Personality Prediction Based on Text Analytics Using Bidirectional Encoder Representations from Transformers from English Twitter Dataset

Joshua Evan Ariyanto, Steven Gerald, Cyrena Tania, and Derwin Suhartono

Department of Computer Science, School of Computer Science, Bina Nusantara University, Jakarta, Indonesia

Ijfis

Abstract

Personality traits can be inferred from a person's behavioral patterns. One example is when writing posts on social media. Extracting information about individual personalities can yield enormous benefits for various applications such as recommendation systems, marketing, or hiring employees. The objective of this research is to build a personality prediction system that uses English texts from Twitter as a dataset to predict personality traits. This research uses the Big Five personality traits theory to analyze personality traits, which consist of openness, conscientiousness, extraversion, agreeableness, and neuroticism. Several classifiers were used in this research, such as support vector machine, convolutional neural network, and variants of bidirectional encoder representations from transformers (BERT). To improve the performance, we implemented several feature extraction techniques, such as N-gram, linguistic inquiry and word count (LIWC), word embedding, and data augmentation. The best results were obtained by fine-tuning the BERT model and using it as the main classifier of the personality prediction system. We conclude that the BERT performance could be improved by using individual tweets instead of concatenated ones.

Keywords: Personality prediction, Twitter, Big Five personality traits, BERT

1. Introduction

Billions of people around the world use social media to interact with others and share information. Through social media, people tend to express their feelings in words and characters. However, it is challenging to infer a person's personality through words and characters on social media.

Personality is a key factor that affects people's behavior. Personality is defined as a characteristic set of behaviors, cognitions, and emotional patterns formed by biological and environmental factors [1]. When encountering a certain situation, a person's reaction, expression, and emotion vary depending on their personality. Research shows that a person's personality information could help management teams decide on a working team because personality can affect work performance [2].

Social media platforms are full of personal information, as it records people's behaviors and interactions. Twitter is one of the largest social media platforms in the world, registering

Received: Mar. 17, 2021

Revised : Jun. 28, 2021

Accepted: Jul. 8, 2021

Correspondence to: Derwin Suhartono

(dsuhartono@binus.edu)

©The Korean Institute of Intelligent Systems

©This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/3.0/>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Figure 1. Twitter 2010-2019 worldwide active monthly user [3].

330 million active users as of 2019 [3] (Figure 1). As a microblogging platform, Twitter enables people to share information and communicate with each other in real time.

In this research, we aim to build a personality prediction system that uses English texts from Twitter as a dataset to predict a user's personality. We used the dataset from a previous research [4], which contains multilingual tweets from 900 users labeled using Big Five personality traits. To build a text classification model, we propose a fine-tuned bidirectional encoding representation from transformers (BERT) as the main model.

2. Related Work

There has been much research related to personality prediction using text datasets. Celli and Lepri [4] compared Big Five Traits and MBTI personality theory from a computing perspective using Twitter datasets. Using a support vector machine (SVM) and AutoWEKA as classifiers, the results show that prediction using MBTI classes achieved better accuracy than the one using Big Five Traits. Although Big Five Traits are much more informative, the variability of the performance also depends on the algorithm used for the prediction. Kazameini et al. [5] presented a deep learning model that outperformed the state-of-the-art on the stream of consciousness Essays dataset. The best model was built using a Bagged-SVM model over BERT word-embedding ensembles. The Bagged-SVM model was implemented using 10 SVM classifiers to predict in parallel. Another study by Aung and Mint [6] predicted personality based on the content of Facebook users. Several classifiers were compared with predict Big Five personality traits, including SVM, convolutional

