Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 7
TPCEE 2024

Logistic Regression Model to Personality Type Prediction Based on the Myers–Briggs Type Indicator

Yunshang Wang*

Department of Computer Science and Technology, Xiamen University Malaysia, Sepang, Selangor, Malaysia

*Corresponding author: CST2209167@xmu.edu.my

Abstract. The Myers-Briggs Type Indicator (MBTI) is a widely-used tool in psychology for determining personality types, playing a crucial role in fields like team building, communication, and personalized marketing. Despite its popularity, accurately classifying MBTI types using machine learning remains a significant challenge. This study focuses on addressing this challenge by exploring the effectiveness of logistic regression in MBTI classification tasks. Two approaches are used: four-times binary classification and multi-class classification. The findings show that while logistic regression performs exceptionally well in binary classification tasks but the accuracy is not good in multi-class classification. Additionally, combining binary classification results yields an overall accuracy that is lower than the direct multi-class classification. These results highlight the limitations of logistic regression in multi-class tasks and suggest the necessity for more advanced models. Future research should focus on improving multi-class classification accuracy, potentially through more complex architectures or hybrid models combining binary and multi-class approaches.

Keywords: Myers-Briggs type indicator; logistic regression; machine learning.

1. Introduction

The Myers-Briggs Type Indicator (MBTI) is an emerging psychology field concept to detect a person's personality quickly [1]. According to the answer to a series of questions, the person's personality can be shown in this indicator. There are 16 types of different MBTI personalities, each consisting of 4 letters. And each letter shows a trend in a certain field. The 4 fields are: mind, energy, nature, strategy.

The mind field can be represented as introverted (I) or extroverted (E), which describes how people obtain energy, such as through the "inner world" or the "outer world."

The energy field can be represented as sensing (S) or intuitive (N), which describes how people process information, whether they pay more attention to the content of the information itself, or obtain more information through divergent thinking.

The nature field can be represented as feeling (F) or thinking (T), which describes the way people make decisions. Whether to make decisions based more on logic and objective facts, or more subjectively and emotionally.

The strategic field can be represented as perceiving (P) or judging (J). This aspect describes people's orientation towards the external world, whether they prefer a structured and certain way of life or a more flexible and adaptable way of life.

Combining the letters of these preferences allows a person's MBTI personality type to be obtained.

Nowadays, a deep understanding of a person's personality has great contributions in many fields, such as psychology and personalized marketing strategies. The MBTI is a useful tool for team building, enhancing communication, and decision-making. So MBTI can be used as a tool more scientifically and effectively to maximize its role, thereby helping to build better interpersonal relationships and promote personalized services [2].

In the field of machine learning, MBTI classification is an important issue in studying classification algorithms. Using machine learning algorithms, models can learn from text data sets to accurately classify personality types in the training set. In existing research, many classifiers have been applied to solve the MBTI classification problem. Some of these studies use multiple classifiers at the same time and compare their accuracy to obtain results [3]. Another type of research is to select a specific classification algorithm for in-depth research at the principle level [4]. This study chose the second research type, using the basic logistic regression classification algorithm to deeply analyze and solve the MBTI classification problem. It aims to establish a high-precision prediction model of MBTI personality types and improve the performance of classification algorithms, thereby providing tools to help make more accurate predictions in multiple fields.

2. Methodology

The logistic regression model is a basic model of the classifier and is selected as the model to deal with this problem. Logistic regression is a widely used classification method with a solid theoretical foundation and relatively simple implementation. It is good at processing binary classification problems and can also be extended to multinomial logistic regression models to effectively solve multi-class classification problems. In addition, logistic regression performs well for linearly separable data and manages over-fitting in the data through appropriate regularization (such as L2 regularization). For these reasons, logistic regression is considered a basic and powerful tool in many classification tasks. It is a reasonable and effective choice for MBTI classification problems.

