MBTI-BASED PERSONALITY PREDICTION FROM TEXT USING MACHINE LEARNING TECHNIQUES

**Punati. Venkata Jahnavi[1], Pulivarthi. Hima Sumana[2],
Shaik. Charishma Kousar[3], Pasupuleti. Himaja[4], Kondru. Jeevan Ratnakar[5]**

[1,2,3,4] *B.Tech, Students. Department of Information Technology,
Vasireddy Venkatadri Institute of Technology, Guntur*
[5] *Assistant Professor, Department of Information Technology,
Vasireddy Venkatadri Institute of Technology, Guntur*

Article DOI: https://doi.org/10.36713/epra14759
DOI No: 10.36713/epra14759

ABSTRACT

Personality prediction research seeks to define and comprehend the nuanced differences in human behavioural tendencies, thought patterns, and emotional expressions. Utilizing an array of methodologies, including psychological assessments, behavioural observations, and computational modelling, researchers aim to anticipate and clarify an individual's distinctive personality traits and characteristics. Natural Language Toolkit (NLTK) approaches are used to preprocess and translate text data into numerical features that can be predicted by machine learning models. The aim of this work is to predict the personality type of an individual linked to their posts and to explore the relevance of the test in the study and categorization of human behaviour using learning models. With the aid of a machine learning model and dataset, the main objective of this research is to determine a person's Myers-Briggs Type Indicator (MBTI) personality type based on their postings. This involves utilizing various methodologies, including psychological assessments and computational modelling, to analyse and classify the unique personality traits and characteristics associated with each MBTI type. The research aims to contribute valuable insights into understanding human behaviour and leveraging machine learning for predictive personality analysis.

KEYWORDS: *Myers-Briggs Type Indicator, Machine Learning Models, Natural Language Toolkit*

1. INTRODUCTION

In the world of psychology, the concept of personality is seen as an important but imprecisely established construct. As a result, psychologists would substantially benefit from developing more specific, objective assessments of current personality models. The majority of research on personality prediction has focused on the MBTI or Big Five personality models, which are the most often observed and experienced personality recognition models in the world. The Big Five personality model is made up of five distinct dimensions: (1) extraversion, (2) agreeableness, (3) conscientiousness, (4) neuroticism, and (5) openness [1]. The MBTI is an introspective self-report examination aiming to reveal certain psychological patterns regarding how people perceive and make decisions about their surroundings. This model detects 16 different types of personalities across four dimensions:

1.Introversion/Extraversion (how one obtains energy),
2.Sensing/Intuition (how one acquires information),
3.Thinking/Feeling (how decisions are made), and
4.Judging/Perceiving (how one portrays oneself to the outer world).

To produce a four-letter test case, a primary alphabet from each category or dimension will be considered, such as ISTP, ENTP, or INFJ. Figure 1 depicts the 16 distinct personality types that result from a combination of the four dimensions mentioned above. For example, ISTJ is the personality type produced by combining Introversion, Sensing, Thinking, and Judging.

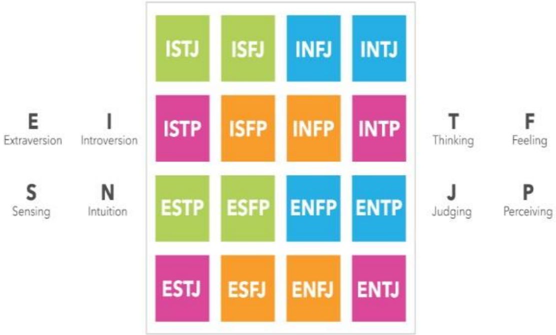


Fig 1. Types of Personality

Personality prediction is a discipline of psychology that uses patterns and emotions to discover and understand individual differences in human behaviour. The MBTI is a self-report introspective exam that identifies unique psychological patterns in how people perceive and make judgments about their environment. The goal of this work is to predict an individual's personality type based on their postings and to investigate the utility of the test in the study and categorization of human behaviour using Learning models. To provide a fuller picture of a person's personality, the MBTI can be used in conjunction with other personality models, such as the Big Five. It can reveal cognitive preferences that other models may not address clearly.

