ICIC Express Letters
Part B: Applications
Volume **13**, Number **3**, March **2022**

ICIC International ©2022 ISSN 2185-2766
pp. **233–240**

TOWARDS PERSONALITY IDENTIFICATION FROM SOCIAL MEDIA TEXT STATUS USING MACHINE LEARNING AND TRANSFORMER

Jefri Tanwijaya[1,*] and Derwin Suhartono[2]

[1]Computer Science Department, BINUS Graduate Program – Master of Computer Science
[2]Computer Science Department, School of Computer Science
Bina Nusantara University
Jl. K. H. Syahdan No. 9, Kemanggisan, Palmerah, Jakarta 11480, Indonesia
[*]Corresponding author: jefri.tanwijaya@binus.ac.id; dsuhartono@binus.edu

Received May 2021; accepted August 2021

Abstract. *Social media gathers a lot of data from its users. These data are seen by the researchers as valuable research materials, one of which is to examine personality identification. This area of research is gaining more popularity among researchers. This study focused on exploring machine learning provided with hyperparameter tuning as well as transformer model as a means of personality identification. Some experiments were conducted in accordance with the Big Five Personality traits. This study used machine learning that is often used in this research topic such as Multinomial Naïve Bayes, Support Vector Machine, Logistic Regression, Random Forest, Gradient Boosting, and eXtreme Gradient Boosting. The results of the study proved that the performance of the transformer surpassed machine learning model with an average accuracy of 62.79%.*

Keywords: Personality identification, Big Five Personality, Machine learning, Transformer

1. **Introduction.** The increasing development of communication technology has resulted in the wider use of social media. One of the social media that has been influenced by the development of communication technology is Facebook. As demonstrated in Figure 1, Facebook has a significant escalation in the number of users every year. In December 2020, there were 2.7 million Facebook users worldwide [1].

On the other hand, Facebook may cause a degradation in mental health for its users [2] such as depression, anxiety, and psychological pressure [3]. Moreover, Facebook gathers a large number of data that can be processed and used as research materials in the field of social sciences, for instance, to investigate people's personality according to their social media activities [4]. Understanding a person's personality can be very useful for a variety of purposes in the field of health and technology, such as early warning based on the last status behavior created, improving the capability of recommendation systems [5] and decision-making processes [6].

A large number of experiments have been carried out in personality identification, one of which is using Artificial Intelligence (AI). Basically, AI is a technique that imitates the intelligence of living things and inanimate objects to solve a problem. One of the methods in AI is machine learning which imitates how humans solve problems [7]. AI and machine learning are now popular in a wide range of research fields such as healthcare, economics, and manufacturing [8], making AI and machine learning promising methods to use. Many variants of machine learning methods have been used for personality prediction, for instance, Multinomial Naïve Bayes, Support Vector Machine, Logistic Regression, Gradient Boosting, and Linear Discriminant Analysis (LDA) [9].

DOI: 10.24507/icicelb.13.03.233

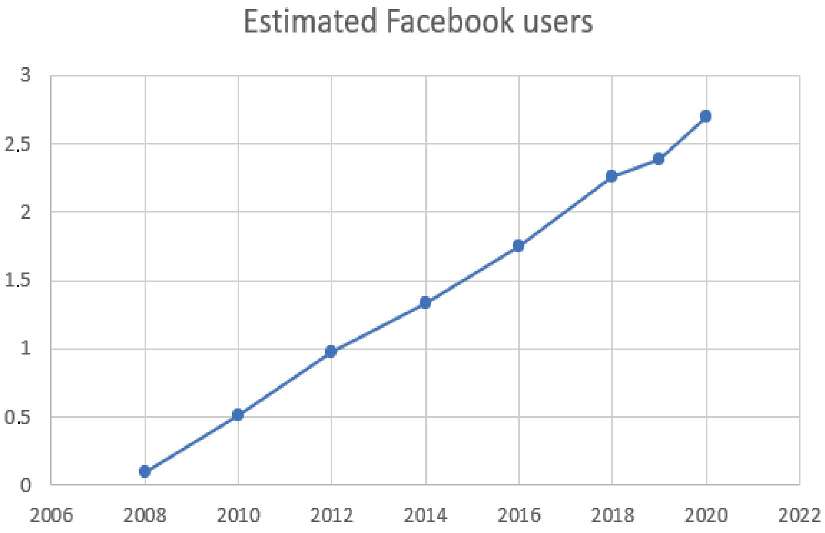

Figure 1. Estimated amount of Facebook users in 2008-2020

However, since machine learning's accuracy needs improvements, deep learning methods appear to be better solutions. Deep learning is a method used to solve complex problems that imitates how the human brain works to obtain the required knowledge [10]. The use of deep learning, such as Multi-Layer Perceptron (MLP), Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), and 1-Dimensional Convolutional Neural Network (CNN-1D) in personality identification is proven to have better performance than machine learning [6].