Logistic regression is performed using the sigmoid function. It plays a vital role in transforming and interpreting the output of the model [5]. The sigmoid function is mathematically expressed as the equation (1):

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

Where e is the base of the natural logarithms (approximately equal to 2.71828), and x is the linear combination of input features. This linear combination is often expressed as the equation (2):

$$x = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \tag{2}$$

In this equation, β_0 is the intercept, β_1, β_2 … β_n are the coefficients for the input features x_1, x_2 … x_n, respectively. The logistic function takes any real-valued number from the linear combination and maps it to a value between 0 and 1. This mapping is crucial because it allows logistic regression to estimate the probability that a given input belongs to a particular class. The S-shaped curve of the sigmoid function in Fig. 1 provides a smooth and continuous probability output. Ensuring that extreme values of the linear combination are squashed towards 0 or 1.

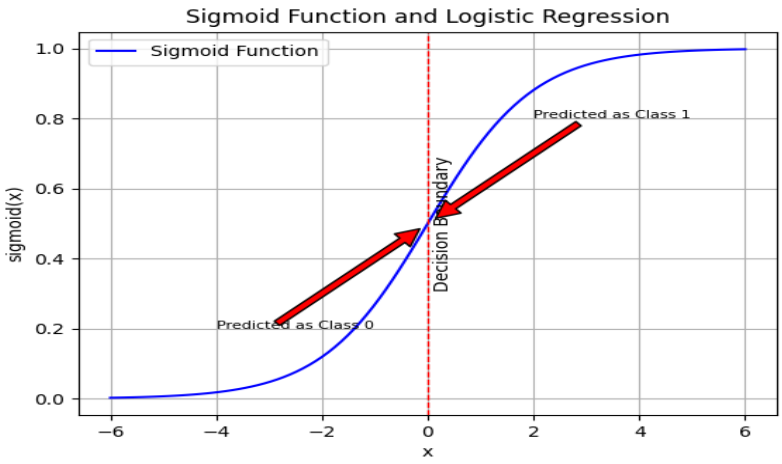

Fig. 1 Sigmoid function and logistic regression

When training a logistic regression model, using optimization algorithms like gradient descent is very important. Logistic regression takes a linear combination of input features and uses the sigmoid function to turn it into a probability value. This conversion is key to interpreting and predicting results. Specifically, when the linear combination of features is input, the sigmoid function maps it to a probability value between 0 and 1. This lets the model predict the likelihood that the input belongs to a particular category, enabling it to make classification decisions [6].

In method selection, two approaches are chosen to analyze the MBTI classification problem. The first approach breaks down the problem into four separate binary classification tasks, which is very suitable for logistic regression. The second approach extends binary logistic regression models to solve the problem using multinomial logistic regression.

2.1. Four-times Binary Classification

Logistic regression has advantages in binary classification tasks, as it can provide a simple model structure and efficient computing performance for classification tasks. Other than that, it can also handle linearly separable data with good generalization and minimal computational overhead. At the same time, the MBTI question has the unique characteristic that it covers four independent domains (Extroverted/Introverted, Sensing/Intuitive, Thinking/Feeling, and Judging/Perceiving). For the personality tests conducted within each domain, Four two-choice questions are finally combined. Therefore, it can be broken down into four binary classification tasks, each task corresponding to producing one letter. This decomposition method takes advantage of logistic regression in binary classification, improving the prediction accuracy of each category. At last, the predicted categories for each letter can be combined using the probabilities output from the logistic regression to arrive at the final MBTI-type prediction.

2.2. Multi-class Classification

In multi-class classification problems, like doing this MBTI classification problem directly without having any conversation, which means that it is treated as 16 different types. Multinomial logistic regression can be used to address the problem in the following steps:

Extending the Linear Model from the equation (2) to get the equation (3) :

$$z_i = \beta_{i0} + \beta_{i1}x_1 + \beta_{i2}x_2 + \cdots + \beta_{in}x_n \tag{3}$$

Softmax Function uses z_i, which is the linear combination for the i-th class:

$$P\,(y = i|x) = \frac{e^{z_i}}{\sum_{j=1}^{k} e^{z_j}} \tag{4}$$

Where $P\ (y=i|x)$ is the probability of x belonging to category i. $\sum_{j=1}^{k} e^{z_j}$ is the sum of the exponentiated scores for all k classes.