The MBTI was developed to explore how people interact with their surroundings, process information, and make decisions. Many people find the MBTI useful for both job and personal development. It can help people understand their communication patterns, work preferences, and learning approaches. Individuals find it more difficult to remember and relate to the Big Five's abstract labels (e.g., high in Neuroticism, low in Agreeableness) than the MBTI's four-letter type code, such as INTJ or ESFP. As a result, the MBTI may be easier to utilize for self-awareness and introspection.

a. Personality Types

Personality is derived from the Latin persona, which means defining a person's behaviours or characteristics [7]. According to [8], a person's particular attitude that distinguishes them from others indicates the essence of their personality. Personality is defined by Hall and Lindzey [9] as "the dynamic organization within the individual of those psychological systems that determine his characteristic behaviour and thought." This strategy determines how each individual uniquely adapts to their circumstances. A person's personality is defined by their sense of self, which influences their behaviour in a unique and dynamic way. This behaviour might alter as a result of experience, education, learning, etc. Setiadi's theory—which holds that personality is the dynamic organization of the system that specifically defines an individual's adaptation to the environment—is made clearer by this viewpoint [10].

The Myers-Briggs Type Indicator (MBTI) was used in this study to predict participants' personality types, providing useful insights for altering organizational culture and task distribution based on popular meta-programs and personality types. As previously stated, an individual's preferences are classified into four categories. Based on the Myers-Briggs Type Indicator®, these categories indicate 16 distinct personality types through various combinations of the personality type key. Figure 2 depicts the 16 personality types that result from the interactions of an individual's preferences. This tool facilitates more informed judgments in organizational contexts by providing a nuanced understanding of personality dynamics.

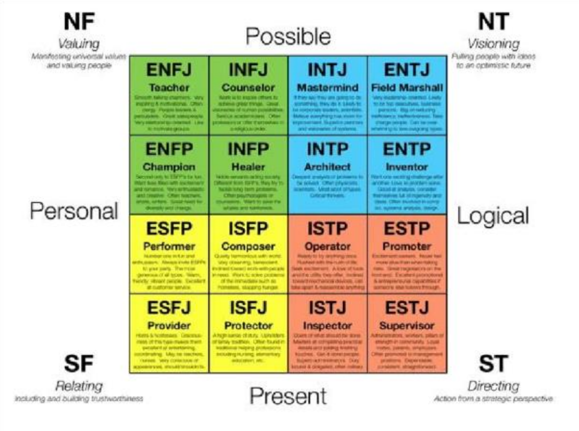


Fig 2. Personality types of MBTI [6]

In this study, the Myers-Briggs Type Indicator® (MBTI) was used to predict participants' personality types. Along with identifying the most common meta-programmes and personality types, this data will enable for changes to the current organizational culture and task distribution. The keywords in Figure 2 correlate to a certain MBTI personality type, and Figure 3 depicts the cognitive processes of each type. The background colour of each kind indicates its major function, while the text colour indicates its supporting duty.

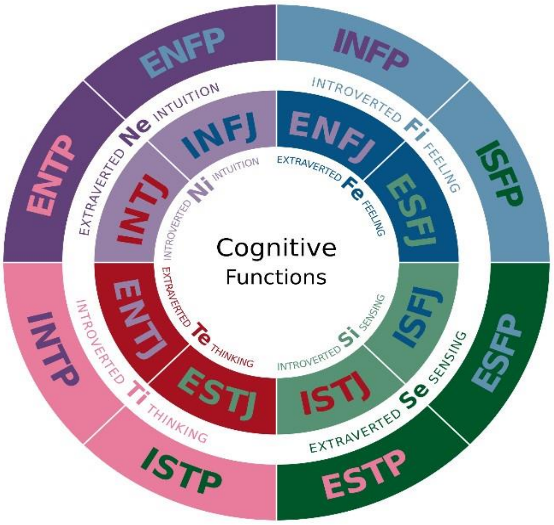


Fig 3. Cognitive Functions of each Personality Type [11]

2. LITERATURE REVIEW

The literature review underscores the increasing significance of social media as a data source for research, stressing the pervasive usage of Twitter and other platforms, especially in Indonesia. It highlights the potential of data mining and related research in fields like named entity recognition, automatic summarization, sentiment analysis, and user profile clustering. This study, however, focuses on personality prediction using social media data—a field that has received less attention in the context of Indonesian language learning. The development of a new Indonesian Twitter dataset with 250 users who have had their personality traits from the Five Factor Model annotated by psychologists is one of the study's contributions. It also introduces the use of XGBoost models for personality prediction, which cover neuroticism, conscientiousness, extraversion, and agreeableness.[1]

