JURNAL EMACS
[Engineering, MAthematics and Computer Science] Vol.6 No.1 January 2024: 1-6
e-ISSN: 2686-2573
DOI: 10.21512/emacsjournal.v6i1.10748

Machine Learning for Predicting Personality Using Facebook-Based Posts

Derwin Suhartono[1*], Marcella Marella Ciputri[2], Stefanny Susilo[3]

[1-3] Computer Science Department, School of Computer Science,
Bina Nusantara University,
Jakarta, Indonesia 11480
dsuhartono@binus.edu; marcella.ciputri@binus.ac.id; stefanny.susilo@binus.ac.id

*Correspondence: dsuhartono@binus.edu

Abstract – *Social media contributes a lot to human life. People can share their thoughts through text, photos, and voice through social media. Information from social media can be useful, including in personality research. Personality can generally be known through personality tests. In this research, personality prediction is formed to determine personality through Facebook posts without using a personality test. We create a model based on big five personality traits using 5 machine learning algorithms: Support Vector Machine (SVM), Multinomial Naive Bayes, Decision Tree, K-Nearest Neighbor, and Logistic Regression. Data augmentation was also used for balancing the dataset value and trained using stratified 10-fold cross-validation. This research yields the highest f1 score on Openness using Multinomial Naive Bayes algorithm of 82.31% and the highest average is 68.62%. So the five supervised Machine Learning algorithms used in this research produced Multinomial Naive Bayes as the best algorithm to predict personality based on big five personality traits from user postings on Facebook.*

Keywords: *Personality Prediction; Big Five Personality; Social Media; Machine Learning; Facebook*

I. INTRODUCTION

In this period, technology has grown rapidly and everyone uses it for daily life. Social media is a digital platform developed with assistance from technology. Everyone can express themselves using social media by contributing text, videos, or images. Some familiar social media are Twitter, Facebook, Instagram, Youtube, Whatsapp, and many more. All social media have the same purpose which is to communicate with people over long distances. Social media users continue to grow every year. The ranking can be seen based on the number of users. Instagram occupies the fourth position while Youtube and Whatsapp occupy the second and third positions. Facebook is the most popular social media that has been used (Statista, 2022). Youtube and Facebook dominate the internet, with 81% and 69% of users (Auxier & Anderson, 2021).

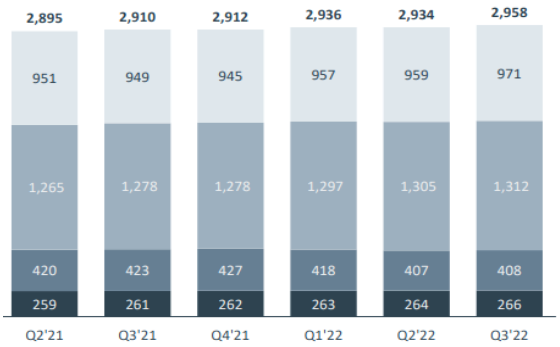

Figure 1. Facebook Monthly Active Users (Source: Meta, 2022)

Facebook users continue to increase annually. On Facebook, people can become friends, engage in one-on-one conversations, and update their statuses. Everyday, a lot of people write regarding their emotions or upload information that the audience can view to convey their feelings. Additionally, there is evidence to support the idea that content created and published on social media user profiles acts as an extension of "one's self" and accurately portrays each user's unique personality rather than highlighting their best qualities (Azucar et. al, 2018). Social media expressions can classify a person's personality and behavior. Every human have different personality and it as a guideline for judging oneself. This happens because

personality influences behaviour, interaction, socialization, establishment, potential, survival, and many more. Personality is very important in the world of work because it can assess performance. Personality can determine the best position to place in each employee. In some cases, it is possible that there are companies that lean more towards personality than skill because basically that personality can turn an ability to achievement. Numerous earlier research (Azucar et. al, 2018), (Suhartono et. al, 2021), (Kunte & Panicker, 2019) that analyzed their nature with their activity on social media substantiate this conclusion. Considering present technology, personality can be determined based on online behavior. Therefore, the goal of this research is to analyze a personality prediction system based on Facebook posts using Machine Learning.