The maximizing Log-Likelihood Function is defined as the equation (5):

$$\text{Log-Likelihood} = \sum_{i=1}^{N} \sum_{j=1}^{k} y_{ij} \log P\,(y = j|x) \tag{5}$$

In order to achieve this goal, various algorithms like Newton-CG, LBFGS, SAG, and SAGA are used for optimization [7]. It makes multinomial logistic regression an ideal choice for the multi-class classification problems of MBTI.

3. Experiments

3.1. Data Set

The data used for training and testing comes from the repository of the Kaggle platform. This text data is saved in a CSV file and mainly comes from the comments section of the Personality Cafe website forum. Based on the ratio of 70% training set and 30% test set, the data is divided into two parts for the next stage of operation:

In more detail, the training data consists of 6072 rows and two columns. Each row corresponds to a user. The posts column represents the user's posts. Posts not only contain plain text content but also include web addresses, emoticons, and something else that won't be used. The type column contains each user's MBTI personality type. Likewise, the test data consists of 2063 rows and 2 columns, each row corresponds to an individual user with the same kind of content in each column. It can also be found that over 90% of posts contain more than 4000 letters. This suggests that individual data entries have sufficient length for accurate MBTI personality classification, reducing the likelihood of inaccuracy due to insufficient words.

For data visualization, two charts are created. First, the 16 MBTI personality types are counted according to the labels in the first column of the data set, and then the data are integrated with the data with the same labels in the data set. A bar chart is used to visually Show the distribution of numbers for each personality type, as shown in Fig. 2.

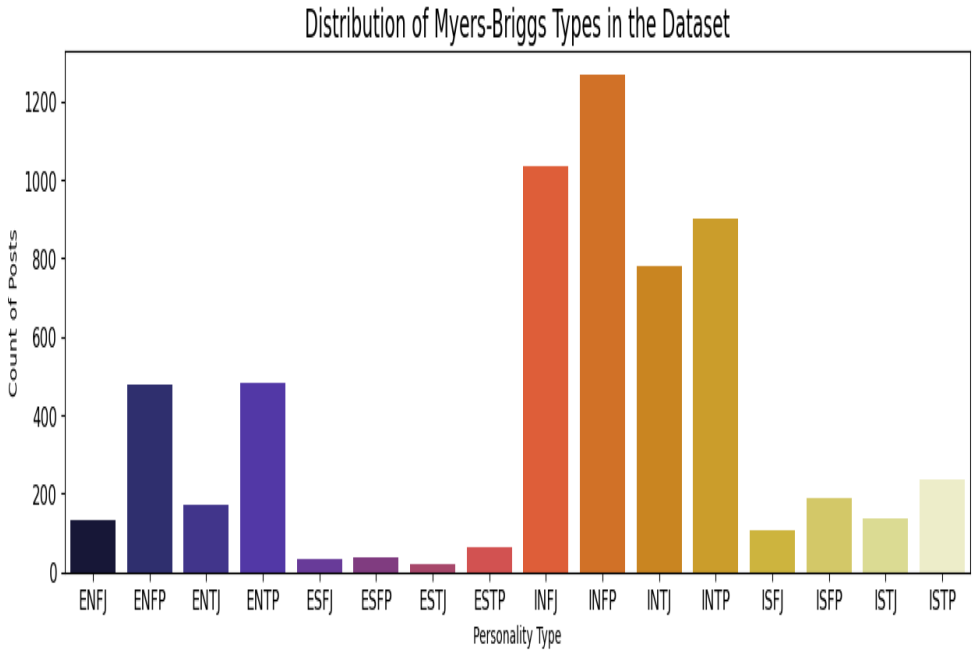

Fig. 2 Distribution of MBTI types in the data set

In Fig. 3's swarm plot, each point represents an observation. This plot intuitively compares the distribution of different personality types by the number of review words. Analyzing variations in word usage and sentence length provides insight into the sentence structure characteristics associated with each personality type [8].

