personality Prediction on Twitter Dataset

Henry Lucky*[1], Ghinaa Zain Nabiilah[2], Nicholaus Hendrik Jeremy[3], Derwin Suhartono[4]

Submitted: 12/11/2022 **Accepted:** 14/02/2023

Abstract: The rapid development of social media has changed the way of interacting and communicating, one of which is using Twitter. Through Twitter, users can express themselves and their feelings directly without limits. It can unconsciously become a medium that reflects one's personality. In conducting personality assessments, the Natural Language Processing (NLP) model can use to predict personality automatically. So, in this study, an experiment was conducted to predict user personality based on the Big Five Personality Traits, especially in Indonesia. Previous research on personality prediction using BERT has provided promising results. However, BERT has drawbacks because it is limited in processing many words. To process information better it requires prediction of personality at the user-level by using all the user's information. Based on this, this research focuses on conducting experiments by proposing the Three Order Ensemble method with the BERT workflow (TOEM-BERT) as a scheme for combining tweets so that tweet data can be used optimally. The testing phase consists of two different experimental scenarios using two types of BERT models: IndoBERT and IndoBERTweet. Parallel test scenarios are carried out using the test set for each model, and linear test scenarios are carried out using the same test set for the entire model. The experiments show that the proposed TOEM-BERT method performs better in all test scenarios by obtaining 78.41% Weighted F1 in the linear test using IndoBERT and 77.84% Weighted F1 in the parallel test using IndoBERTweet.

Keywords: *Big Five, Personality prediction, IndoBERT, IndoBERTweet, Indonesian Twitter, Ensemble model*

1. Introduction

Social media has revolutionized the way of interacting, especially in communication, to connect with various parties online. A wide selection of social media platforms encourages social media interactions without boundaries. One of the popular social media platforms used is Twitter. Twitter has around 486 million users and is one of the world's most active social media platforms [1]. Twitter users in Indonesia also continue to increase. In 2022 more than 19 million Indonesians are active Twitter users [2]. Using Twitter makes it possible for users to exchange information, share thoughts, or even share personal stories to build new relationships and preserve existing ones. They also can express their thoughts and feelings directly at any time. So, the information produced and shared by social media users is considered a reflection of the self that reflects the true personality [3]. Specifically, changes in human interaction with social media make it possible for research in personality predictions that include psychological characteristics and user behavior based on information shared on social media [4]. Personality is a collection of characteristics and patterns that reflect individual behavior [5]. Personality information can be applied in various fields, such as online marketing, employee recruitment, personal recommendation, and counseling guidance. Several models describe personalities, such as MBTI, DISC, and Big Five. In the Big Five personality theory, personality is grouped into five factors: openness, conscientiousness, extraversion, agreeableness, and neuroticism. These five factors have also been shown to be significantly related to user behavior on social media [6].

Personality assessment is usually carried out using psychological tests with the help of specialists. However, this requires a long time and is expensive. As a solution, research in the field of Natural Language Processing (NLP) can be used to analyze the relationship between language use and personality in developing models that can automatically detect individual personality traits based on text shared on social media. Research conducted by [7] using several machine learning models such as Naïve Bayes, Neural Network, and SVM with TF-IDF, LIWC, EmoSenticNet, and ConceptNet as Feature Vector Generation. TF-IDF is considered unsuitable for data originating from Twitter because they cannot recognize the

[1]*Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta – 11480, INDONESIA*
ORCID ID : 0000-0002-4233-0409
[2]*Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta – 11480, INDONESIA*
ORCID ID : 0000-0001-7638-7449
[3]*Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta – 11480, INDONESIA*
ORCID ID : 0000-0003-3242-365X
[4]*Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta – 11480, INDONESIA*
ORCID ID : 0000-0002-3271-5874
** Corresponding Author Email: henry.lucky@binus.ac.id*

meaning and semantics of words simultaneously, so the research conducted by [8] added the word embedding Glove method with the XGBoost classification method. Glove is considered more suitable for handling complex and dynamic Twitter data. Approaches to personality prediction usually use one architecture to a certain extent. In addition, the architecture of the machine learning model also needs to be improved in retaining information from the previous word, and the resulting context does not pay attention to the order of the sentence. To overcome these limitations, the research conducted by [9] uses deep learning models to predict personality at the user-level. A hierarchical hybrid model based on a self-attention mechanism and a Convolutional Neural Network (CNN) combined with embedded multi-head self-attention and Bidirectional Long Short-Term Memory (Bi-LSTM) with native word embedding modules. The goal is to create feature diversity by making the model extract more semantic information horizontally and vertically.

The development of the Deep Learning model in the field of NLP resulted in the success of the bidirectional technique, thus inspiring the creation of various language models that adapt the transformer architecture. One example is Bidirectional Encoder Representation from Transformers (BERT) [10]. BERT is known for its ability to solve various problems related to text and NLP. BERT uses the Encoder part of the Transformer, which is designed with bidirectional (left-right and right-left) representation of text together and combines Mask Language Model (MLM) with Next Sentence Prediction (NSP) together. Thus, BERT is considered one of the best methods for understanding texts, especially in complex contexts [10]. Due to its better ability to understand text, several studies have also started using BERT for personality prediction. As in the research conducted by [11], which uses tweets from various languages. The BERT model is proven to have optimal results in this study, with an F1 score of 0.97 on Dutch tweet data. In exploring further, the process of analyzing information based on text delivered via tweets, research conducted by [12] uses BERT architecture and emotional features (SenticNet5) as feature extraction classified with CNN. The combined results of all these approaches produced an accuracy value of 0.9251.

