Identifying 16 Personalities Based on Daily Writing by Random Forest and Logistic Regression

Bowen Xiao*

Shanghai Pinghe School, Shanghai, China

*Corresponding author: xiaobowen@shphschool.com

Abstract. The Myers-Briggs Type Indicator (MBTI) is a self-reported personality test based on psychological theory proposed by Carl Gustav Jung, elaborated and developed by Katharine Cook Briggs and Isabel Briggs Myers. It separates human's personality into 4 dimensions: mind, energy, nature and tactics, 2 directions are provided in each dimension, and it results 16 distinct personalities. Through the exquisite figures design of each personality and extensive advertisement, the test has immediately attracted a huge amount of participants among the youth. Finally, the MBTI test gradually creates a new trend of labeling and judging people according to their personality. Current studies have predicted the personality through linear SVM, Neural network, and K-means. This study devotes to utilize Logistic Regression model and Random forest model to predict based on the daily conversation. As a result, the accuracy that the author gets is relatively high comparing with similar works, and the best accuracy of 4 dimensions (Mind, Energy, Nature, Tactics) in author's model reaches 78.67%, 85.82%, 58.73% and 63.0%.. The purpose of this study is examining the practicability of implementing machine learning in the psychological field, meanwhile, extends the methods of predict one's personality. Studies in this field can be further used to help analyze a people, and assist the doctors to treat psychological problems.

Keywords: Machine Learning, MBTI, Personality Test, Logistic Regression, Random Forest.

1. Introduction

The Myers-Briggs Type Indicator (MBTI) is a self-reported personality test based on the psychological theory proposed by Carl Gustav Jung, elaborated and developed by Katharine Cook Briggs and Isabel Briggs Myers. The essence of C.G.Jung's theory is the apparently random variation in human behavior is orderly and consistent, being due to basic differences in the ways individuals prefer to use their perception and judgment.[1] Different with the Big Five personality test[2] that received more recognition in the psychological field, C.G.Jung proposed the four most important functions of personality: mind, emotion, intuition, and sense. These four traits developed into the 4 dimensions of today's MBTI test: mind, energy, nature, and tactics. The variance of behavior in 4 dimensions is concluded into 2 directions representing divergent preferences of individuals, and with the combination of 4 dimensions, 16 distinct personalities emerged.[3] 1. Extraverted (E) or Introverted (I) 2. Sensing (S) or Intuition (N) 3. Thinking (T) or Feeling (F) 4. Judging (J) or Perceiving (P)

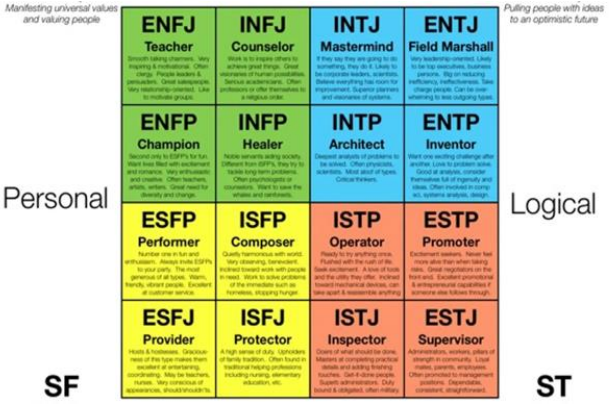

Figure1. Illustration of 16 Personalities types

From the author's research, 4 articles are found that have a related topic. On April the ninth 2020, an article(MBTI) named "Machine Learning Approach to Personality Type Prediction Based on the Myers–Briggs Type Indicator"[4] in the magazine called MDPI used the extreme gradient boosting and recurrent neuron network models, meanwhile, they concluded from the earlier research that methods like SVM, K-means and Naive Bayes are used before. In an article (BERT)[5] published on May the seventeenth 2021, Hans Christian and his partners utilized BERT, RoBERTa, and XLNet models to predict personality. In "A Comparative Study of Different Classifiers for Myers-Brigg Personality Prediction Model"(Comparative Study)[6], the authors directly utilized Naive Bayes, Random Forest and Logistic regression models to classify 16 personalities instead of dividing them into 4 dimensions. In addition, in the article "MBTI Based Personality Prediction of a User Based on Their Writing on Social Media"(Linear SVM)[7], the group achieved the prediction through the linear SVM.