neural networks (CNN), XGBoost, multi-layer perceptron, and mean absolute error (MAE). Moreover, the linguistic features of the different approaches were also applied when building the prediction system. The best performance was achieved using the CNN model with the addition of linguistic inquiry and word count (LIWC) linguistic features and feature selection using the Pearson correlation coefficient. Rahman et al. [7] proposed a CNN approach by implementing several linguistic features, such as Mairesse baseline feature set, NRC Emotion Lexicon (EmoLex), and Word2vec word embedding. Activation functions such as sigmoid, tanh, and leaky rectified linear units (ReLUs) were compared using the proposed model with the best performance achieved using leaky ReLU. Leonardi et al. [8] presented a multilingual personality trait estimator using the Transformers model by exploiting its capabilities of working at the sentence-level model. The model was built using sentence embeddings from the transformer encoder model and used as the input of a neural network model that performs a continuous regression. The reported result shows a better score than the previous state-of-the-art methods using the myPersonality dataset. An approach by Carducci et al. [9] to predict personality traits from tweets by building a model trained using Facebook text datasets. The model was built using an SVM classifier and Fast-Text word embedding, and then trained with hyperparameter tuning to find the best model configuration. BERT are an open-source language representation model developed by Google in 2018. BERT is designed to pre-train deep bidirectional representations from the unlabeled text by jointly conditioning on both the left and right contexts in all layers [10]. In this manner, BERT considers the same word within a different sentence to have a different meaning.

3. Dataset

The dataset used was obtained from previous studies [4]. It consists of 900 Twitter users with binary Big Five labels as y (high) or n (low) for each of the five personality traits (Table 1). Each user is given various Big Five personality tests, ranging from 10 to 44 items for each test. For a maximum of 40 tweets, data from each user were collected through Twitter advanced search queries. The dataset contains multilingual tweets in languages such as English, Italian, Spanish, and Dutch. For each user, the collected tweets were concatenated without a delimiter. The majority of sampled tweets were more than 17, with 794 of 900 users. Thus, in this study, we only used records with more than 17 sampled tweets. In addition, some

Table 1. Dataset label distribution

Trait	Original		Final	
	High	Low	High	Low
AGR	456	529	401	390
CON	529	369	468	323
EXT	457	441	420	371
NEU	449	449	395	396
OPN	477	421	401	390

AGR, agreeableness; CON, conscientiousness; EXT, extraversion; NEU, neuroticism; OPN, openness.

records with tweets consisting of only links and repeated words were discarded, resulting in final datasets consisting of only 791 users.

4. Methodology

Figure 2 presents an overview of the experiment. The dataset was preprocessed into ten different scenarios. Each scenario was tailored to suit specific classifiers. To validate the model during training, stratified 10-fold cross-validation was used. The classifiers were then evaluated in terms of accuracy, precision, recall, and F1 score. To utilize BERT as a classifier in some scenarios, we translated the multilingual dataset into English.

4.1 Preprocessing

Several preprocessing steps need to execute before extracting features from the tweets data.

1) Number of tweets filtration

We removed text data from less than 17 sampled tweets from the dataset. Bot-generated tweet data that consist of the same recurrence sentence and URL are removed as well.

2) Multilingual text translation

The multilingual text in the dataset was translated into English for some experiment scenarios.

3) Text cleaning

We performed several text cleaning methods to ensure that the model learned appropriate data. Several steps were performed, including retweet removal, URL removal, removal of more than three consecutive characters,

changing the username to “user” token, changing numbers to “number” token, and acronym elaboration (such as changing “I’ve” to “I have”).

4.2 Feature Extraction

We applied several feature extraction methods to find the best linguistic features that suit the data and model.

1) N-gram

As stated in [10], multiple languages are much more comparable through character N-grams. Thus, character N-grams were extracted from the multilingual dataset. For the English-translated dataset, word N-grams were extracted. All N-grams were associated with their TF-IDF values.

2) Linguistic inquiry and word count

Figure 2. Experiment flowchart for each single scenario.

In some scenarios, 88 features of the LIWC were extracted from the dataset. These features map words to the psycholinguistic categories as a ratio.

3) Word embedding

We apply word embedding to every deep learning approach. It is used to convert text into a meaningful vector of numbers. For CNN, 300 dimensions global vectors (GloVe) [11] pre-trained word embeddings. For the BERT model, we used WordPiece word embedding.

