Machine Learning-Based Personality Prediction Using the Myers-Briggs Type Indicator

Yi Luo *

School of Information Science and Engineering, Shandong Agricultural University, Tai'an, Shandong, China

* Corresponding Author Email: noc@sdau.edu.cn

Abstract. In the digital age, artificial intelligence and machine learning are becoming integrated into all aspects of life. The Myers-Briggs Type Indicator (MBTI) is widely utilized in career planning, team building, and counseling. Traditional MBTI tests require active participation and can be time-consuming and subjective. However, using machine learning to predict MBTI types represents a novel research direction. This study explored the application of machine learning to predict an author's MBTI type based on text data, such as social media posts. A successful text-based classifier was developed, establishing a linguistic foundation for understanding MBTI and personality traits. The research utilized an MBTI dataset derived from posts on the Personality Café forum, along with their corresponding types. The dataset underwent rigorous preprocessing to ensure accuracy and consistency. Two models were trained on the processed dataset: random forest and logistic regression. These models effectively manage complex datasets and generate accurate predictions. The study concluded that predicting personality traits from text data using machine learning is feasible. This finding has significant implications for personality theory, paving new pathways for understanding and applying insights from assessments. This research highlights the significant potential of machine learning in enhancing our comprehension of personality traits and establishes a solid foundation for future studies in this field.

Keywords: Machine learning; Personality prediction; Logistic regression.

1. Introduction

In the realm of psychology, individual characteristics is regarded as a powerful yet vaguely defined concept. Psychologists would greatly benefit from more concrete, data-driven evaluations of existing personality models [1]. This study aims to explore one specific personality model: the MBTI. The object of this paper is to utilize machine learning methodologies to create a predictive model that processes textual information, like social media entries, and forecasts the writer's MBTI personality category. Successfully creating such a classifier of this nature would offer a solid linguistic basis for comprehending MBTI and the broader spectrum of personality.

Furthermore, developing an accurate text-based classifier could have significant implications for the scope of psychology, considering the intricate relationships between natural language processing, various machine learning algorithms, and personality types.

In this paper, the forecasting of personality characteristics is successfully achieved via the analysis of users' submissions of cover letters, poems, and short essays.

The object of this paper is to utilize a machine learning (ML) algorithm to anticipate individual personality categories [2]. To achieve this, the MBTI dataset was utilized, which comprises posts from the Personality Café forum along with corresponding personality types based on the MBTI. The object of the paper is to make use of the machine learning algorithm to predict individual personality types [2]. Therefore, this paper used the MBTI dataset, which contains the posts in the Personality Café forum and homologous personality types based on the frame of the MBTI.

2. Method

The methods used in this paper are random forest and logistic regression.

2.1. Data pre-processing

Data pre-processing is an essential procedure in this paper to prepare raw data for further processing. It essentially involves transforming raw, disorganized data into a form that is understandable and usable [3]. The primary objective of data pre-processing is to ensure that the quality of the data is sufficient for effective and efficient data mining and subsequent analytical tasks.

1. All the text is transformed into lower case.
2. URLs, mentions, special characters, and stop words are removed.
3. The process continues with extracting the stem and reducing its morphology.
4. The Term Frequency-Inverse Document Frequency (TF-IDF) technique is utilized for vectorizing the text.
5. Dealing with unbalanced data involves addressing the issue that some personality types in the MBTI dataset have significantly fewer samples than others. This is managed by using the undersampling technique, the oversampling technique, and the SMOTE technique to balance the data [4].

The fundamental concept of TF-IDF is predicated on the principle that a word or phrase which frequently occurs in one document (indicated by a high term frequency, or TF) but seldom appears in other documents, possesses strong discriminative power and is thus well-suited for classification purposes [4].

The greater the TF-IDF value, the more significant the word's importance within the document [4]. The formula is expressed as:

$$TF - IDF = TF \times IDFTF - IDF = TF \times IDF \quad (1)$$

Where TF = maximum word frequency in a document; IDF = log number of documents that contain the word Total number of documents [5].

