MBTI Personality Prediction Using Only What People Post in Online Forums

Zipeng Guo

Virginia Polytechnic Institute and State University, Blacksburg VA 24060, USA
zipengg@vt.edu

Abstract. The Myers-Briggs Type Indicator (MBTI) is a well-known personality assessment designed to classify individuals into different personality types. It is a widely used personality test that helps people understand others' cognitive, decision-making, and behavioral preferences. The assessment attempts to gain insight into how people think, behave, communicate, and interact with the world around them and with the world around them. The MBTI focuses on four dimensions: thinking orientation, the way people get information, the way people make decisions, and the way people live their lives. The results of this study can help people better understand their tendencies to think and behave in a better way, better develop themselves, understand different people's preferences in choosing and adapting to careers, help us plan our careers, find careers that make use of our talents, and identify whether the chosen career is right for us. In this study, a new method is created to infer an individual's MBTI label based on people's online postings. The performance of the method proposed in this study was compared with other existing methods, and the results showed better accuracy and reliability.

Keywords: Myers-Briggs Type Indicator; Personality; Prediction; Logistic Regression; extreme Gradient Boosting.

1 Introduction

MBTI, often The Myers-Briggs Type Indicator, is the most often popular personality test designed to classify individuals into different personality types. It is a widely used personality test that helps people understand their and others' preferences in cognition, decision-making, and behavior. It was created during World War II by Katherine C. Briggs and her daughter, Isabel Briggs-Myers. They intended to develop a useful tool that people could use to better understand themselves and other people, drawing on Carl Jung's theories as inspiration. The aim is to reveal how people view the world and process information. The assessment seeks insight into how people think, behave, communicate, and interact with the world around them, and interaction with the world around them. based on the ideas of Carl Jung, the MBTI divides personalities into 16 types. It focuses on four dimensions: thinking orientation, the way people get information, the way people make decisions, and the way people live [1]. By understanding these dimensions, people can learn about personality traits, such as

P. Kar et al. (eds.), *Proceedings of the 2023 International Conference on Image, Algorithms and Artificial Intelligence (ICIAAI 2023)*, Advances in Computer Science Research 108,
https://doi.org/10.2991/978-94-6463-300-9_60

whether people tend to interact with others, pay attention to practical details, look at the big picture, and much more [2].

MBTI not only helps us to understand ourselves better but also improves the understanding of how others behave and communicate. The MBTI is often utilized in various contexts, such as marital counseling, group facilitation, career guidance, and personal growth. It can help individuals gain self-awareness, understand their communication and decision-making styles, and appreciate the diversity of others [3]. However, it is essential to note that people must recognize that the MBTI has its limitations. It should be viewed as a tool for self-reflection, for informational purposes only, rather than a definitive measure of personality. The attempt to build a model to predict a person's MBTI personality type is modeled in Python through machine learning techniques. The four major personality types are shown in Figure 1.

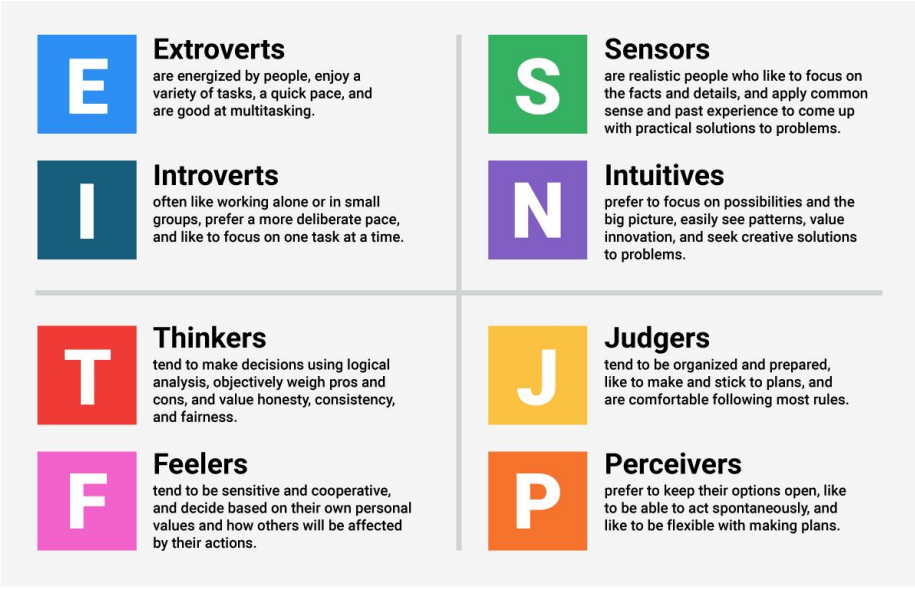

Fig. 1. MBTI Type[4]