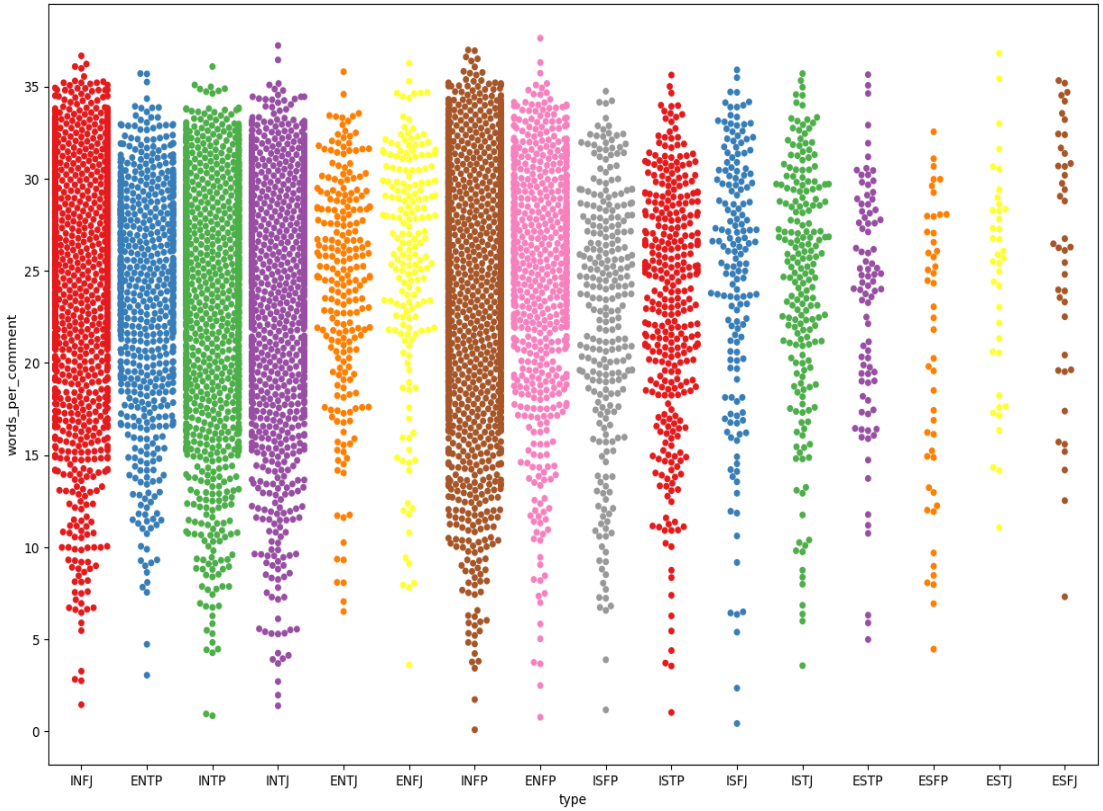

Fig. 3 Words per comment for each personality type

3.2. Data Preprocessing

3.2.1. Lowercasing

With the purpose of improving the accuracy and consistency of the text-based data. All the words in the text are converted into lowercase. This step is significant because it allows both the uppercase and the lowercase word to be seen as the same word, which the different situations or places may cause.

3.2.2. Selective Word Removal

It is necessary to do some word removal. Initially, the existing URL and the data point containing links to websites should be removed since they contribute nothing to the classification. Then, the words in the data should be as meaningful as possible, so that the stop words like ‘for’, ‘them’, ‘you’ etc. can be removed by using the NLTK library. Furthermore, the numbers, extra spaces, and special characters (such as vertical bars ' | ' hyphens' - ', and commas ', ').

3.2.3. Lemmatization

To standardize the text, the Python library nltk. stem.WordNetLemmatizer is used to lemmatize the text. This is a process to transform the various inflected forms back to their common root word with one shared meaning, for instance, ‘gone’, ‘going’, and ‘went’ are all converted to "go".

3.2.4. Tokenization

In natural language processing(NPL), machine learning models and algorithms need to receive an understanding-friendly form of data to deal with it efficiently. Because of that, tokenization is a crucial step in data preprocessing that aims to divide text-based data into a smaller unit. The TreebankWordTokenizer in the NLTK library is used to perform this problem [9]. For example, for a post "This is an example post, demonstrating tokenization.", the tokenization process will generate a list of words: ['This', 'is', 'an', 'example', 'post', ',', 'demonstrating', 'tokenization', '.'].