There are various personality classifications that can be used, such as Myers-Briggs Type Indicator (MBTI), Big Five Personality Traits, and Dominance Influence Steadiness Conscientiousness (DISC). The most popular personality models are MBTI and Big Five Personality Traits. Since the MBTI type is a 4-letter coding (for example; INTJ), it is simpler to gather gold-standard labeled data about MBTI than about Big Five which makes MBTI more popular now than it was previously (Celli & Lepri, 2018). From our literature review, the use of Big Five Personalities these days also increases. We chose the Big Five Personality Traits as our personality model for this research. Big Five Personality Traits consists of five traits that are usually called as OCEAN (Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism).

Machine learning has contributed a lot to research, especially in personality prediction. There are several machine learning algorithms that have been used for personality prediction systems from previous research, such as traditional machine learning or even deep learning implementation with some additional features. Additional features such as LIWC (Linguistic Inquiry and Word Count) and SNA (Social Network Analysis) are examples of famous features for personality prediction.

Several earlier studies (Kunte & Panicker, 2019), (Tandera et. al, 2017) make use of the well-known dataset like MyPersonality dataset. MyPersonality dataset consists of 250 Facebook users, 9,918 records, Facebook posts in raw texts, and network features. This dataset is labeled with the Big Five Personality, making it relevant for this research. As the dataset contains some imbalanced values, we made some adjustments to the data with data augmentation. Facebook post text will undergo preprocessing for improved quality. This research implements five machine learning algorithms, such as Support Vector Machine (SVM), Multinomial Naive Bayes, Decision Tree, K-Nearest Neighbor, and Logistic Regression as the model classifier and integrates Stratified K-Fold Cross Validation for improved outcomes.

Big Five Personality which includes Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism (OCEAN) served as the personality model for our experiment (Rosen & Kluemper, 2008). Example of Big Five Personality Traits is presented in Figure 2 below.

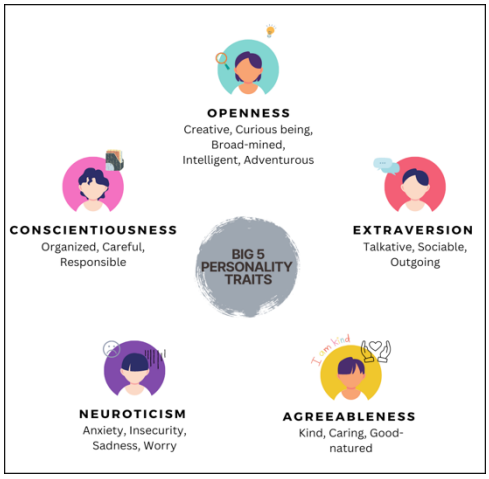

Figure 2. Big Five Personality Description

There have been numerous personality prediction studies, such as building personality prediction for Indonesian users from Twitter dataset (Jeremy et. al, 2019), (Suhartono et. al, 2017), personality prediction from multiple social media (Suhartono et. al, 2021), or personality prediction with different personality types (Ontoum & Chan, 2022). Similar to our experiments, a previous study (Tandera et. al, 2017), (Aung & Myint, 2019) used MyPersonality as their dataset. In previous research (Kunte & Panicker, 2019), they implemented 3 machine learning, including Naive Bayes, SVM, and KNN using TF-IDF with the highest accuracy achieved by Naive Bayes. Another research (Tandera et. al, 2017), implemented traditional Machine Learning, including Naive Bayes, Support Vector Machine (SVM), Logistic Regression, Gradient Boosting, and Linear Discriminant Analysis (LDA) to be compared with Deep Learning. It also implements some features, such as LIWC, SPLICE (Structured Programming for Linguistic Cue Extraction), SNA, and also resampling for data balancing. In this research, we focus on using five machine learning algorithms using TF-IDF, synonym replacement for data balancing to compare each technique and also implement stratified k-fold cross-validation instead of k-fold cross-validation.

Stratified k-fold cross-validation is similar to the k-fold cross-validation but different in separating data for each group. K-fold cross-validation randomly divides the data into groups which can lead to an unbalanced distribution of data by class. It is said that stratified k-fold cross-validation guarantees each class will be distributed equally on every fold (Widodo et. al, 2022). This signifies that the data will not be divided randomly, but rather fairly and will provide more stable outcomes for each fold.

The research mentioned previously (Tandera et. al, 2017) achieves Support Vector Machine and Logistic Regression for the traditional machine learning classification with their highest accuracy of 70.4% for Openness personality. There is also a possibility that the results are influenced by other factors, such as data imbalance, preprocessing, etc.

II. METHODS

This research is a literature review articles to answer the main problems. Methodology scheme is presented in Figure 3 below.