In adapting to the various needs of NLP tasks, the BERT model continues to be developed. One of them is the existence of IndoBERT, as the development of the BERT model with a certain language, Indonesian. The development of the IndoBERT model was also carried out by creating IndoBERTweet, which is the first large-scale pre-trained model specifically for Indonesian Twitter data [13]. Research related to personality prediction using Indonesian has also started using BERT or IndoBERT, such as research conducted by [14] using a combination of BERT and IndoBERT to get better prediction results. Previous studies have typically used a tweet-level approach to build personality prediction models. Since the number of word tokens that can be processed by the BERT model is limited to 512 words, tweet data with a certain number of words is usually processed for extraction. This does not adequately describe the overall interactions and activities of social media users because Twitter users usually have a lot and varied information. Personality prediction at the user-level using more Twitter user information is carried out to overcome these limitations to improve the model classification results. Previous research [15] has tried to make a user-level personality prediction using the Twitter dataset. However, they only took 25 last tweets of the user because of the limitation on the BERT model, which caused the model to not use the full potential of the dataset and gave inferior results to previous works with the same dataset [16]–[19].

A technique that may be able to use the extracted information from the dataset to the maximum and gives better results at once is ensemble. Reference [20] proposed the Bagged SVM over the BERT Word Embedding Ensembles method. This approach divides the dataset into several subs and the text into several parts according to the sentence order and period. Then each of these sub-data was extracted using the BERT model and the SVM classification model. The final step of this approach is to use ensemble majority voting to select user personality labels, which is a bagging method. Ensemble techniques are used to find the best solution from several algorithm models by combining them. Based on this, this study aims to build a personality prediction model at the user-level using a simple ensemble approach. We proposed Three-Order Ensemble Model that uses the BERT model as the foundation (TOEM-BERT). The classification process used a multi-label approach because the dataset classifies users with one or more big five personality types. TOEM-BERT is inspired by the BART [21] training scheme, which changes sentences by adding, deleting, or shuffling words to get better results. TOEM-BERT divided the dataset into three parts based on the order of tweets to extract more information from users' tweets. It then applied the ensemble bagging technique, which combines the outputs from each model and then calculates the average of all the model outputs to predict the final prediction label to obtain better model prediction results.

2. Related Works

Personality by itself is an ever-growing field of research. Personality has been categorized into traits which in the collective is called personality inventory [22]. In the field of personality prediction, numerous personality inventories have been used to identify users' personalities, such as MBTI [23], [24], Big Five [16], [19], and other inventories such as dark triad [25]. Various modalities are also explored, with social media content being the most popular due to its

abundance in the dataset [26]–[29]. Other modalities include essay [30], [31], online forum [32], [33], email [34], and non-linguistic modal [35], [36].

An approach to identify a user's personality is to identify each trait on its own, meaning that a model is created for each trait [24]. For the Big Five, this means traits are split into five labels, and each trait is classified by its own classifier [37]. This approach is supported by a theory that traits are independent to one another [38]. However, it is possible to build a model that outputs each trait's qualities all at once by building a multi-label model. A multi-label BERT-based classifier was able to perform on par with single-label prediction [12]. It is also observed that the multi-label model performs better on MBTI than the Big Five. However, looking at the Big Five result, it is shown that the difference between single-label and multi-label models is rather insignificant. Another study investigating the performance difference between single-label and multi-label models on shallow-level machine learning shows a similar conclusion [39].

The ensemble model has been proven to leverage the language model performance, both at shallow-level [40] and deep-level [41]. In personality prediction, [42] uses meta-learning on both essay and Facebook posts, using trigram as their feature. However, the study treats the traits as five single-labels. [19] uses gradient boosted tree known as XGBoost. Another approach is to utilize bagged SVM on the BERT model [20]. The study also uses Mairesse features to enrich the feature. Again, the study treats each trait independently. [43] uses ensemble to investigate which modality works best to identify each trait in the Big Five between text, audio, and video. In relevance to the text model, the study discovers that the text modal does not carry as much important information for extraversion and openness. Although the text modal carries good information for agreeableness and conscientiousness, the significance is still very lower than video and audio. However, crosschecking the result with previous studies show that openness still can be correctly identified better than agreeableness and conscientiousness [20], [42], [44].

3. Methodology

This section explains the workflow of this study to reach the research objectives, which start from data collection, preprocessing, feature extraction, modeling and fine-tuning, and evaluation. The proposed TOEM-BERT workflow is shown in Fig. 1. The first step was to collect a dataset that would be used in the training and model evaluation. Then, preprocessing of the dataset, which split the dataset into three variants, was performed. The datasets are then used to fine-tune BERT models and evaluated with the F1 score for each trait, Macro F1, and Weighted F1 score for each model.