4) Analysis of variance (ANOVA)

In some scenarios, ANOVA was used as the univariate feature selection method. The top 2,000 best F-value features were selected.

5) Text split

The BERT classifier had a maximum input sequence length of 512. To handle this, in some scenarios, the concatenated tweets of each user were split into single tweets. The binary labels of the concatenated tweets were then spread to all the single tweets.

6) Data augmentation

Easy data augmentation (EDA) was used to increase the number of training data in some scenarios. EDA yields strong results for smaller datasets [12].

4.3 Classifier

Various classifiers were tested and compared. For the BERT classifier, we compared the fine-tuning approach and the feature-based approach [13]. The classifiers were SVM, CNN, BERT, BERT-Logistic Regression (BERT-LR), and BERT-SVM.

The WordPiece embedding tokenized the concatenated tweets into token sequences of various lengths. Of the 791 concatenated tweets, 525 had fewer than 400 tokens (Figure 3). Thus, for the configuration of a maximum input sequence in the BERT variants, a length of 400 was used.

The text split method separates the concatenated tweets into individual tweets for each user. For scenarios that included the text split method, the WordPiece embedding tokenized tweets have mostly around 80 tokens (Figure 4). Overall, 7,547 of the 7,690 total tweets had less than 80 tokens. Therefore, for BERT variants that used text splits, a maximum input sequence length of 80 was used.

In this research, we experimented with ten different scenarios to find the best method for use in the classifiers. The scenarios

Figure 3. Cumulative frequencies of the WordPiece tokenized concatenated tweets.

Figure 4. Cumulative frequencies of the WordPiece tokenized the tweets.

have different preprocessing methods. Table 2 describes the ten experimental scenarios with their corresponding classifiers.

The previous research conducted on the same dataset is described as Scenario FS0 with SVM and AutoWEKA Classifier from Table 2. Apart from FS0, this research experimented on different text preprocessing, feature extractions, classifiers, and data manipulation methods (as stated in Section 4.2). We focused highly on the BERT classifier, which used different text preprocessing methods, as stated in Section 4.1, and the feature extraction method included in the BERT model.

Table 2. Experimented feature set (scenario) from dataset

Name	Description	Classifier
FS0	(previous research, multilingual) Binary character N-gram, LIWC, metadata	SVM, CNN
FS1	(multilingual) TF-IDF character N-gram (N=3,4,5)	
FS2	TF-IDF Word N-gram (N=1,2,3)	
FS3	TF-IDF Word N-gram (N=2)	
FS4	TF-IDF Word N-gram (N=1,2,3), ANOVA	
FS5	TF-IDF word N-gram (N=1,2,3), LIWC, ANOVA	CNN, BERT, BERT-LR, BERT-SVM
FSW	Text	
FSWA	Text + Data Augmentation	
FSWS	Text (split)	
FSWSA	Text (split) + Data Augmentation	

5. Results

All predictive models were evaluated using the F1 measure metric, but we also included accuracy, recall, and precision metrics in our research. Each predictive model was trained into five specific models to predict each trait of the Big Five traits individually. Furthermore, the five trait-specific model results are averaged to be used as a comparison in Tables 3 and 4.

The fine-tuned BERT classifier had the highest F1 score among all classifiers (Table 4). Table 3 shows the average of each metric: accuracy (A-AVG), precision (P-AVG), recall (R-AVG), and F1 score (F-AVG). The bold value represents the best scenario for the fine-tuned BERT model.

Based on Table 3, the fine-tuned BERT achieved the highest score in all metrics in the FSWS scenario, which uses the text split and word embedding. Moreover, the text split method increased the performance of the fine-tuned BERT model. The performance of FSWS is higher than that of FSW, and FSWSA

Table 3. Evaluation metric result of the fine-tuned BERT variants

Scenario	A-AVG	F-AVG	P-AVG	R-AVG
FSW	0.55	0.55	0.56	0.59
FSWA	0.536	0.55	0.55	0.55
FSWS	0.60	0.59	0.60	0.646
FSWSA	0.58	0.55	0.57	0.57

The average was calculated for all five trait scores.