3.3. Model Implementation

TfidfVectorizer(TFIDF) is chosen to transform the Data. Among various common text processing tools, unlike CountVectorizer which simply counts the number of times each word appears in the document, TFIDF is chosen because of its unique properties. TFIDF Vectorizer introduces the concept of "inverse document frequency part" (IDF) based on word frequency. The function of IDF is to reduce the weight of words that appear frequently in all documents, thereby highlighting those words that are more representative of a certain document [10].

During the four-times binary classification experiment, each personality type is converted into multiple binary classification labels and it can transform complex personality types into simple binary classification problems. Then, the TFIDF vectorization method performs feature extraction on the text data for the previously classified training set and test set. Next, grid search is used to tune the parameters of the logistic regression model to optimize model performance. Grid search finds the parameter values that give the model the best performance by iterating through all possible combinations of parameters. The parameters tuned include regularization parameter C and penalty type differently in different trails. After training is completed, the evaluation of the model's performance is shown in predicting the results of the test set. Specific evaluation indicators include accuracy and confusion matrix.

The same procedure is used during the Multi-class classification experiment except for the first step. Instead of converting it into multiple binary classification labels, it is treated as 16 labels directly. Also, the accuracy and confusion matrices are generated for further comparison and interpretation.

3.4. Result

In model performance evaluation, accuracy is a key metric. To fully demonstrate the performance of our logistic regression model, Table 1 details the accuracy results of the two models on the test set.

Through this table, the intuitive comparison of the accuracy can be conducted between the accuracy of each binary classification and the accuracy in the multi-class classification experiment.

Table 1. The accuracy result of the two models on the data set

Type	Accuracy
E/I	85.8%
N/S	89.3%
T/F	86.2%
P/J	79.1%
Multi-class	62.0%

To further analyze the classification effect of the model, the confusion matrix is drawn. Fig. 4 and Fig. 5 show more detailed classification performance details of the two analysis methods and provide an in-depth understanding of the model's performance in different categories.