3.1. Data and Preprocessing

The dataset used in this paper is taken from [16], which is Indonesian Twitter tweets and the user profile data consisting of 508 users with 46,831 tweets in total. For each user, up to a maximum of 100 tweets were collected in the data-gathering process. The dataset has five classes representing the Big Five personality traits. Each class has its own label representing the affinity between the tweets and the personality traits, which is written with either "High" or "Low". Three psychology experts labeled the dataset with a voting system: for each trait of one user, the label that gets the most vote by experts becomes the label in the dataset. The label distribution is shown in Table 1. As we can see, there is an imbalance in Conscientiousness and Extroversion traits, while other traits are balanced.

Table 1. Label distribution of the dataset

	High	*Low*
Openness	272 (53.5%)	236 (45.5%)
Conscientiousness	131 (25.8%)	377 (74.2%)
Extroversion	363 (71.5%)	145 (28.5%)
	278	230

The initial text preprocessing methods that we use in this paper differed from the original dataset [19], which comprised text preprocessing methods for machine learning algorithms. As BERT is used as our main model, we only removed the use of URLs, symbols, and emoticons contained in tweets. Lowercasing is also done to normalize the text. For a user-level personality prediction

As the maximum token for BERT is 512 tokens and the concatenated tweets data has an average of 1603 tokens per user, we propose three data ordering schemes to concat the tweets to make use of all tweets data, in contrast to [15] that only use the 25 last tweets of the user in the dataset. The tweets are ordered based on the tweets' date. As in Fig. 2, the concatenation schemas are: (1) ascending order (AO): the tweets are ordered starting from the oldest tweet that has been posted by the user in the dataset. This is the simplest method; hence we will use it as the baseline. (2) descending order (DO): the tweets are ordered starting from the newest tweet N that has been posted by the user in the dataset. (3) random order (RO): the tweets are ordered randomly with a random state of 42. With this random order, hopefully, the variance of the dataset will be increased as it used the tweets that may not be covered in ascending or descending order, thus creating a more robust model.

3.2. Proposed Method

3.2.1. Deep Learning Method

For the classifier, two Indonesian BERT models that have been pre-trained with Twitter data are used. The first is IndoBERT [45] which was trained in two phases for 1M and 68k steps. In the first phase, it was pre-trained with 128 tokens, while in the second phase, it was pre-trained with 512 tokens. The model was pre-trained on the Indo4B dataset, consisting of 3.6B words from various sources, including the Twitter dataset. The second is IndoBERTweet [13] which follows the same pre-training procedure as the original BERT model [10]. The only difference is that the maximum length is set to 128 tokens only. The model was pre-trained on the Indonesian tweets dataset, which consists of 26M tweets with 409M word tokens. The dataset was taken with 60 keywords covering four main topics: economy, health, education, and government. Both models are standard BERT base models, a transformer encoder with 12 hidden layers (dimension=768), 12 attention heads, and 3 feed-forward hidden layers (dimension=3,072). The only differences are (1) the maximum length; IndoBERT has 512 tokens while IndoBERTweet has 128 tokens and (2) the embedding size; IndoBERT's is 50,000 while IndoBERTweet's is 31,923.

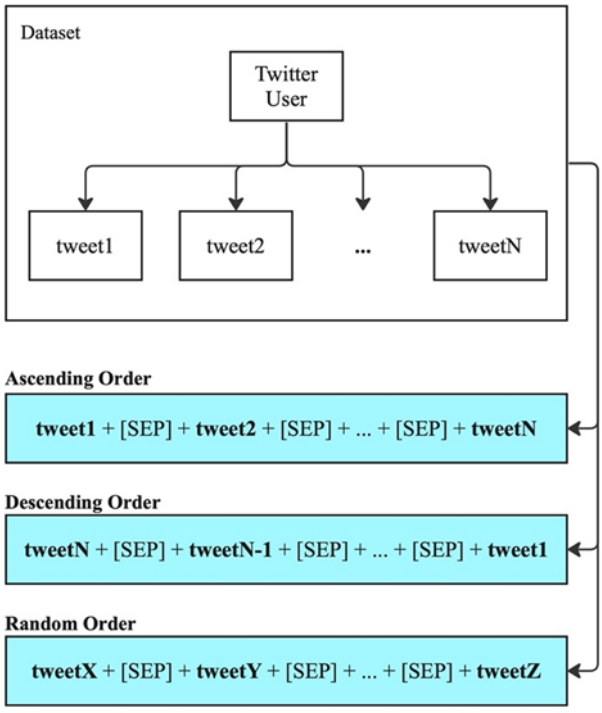

Fig. 2. Proposed Tweets Concatenation Method

3.2.2. Multi-label Classification

In contrast to the previous research with the same dataset [15]–[19], which used five different binary classifiers to predict each trait from the Big Five personality traits, in this paper, we use a multi-label classification model, following [46]–[48]. In fact, using the multi-label classification model is more sensible for the dataset, as the training is only done once to predict big five traits, while training using five different classifiers is done five times to predict big five traits. For the multi-label classification model, a feed-forward neural network is added on top of each BERT models with a sigmoid layer and Binary Cross Entropy (BCE) as the loss function:

$$BCE = -\frac{1}{N}\sum_{i=0}^{N} y_i \cdot \log\big(\sigma(\hat{y}_i)\big) + (1 - y_i) \cdot \log\big(\sigma(1 - \hat{y}_i)\big) \quad (1)$$

where N is the amount of data; y_i and $\hat{y}_i$ are the true label and the predicted label respectively: σ indicates the sigmoid layer.