This research will use the MyPersonality dataset. In identifying the personality, several machine learning algorithms that are often used in this research topic and transformers were used. The machine learning was given several scenarios such as providing feature selection using Chi-square, feature extraction using TF-IDF and hyperparameter tuning using the grid search method. This paper contributes to the exploration of machine learning algorithms given hyperparameter tuning and transformers which are still rarely used in this particular topic. In addition, the results of this study indicate that transformers have better performance than machine learning.

2. **Related Work.** Many studies have been conducted in the area of personality prediction. This study made use of a dataset labeled with the Big Five Personality traits, namely Openness (OPN), Conscientiousness (CON), Extraversion (EXT), Agreeableness (AGR), and Neuroticism (NEU). These five personalities describe the basic personality of humans [11]. In personality prediction, there are many components in social media that can be observed as the characteristics of features, such as status, videos, and user habits. However, this research only focused on text data for personality identification.

There are several social media that are commonly used in personality identification research, for instance, Facebook, Twitter, Blogs, and YouTube. There is a study which used Twitter profile data as a feature set which were further analyzed using the ZeroR machine learning method and the Gaussian Process [12]. Another study on personality identification used dataset from Facebook, Twitter, and YouTube altogether with a cross-media learning approach to conduct personality analysis on social media. The study aimed to answer several questions in personality identification problems, such as determining whether these problems should be treated as multi-label prediction problems or separate predictions, determining good features, and discovering what reduces the accuracy after using the same model on different social media [6].

Along with the development of research in the area of personality prediction, the use of machine learning and deep learning has become increasingly popular. There is a research that compared machine learning (Naïve Bayes, Support Vector Machine (SVM), Logistic Regression, Gradient Boosting and Linear Discriminant Analysis (LDA), eXtreme Gradient Boosting (XGBoost)) performance to deep learning (Multi-Layer Perceptron (MLP), Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), 1-Dimensional Convolutional Neural Network (CNN-1D), LSTM + CNN-1D). This research used several scenarios, such as resampling, feature selection, and several kinds of feature extraction such as Linguistic Inquiry and Word Count (LIWC), Structured Programming for Linguistic Cue Extraction (SPLICE), and Social Network Analysis (SNA). The results proved that MLP demonstrates the highest average accuracy for MyPersonality dataset with an accuracy of 70.78%, followed by CNN-1D with an accuracy of 63.84%. While in the manually collected dataset, LSTM + CNN-1D indicates the best accuracy of 74.17%, followed by MLP with an accuracy of 73.87% [10].

There were also researches that attempted to use XGBoost which draws on the correlation between each feature set. This research reported that XGBoost demonstrates the best average accuracy of 74.2% [13]. In another study, the combination of LIWC feature extraction was carried out by which each status on the dataset was trained on Google news pre-trained word2vec with 300 dimensions. The research utilized CNN to predict the personality. For Openness trait, the proposed method has the highest accuracy of 76% compared to other methods [14]. Based on the aforementioned researches, hyperparameter tuning is rarely performed. Therefore, this research attempted to use hyperparameter tuning and transformer as a means of examining the personality prediction.

Another recent study uses two datasets, namely MyPersonality as training data and Twitter dataset from the NetMiner tool as testing data. This study took machine learning approaches such as Logistic Regression, Multinomial Naïve Bayes, Gaussian Naïve Bayes, and Random Forest. Meanwhile, the proposed model in the study uses both deep sequential neural network and multi-target regression. It was found that the proposed model obtained an average accuracy of 78% [15]. In addition, another study took the approach of Random Forest, XGBoost and AdaBoost with the addition of Penalized SVM and Up-Sampled-Logistic Regression as a baseline. This study used a pre-labeled tweets dataset and adapted an open vocabulary approach as a feature extraction. It was found that Adaboost had the best accuracy with 81%, but in terms of recall, XGBoost had the best recall of 58%. In that study personality prediction is used as cyber violence detection in social media [16]. Personality identification can also be used as a first alarm if there are any indications of depression, improving the capabilities of recommendation systems [5] and decision-making processes [6].