The existing research on the relationship between personality recognition and mental health disorders, with a particular focus on borderline personality disorder (BPD), can be examined in the literature review conducted for this study. It ought to look into earlier studies on personality identification from social media data and how it's used to evaluate mental health. The review should also examine the techniques and machine learning algorithms used in personality classification, as well as how well they identify personality traits. It should also look at research on personality assessments' role in the early identification of mental health conditions and investigate the advantages of associating personality traits with conditions like borderline personality disorder (BPD). Furthermore, it is critical to take into account any gaps or restrictions in the previous research and how the current study attempts to address and contribute to these areas.[2]

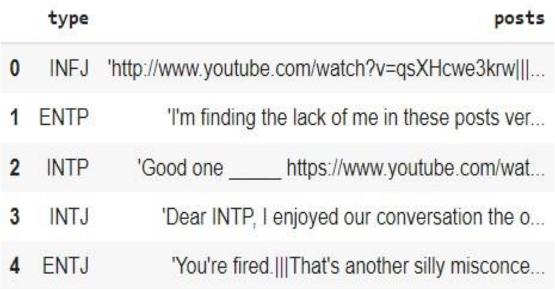
In particular, the research on automated personality prediction in the context of social media data analysis can be explored in the literature review for this study. It should highlight the methods and machine learning algorithms used in personality prediction models and go over earlier research that looked at the relationship between language use on social media and personality traits. In addition, given the high level of social media usage in places like Indonesia, the review should look at the geographical and linguistic diversity of these studies, paying special attention to the dearth of research in languages like Bahasa Indonesia. It's critical to take into account the Five Factor Model, its applicability in personality prediction, and the possibility of cross-cultural differences in these relationships. The importance of the study's contributions such as the creation of a personality prediction system for Bahasa Indonesia, the assessment of the system's precision, and the comparison of machine learning algorithms in this particular linguistic and cultural setting should also be covered in the literature review.[3]

The development of personality prediction techniques in the context of social media and online behaviour analysis could be the main topic of this study's literature review. The article ought to delve into the diverse machine learning algorithms and personality models employed in prior studies, with a particular focus on the potential of Artificial Neural Networks as a method for personality prediction. It should also emphasize how important social media sites like Facebook are as a rich data source for comprehending online behaviours and personality traits because of their enormous user base and abundance of social interactions. The review may also take into account the relevance of various personality models in this predictive context, including the MBTI, DISC, and Big Five. Contributions of the study include examining how user behaviour and personality relate to each other on social networks. The utilization of KNN, Artificial Neural Networks, and Logistic Regression should be portrayed as novel approaches to improving personality prediction using machine learning methods.[4]

3. METHODOLOGY

a. Data Collection

In order to predict the personality of the data in the dataset, the data must be cleaned and removed of various punctuation marks and emotions. A huge number of emoticons were cleaned by rounding off various redundant terms in the dataset, and multiple words were cleaned using the processing library's regular expression. A variety of punctuations are also included in the collection. The dataset provided has been cleansed with appropriate data for prediction and, more importantly, implementation for personality type analysis. This study takes use of Kaggle's publicly available Myers-Briggs personality type dataset. The dataset contains 8675 observations, each of which displays the author's four-letter Myers-Briggs personality type as well as the raw text of their previous 50 posts. Figure 4 displays two columns with the name's kind and posts. The 16 personality types are listed in the type column, while raw text is listed in the posts column.



	type	posts
0	INFJ	'http://www.youtube.com/watch?v=qsXHcwe3krw|||...
1	ENTP	'I'm finding the lack of me in these posts ver...
2	INTP	'Good one _____ https://www.youtube.com/wat...
3	INTJ	'Dear INTP, I enjoyed our conversation the o...
4	ENTJ	'You're fired.|||That's another silly misconce...

Fig 4. Data in Dataset

b. Data Visualization

Data visualization can aid in the identification of patterns, trends, and outliers in data. Matplotlib, a Python plotting library, and Seaborn, a Python data visualization library, were used to preview the data. To determine the frequency of each personality type, the value counts () technique is utilized. Figure 5 depicts the frequency of each MBTI personality type. The dataset has a clear imbalance between the various classifications. We notice that some personality types have substantially more data than others.