Table 4. Accuracy and F1 score metric result of all the classifiers

Model	A-AVG	F-AVG
Celli_SVM	0.591	-
Celli_AutoWEKA	0.67	-
SVM	0.624	0.576
CNN	0.533	0.58
BERT	0.60	0.59
BERT-LR	0.564	0.532
BERT-SVM	0.606	0.516

The average was calculated for all five trait scores.

is higher than that of FSWA.

Although data augmentation could increase the amount of dataset, the results showed that it decreased the performance of the fine-tuned BERT. The performance at FSWA is lower than that of FSW, and FSWSA is lower than that of FSWS.

For each classifier, several scenarios were tested and compared. Table 4 lists the best performances achieved for each classifier. The bold values represent the highest values for each metric. In addition, Celli_SVM and Celli_AutoWEKA are the previous research performances conducted on the same dataset. The average accuracy (A-AVG) was used as an evaluation metric based on previous research comparison purposes. The fine-tuned BERT, BERT-SVM, and SVM performed better than the Celli_SVM model, but worse than the Celli_AutoWEKA model.

The F1 score was used as the main evaluation metric to avoid the accuracy paradox. The highest value of the average F1 score (F-AVG) was achieved using the fine-tuned BERT model. However, the feature-based approach of BERT, which is BERT-LR and BERT-SVM, performed poorly in comparison with fine-tuned BERT and SVM. This indicates that, for the current dataset, the fine-tuned approach is better than the feature-based approach BERT. The fine-tuned BERT successfully surpassed the performance of CNN with GloVe word embeddings and SVM.

The fine-tuned BERT model was chosen as the main model owing to the best F1 score evaluation metric result.

5.1 Working Procedure

The final working procedure for predicting personality is based on the fine-tuned trained BERT model, as shown in Figure 5. The method is composed of four chronological main steps: text preprocessing, text tokenization, feature extraction, and classification. The text preprocessing method removes URLs and

retweets, changes usernames and digits to tokens, and elaborates acronyms. The preprocessed tweets were then tokenized using the trained BERT tokenizer from the experiment. The maximum token length used was 80, based on the previous dataset on which the BERT was trained. Feature extraction was performed by trained BERT, which resulted in WordPiece embeddings. The WordPiece embeddings were then passed on the trained single-layer perceptron to classify each trait from each tweet as a binary value. To achieve a single binary class for all tweets, a voting mechanism was used to choose the dominant classified binary class. There are five fine-tuned trained BERT models for each specific Big Five traits. Finally, there were five binary classes from each trait that represent the five dimensions of personality expressed in the analyzed tweets.

Figure 5. Working procedure of the proposed method.

6. Conclusion

The aim of this study was to find the best approach to predict personality based on the Big Five personality theory using Twitter as our data source. To do so, we compared several classifiers using preprocessed feature sets. The classifiers used in this research were SVM, CNN, and BERT. Based on the results, the deep learning approach performed better with the BERT model and attained the highest performance score. In the BERT model, applying the text split method to the feature set increased the performance score, while applying data augmentation reduced the performance score. For this dataset, using pre-trained BERT with a fine-tuning approach yields a better result compared with the feature-based approach. The LIWC linguistic feature was also applied, and it showed a better result compared with the feature sets that did not use it. The insufficient number of datasets is the main limitation of this study, as a deep learning model requires a large number of datasets. In future works, it is advisable to implement a semi-supervised approach to obtain a larger dataset or choose a larger dataset in advance.

Conflict of Interest

No potential conflict of interest relevant to this article was reported.

Acknowledgement

We would like to express our deep gratitude to Fabio Celli, Ph.D., for providing the dataset and constructive advice on this research.