3. **Methodology.** This research is divided into several parts. Research planning explains how the problem was raised into a research material. Data preparation and model initialization explain how initialization was needed to start and build a research model. Training model section explains how the model processes preprocessed data and how the machine learning and transformer is evaluated. The tuning parameter section was tested on the parameters until the best parameter was obtained. Finally, evaluation was carried out on the proposed model.

Figure 2 elaborates the workflow of our research. In the first stage, research planning, identification of problems found in the personality prediction area was carried out. Approaches needed to overcome these problems were determined. Literature studies regarding concepts and models were also collected at this stage.

In the second stage, data collection and initialization of components and models were carried out. MyPersonality dataset was downloaded. The dataset was analyzed to get the characteristics such as the amount of data, the number of classes, and the distribution of

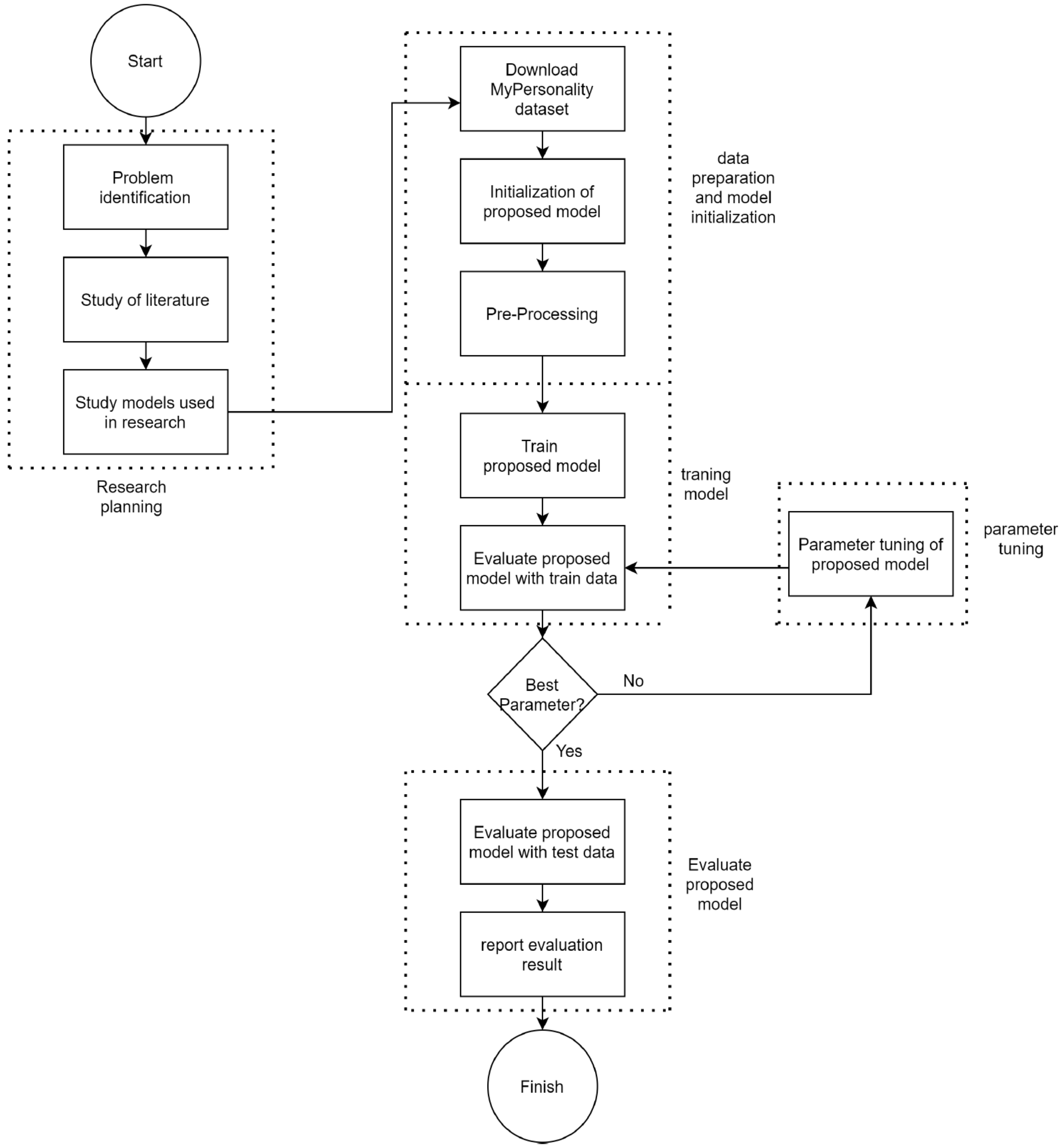
Start
Problem identification
Study of literature
Study models used in research
Research planning
Download MyPersonality dataset
Initialization of proposed model
Pre-Processing
data preparation and model initialization
Train proposed model
Evaluate proposed model with train data
traning model
Parameter tuning of proposed model
parameter tuning
Best Parameter?
No
Yes
Evaluate proposed model with test data
report evaluation result
Evaluate proposed model
Finish