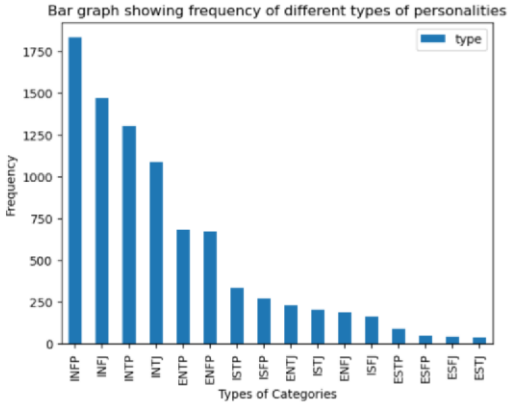


Fig 5. Number of Posts for each Personality Type

C. Preprocessing

Some word reduction was necessary because the data came from an online forum where the writers were only allowed to express themselves via text. The primary reason for this was that the sample's non-uniform MBTI type distribution did not match the actual MBTI type proportions. Ultimately, it has been determined that this is because of the information gleaned from an online discussion forum created specifically for personality type debate, where posts frequently repeated the MBTI traits. This may potentially affect the accuracy of the model. Consequently, all of the characters were maintained in lowercase. The subsequent information was deleted:

1. URLs
2. Non-English letter phrases (such as +, -, etc.),
3. Use the NLTK library to stop words (such as commonly used words like "a," "an," "the," etc.),
4. MBTI profile strings (such INFP, ESTJ, etc.) to prevent model manipulation when trying to pinpoint specific MBTI mentions.

In the end, the NLTK package was used to lemmatize the text in order to increase the dataset's significance and relevance. Data must first be cleaned, transformed, and prepared in order to be used for analysis or model creation.

d. Feature Extraction

The process of predicting a person's personality from text involves examining their written or spoken communication to glean important details about their disposition, conduct, and character. Preprocessing and feature extraction techniques like text tokenization, stop-word removal, and the application of complex vectorization algorithms like TF-IDF or word embeddings are usually the first steps in this process. Following their transformation, these text features are fed into machine learning models, which are trained to identify particular personality traits based on language patterns and writing style. This methodology offers a multitude of pragmatic implementations, ranging from customizing user experiences and content recommendations to supporting mental health practitioners in comprehending their patients more thoroughly.

e. Classification Models

In this stage, classification techniques including Logistic Regression (LR), Support Vector Machine (SVM), Gaussian NB, Random Forest, and XG-Boost are utilized. The category in which it falls is Supervised Learning. In addition to being split up into four binary classification jobs, the classification task is separated into sixteen classes. The MBTI type is made up of four binary groupings, which explains this. So, four well-known binary classifiers that each focus on a different personality trait have been trained.

First, the posts are processed and sanitized. Next, two crucial transformations are used to produce the target labels (Y) and input features (X). Applying Count Vectorizer and TF-IDF Transformer results in the creation of X, which stands for the cleaned posts. By encoding them as binary values (1 for presence, 0 for absence), Count Vectorizer essentially builds a matrix that counts the number of unique words that appear on each page. The Term-Frequency-Inverse Document Frequency (TF-IDF) Vectorizer, on the other hand, uses the Count Vectorizer output to create a matrix. Y stands for the binarized personality type markers. MBTI dichotomies (E/I, S/N, T/F, and J/P) are the four columns that correlate to the different personality trait categories. The classification algorithm is then given the X and Y matrices.

The X and Y matrices are then passed to the classification algorithm. These algorithms are supervised learning techniques that forecast personality traits based on features derived from text input. The model can predict each personality attribute separately. The dataset is divided between training and testing sets using the sklearn library's train_test_split() function. The training set receives around 80% of the data, with the remaining 20% going to the test set. On the basis of the training data, distinct classification models for each type of personality trait are constructed. Once trained, these models are tested against test data to see how well they perform.

4.SYSTEM ARCHITECTURE

We used an organized approach in our MBTI dataset personality prediction process. To assure data quality, pretreatment and cleaning were done first. Tokenization was then applied to divide the text into meaningful chunks. Lemmatization was used to guarantee textual uniformity while feature extraction techniques were used to obtain pertinent information. We then used classification models to predict the four MBTI dichotomies (E/I, S/N, T/F, and J/P), such as XG Boost, SVM, and Logistic Regression. TF and Random Forest were used as feature selection techniques to improve accuracy. These models then enabled insights from the text data and aligned with the MBTI framework by using the specified attributes to predict MBTI personality types from specific text inputs.