3.2.3. Simple Ensemble

As mentioned in the previous section (3.1), the dataset is split with three different methods, resulting in three different datasets. Each dataset is trained using the BERT model, resulting in three different models. To make use of more data, a simple ensemble method based on bagging [49] using those three models is applied. With the method, the outputs from the three models are averaged in the testing phase. The equation for the ensemble output is:

$$\hat{y} = \sigma\left(\frac{1}{3} \cdot \left(BERT_1(x_1) + BERT_2(x_2) + BERT_3(x_3)\right)\right) \quad (2)$$

where $\hat{y}$ is the predicted label; *BERT* indicates the multi-label classification model using BERT as the base; x indicates the input data for the model, which consist of the (1) AO; (2) DO; and (3) RO.

3.3. Experimental Setup

The dataset is split into train, validation, and test split with a ratio of 7:1:2 respectively. Hereby, the experiment is also split into the training and validation phase and the testing phase. Experiments on all phases are done using 1×RTX3090 GPU (24GB). All splits and experiments are done with a random state of 42.

3.3.1. Training and Validation Phase

Training is done using IndoBERT [50] (indobenchmark/indobert-base-p2) and IndoBERTweet [13] (indolem/indobertweet-base-uncased) model with newly initialized feed-forward layer on top of them using the loss function in equation (1). Each model is trained individually on the three concatenation order data with a batch size of 4 to make sure the training has enough number

Table 2. Results of Parallel Testing

Traits	*Baseline*		*Proposed Model*					
	IndoBERT AO	***IndoBERTweet AO***	***IndoBERT DO***	***IndoBERT RO***	***IndoBERTweet DO***	***IndoBERTweet RO***	***TOEM IndoBERT***	***TOEM IndoBERTweet***
Openness	**75.23**	76.80	69.92	73.04	68.91	71.87	70.91	74.81
Conscien-tiousness	73.08	73.47	63.64	70.83	62.75	59.02	**75.56**	70.37
Extraversion	85.88	85.71	85.23	85.19	85.06	**86.59**	86.39	85.23
Agreeable-ness	80.00	79.31	80.00	79.31	81.30	**83.46**	80.00	82.26
Neuroticism	62.14	64.76	**71.29**	66.67	64.58	62.07	64.37	67.35
Macro F1	75.26	**76.00**	74.01	75.01	72.52	72.60	75.44	**76.00**
Weighted F1	76.97	77.60	76.34	76.72	75.13	75.91	76.83	**77.84**

Table 3. Results of Linear Testing

Traits	*Baseline*		*Proposed Model*					
	IndoBERT AO	***IndoBERTweet AO***	***IndoBERT DO***	***IndoBERT RO***	***IndoBERTweet DO***	***IndoBERTweet RO***	***TOEM IndoBERT***	***TOEM IndoBERTweet***
Openness	75.23	**76.80**	72.07	76.79	73.87	73.53	75.63	76.56
Conscien-tiousness	73.08	73.47	73.91	**76.00**	69.77	66.67	**76.00**	65.38
Extraversion	85.88	85.71	86.71	87.12	**87.21**	87.06	85.71	85.71
Agreeable-ness	80.00	79.31	76.19	78.12	76.79	78.79	**82.54**	80.00
Neuroticism	62.14	64.76	**67.39**	60.87	56.10	64.58	65.35	64.58
Macro F1	75.26	76.01	75.25	75.78	72.75	74.13	**77.05**	74.45
Weighted F1	76.97	77.60	76.68	77.31	75.00	76.51	**78.41**	76.90

of steps, resulting in six models in total. For each model, hyperparameter tuning is conducted using a bayesian optimization search algorithm [51] to search for the optimum hyperparameter faster. For hyperparameter search, we use a variation of learning rate: 1e-5, 2e-5, 5e-5; weigh decay: 1e-4, 1e-3, 1e-2; and epochs: 8 and 16 on each model. For IndoBERT, a learning rate of 2e-5, weight decay of 1e-3, and epochs of 8 are used. Meanwhile, for IndoBERTweet, a learning rate of 2e-5, weight decay of 1e-4, and epochs of 16 are used. Models are trained with AdamW optimizer [52] with a linear scheduler. Validation and model checkpointing are done for every epoch, and the best model with minimum validation loss is stored.

As the amount of data is small, deep learning model tends to have unstable training. To deal with unstable training, the training for each model and each data are conducted three times to find the minimum validation loss with the optimal hyperparameters.

3.3.2. Testing Phase

In this phase, the best model checkpoints are tested against the test set. The testing phase consists of two different scenarios to estimate the proposed model performance thoroughly. F1-score is used to estimate the model performance for each personality trait, while Macro F1 score is used to estimate the model's overall performance, following [14]–[19]. In addition, as there is data imbalance for Conscientiousness and Extraversion traits, Weighted F1-score is also used to better estimate the model's overall performance on imbalanced data.