Figure 2. Research workflow

data in each class. After the analysis was completed, machine learning and transformer models were initialized. Machine learning approaches which are often used in text classification and personality prediction problems such as Multinomial Naïve Bayes, Support Vector Machine (SVM), Logistic Regression, Random Forest, Gradient Boosting, and eXtreme Gradient Boosting were employed in this research.

Furthermore, the data went through several preprocessing processes, such as changing letters to lowercase, deleting Uniform Resource Locator (URL), removing punctuation, and democratization. Moreover, feature extraction was carried out using the Term Frequency-Inverse Document Frequency (TF-IDF). Once extracted, the data became a feature vector which was then used for feature selection. Feature selection aimed to get the best features so that it can produce a good classification model. In addition, the feature selection could also reduce the dimensionality problem so that it could reduce resource usage when processing the data. The data were then divided into two parts, namely training data and testing data.

The proposed model was trained and the performance results were calculated. If the performance was not good enough, parameter tuning would be employed. This stage replaced the old model parameters with the new ones in hope that these parameters could improve the model's performance. The model with new parameters was returned to the training stage and re-evaluated. This was done continuously until stopping conditions were met, for instance, when convergent performance and maximum iteration were achieved. The final stage was the evaluation of the proposed model. Performance results with testing data were reported as an evidence of testing the proposed model.

3.1. **Dataset.** The dataset used in this study is MyPersonality which contains Facebook status. This dataset is commonly used in personality prediction research because it has been labeled with the Big Five Personality traits. MyPersonality dataset contains 9,918 different states that have been labeled.

Table 1 contains the number of distributions for each class in the dataset. Openness has the highest number of statuses with 7,370 posts, followed by Agreeableness (5,268 posts), Conscientiousness (4,556 posts), Extraversion (4,210 posts). Class with the smallest number of posts is Neuroticism with 3,717 posts. Based on the total distribution, it can be concluded that there is an imbalance in the distribution of the class, especially in the Neuroticism class and the Openness class.

Table 1. Number of statuses in each class

	EXT	NEU	AGR	CON	OPN
Total	4,210	3,717	5,268	4,556	7,370

3.2. **Preprocessing.** Preprocessing is the most important stage before the model performs classification. There are several steps in preprocessing, that is, changing letters to lowercase and removing URLs, punctuation, and numbers. Social media status often contains URLs, emojis, and abbreviations so that additional preprocessing is required. It covers extending abbreviations, acronyms, initials, slang and removing user tags, hashtags, and emojis. This stage also added wrong word justification, such as typos and repeated letters.

3.3. **Feature extraction.** This study used TF-IDF as a feature extraction. TF-IDF algorithm assigns weight to words based on the number of word occurrences in a document. With this weighting, it was hoped that TF-IDF could extract features that were considered important in the document.

3.4. **Feature selection.** Feature selection plays an important role in the classification because of its big impact, that is, increasing the speed and accuracy of the classification model. This is due to the fact that feature selection can reduce the dimensions of the data used and select the best features so that the classification model could be more optimal. In this study, Chi-square was employed to select features.

4. **Results and Discussion.** As previously stated, this study used machine learning methods and transformer. The machine learning models used were Multinomial Naïve Bayes, Support Vector Machine (SVM), Logistic Regression, Random Forest, Gradient Boosting, and eXtreme Gradient Boosting. Machine learning models passed several trial scenarios and were then compared with transformer to get the best performance.

4.1. **Experiment.** Experimental scenarios were only given to machine learning because deep learning performed different features engineering at the time the training was carried out. First, validation was carried out on the dataset by means of 10-fold cross validation. The results of the 10-fold cross validation were divided into 80% for training data and 20% for testing data. To discover the correlation between words, n-gram was used. In this case, trigram was employed. This value was obtained from the trials conducted.