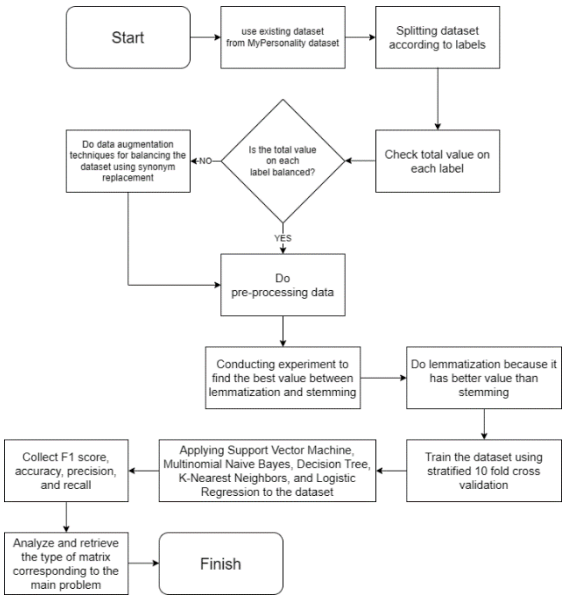

Figure 3. Proposed Methodology

As illustrated in figure 3, we begin by using the existing dataset from MyPersonality based on Facebook posts. This dataset will be split according to the labels from big five personality traits then check the total value on each label. The value of the dataset is ensured to be balanced to avoid data bias by calculating the value of positive and negative on each label. After implementing data augmentation techniques using synonym replacement for imbalance dataset, this data will be pre-processed to improve its quality before being utilized for model classification. After preprocessing was done, this data will be trained using stratified 10-fold cross-validation then applying five supervised machine learning as our model classification and collect F1 score, accuracy, precision, and recall as type of matrix then analyze and retrieve it to get the best value on this research.

2.1 Dataset

In this research, we use MyPersonality as our dataset that consists of 250 Facebook users, 9,918 records, Facebook posts in raw texts and network features like network size, betweenness centrality, brokerage, and transitivity but for this research, network features is removed because in this research only need take the text data. MyPersonality dataset is given a personality label based on the Big Personality Traits model. Every label in every raw text from the dataset has values of n and y, where n indicates a negative value and y indicates a positive value. Total value for each label is presented in Table I below.

Table I. Total value of MyPersonality dataset

	OPN	CON	EXT	AGR	NEU
n	2,547	5,361	5,707	4,649	6,200
y	7,370	4,556	4,210	5,268	3,717

2.2 Data Augmentation

Based on Table I, each label has an imbalance value so the dataset has to be balanced using data augmentation techniques, for example Synonym Replacement (Kobayashi, 2018) which is one of the numerous data augmentation techniques (Wei & Zou, 2019). Imbalanced data in the classification task, can lead to accuracy values and low recall rates from a small number of samples (Liu et. al, 2020).

The dataset has to be separated into a single label with positive and negative value then identify the label's highest value and determine the necessary amount of data. For example, the dataset with EXT label has the highest n value is 5,707 while the value of y is 4,210 so the positive value has to be increased to about 1,497 data by extracting and translating up to 1,497 synonyms. Data addition can be greater than the target and in this research, the value of y is greater than n or equals. The data augmentation will be performed in Python using the NLPAUG library and SynonymAug function.

After balancing the data with synonym replacement techniques, remove duplicate data to prevent data from overloading. Total value for each label with data augmentation techniques is presented in Table II.

Table II. Total value for each label with data augmentation techniques of MyPersonality dataset.

	OPN	CON	EXT	AGR	NEU
n	7,286	5,341	5,682	5,250	6,176
y	7,340	5,384	5,745	5,250	6,197

2.3 Pre-processing

The dataset needs to be processed for Machine Learning to classify it. The pre-processing step is needed so the data becomes clean and has accurate information. Pre-processing steps consist of removing hyperlinks, characters, numbers, whitespaces, emojis, lowering case, lemmatization, deleting duplicates, and removing stopwords. Pre-processing steps like removing hyperlinks, characters, text emoticon, numbers, and whitespaces were done using regular expression (regex) while lemmatization and stopwords were done using NLTK library. All data must be ensured in English and using lemmatization rather than stemming for this research because stemming has possibility of either over-stemming which is belonging to the same stem but has a very different meaning or under-stemming which is belonging to the same conceptual group are converted to two different stems or roots (Balakrishnan & Lloyd-yemoh, 2014). Pre-processing steps is presented in Figure 4 below.