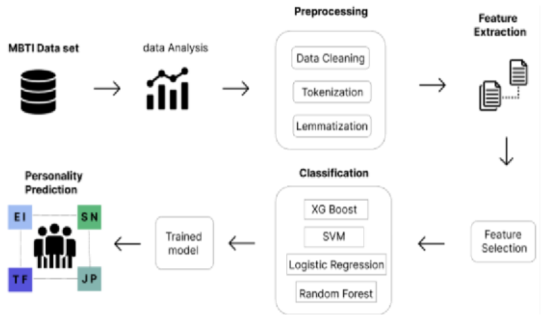


Fig 6. System Architecture of MBTI-BASED PERSONALITY PREDICTION

5.RESULTS

After conducting a comprehensive comparison of different classification algorithms, which included Gaussian Naive Bayes, Random Forest, Support Vector Machine (SVM), Logistic Regression, and XGBoost, we visualized the results in Figure 7, a bar plot. Notably, XGBoost emerged as the topper forming algorithm in terms of accuracy among the options considered.

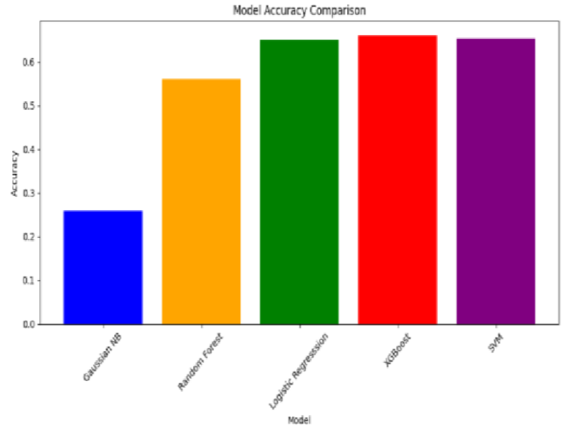


Fig 7. Comparison of Algorithms

In the first category of Introversion (I)/Extroversion (E), Extroversion (E) has a much broader distribution than Introversion (I). In the second category, Intuition (N)/Sensing (S), Sensing (S) has a significantly higher distribution than Intuition (N). Furthermore, Figure 8 shows that for the third category, Thinking (T)/Feeling (F), the distribution of Thinking (T) is slightly higher than the distribution of Feeling (F). In conclusion, there is a greater dispersion of Judging (J) than Perceiving (P).

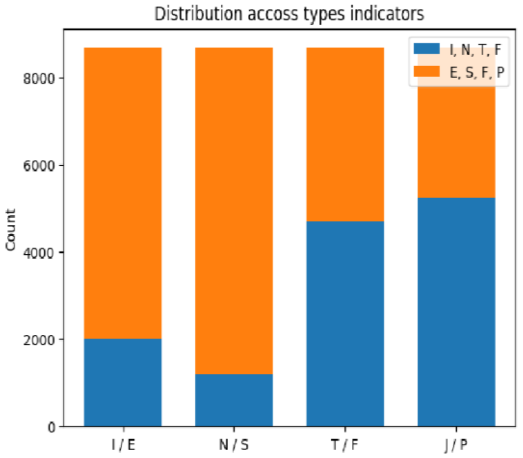


Fig 8. Distribution across MBTI-type Indicators

The strength of the links and variables can be gauged using the Pearson correlation coefficient. The correlation between every random variable (Xi) and every other value in the table (Xj) in a correlation matrix can be used to determine which pairs have the highest correlation. To determine the significance of the link between two variables, one must find the coefficient value, which might vary from 1.00 to 1.00. The correlation coefficient between personality type identifiers is displayed in Figure 9.