References

- [1] P. J. Corr and G. Matthews, *The Cambridge Handbook of Personality Psychology*. Cambridge, UK: Cambridge University Press, 2009.
- [2] R. Lambiotte and M. Kosinski, "Tracking the digital footprints of personality," *Proceedings of the IEEE*, vol. 102, no. 12, pp. 1934-1939, 2014. <https://doi.org/10.1109/JPROC.2014.2359054>
- [3] Statista, "Number of monthly active Twitter users worldwide from 1st quarter 2010 to 1st quarter 2019," 2021 [Online]. Available: <https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/>.

- [4] F. Celli and B. Lepri, "Is big five better than MBTI? A personality computing challenge using Twitter data," in *Proceedings of the 5th Italian Conference on Computational Linguistics (CLiC-it)*, Torino, Italy, 2018.
- [5] A. Kazameini, S. Fatehi, Y. Mehta, S. Eetemadi, and E. Cambria, "Personality trait detection using bagged SVM over BERT word embedding ensembles," in *Proceedings of the 4th Widening Natural Language Processing Workshop*, Seattle, WA, 2020.
- [6] Z. M. M. Aung and P. H. Myint, "Personality prediction based on content of Facebook users: a literature review," in *Proceedings of 2019 20th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)*, Toyama, Japan, 2019, pp. 34-38. <https://doi.org/10.1109/SNPD.2019.8935692>
- [7] M. A. Rahman, A. Al Faisal, T. Khanam, M. Amjad, and M. S. Siddik, "Personality detection from text using convolutional neural network," in *Proceedings of 2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT)*, Dhaka, Bangladesh, 2019, pp. 1-6. <https://doi.org/10.1109/ICASERT.2019.8934548>
- [8] S. Leonardi, D. Monti, G. Rizzo, and M. Morisio, "Multilingual transformer-based personality traits estimation," *Information*, vol. 11, no. 4, article no. 179, 2020. <https://doi.org/10.3390/info11040179>
- [9] G. Carducci, G. Rizzo, D. Monti, E. Palumbo, and M. Morisio, "Twitpersonality: computing personality traits from tweets using word embeddings and supervised learning," *Information*, vol. 9, no. 5, article no. 127, 2018. <https://doi.org/10.3390/info9050127>
- [10] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "Bert: pre-training of deep bidirectional transformers for language understanding," 2018 [Online]. Available: <https://arxiv.org/abs/1810.04805>.
- [11] C. Lecluze, L. Rigouste, E. Giguët, and N. Lucas, "Which granularity to bootstrap a multilingual method of document alignment: character N-grams or word N-grams?," *Procedia-Social and Behavioral Sciences*, vol. 95, pp. 473-481, 2013. <https://doi.org/10.1016/j.sbspro.2013.10.671>
- [12] J. Pennington, R. Socher, and C. D. Manning, "Glove: global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, 2014, pp. 1532-1543.
- [13] J. Wei and K. Zou, "Eda: easy data augmentation techniques for boosting performance on text classification tasks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China, 2019, pp. 6381-6387.

Joshua Evan Arijanto is an undergraduate at Bina Nusantara University (BINUS). His research interests include natural language processing and personality prediction.
E-mail: joshua.arijanto@binus.ac.id

Steven Gerald is an undergraduate at Bina Nusantara University (BINUS). His research interests include natural language processing and personality prediction.
E-mail: steven.geraldy@binus.ac.id

Cyrena Tania is an undergraduate at Bina Nusantara University (BINUS). Her research interests include natural language processing and personality prediction.
E-mail: cyrena.tania@binus.ac.id

Derwin Suhartono is faculty member of Bina Nusantara University, Indonesia. He got his Ph.D. degree in computer science from Universitas Indonesia in 2018. His research fields are natural language processing. Recently, he is continually doing research in argumentation mining and personality recognition. He actively involves in Indonesia Association of Computational Linguistics (INACL), a national scientific association in Indonesia. He has his professional memberships in ACM, INSTICC, and IACT. He also takes role as reviewer in several international conferences and journals.
E-mail: dsuhartono@binus.edu