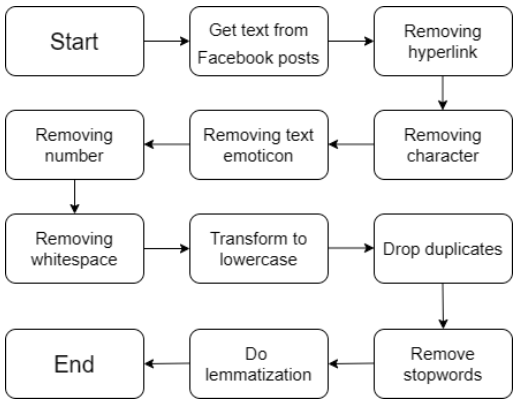

Figure 4. Pre-processing steps

2.4 Model Classification

In this step, five supervised machine learning were used such as Support Vector Machine (SVM), Multinomial Naive Bayes (MNB), Decision Tree (DT), K-Nearest Neighbor (KNN), and Logistic Regression (LR) for personality prediction. In this step, stratified k-fold cross-validation is implemented in Python using the sklearn library. The number of k is 10 and TF-IDF statistical measures were used to determine the importance of a word in the document (Suhartono et. al, 2016).

III. RESULTS AND DISCUSSION

A comparison between lemmatization and stemming was conducted and it was found that lemmatization is higher than stemming. The percentage is shown in table III and IV.

Table III. Percentage value F1 Score using lemmatization

Algorithm	Big Five Personality Traits				
	OPN (%)	CON (%)	EXT (%)	AGR (%)	NEU (%)
SVM	81.66	61.15	63.19	62.13	67.21
MNB	**82.31**	61.86	65.32	65.07	68.55
DT	74.96	53.01	56.46	54.66	59.00
KNN	70.16	49.02	46.42	47.48	40.76
LR	80.24	61.30	62.85	61.30	66.71

In this research lemmatization was implemented as a part of preprocessing step because the result of each classification model and label is higher than the others so it can be called better too. It can be seen from figure 3 with OPN label has four model classifications such as Support Vector Machine, Multinomial Naive Bayes, Decision Tree, and Logistic Regression which have a higher value in lemmatization while K-Nearest Neighbor has a low value. This scenario states that choosing lemmatization improves quality over stemming.

Table IV. Percentage value F1 Score using stemming

Algorithm	Big Five Personality Traits				
	OPN (%)	CON (%)	EXT (%)	AGR (%)	NEU (%)
SVM	80.90	61.17	62.33	61.53	66.30
MNB	**81.11**	61.63	64.78	64.70	67.54
DT	74.24	53.16	56.23	54.29	59.70
KNN	74.81	53.70	56.64	54.88	59.20
LR	79.75	61.12	62.18	60.75	66.48

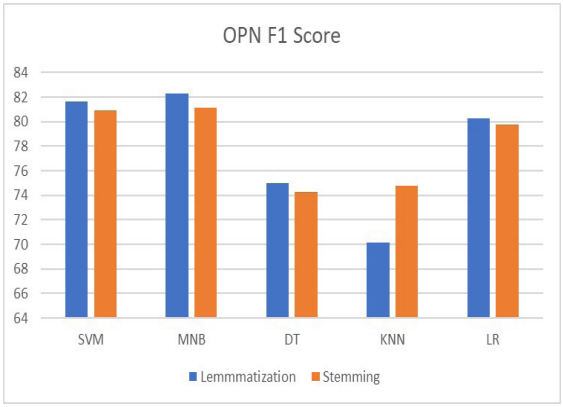

Figure 3. Performance comparison between lemmatization and stemming based on F1 Score with OPN label

F1 score, precision, accuracy, and recall are the popular types of matrices for evaluating the effectiveness of machine learning algorithms. The percentage will be shown in table V, VI, VII, and VIII.

Table V shows the percentage result of the F1 score where the highest percentage is 82.31% obtained from Openness (OPN) using Multinomial Naive Bayes (MNB) while the lowest percentage is 40.76% obtained from Neuroticism (NEU) using K-Nearest Neighbors (KNN). The highest average from all labels is 68.62% using Multinomial Naive Bayes (MNB) while the lowest average is 50.77% using K-Nearest Neighbors (KNN).