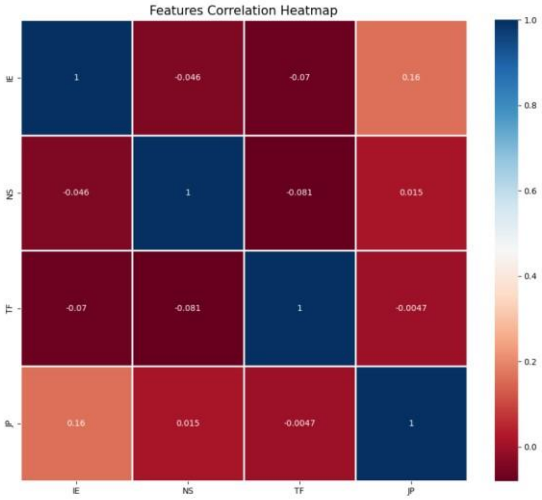


Fig 9. Feature Correlation Heatmap Between Personality Type Indicators

In order to ensure the accuracy of the input data, we include a vital validation stage in our work. The provided input text is validated by this validation process. To maintain the integrity of our study, the system asks the user to supply valid text when it detects an erroneous input. The algorithm predicts one of the 16 different personality types as the ultimate result when it receives valid data. Our research's personality type predictions are accurate and relevant because of this stringent validation process, which acts as a cornerstone.

6. CONCLUSION

We built a machine learning classifier to predict MBTI personality types using Python modules such as Pandas, NumPy, NLTK, Seaborn, Matplotlib, and Sklearn. The efficiency of three classification algorithms was investigated: logistic regression, SVM, and XGBoost. Notably, the XGBoost model attained a multi-class accuracy of approximately 66.05%. To increase accuracy even further, we investigated the use of binary classifications independently for each of the four personality dichotomies. This strategy was helpful in refining the prediction process. Our findings emphasize the need of acquiring precise data, particularly across diverse networks, and employing advanced data processing tools to forecast personality types. The proposed method, which leverages the Myers-Briggs Type Indicator, provides a robust framework for categorizing personality types into sixteen distinct patterns covering four major dichotomies: The four personality traits are (1) introversion - extroversion, (2) sensing - intuition, (3) thinking - feeling, and (4) judging - perceiving. This work adds to the growing body of knowledge in the field of personality prediction by emphasizing the need of precise data collection and the potential for more accurate personality categorization.

7. FUTURE ENHANCEMENT

Consider incorporating additional data modalities, such as audio, visuals, or user behaviour, to develop a multi-modal approach to personality prediction. Combining textual and other sorts of data may provide a more comprehensive knowledge of personality. Investigate the possibilities of predicting personality types in different languages. Extend the model to accommodate multilingual datasets and study how linguistic and cultural variables affect personality prediction across languages.

8. REFERENCES

1. *Ong, V., Rahmanto, A. D., Williem, W., Jeremy, N. H., Suhartono, D., & Andangsari, E. W. (2021). Personality Modelling of Indonesian Twitter Users with XGBoost Based on the Five Factor Model. International Journal of Intelligent Engineering & Systems, 14(2).*
2. *Siva Shanmugam, G., Choudhary, D., Anand, H., & Xavier, C. Application of machine learning using MBTI classification to understand the borderline disorder.*
3. *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.*
4. *Karnakar, M., Rahman, H. U., Santhosh, A. J., & Sirisala, N. (2021, October). Applicant personality prediction system using machine learning. In 2021 2nd global conference for advancement in technology (GCAT) (pp. 1-4). IEEE.*
5. *Amirhosseini, M. H., & Kazemian, H. (2020). Machine learning approach to personality type prediction based on the myers–briggs type indicator®. Multimodal Technologies and Interaction, 4(1), 9.*
6. *Tieger, P.D.; Barron-Tieger, B. Do What You Are: Discover the Perfect Career for You through the Secrets of Personality Type, 4th ed.; Sphere: London, UK, 2007.*
7. *Darsana, M. The influence of personality and organisational culture on employee performance through organisational citizenship behaviour. Int. J. Manag. 2013, 2, 35–42.*
8. *Alwi, H.; Sugono, D.; Adiwirmata, S. Kamus Besar Bahasa Indonesia; Balai Pustaka: Jakarta, Indonesia, 2003.*
9. *Hall, C.; Lindzey, G. Theories of Personality, 2nd ed.;Wiley: New York, NY, USA, 1970.*
10. *Setiadi, N.J. Perilaku Konsumen Konsep Dan Implikasi Untuk Strategi Dan Penelitian Pemasaran; Prenada Media:Jakarta, Indonesia, 2003.*
11. *Beech, J. The Cognitive Functions of each Personality Type. Available online: https://siteassets.pagecloud.com/greeleymosaic/downloads/Myers-Briggs-ID-7fbebb3b-f94d-468ace4f-7c488703c102.pdf (accessed on 19 September 2018).*