There were several test scenarios for all machine learning models. All scenarios were added one by one to be processed so that the accuracy of machine learning slowly increased. The stage began with the use of feature selection using Chi-square. Some adjustments were needed so that the result of the feature selection could be as expected. After that, the provision of TF-IDF was carried out as feature extraction.

In the final stage, hyperparameter tuning was added to the machine learning models. Hyperparameter tuning is a combination of parameters that controls the learning process in the model. Just like feature selection, hyperparameter tuning is very important in machine learning models before the training process is carried out. Doing a hyperparameter tuning can minimize loss function and give good results in the classification. Hyperparameter tuning used in this study was grid search. Grid search is a technique that manually searches for specific subsets within the hyperparameter search space. Table 2 presents the scenario used in this study.

Table 2. Experimental machine learning scenarios

Scenarios	Hyperparameter tuning		TF-IDF		Feature selection	
	No	Yes	No	Yes	No	Yes
1	√		√		√	
2	√		√			√
3	√			√	√	
4	√			√		√
5		√	√		√	
6		√	√			√
7		√		√	√	
8		√		√		√

4.2. **Results.** Table 3 provides the information of the best accuracy in each model of a certain class. The machine learning method has succeeded in obtaining an average accuracy of 61% and transformer with an average accuracy of 62.79% in examining personality prediction on MyPersonality dataset. Based on the results, it was obtained that scenario 8, in which hyperparameter tuning, feature extraction, and feature selection were used, produced the highest accuracy compared to others.

Table 3. Classification results of machine learning and transformer

Scenarios	Big Five Personality					Average
	EXT	NEU	AGR	CON	OPN	
Multinomial Naïve Bayes	58.06%	63.14%	54.06%	55.08%	74.79%	61.02%
Support Vector Machine	58.17%	63.29%	55.23%	54.35%	74.90%	61.18%
Logistic Regression	58.07%	63.23%	54.11%	55.08%	75.03%	61.10%
Gradient Boosting	58.12%	63.17%	54.11%	55.18%	75.0%	61.11%
Random Forest	58.67%	62.86%	54.53%	56.27%	74.87%	61.44%
XGBoost	59.13%	63.82%	54.30%	55.11%	74.7%	61.41%
Transformer	58.44%	63.22%	58.2%	58.66%	75.46%	62.79%

In the first scenario, machine learning was operated without using any additions. The highest average accuracy was 60.81%, achieved by XGBoost. In scenarios 2, 3 and 4, trials were carried out one by one starting from feature selection, feature extraction, and hyperparameter tuning. XGBoost resulted in the highest accuracy of 61.74%, 61.22%, and 60.36%, followed by Random Forest of 60.23%, 60.37%, and 60.13% respectively.

The results of scenario 8 indicate that transformer has the highest accuracy with an accuracy of 62.79%. However, in each machine learning model there was no significant difference in accuracy. For instance, Random Forest managed to get an average accuracy of 61.44%, followed by XGBoost with an accuracy of 61.41%. The highest accuracy in terms of class was achieved by Openness with an accuracy of 75.46% using transformer.

Based on the results of the analysis, the machine learning performance did not make up any significant difference in accuracy and was quite stable between 61%. As for the accuracy of each class, the Openness class had the highest average accuracy of 74.96%. The use of transformers succeeded in increasing the average accuracy by 62.79%, which outperformed all accuracy from machine learning.

This study found that each machine learning and transformer model still cannot handle all classes in the Big Five Personality, but the use of transformer managed to produce better accuracy than the machine learning method. The results of this study may differ from other studies due to differences in the parameters and processing data used.

4.3. **Discussion.** This research examined personality prediction using machine learning and transformer. The data were processed through preprocessing, feature extraction, and feature selection. Machine learning was optimized using hyperparameter tuning with grid search method. This study found that hyperparameter tuning in machine learning can increase the accuracy quite significantly, but the accuracy among machine learning methods did not have a significant difference. The results of these findings are not much different from the previous study [9]. On the other hand, transformer managed to provide better accuracy than all machine learning methods with an average accuracy of 62.79%.

One of the limitations in this study is the use of feature extraction variations such as LIWC and SPLICE. The use of different feature extractions was expected to provide better accuracy than TF-IDF [17]. In addition, the transformer model also needs to be optimized. Future research can consider the use of other deep learning which is expected to provide better accuracy without considering feature engineering problems and the use of evolutionary algorithms to perform hyperparameter tuning.