3.3.2.1. Parallel Testing

In this scenario, the testing is done with each model's respective test set. The model that is trained on ascending order data will be tested on ascending order test set, the model that is trained on descending order data will be tested on descending order test set, and the model that is trained on random order data will be tested on the random order test set. The performance will reflect the direct performance of

each model and concatenation order data. The proposed ensemble method in equation (2) is also used in this scenario for IndoBERT and IndoBERTweet models.

3.3.2.2. Linear Testing

In this scenario, the testing is done on all models only with one kind of test set to estimate the knowledge transfer ability between the model and concatenation order data. The ascending order test set is chosen as it is our baseline. As for the ensemble method, there is a slight change to equation (2). The equation used here is:

$$\hat{y} = \sigma\left(\frac{1}{3} \cdot \left(BERT_1(x_1) + BERT_2(x_1) + BERT_3(x_1)\right)\right) \quad (3)$$

Notice that the only input used for all models is the ascending order test set.

4. Results and Discussion

The multi-label classification result with parallel testing is shown in Table 2 on F1 score metric. The bolded scores are the highest score for each row (trait). The TOEM-IndoBERTweet gained the best result in terms of Macro F1 and Weighted F1 score with 76% and 77.84% respectively, meaning that the proposed model utilizing IndoBERTweet has the highest overall performance compared to other models. Interestingly, IndoBERTweet AO also has the same Macro F1 of 76%, meaning that the Three order ensemble model didn't increase the performance for Macro F1. However, weighted F1 is increased by 0.24%, which implies that our proposed model with IndoBERTweet deals with imbalanced data better. In contrast, TOEM-IndoBERT successfully increased the performance from the baseline IndoBERT AO by 0.18% on Macro F1. However, there is a drop of 0.14% on weighted F1. This behavior happens because there is a significant drop in the Openness trait by 4.32%, even though TOEM-IndoBERT successfully increased the score on other traits.

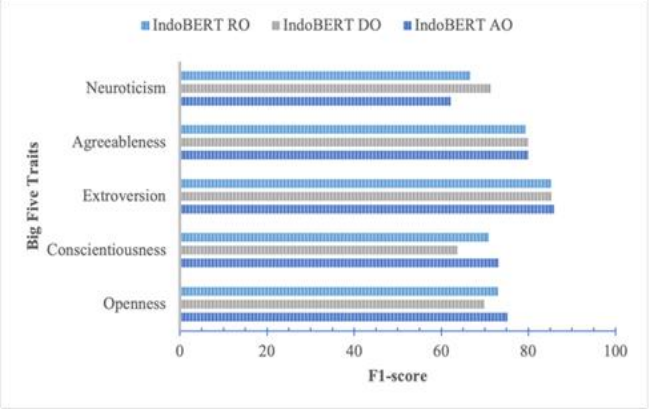

Fig. 3. Comparison of individual three order models on parallel testing using IndoBERT

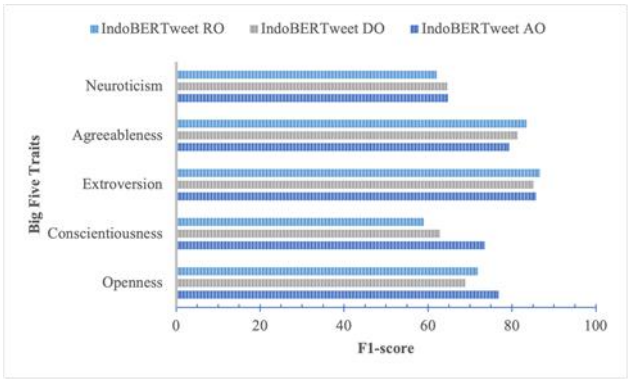

Fig. 4. Comparison of individual three order models on parallel testing using IndoBERTweet

Afterward, we discuss the results of individual three-order models. The comparison of IndoBERT models is shown in Fig. 3, and the comparison of IndoBERTweet models is shown in Fig. 4. As we can see in Fig. 3, the DO and RO scheme only overpower the baseline AO scheme on Neuroticism trait. On Agreeableness and Extroversion, the three models almost have balanced results, while on Conscientiousness and Openness, a significant drop occurs in the DO scheme. Overall, using IndoBERT, the DO and RO models didn't increase the performance from the baseline AO model. As for IndoBERTweet in Fig. 4, the RO model can overpower the performance of baseline AO on Agreeableness and Extroversion traits. However, the overall performance using the DO and RO models also didn't increase the performance as there is a significant drop in Conscientiousness trait.

However, in parallel testing, each model is tested using the different order data according to the training set, meaning that we only compare the individual model performance instead of comparing how well the model performs against each other. Therefore, linear testing is performed where all individual models and ensemble models are tested only with the baseline AO test set. The result from linear testing is presented in Table 3. Our proposed TOEM still has the best performance on Macro F1 and Weighted F1, with scores of 77.05% and 78.41%, respectively, while using IndoBERT. TOEM-IndoBERT also has the highest F1-score on Conscientiousness and Agreeableness traits. In contrast, TOEM-IndoBERTweet failed to achieve better results than baseline IndoBERTweet AO in terms of Macro F1 and Weighted F1 as a significant drop in Conscientiousness trait is spotted (-8.09%).