In this study, the author devoted himself to classifying 16 distinct personalities through machine learning depending on the information they posted online. On the kaggle.com [8], a data set containing 8600 specimens of different personalities and the corresponding information they posted online is provided. Since each dimension has 2 directions, logistic regression, a model that can perform great while classifying 2 clusters, is utilized in the study to determine one's personality in varied dimensions. Before fitting the matrix, and transforming the texts into data, it has to overcome the natural language process. The study implements machine learning to test one's personality with daily messages, which is quite scarce in the investigations of the psychological field. The result of the study would examine the practicability of machine learning in the psychological field and explore more methods for finding the personality. Finally, best accuracy of the author's model reaches 78.67%, 85.82%, 58.73% and 63.0% in Mind, Energy, Nature and Tactics. Furthermore, the study digs into the discussion of factors that may influence the results.

2. Method

2.1. Data processing

2.1.1 Removing the useless information

A crucial step in approaching the model is firstly filtering the useless information. Obscuring the first five specimens in the posts list, the author found out each specimen contains the website sign (https://www.), numbers, punctuation, and the users' sign(@....). In our coding, the website sign is replaced by 'urlweb', and the users' sign is replaced by 'user'. All other unrelated information (punctuation and numbers) is deleted. Eventually, each post alters into a list only containing words.

2.1.2 Tokenization and Lemmatization

Tokenization is a word scanner imported in Python, it can separate a sentence or a paragraph into a list of separated words, meanwhile marking them as a string. Lemmatization is a crucial step in natural language processing, it can return the words to their original tense. Thus, the tokens are generalized to lemmas, which ascend the similarity in each post.

2.1.3 CounVector and TF-IDF Transformer

In the official document on scikit-learn.org [9], CountVector is defined as a package that can convert a collection of text documents to a matrix of token counts and produce a sparse representation of the counts using script. sparse.csr matrix. In other words, each word in the text that is applied to Count Vector would be assigned with one number to represent it. After applying the Count Vector, the TF-IDF Transformer is used to calculate the term frequency and the inverse document frequency. TF-IDF is a weighted technique utilized in information retrieval and text mining. The importance of a word ascends with the increasing of times it appears in the text but descends with the increasing of times it appears in the corpus. Thus, after applying TF-IDF Transformer, it can automatically filter

the words that have a small weight but store the words with relatively high weight. Finally, the data conducted as the x-axis in the logistic regression model is completely processed.

2.2. Model

2.2.1 Logistic Regression Model

After applying the Count Vector, the TF-IDF Transformer is used to calculate the term frequency and the inverse document frequency. TF-IDF is a weighted technique utilized in information retrieval and text mining. The importance of a word ascends with the increasing of times it appears in the text but descends with the increasing of times it appears in the corpus. Thus, after applying TfidiTransformer, it can automatically filter the words that have a small weight but store the words with relatively high weight. Finally, the data conducted as the x-axis in the logistic regression model is completely processed.

$$f(x) = \frac{1}{1+e^{x}} \tag{1}$$

In the coding process, the function is assumed as:

$$h(x) = \frac{1}{1+e^{(w^{T}x)}} \tag{2}$$

In equation(2), w is a matrix of the undetermined parameter that helps the prediction made of x-axis data to be more accurate. Solving the w is a crucial step in the logistic regression model. For solving the parameter, the loss function which measures the Mean Square Error (MSE)is always used, the smaller the loss function is, the more accurate the parameters are. Gradient descending is always utilized to find the minimum value of the loss fun1ction, thus solving the parameters.

$$J(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}(h(x_i) - y_i)^2 \tag{3}$$

Utilizing the function: get dummies in the pandas package, each personality is simply separated into 4 dimensions and represented by 1 and 0. The data that act as the y-axis have been completely processed. Fitting into the model, the final prediction is achieved.