The history of the development of the MBTI test is relatively simple, divided into roughly four phases: First, the initial stage (1943-1956): Myers and Briggs first released the MBTI test in 1943, but the test had a more straightforward format, containing only 93 questions and no standardized scoring system. Second one is increased assessment and standardization (1957-1980): In 1957, Myers and Briggs increased the number of test questions to 126 and established a standardized scoring system to ensure the accuracy and reliability of the test. At this time, the MBTI test gradually became widely used and was considered a reliable personality assessment tool[5]. The third one is Clinical application (1981-1990): In 1981, the MBTI test began to be widely used in clinical psychology to help doctors and therapists better

understand the personality types of their patients and provide more accurate treatment plans. The last one is Development and innovation (1991-present): Nowadays, the MBTI test has been continuously developed and innovated. Today, the MBTI test has become a widely used personality assessment tool in psychology and recruitment, training, leadership development, and other fields.

The Psychological Types written by Jung inspired Katherine C. Briggs and her daughter, Isabel Briggs-Myers together divided the MBTI into four dimensions[6]. The four dimensions correspond to four criteria, and each person's personality falls on one of them. Wherever this criterion is close, it means that the person has a preference for it. They are the source of drive: introversion (I) - extroversion (E), the Intuition (N) and feeling (S) are ways of getting information, thinking (T) and feeling (F) are ways of making choices, and perception (P) and judgment (J) are ways of approaching ambiguity. The source dimension of the drive is the one that relatively distinguishes individuals the most. Introversion (I) biased individuals will focus more on the inner world; they are immersed in their world alone, cannot help themselves, and are relatively calm. Extroverted (E) people are more focused on the external world; they like to pour a lot of energy into getting along with others and are relatively impulsive. The way of receiving information is divided into intuition and feeling. The former focuses on abstract concepts and ignores the immediate material objects. They rely on instinct rather than repeated thought. The latter is the opposite, with physical objects being the things that get much attention. These people spend more time on the details and always keep their sanity and love for things. The way decisions are made is divided into thinking and emotions. This dimension indicates the individual's approach to making decisions and drawing conclusions, whether objective, logical reasoning or subjective emotions and values. Emotional individuals expect their emotions to be consistent with those of others, and they make decisions based on what is important to them and others; their rational judgments are based on their values. Individuals make decisions through objective, impersonal logical analysis of situations, focusing on cause and effect relationships and seeking objective scales of facts, and are therefore less influenced by personal feelings. The last dimension is perceptual and judgmental. This dimension is mainly reflected in the individual's lifestyle. Judgmental individuals tend to live openly and naturally, that is, in an orderly and planned way, and they prefer to solve problems independently. Perceptual individuals, on the other hand, prefer to live in a fixed way, i.e., perceptual experience, and they are constantly gathering information to keep their lives flexible and natural. They strive to keep events open and enjoy the feeling of going with the flow. As shown in Figure 2, the respective correspondence is shown.

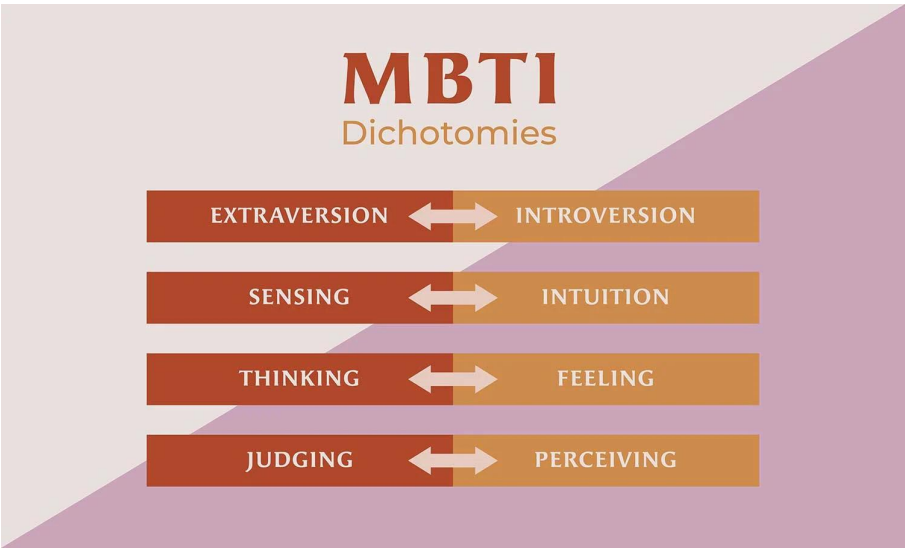

Fig. 2. Correspondence[7]