We compare the individual models' performance with the parallel testing setting in Fig. 5 and Fig. 6. On Fig. 4 using IndoBERT, the DO and RO models outperformed the baseline AO on four out of five traits. The DO model outperformed baseline AO on Neuroticism with an improvement of 5.25%. Meanwhile, the RO model outperformed the baseline AO on Openness,

Conscientiousness, and Extroversion with an improvement of 1.56%, 2.92%, and 1.24%, respectively. Overall, the TOEM-IndoBERT achieved better results than the baseline AO. In Fig. 5 using IndoBERTweet, the DO and RO models only outperformed the baseline AO on Extraversion trait by 1.5% and 1.35%, respectively. Meanwhile, for the other four traits, the baseline AO still outperformed the DO and RO models.

The difference in results on IndoBERT and IndoBERTweet between parallel testing and linear testing indicates that IndoBERTweet has better individual models and therefore is better at dealing with the tweets data that has the same pattern as the training data. Meanwhile, IndoBERT

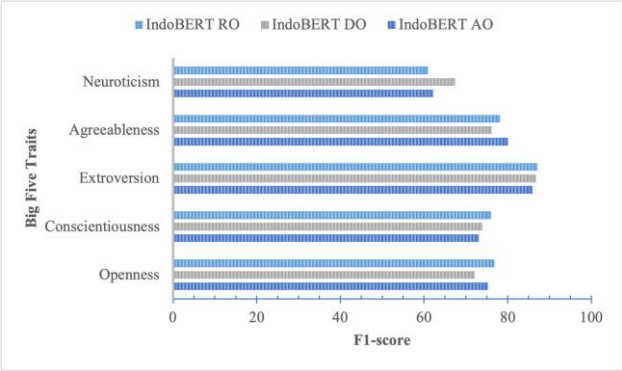

Fig. 5. Comparison of individual three order models on linear testing using IndoBERT

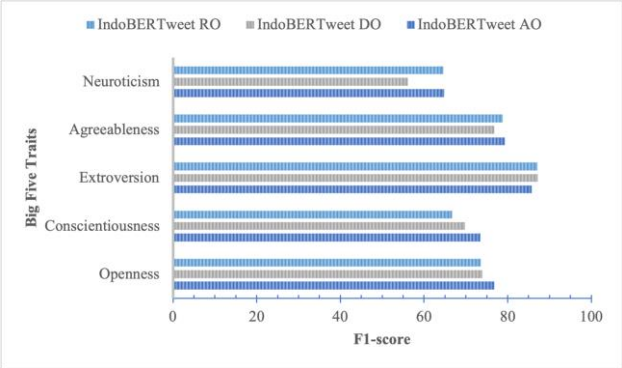

Fig. 6. Comparison of individual three order models on linear testing using IndoBERTweet

performs better in linear setting, indicating that it has better knowledge transferability than IndoBERTweet. We hypothesize the behavior is caused by the fact that IndoBERT was pre-trained with various data sources such as website articles, Wikipedia, Twitter, etc., which makes it have more knowledge than IndoBERTweet, which only pre-trained on tweets data. To top it off, our proposed TOEM-BERT achieved better results either on parallel testing or linear testing. The only limitation is that our proposed method still doesn't use all the users' tweets in the dataset.

5. Conclusion

In this paper, we proposed TOEM-BERT for user-level Big Five personality prediction. The proposed method consists of tweets concatenation method to represent user content better than using a single tweet to represent the user's personality and using a simple ensemble to gain better results. Based on our experiments, TOEM-BERT successfully outperformed the baseline models. For parallel testing, TOEM-IndoBERTweet performs better than the baseline IndoBERTweet AO. Although both models yield the same score for Macro F1, TOEM-IndoBERTweet yields a better score on Weighted F1 by 0.24%. For linear testing, TOEM-IndoBERT outperforms the baseline on both Macro F1 and Weighted F1 by 1.04% and 0.81%, respectively. For the next study, we will investigate the long input language models and other schemes to fully use all of the users' tweets to better predict their personalities.

Author contributions

Henry Lucky: Conceptualization, Methodology, Experiments, Writing-Original draft preparation. **Ghinaa Zain Nabiilah:** Visualization, Writing-Original draft preparation, Validation. **Nicholaus Hendrik Jeremy:** Data curation, Writing-Original draft preparation, Validation. **Derwin Suhartono:** Investigation, Writing-Reviewing, and Editing, Validation.

Conflicts of interest

The authors declare no conflicts of interest.

References

[1] S. Kemp, "Twitter Statistics and Trends," *datareportal.com*, 2022.

[2] S. Kemp, "Digital 2022: Indonesia," *datareportal.com*, 2022.

[3] R. Buettner, "Predicting user behavior in electronic markets based on personality-mining in large online social networks: A personality-based product recommender framework," *Electronic Markets*, vol. 27, no. 3, pp. 247–265, Aug. 2017, doi: 10.1007/s12525-016-0228-z.

[4] M. D. Back *et al.*, "Facebook profiles reflect actual personality, not self-idealization," *Psychol Sci*, vol. 21, no. 3, pp. 372–374, 2010, doi: 10.1177/0956797609360756.