Table V. Percentage of F1 Score Machine Learning classification result by using myPersonality dataset

Algorithm	Big Five Personality Traits					Average (%)
	OPN (%)	CON (%)	EXT (%)	AGR (%)	NEU (%)	
SVM	81.66	61.15	63.19	62.13	67.21	67.07
MNB	**82.31**	61.86	65.32	65.07	68.55	**68.62**
DT	74.96	53.01	56.46	54.66	59.00	59.62
KNN	70.16	49.02	46.42	47.48	40.76	50.77
LR	80.24	61.30	62.85	61.30	66.71	66.48

Table VI shows the percentage result of accuracy where the highest percentage is 81.40% obtained from Openness (OPN) using Multinomial Naive Bayes (MNB) while the lowest percentage is 52.91% obtained from Agreeableness (AGR) using K-Nearest Neighbors (KNN). The highest average from all labels is 69.22% using Multinomial Naive Bayes (MNB) while the lowest average is 56.86 using K-Nearest Neighbors (KNN).

Table VI. Percentage Accuracy Machine Learning classification result by using myPersonality dataset

Algorithm	Big Five Personality Traits					Average (%)
	OPN (%)	CON (%)	EXT (%)	AGR (%)	NEU (%)	
SVM	79.92	62.00	64.49	61.31	69.15	67.37

MNB	**81.40**	63.11	65.28	65.07	71.22	**69.22**
DT	73.67	56.19	58.52	56.00	61.92	61.26
KNN	66.36	53.16	55.00	52.91	56.88	56.86
LR	78.25	61.49	63.80	60.87	67.41	66.36

Table VII shows the percentage result of precision where the highest percentage is 80.16% obtained from Openness (OPN) using Multinomial Naive Bayes (MNB) while the lowest percentage is 53.07% obtained from Conscientiousness (CON) using K-Nearest Neighbors (KNN). The highest average from all labels is 69.03% using Multinomial Naive Bayes (MNB) while the lowest average is 58.60% using K-Nearest Neighbors (KNN).

Table VII. Percentage Precision Machine Learning classification result by using myPersonality dataset

Algorithm	Big Five Personality Traits					Average (%)
	OPN (%)	CON (%)	EXT (%)	AGR (%)	NEU (%)	
SVM	76.65	62.26	65.62	61.15	71.14	67.36
MNB	**80.16**	63.73	65.28	61.05	74.93	**69.03**
DT	73.02	56.85	59.54	57.04	62.96	61.88
KNN	64.30	53.07	57.58	53.94	64.10	58.60
LR	74.96	61.33	64.55	60.98	67.71	65.90

Table VIII shows percentage results of recall where the highest percentage is 87.38% obtained from Openness (OPN) using Support Vector Machine (SVM) while the lowest percentage is 30.10% obtained from Neuroticism (NEU) using K-Nearest Neighbors (KNN). The highest average from all labels is 68.59% using Multinomial Naive Bayes (MNB) while the lowest average is 47% using K-Nearest Neighbors (KNN).

Table VIII. Percentage Recall Machine Learning classification result by using myPersonality dataset

Algorithm	Big Five Personality Traits					Average (%)
	OPN (%)	CON (%)	EXT (%)	AGR (%)	NEU (%)	
SVM	**87.38**	60.11	60.96	63.16	63.72	67.07
MNB	84.60	60.11	65.38	69.68	63.20	**68.59**
DT	77.05	49.74	53.70	52.51	55.56	57.71
KNN	77.45	45.80	39.15	42.51	30.10	47.00
LR	86.35	61.29	61.25	61.64	65.79	67.26

Based on the findings of data acquired from all four matrices, including their highest and lowest percentages, the accuracy matrix has the highest value of 69.22%, but the F1 score data will be used because false negatives and false positives are crucial in this research (Davis & Maiden, 2021). Although the accuracy value is the highest, there's a possibility of being exposed to the accuracy paradox which is not good for classification. For example, if the dataset is imbalanced, it is possible that the accuracy value remains high, but data bias will occur because it only considers one side.

IV. CONCLUSION

Based on the results of this research conducted with five supervised Machine Learning classification models to predict personality with big five personality traits from Facebook posts give the highest F1 Score using Multinomial Naive Bayes (MNB) algorithm. The implementation using TF-IDF, Lemmatization, synonym replacement, and using stratified k-fold cross-validation works well in this research. It can be said that the best classification model used for personality prediction according to this research is Multinomial Naive Bayes (MNB) with F1 Score as the benchmark.

REFERENCES

Aung, Z. M. M., & Myint, P. H. (2019, July). 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) (pp. 34-38). IEEE. doi: 10.1109/SNPD.2019.8935692.