The main tasks required are Data cleaning, Data processing, and Data converting. Systems engineering and autonomous control both depend on data processing. All facets of social production and social life involve data processing. The expansion and depth of data processing technology's applications, along with their technological advancement, have had a significant impact on how society has developed over time. The next step is data cleaning and processing. The data needed for the research are the remarks people post on the network, so it is necessary to remove the other parts to improve data quality, ensure data integrity and improve the efficiency of data analysis. Data cleaning is also one of the essential steps in data analysis and one of the necessary steps to ensure accurate and reliable analysis results. After this step is finished, the data can be processable.

2 Methods

2.1 Logistic Regression

It is a quantitative framework that predicts binary or categorical outcomes. It is a type of regression analysis that is commonly used for classification problems, where the dependent variable (the outcome) is categorical and can take only two values, such as "yes" or "no." The likelihood that a given observation belongs to a certain category or class is estimated via logistic regression. Using the logistic function in regression, commonly referred to as the sigmoid function, to predict the connection between independent variables and the likelihood that an event will occur is known as "logistic" regression. In the logistic regression model, it is presumable that the log-odds (logarithm of the odds) of the dependent variable have a linear relationship with

the independent variables. The logistic function is used to translate the log odds, mapping them to a range between 0 and 1, representing the likelihood that the incident will happen. Maximum likelihood estimation, which looks for parameter optimum settings to increase the likelihood of witnessing the provided value, is used to estimate the model parameters in logistic regression.

Both categorical and continuous independent variables are supported. It is widely used in various fields, including medicine, social sciences, marketing, finance, and machine learning, for tasks such as predicting the likelihood of disease occurrence, customer churn, credit default, sentiment analysis, and more. Logistic regression, it should be noted, makes a linear assumption about how the independent variables and log odds are related. If the connection is non-linear, more complex models like polynomial logistic regression or other non-linear classification algorithms may be more appropriate.

2.2 Linear Support Vector Classification

For binary classification tasks, linear support vector classification is a popular supervised learning approach, often known as linear SVM (Support Vector Machine).[8]. It belongs to the family of maximum margin classifiers, which means it seeks to discover the optimal decision boundary that maximally divides information into distinct groups. Finding a hyperplane in the feature space that effectively divides the two classes is the main goal of Linear SVM. The term "linear" indicates that this algorithm assumes a linear decision boundary, which is a straight line in two dimensions or a hyperplane in higher dimensions.

The key notion of support vector machines (SVM) is finding the hyperplane that optimizes the margin, where the margin is measured as the distance between the hyperplane and the data points that are closest to each class. Setting the decision boundary relies heavily on support vectors are the data points that are located in the closest proximity to the hyperplane.

In Linear SVM, the algorithm seeks to locate the hyperplane that is most suitable via the process of optimizing a problem. The goal is to make the classification error as little as possible while increasing the margin as much as possible. By maximizing the margin, Linear SVM can achieve better generalization and improved performance on unseen data. The capacity of Linear SVM to effectively handle high-dimensional data is one of its benefits. It is ideal for a variety of applications since it performs especially well when there are more characteristics than samples. Additionally, Linear SVM is less prone to overfitting compared to some other complex models.

To make predictions, in Linear SVM, the new data points are categorized into classes according to the side of the decision boundary they fall on. The algorithm uses the sign of the discriminant function to classify the points into one of the two classes.

2.3 Stochastic Gradient Descent

It is an optimization algorithm commonly used in deep learning for training models[9]. It is particularly suited for large datasets because it performs updates on individual training examples or small subsets of data, making it computationally efficient.

The key idea behind SGD is to iteratively update the model's parameters in the direction that minimizes a specified loss function. Instead of considering the entire training set in each iteration, SGD randomly selects a single training example or a mini-batch of illustrations of how to calculate the gradient and update the parameters. The process of updating the parameters involves taking small steps in the direction with the steepest grade descent pertaining to the loss function with regard to the now active parameter values. This helps the algorithm converge toward an optimal set of parameters that minimize the loss.

SGD is widely used in training various machine learning models, such as linear models and support vector machines, are widely utilized in various domains. It is typically used in conjunction with other techniques, such as learning rate schedules and momentum, to improve convergence and achieve better performance.

2.4 Random Forest

The aforementioned technique is a versatile ensemble learning method commonly employed in the field of machine learning, serving purposes in both jobs including categorization and regression analysis. It is composed of a large number of decision trees, each of which is trained each time using a fresh random subset of the data and characteristics. After then, the forecasts obtained from each of the different trees are integrated to provide the ultimate forecast.