5. **Conclusions.** The accuracy value obtained using machine learning was not good enough since there were classes that still had low accuracy. Based on the data obtained, it is very important to carry out the process of feature selection, feature extraction, and hyperparameter tuning. It is also necessary to develop hyperparameter tuning and text processing to produce good accuracy. However, by using the deep learning model, the transformers were proven to be able to beat the average accuracy of all the machine learning models tested with an accuracy of 62.79% and get the highest accuracy on the Openness class with an accuracy of 75.46%.

In the future study, deep learning will be used and the use of evolutionary algorithms will be considered to perform hyperparameter tuning so that parameter determination is not done manually.

REFERENCES

[1] Statista, *Facebook: Number of Monthly Active Users Worldwide 2008-2020*, https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/, Accessed on February 2, 2021.

[2] H. H. S. Kim, The impact of online social networking on adolescent psychological well-being (WB): A population-level analysis of Korean school-aged children, *International Journal of Adolescence and Youth*, vol.22, no.3, pp.364-376, 2017.

[3] B. Keles, N. McCrae and A. Grealish, A systematic review: The influence of social media on depression, anxiety and psychological distress in adolescents, *International Journal of Adolescence and Youth*, vol.25, no.1, pp.79-93, 2020.

[4] M. Kosinski, S. C. Matz, S. D. Gosling, V. Popov and D. Stillwell, Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines, *American Psychologist*, vol.70, no.6, 2015.

[5] A. V. Prando, F. G. Contratres, S. N. A. de Souza and L. S. de Souza, Content-based recommender system using social networks for cold-start users, *KDIR*, pp.181-189, 2017.

[6] G. Farnadi, G. Sitaraman, S. Sushmita, F. Celli, M. Kosinski, D. Stillwell, S. Davalos, M. C. Moens and M. De Cock, Computational personality recognition in social media, *User Modeling and User-Adapted Interaction*, vol.26, no.2, pp.109-142, 2016.

[7] N. Kühl, M. Goutier, R. Hirt and G. Satzger, Machine learning in artificial intelligence: Towards a common understanding, *arXiv.org*, arXiv: 2004.04686, 2020.

[8] R. Cioffi, M. Travaglioni, G. Piscitelli, A. Petrillo and F. De Felice, Artificial intelligence and machine learning applications in smart production: Progress, trends, and directions, *Sustainability*, vol.12, no.2, 2020.

[9] T. Tandera, D. Suhartono, R. Wongso and Y. L. Prasetio, Personality prediction system from Facebook users, *Procedia Computer Science*, vol.116, pp.604-611, 2017.

[10] I. Goodfellow, Y. Bengio, A. Courville and Y. Bengio, *Deep Learning*, MIT Press, Cambridge, 2016.

[11] B. De Raad and B. Mlacic, Big five factor model, theory and structure, *International Encyclopedia of the Social & Behavioral Sciences*, no.2, 2015.

[12] J. Golbeck, C. Robles, M. Edmondson and K. Turner, Predicting personality from Twitter, *2011 IEEE 3rd International Conference on Privacy, Security, Risk and Trust and 2011 IEEE 3rd International Conference on Social Computing*, pp.149-156, 2011.

[13] M. M. Tadesse, H. Lin, B. Xu and L. Yang, Personality predictions based on user behavior on the Facebook social media platform, *IEEE Access*, no.6, pp.61959-61969, 2018.

[14] C. Yuan, J. Wu, H. Li and L. Wang, Personality recognition based on user generated content, *2018 15th International Conference on Service Systems and Service Management*, pp.1-6, 2018.

[15] N. Aslam, K. M. Khan, A. N. S. Munir and J. Nadeem, Analysis of personality assessment based on the five-factor model through machine learning, *International Journal of Scientific & Technology Research*, vol.10, 2021.

[16] R. Zarnoufi and M. Abik, Big five personality traits and ensemble machine learning to detect cyberviolence in social media, in *Innovation in Information Systems and Technologies to Support Learning Research. EMENA-ISTL 2019. Learning and Analytics in Intelligent Systems*, M. Serrhini, C. Silva and S. Aljahdali (eds.), Cham, Springer, 2019.

[17] A. ElMessiry, Z. Zhang, W. O. Cooper, T. F. Catron, J. Karrass and M. P. Singh, Leveraging sentiment analysis for classifying patient complaints, *Proc. of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pp.44-51, 2017.