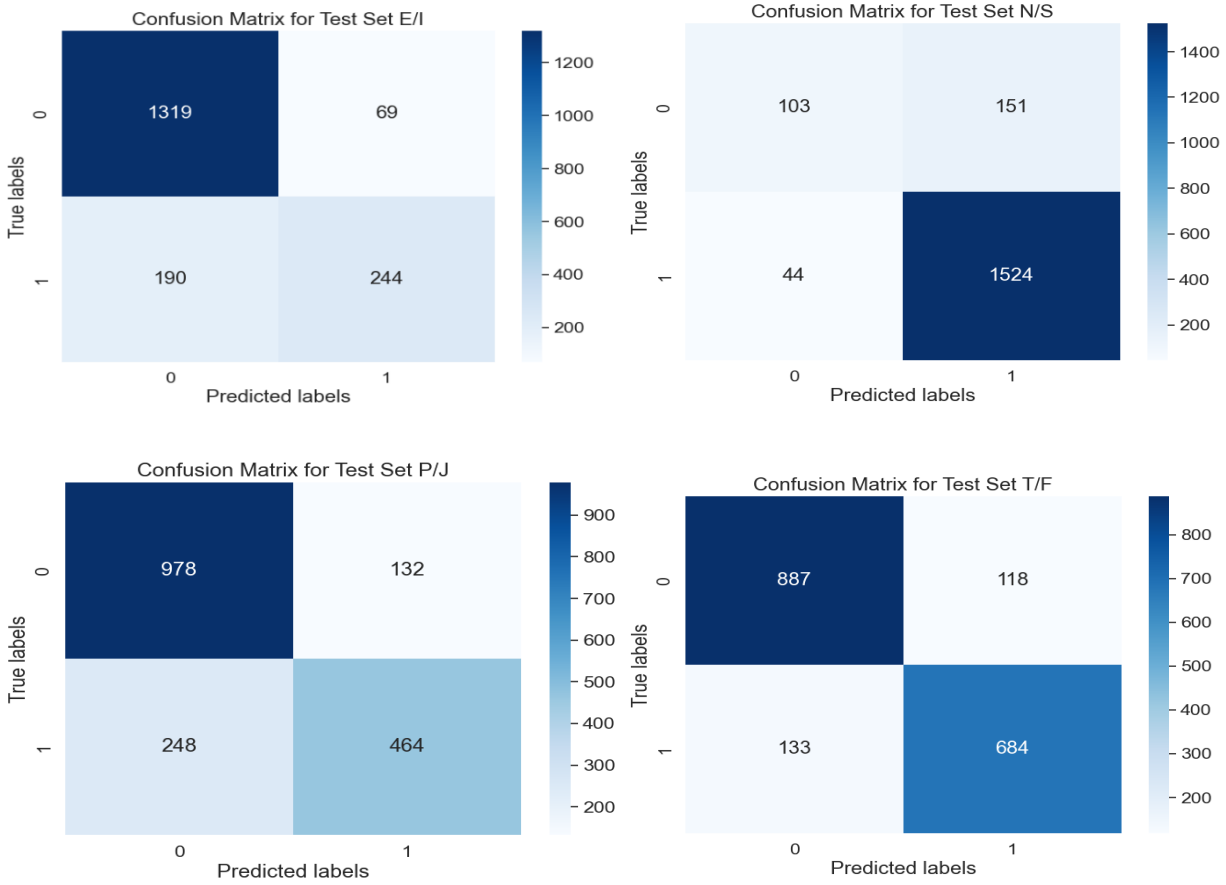

Fig. 4 The confusion matrix of four-times binary classification

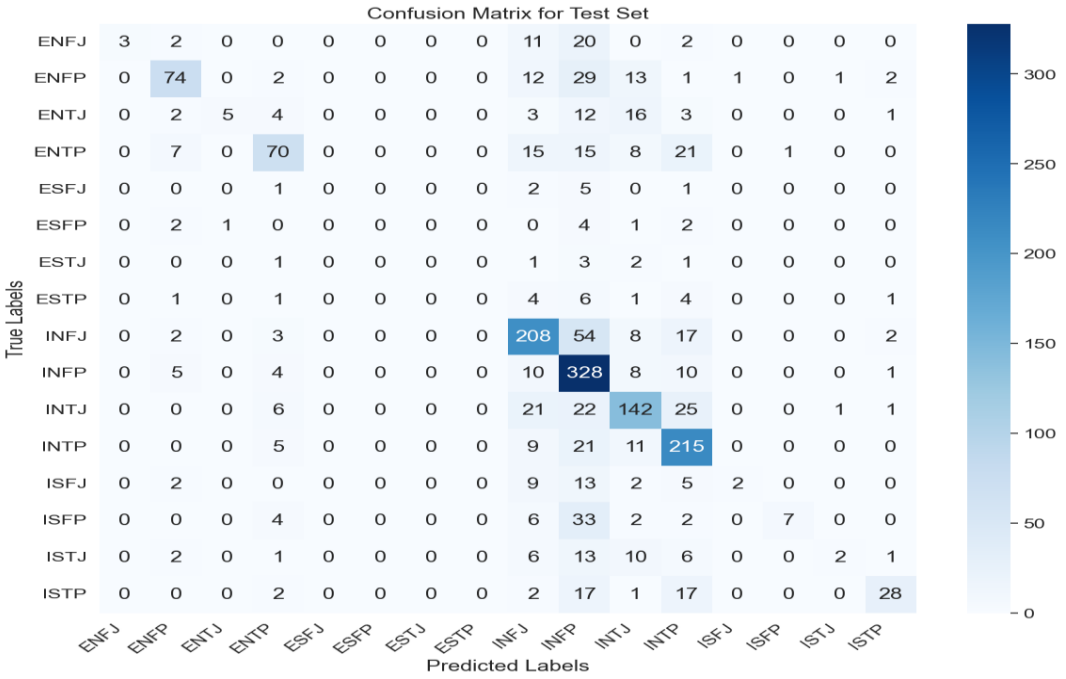

Fig. 5 The confusion matrix of multi-class classification

In addition to the quantitative analysis, Fig. 6 of word clouds is generated to illustrate each letter in MBTI's evaluation standard. The word cloud chart shows the words that appear more frequently in the text data, and these words have higher weights when constructing the TF-IDF feature matrix. Provides a valuable visual reference for further text analysis and model optimization.