A stronger and more reliable model is created by combining the predictions of numerous weak learners (decision trees) in Random Forest. In Random Forest, the word "forest" refers to the group of decision trees that collaborate to provide forecasts.

2.5 Extreme Gradient Boosting

The advanced machine learning method known as it is a member of the family of gradient boosting algorithms. It is intended to enhance the performance of traditional gradient-boosting algorithms by addressing their limitations and incorporating additional features[10].

XGBoost is particularly known for its exceptional predictive power and efficiency. It is applied in various machine learning competitions and is considered a go-to algorithm for many data scientists and practitioners. An indicator of the relative weighting of each feature in the model's predictions is provided by XGBoost as a measure of feature relevance. This knowledge may be used to choose features, understand the relationships between features, and identify the most influential factors.

Overall, XGBoost is a cutting-edge algorithm that pushes the boundaries of gradient-boosting methods. Its innovative features, efficient implementation, and exceptional predictive capabilities have made it a valuable tool in the machine learning toolkit.

3 Results

By dividing the dataset into a test set and a training set of 70% and a test set of 20%, and 10% as the validation set. Accuracy rate and F1 score evaluation are used to

evaluate model performance. Table 1 is a comparison of model results. The random forest has an accuracy rate is 86.85, and the F1 score reaches 83.97. Logistic Regression has a percentage of the correctness of 82.33 and an F1 rating of 77.20. Linear Support Vector has a percentage of correctness of 83.72 and an F1 rating of 84.25. Extreme Gradient Boosting has a percentage of correctness of 84.77% and an F1 rating of 81.23%. Table 1 shows the performance of these four types.

Random Forest got the best performance from Table 1. below. Logistic regression maps the input characteristics to the probability space through linear regression and logical function, but its disadvantage is that it has a limited effect on complex nonlinear problems. SVM separates samples by searching for the maximum interval Hyperplane. The disadvantage is that it is slow to process large-scale data sets and high-dimensional features. Therefore, these two algorithms perform poorly in handling the complex problems presented in this article.

Table 1. Comparison of model results

Model	F1	Accuracy
Random Forest	83.97%	86.85%
logistic Regression	83.97%	86.85%
Linear Support Vector	84.25%	83.72%
Extrem Gradient Boosting	84.77%	81.23%

4 Conclusion

This paper uses four different models for training and testing, including Logistic Regression, Linear Support Vector, Random Forest, and Extreme Gradient Boosting. Random Forest performs best. Due to many features in the dataset, A scale for evaluating features relevance is what Random Forest produces, indicating the proportion weights each feature contributes to the model's predictions. This knowledge aids in the choice of features, the recognition of the most important variables, and the development of an understanding of the underlying connections in

the data. By considering feature importance, Random Forest can focus on the most relevant features, leading to improved performance.

References

1. An introduction to MBTI, https://contentoo.com/blog/an-introduction-to-mbti/, last accessed 2023/7/2.
2. MBTI® basics, https://www.myersbriggs.org/my-mbti-personality-type/mbti-basics/home.htm, last accessed 2023/7/2.
3. Amirhosseini, M.H., Kazemian, H.: Machine learning approach to personality type prediction based on the myers–briggs type indicator®. Multimodal Technol. Interact. 4(1), e9 (2020).
4. Darsana, M.: The influence of personality and organisational culture on employee performance through organisational citizenship behaviour. Int. J. Manag. 2, 35–42 (2013).
5. The history of the MBTI® assessment, https://eu.themyersbriggs.com/en/tools/MBTI/Myers-Briggs-history, last accessed 2023/7/2.
6. Myers-briggs type indicator (MBTI) personality, https://drjosephhammer.com/resources/systematic-career-exploration-approach-scea/step-3-narrow-your-occupations-roster/myers-briggs-type-indicator-mbti-personality/, last accessed 2023/7/2.
7. How the Myers-Briggs type indicator works: 16 personality types, https://www.simplypsychology.org/the-myers-briggs-type-indicator.html, last accessed 2023/7/2.
8. Support vector machine — Introduction to machine learning algorithms, https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47, last accessed 2023/7/2.
9. Paireder, T., Motz, C., Huemer, M.: Normalized stochastic gradient descent learning of general complex-valued models. Electron. Lett. 57(12), 493–495 (2021).
10. Raj, S.N.M., Joshua, E.S.N., Swathi, K., Neeraja, S., Bhattacharyya, D.: Analyzing comments on social media with XG boost mechanism. In: Machine Intelligence and Soft Computing, pp. 33–37. Springer Nature Singapore, Singapore (2022).