[5] H. A. Schwartz and L. H. Ungar, "Data-Driven Content Analysis of Social Media: A Systematic Overview of Automated Methods," *Annals of the American Academy of Political and Social Science*, vol. 659, no. 1, pp. 78–94, May 2015, doi: 10.1177/0002716215569197.

[6] R. M. Bergner, "What is personality? Two myths and a definition," *New Ideas in Psychology*, vol. 57. Elsevier Ltd, Apr. 01, 2020. doi: 10.1016/j.newideapsych.2019.100759.

[7] S. Bharadwaj, S. Sridhar, R. Choudhary, and R. Srinath, "Persona Traits Identification based on Myers-Briggs Type Indicator(MBTI) - A Text

Classification Approach," *International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, 2018.

[8] P. Kumar and M. L. Gavrilova, "Personality Traits Classification on Twitter."

[9] X. Wang, Y. Sui, K. Zheng, Y. Shi, and S. Cao, "Personality classification of social users based on feature fusion," *Sensors*, vol. 21, no. 20, Oct. 2021, doi: 10.3390/s21206758.

[10] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 2019, vol. 1, pp. 4171–4186.

[11] V. G. dos Santos and I. Paraboni, "Myers-Briggs personality classification from social media text using pre-trained language models," Jul. 2022, doi: 10.3897/jucs.70941.

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

[13] F. Koto, J. H. Lau, and T. Baldwin, "IndoBERTweet: A Pretrained Language Model for Indonesian Twitter with Effective Domain-Specific Vocabulary Initialization," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 10660–10668.

[14] Kelvin, I. S. Edbert, and D. Suhartono, "UTILIZING INDOBERT IN PREDICTING PERSONALITY FROM TWITTER POSTS USING BAHASA INDONESIA," *ICIC Express Letters*, vol. 17, no. 1, pp. 123–130, Jan. 2023, doi: 10.24507/icicel.17.01.123.

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

[16] N. H. Jeremy, C. Prasetyo, and D. Suhartono, "Identifying personality traits for Indonesian user from twitter dataset," *International Journal of Fuzzy Logic and Intelligent Systems*, vol. 19, no. 4, pp. 283–289, 2019, doi: 10.5391/IJFIS.2019.19.4.283.

[17] H. Christian, D. Suhartono, A. Chowanda, and K. Z. Zamli, "Text based personality prediction from multiple social media data sources using pre-trained language model and model averaging," *J Big Data*, vol. 8, no. 1, pp. 1–20, 2021, doi: 10.1186/s40537-021-00459-1.

[18] G. Y. N. N. Adi, M. H. Tandio, V. Ong, and D. Suhartono, "Optimization for Automatic Personality Recognition on Twitter in Bahasa Indonesia," *Procedia Comput Sci*, vol. 135, pp. 473–480, 2018, doi: 10.1016/j.procs.2018.08.199.

[19] V. Ong, A. D. S. Rahmanto, W. Williem, N. H. Jeremy, D. Suhartono, and E. W. Andangsari, "Personality Modelling of Indonesian Twitter Users with XGBoost Based on the Five Factor Model," *International Journal of Intelligent Engineering and Systems*, vol. 14, no. 2, pp. 248–261, 2021, doi: 10.22266/ijies2021.0430.22.

[20] A. Kazameini, S. Fatehi, Y. Mehta, S. Eetemadi, and E. Cambria, "Personality Trait Detection Using Bagged SVM over BERT Word Embedding Ensembles," pp. 1–4, 2020, [Online]. Available: http://arxiv.org/abs/2010.01309

[21] M. Lewis *et al.*, "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 7871–7880.

[22] R. R. McCrae and P. T. Costa Jr, "The five-factor theory of personality.," 2008.

[23] M. H. Amirhosseini and H. Kazemian, "Machine learning approach to personality type prediction based on the myers–briggs type indicator®," *Multimodal Technologies and Interaction*, vol. 4, no. 1, p. 9, 2020.

[24] K. A. Nisha *et al.*, "A Comparative Analysis of Machine Learning Approaches in Personality Prediction Using MBTI," in *Computational Intelligence in Pattern Recognition*, Springer, 2022, pp. 13–23.

[25] C. Sumner, A. Byers, R. Boochever, and G. J. Park, "Predicting dark triad personality traits from twitter usage and a linguistic analysis of tweets," in *2012 11th international conference on machine learning and applications*, 2012, vol. 2, pp. 386–393.

[26] Z. M. M. Aung and P. H. Myint, "Personality Prediction Based on Content of Facebook Users:A Literature Review," in *2019 20th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)*, 2019, pp. 34–38. doi: 10.1109/SNPD.2019.8935692.

[27] D. Markovikj, S. Gievska, M. Kosinski, and D. Stillwell, "Mining facebook data for predictive personality modeling," in *Proceedings of the International AAAI Conference on Web and Social Media*, 2013, vol. 7, no. 2, pp. 23–26.

[28] N. de Ven, A. Bogaert, A. Serlie, M. J. Brandt, and J. J. A. Denissen, "Personality perception based on LinkedIn profiles," *Journal of Managerial Psychology*, 2017.