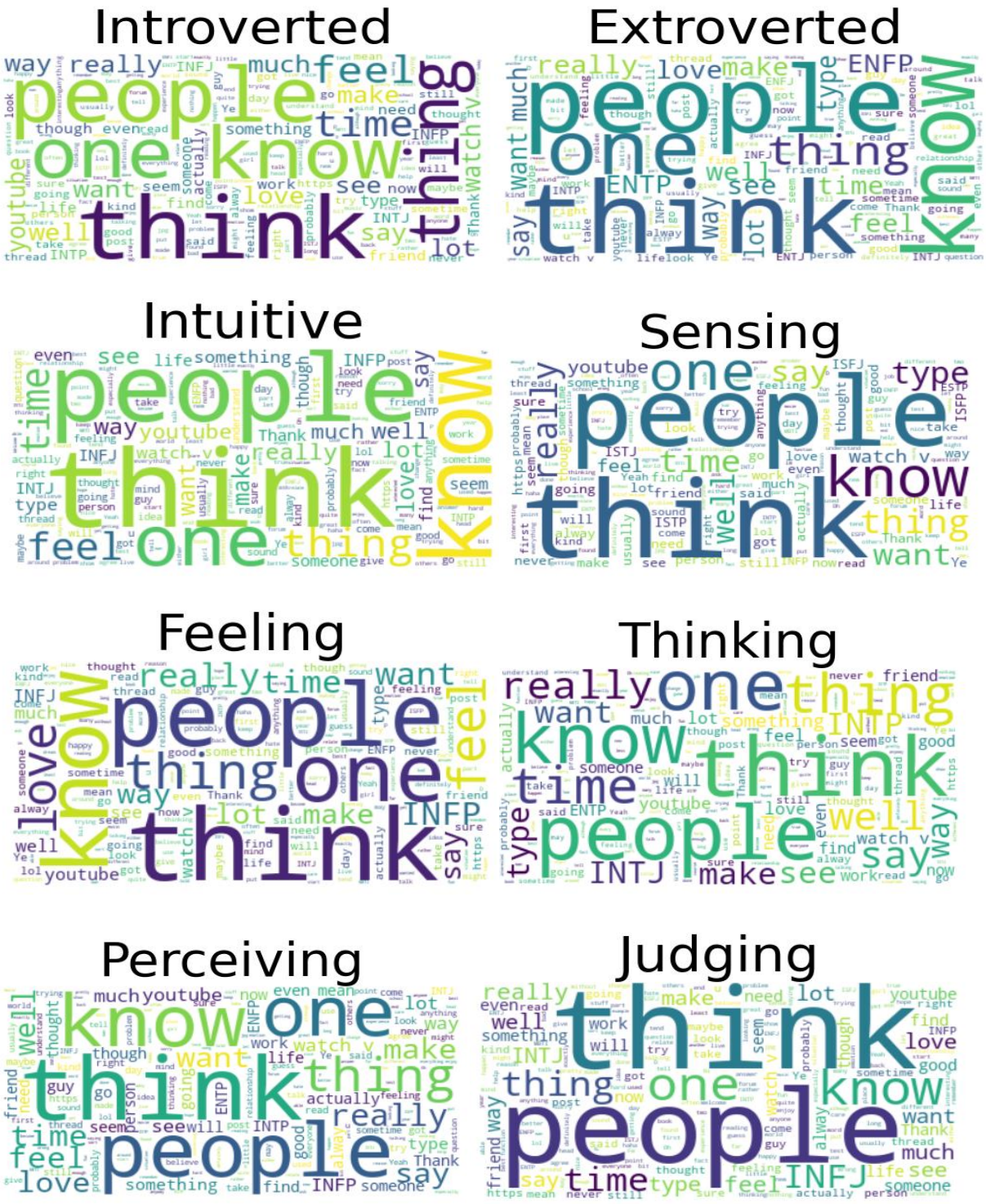

Fig. 6 Word clouds of each letter in MBTI's evaluation dimension

4. Conclusion

This study evaluates the performance of logistic regression models in different classification tasks. The results show that the model achieves significant success in binary classification with very high accuracy. For example, the accuracy of the I/E dimension is 85.8%, the N/S dimension is 89.3%, the T/F dimension is 86.2%, and the P/J dimension is 79.1%. These results demonstrate that logistic regression is very capable of identifying individual binary classification dimensions. However, when the problem is transformed into a multi-class classification, the model's accuracy drops significantly to only 62%. This decrease may be due to the increasing complexity of multi-class classification tasks. In order to better understand the relationship between binary classification results and multi-category

classification results, the overall accuracy rate after combining the binary classification results is calculated to be 52.8%, which can be approximated by multiplying the individual binary classification accuracies. The accuracy of 52.8% is lower than the 62% accuracy of multi-class classification. This result shows that although the model performed well on a single binary classification task, these results failed to provide higher accuracy than direct multi-class classification when combined together. Therefore, despite achieving high accuracy in the binary classification task, merging these results does not improve the overall classification performance and is worse than directly performing multi-class classification.

The main contribution of this study is to give two models to solve the MBTI classification problem, and also reveal the advantages of logistic regression in binary classification problems and the limitations in multi-class classification tasks. It points out the advantages of directly performing multi-class classification when dealing with multi-dimensional classification problems. This finding has important guiding significance for model design and selection of classification strategies.