2.2.2 Random Forest Model

In the past related work, the methodologies of Neuron Network (CNN), Logistic Regression, SVM, and Extreme Gradient Boosting were used, however, a relatively new method of ensemble learning has never been implemented in a similar field. Thus, the author decided to apply the random forest model which is a typical bagging model in ensemble learning. Bagging equals boostrap aggregating, which is one kind of method of sampling including returning the value. In other words, it is composed of several basic models.

For Random Forest model, it is constructed with a group of decision trees that have the same target, and the final prediction is processed from the result by averaging or voting. To ensure the diversity of the random forest, each decision tree should be varied, which is achieved by picking the sample randomly and choosing randomly 60% of traits [10]. Since it is composed of several different decision tree models, the random forest seems to be more flexible than logistic regression.

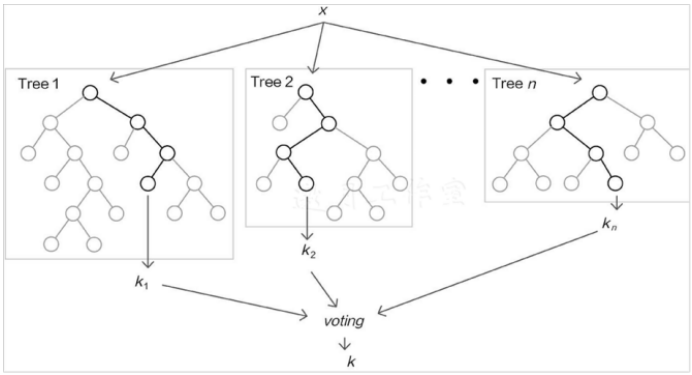

Figure 2: The Logic of Random Forest

Meanwhile, altering the n estimator in the Random Forest model can change the maximum iteration. Higher the times of iteration are, more decision trees there are, as a result, more accurate the prediction is. However, the improvement made by increasing the n estimator each time reduces and the time consumed increase, thus, finding a balance is crucial to the model.

3. Results

In the methodology section, the author has already utilized 1 and 0 to describe each personality, meanwhile, the posts have been transformed into Tf-idf vectors. Fitting the logistic regression model and Random Forest into 4 dimensions, we can achieve the prediction of the test set. To examine our model, the confusion matrix is utilized to output the accuracy of the algorithm.

Having an insight into the accuracy, the score of Mind-logistic regression and Energy-logistic regression is much higher than the other two logistic regression models. The reasons that affect the difference in each model would be profoundly discussed in the next section.

Table 1. Results of two models.

Model	Mind(I/E)	Energy(N/S)	Nature(T/F)	Tactics(J/P)
Logistic Regression	**78.27%**	**85.82%**	**58.04%**	**61.44%**
Random Forest(300 n estimator)	**78.67%**	**85.82%**	**58.73%**	**63.0%**

4. Discussion

4.1. Accuracy Variation

The logistic regression of Mind and Energy is great, by contrast, the performance of Nature and Tactics logistic regression is relatively inaccurate. Drawing the diagram that shows the distribution of population in each dimension, it is surprising that Mind and Energy's populations tend to be unbalancing distributed, by contrast, the distributions of Nature and Tactics are more balancing in Figure 4. However, the dataset with high unbalance surprisingly performs better. After considering this issue, the author makes the hypothesis that since logistic regression is a dichotomy, the unbalance dataset provides more information for one cluster in the model, as a result, the accuracy of that cluster ascended, thus, other specimens that are exclusive in that cluster would be automatically recognized as another cluster. Briefly, the unbalance of the data set would not have an absolute influence on the accuracy of the model. With a huge base of data, it may even provide a positive influence.