[29] F. Piedboeuf, P. Langlais, and L. Bourg, "Personality extraction through LinkedIn," in *Canadian Conference on Artificial Intelligence*, 2019, pp. 55–67.

[30] Y. Mehta, S. Fatehi, A. Kazameini, C. Stachl, E. Cambria, and S. Eetemadi, "Bottom-up and top-down: Predicting personality with psycholinguistic and language model features," in *2020 IEEE International Conference on Data Mining (ICDM)*, 2020, pp. 1184–1189.

[31] K. El-Demerdash, R. A. El-Khoribi, M. A. I. Shoman, and S. Abdou, "Deep learning based fusion strategies for personality prediction," *Egyptian Informatics Journal*, vol. 23, no. 1, pp. 47–53, 2022.

[32] M. Gjurković and J. Šnajder, "Reddit: A gold mine for personality prediction," in *Proceedings of the second workshop on computational modeling of people's opinions, personality, and emotions in social media*, 2018, pp. 87–97.

[33] E. J. Choong and K. D. Varathan, "Predicting judging-perceiving of Myers-Briggs Type Indicator (MBTI) in online social forum," *PeerJ*, vol. 9, p. e11382, 2021.

[34] J. Shen, O. Brdiczka, and J. Liu, "Understanding email writers: Personality prediction from email messages," in *International conference on user modeling, adaptation, and personalization*, 2013, pp. 318–330.

[35] B. Ferwerda and M. Tkalcic, "Predicting users' personality from instagram pictures: Using visual and/or content features?," in *Proceedings of the 26th conference on user modeling, adaptation and personalization*, 2018, pp. 157–161.

[36] E. Harris and A. C. Bardey, "Do Instagram profiles accurately portray personality? An investigation into idealized online self-presentation," *Front Psychol*, vol. 10, p. 871, 2019.

[37] A. C. E. S. Lima and L. N. de Castro, "A multi-label, semi-supervised classification approach applied to personality prediction in social media," *Neural Networks*, vol. 58, pp. 122–130, 2014, doi: https://doi.org/10.1016/j.neunet.2014.05.020.

[38] J.-M. Dewaele, "Personality: Personality traits as independent and dependent variables," in *Psychology for language learning*, Springer, 2012, pp. 42–57.

[39] G. Farnadi *et al.*, "Computational personality recognition in social media," *User Model User-adapt Interact*, vol. 26, no. 2, pp. 109–142, 2016.

[40] O. Sagi and L. Rokach, "Ensemble learning: A survey," *Wiley Interdiscip Rev Data Min Knowl Discov*, vol. 8, no. 4, p. e1249, 2018.

[41] M. A. Ganaie, M. Hu, and others, "Ensemble deep learning: A review," *arXiv preprint arXiv:2104.02395*, 2021.

[42] B. Verhoeven, W. Daelemans, and T. de Smedt, "Ensemble methods for personality recognition," in *Proceedings of the International AAAI Conference on Web and Social Media*, 2013, vol. 7, no. 2, pp. 35–38.

[43] O. Kampman, E. J. Barezi, D. Bertero, and P. Fung, "Investigating audio, visual, and text fusion methods for end-to-end automatic personality prediction," *arXiv preprint arXiv:1805.00705*, 2018.

[44] Z. R. Samani, S. C. Guntuku, M. E. Moghaddam, D. Preo\c{t}iuc-Pietro, and L. H. Ungar, "Cross-platform and cross-interaction study of user personality based on images on Twitter and Flickr," *PLoS One*, vol. 13, no. 7, p. e0198660, 2018.

[45] B. Wilie *et al.*, "IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding," *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pp. 843–857, 2020.

[46] Y. Zhang, D. Miao, Z. Zhang, J. Xu, and S. Luo, "A three-way selective ensemble model for multi-label classification," *International Journal of Approximate Reasoning*, vol. 103, pp. 394–413, 2018, doi: 10.1016/j.ijar.2018.10.009.

[47] W. Farlessyost, K. R. Grant, S. R. Davis, D. Feil-Seifer, and E. M. Hand, "The effectiveness of multi-label classification and multi-output regression in social trait recognition," *Sensors*, vol. 21, no. 12, pp. 1–15, 2021, doi: 10.3390/s21124127.

[48] P. Kamtar, D. Jitkongchuen, and E. Pacharawongsakda, "Multi-label classification of employee job performance prediction by disc personality," *ACM International Conference Proceeding Series*, pp. 47–52, 2019, doi: 10.1145/3366650.3366666.

[49] J. Risch and R. Krestel, "Bagging BERT models for robust aggression identification," in *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, 2020, pp. 55–61.

[50] B. Wilie *et al.*, "IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding," *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pp. 843–857, 2020, [Online].

Available: https://www.aclweb.org/anthology/2020.aacl-main.85

[51] S. Paul, "Bayesian Hyperparameter Optimization - A Primer on Weights & Biases," *www.wandb.ai*, 2020. https://wandb.ai/site/articles/bayesian-hyperparameter-optimization-a-primer (accessed Jan. 02, 2023).

[52] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," in *International Conference on Learning Representations*, 2018.