In the future, further research can be devoted to improving the performance of multi-class classification models. Specifically, more complex model architectures, such as deep learning models, or more contextual information can be introduced to enhance classification accuracy. In addition, studying how to effectively combine the advantages of binary classification and multi-class classification and develop hybrid models is also a direction worth exploring. With these improvements, better classification results are expected to be achieved when dealing with complex multi-dimensional data.

References

[1] Lee Hyejin, Shin Yoojin. A Study on MBTI Perceptions in South Korea: Big Data Analysis from the Perspective of Applying MBTI to Contribute to the Sustainable Growth of Communities. Sustainability, 2024, 16(10): 4152.

[2] Zhuo Anni. Application of the MBTI personality test in the workplace. Knowledge Economy, 2009, (2): 5-5.

[3] Villegas-Ch. William, Erazo Daniel Mauricio, Ortiz-Garces Iván, Gaibor-Naranjo Walter, Palacios-Pacheco Xavier. Artificial intelligence model for the identification of the personality of Twitter users through the analysis of their behavior in the social network. Electronics, 2022, 11(22): 3811.

[4] Amirhosseini Mohammad Hossein, Kazemian Hassan. Machine Learning Approach to Personality Type Prediction Based on the Myers–Briggs Type Indicator®. Multimodal Technologies and Interaction, 2020, 4(1): 9.

[5] Agarwal Devesh, Karthikeyan M. Personality Prediction Using Machine Learning. International Research Journal of Modernization in Engineering Technology and Science, 2022.

[6] Zaidi Abdelhamid. Mathematical justification on the origin of the sigmoid in logistic regression. Central European Management Journal, 2022, 30(4): 1327-1337.

[7] Chin Xin Yee, Han Yang Lau, Zhi Xin Chong, Man Pan Chow, Zailan Arabee Abdul Salam. Personality prediction using machine learning classifiers. Journal of Applied Technology and Innovation, 2021, 5(1): 1.

[8] Chaudhary Shristi, Singh Ritu, Hasan Syed Tausif, Kaur Inderpreet. A Comparative Study of Different Classifiers for Myers-Brigg Personality Prediction Model. International Research Journal of Engineering and Technology (IRJET), 2018, 5(5): 1.

[9] Ryan Gregorius, Katarina Pricillia, Suhartono Derwin. Mbti personality prediction using machine learning and smote for balancing data based on statement sentences. information, 2023, 14(4), 217.

[10] Kumar Vipin, Subba Basant. A TfidfVectorizer and SVM based sentiment analysis framework for text data corpus. 2020 National Conference on Communications (NCC), 2020, 1-6.