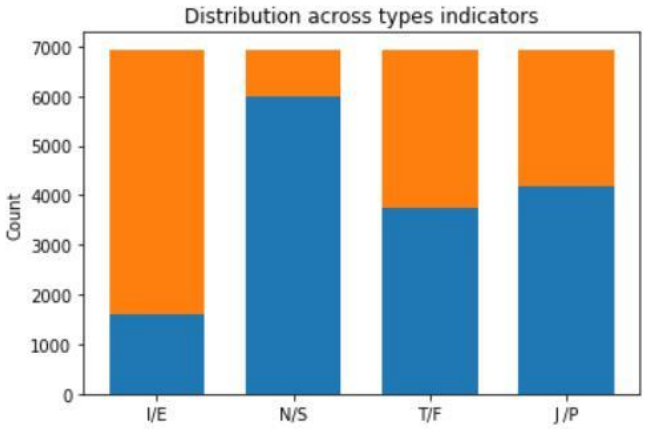

Figure 3: Distribution across types indicators

4.2. Noise

Utilizing the word cloud package in Python, we could print a figure of the word cloud of each personality. The author observed that the words "think", "people" and "one" are the first three words with the highest repetition. The author firstly thought the words would cause ambiguity to the computer. However, after removing all "think" in the text, the whole accuracy shows a decrease; after deleting all "people" in the text, the whole accuracy shows a slight increase; after removing all "one" in the text, the effect of accuracy is distinct for the varied logistic regression model and Random Forest model.

Table 2. Results of deleting specific words

Model	Mind	Energy	Nature	Tactics
LR(deleting "think")	74.8%	85.82%	54.41%	53.31%
LR(deleting "people")	78.33%	85.82%	59.54%	62.54%
RF(deleting "think")	73.83%	85.82%	54.06%	51.76%
RF(deleting "people")	78.39%	85.82%	59.37%	62.59%

Thus, according to the concept of Tf-idf transformer, the noise in the text would be words that have high and similar times of repetition in each personality. Although "Think" has a high repetition, it differs from 16 personalities, thus, it cannot be defined as noise, and the boundary of noise should be more profoundly studied.

4.3. Result Comparison

The author compares his accuracy with the results of 4 articles mentioned before. However, differences are presented in distinct investigation. For instance, BERT implemented the model in Big-five instead of MBTI, thus, its result has no comparability, and Comparative Study did not separate the MBTI into 4 dimensions.

Table 3. Results Comparison

Method	Mind	Energy	Nature	Tactics
MBTI	78.17%	86.06%	71.70%	65.70%
Linear SVM	58.9%	32.9%	30.7%	60.2%
Ours	78.67%	85.82%	58.73%	63.0%

5. Conclusion

In this paper, the author completes the 16 personalities prediction based on daily writing by Logistic Regression model and Random Forest model, and the best accuracy of 4 dimensions (Mind,

Energy, Nature, Tactics) reaches 78.67%, 85.82%, 58.73% and 63.0%. Comparing with recent studies in the similar field, the author finds that his algorithms have a satisfying performance. However, improvements are remained. During the discussion part, the author discovers that the accuracy of the model is related with the distribution of the data, and makes a hypothesis. Meanwhile, the author tries to identify the noise in the data, he concludes noise as the words that have high and similar times of repetition in each personality. Furthermore, the clear boundary of noise is not defined, and several improvements can be made.

References

[1] Resource from https: //www.myersbriggs.org

[2] L. R. Goldberg, "The Structure of Phenotypic Personality Traits," American Psychologist, vol. 48, pp. 26-34, 1993.

[3] C. G. Jung. (1921). Psychological Types. [Online].Available: http://ccmps.net/jung/types.pdf

[4] 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.Ma Kunlong. Short term distributed load forecasting method based on big data. Changsha: Hunan University, 2014.

[5] Christian, H., Suhartono, D., Chowanda, A., & Zamli, K. Z. (2021). Text based personality prediction from multiple social media data sources using pre-trained language model and model averaging. Journal of Big Data, 8(1), 1-20.

[6] Amjady N. Short-term hourly load forecasting using time series modeling with peak load estimation capability. IEEE Transactions on Power Systems, 2001, 16(4): 798-805.

[7] MBTI Based Personality Prediction of a User Based on Their Writing on Social Media. International Journal of Engineering and Advanced Technology(1). doi:10.35940/ijeat.a1523.109119.

[8] Resource from https://www.kaggle.com/competitions/edsa-mbti/overview/evaluation

[9] Resource from https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html

[10] Resource from https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html?highlight=randomforest#sklearn.ensemble.RandomForestClassifier