Auxier, B., & Anderson, M. (2021). Social media use in 2021. Pew Research Center, 1, 1-4.

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., doi: https://doi.org/10.1016/j.paid.2017.12.018.

Balakrishnan, Vimala and Lloyd-Yemoh, Ethel (2014). Stemming and lemmatization: A comparison of retrieval performances. In: Proceedings of SCEI Seoul Conferences, 10-11 Apr 2014, Seoul, Korea.

Celli, F., & Lepri, B. (2018, December). Is big five better than MBTI? A personality computing challenge using Twitter data. In Proceedings of the Fifth Italian Conference on Computational Linguistics CLiC-it (Vol. 2018, pp. 93-98).

Christian, H., Agus, M. P., & Suhartono, D. (2016). Single document automatic text summarization using term frequency-inverse document frequency (TF-IDF). ComTech: Computer, Mathematics and Engineering Applications, 7(4), 285-294.

Christian, H., Suhartono, D., Chowanda, A., and Zamli, K., 2021. "Text based personality prediction from multiple social media data sources using pre-trained language model and model averaging," J. Big Data, vol. 8, May 2021, doi: 10.1186/s40537-021-00459-1.

Jeremy, N. H., Prasetyo, C., & Suhartono, D. (2019). Identifying personality traits for Indonesian user from twitter dataset. International Journal of Fuzzy Logic and Intelligent Systems, 19(4), 283-289. doi: 10.5391/IJFIS.2019.19.4.283.

K. Davis and R. Maiden, "The Importance of Understanding False Discoveries and the Accuracy Paradox When Evaluating Quantitative Studies," Stud. Soc. Sci. Res., vol. 2, no. 2, p. p1, 2021, doi: 10.22158/sssr.v2n2p1.

Kobayashi, S. (2018). Contextual augmentation: Data augmentation by words with paradigmatic relations. arXiv preprint arXiv:1805.06201. doi: 10.18653/v1/N18-2072.

Kunte, A. V., & Panicker, S. (2019, October). Analysis of machine learning algorithms for predicting personality: brief survey and experimentation. In 2019 global conference for advancement in technology (GCAT) (pp. 1-5). IEEE. doi: 10.1109/GCAT47503.2019.8978469.

Liu P., Wang X., Xiang C., and Meng W., 2020. "A Survey of Text Data Augmentation," in 2020 International Conference on Computer Communication and Network Security (CCNS), pp. 191–195. doi: 10.1109/CCNS50731.2020.00049.

Meta, 2022. "Meta Earnings Presentation Q3 2022". https://s21.q4cdn.com/399680738/files/doc_financials/2022/q3/Q3-2022_Earnings-Presentation.pdf

Ong, V., Rahmanto, A. D., Suhartono, D., Nugroho, A. E., Andangsari, E. W., & Suprayogi, M. N. (2017, September). Personality prediction based on Twitter information in Bahasa Indonesia. In 2017 federated conference on computer science and information systems (FedCSIS) (pp. 367-372). IEEE.

Ontoum, S., & Chan, J. H. (2022). Personality Type Based on Myers-Briggs Type Indicator with Text Posting Style by using Traditional and Deep Learning. arXiv preprint arXiv:2201.08717. [Online]. Available: http://arxiv.org/abs/2201.08717

Rosen, P. A., & Kluemper, D. H. (2008). The impact of the big five personality traits on the acceptance of social networking website. AMCIS 2008 proceedings, 274.

Statista, 2022. "Instagram users worldwide 2025 | Statista". https://www.statista.com/statistics/183585/instagram-number-of-global-users/

Tandera, T., Suhartono, D., Wongso, R., & Prasetio, Y. L. (2017). Personality prediction system from facebook users. Procedia computer science, 116, 604-611. doi: https://doi.org/10.1016/j.procs.2017.10.016.

Wei J. and Zou K., 2019. "EDA: Easy data augmentation techniques for boosting performance on text classification tasks," EMNLP-IJCNLP 2019 - 2019 Conf. Empir. Methods Nat. Lang. Process. 9th Int. Jt. Conf. Nat. Lang. Process. Proc. Conf., pp. 6382–6388. doi: 10.18653/v1/d19-1670.

Widodo, S., Brawijaya, H., & Samudi, S. (2022). Stratified K-fold cross validation optimization on machine learning for prediction. Sinkron: jurnal dan penelitian teknik informatika, 7(4), 2407-2414. doi: 10.33395/sinkron.v7i4.11792.