This technology is widely used in search engines, text mining, information retrieval, and other fields to help find keywords in document collections or corpora [5].

Due to its effectiveness in weighting words for machine learning algorithms in natural language processing, separate classifiers have been trained for this purpose [6]. In this process, both count vectorize and TF-IDF vectorize were applied to convert the model into a vector form, and words appearing with frequencies between 10% and 70% were filtered out [7].

2.2.1 Random forest

The random forest predict model is an integrated learning method [8]. It can improve the accuracy of predictions by combining multiple decision trees. Fig. 1 of random forest shows the establishment process of random forest. The literal description is as follows:

1. The Bootstrap (sampling with replacement) method is used from the original dataset, which is transformed to yield N additional training datasets.
2. A decision tree is built for each new training data set, resulting in N decision trees.
3. When splitting at each node of each decision tree, only M features are selected for comparison (where M is less than or equal to the total number of features), and the best feature is chosen for the split [9].
4. Finally, the N decision trees are combined to form a random forest prediction model.

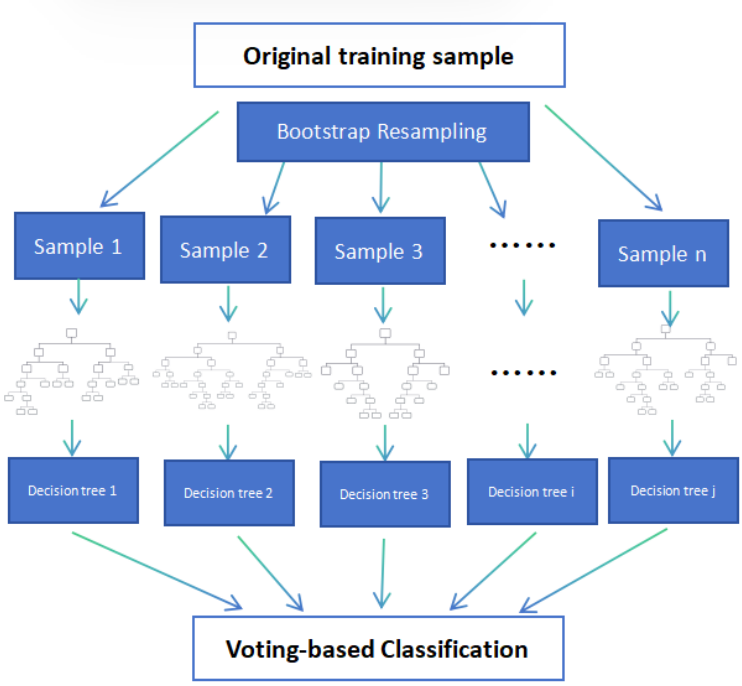

Fig. 1 Random Forest

The advantages of the Random Forest prediction model are as follows:

1. Random Forest model can handle high-dimensional data with many features and does not require feature selection, allowing it to assess the importance of the features.

2. It is capable of evaluating the interaction between different features.

3. The model can effectively process nonlinear data.

4. The training speed is relatively quick, making it amenable to parallelization methods. The model is insensitive to missing values and outliers.

In a word, the Random Forest predict model is a efficacious and flexible machine learning algorithm, and it is suitable for all kinds of prediction tasks. It raises the accuracy of prediction through the combination of multiple decision trees, and it has many advantages, such as handling high-dimensional data, nonlinear data, and missing values.

2.2.2 Logistic regression

The logistic regression prediction model is a machine learning algorithm to handle binary classification [4]. It can predict the probability of an event occurring through the relationship between data features and target variables [5]. The output of logistic regression is a probability value between 0 and 1. It usually sets a threshold value (such as 0.5). When the prediction probability is greater than the threshold, the class is positive, and otherwise, the class is negative [10]. The core of the logistic regression prediction model is a logical function (also known as Sigmoid function), whose expression is:

$$h(x) = \frac{1}{1+e^{-(\omega x+b)}} \tag{2}$$

Where h(x) represents the prediction probability, indicating the degree of confidence that the input feature x belongs to a particular class.

x is the input feature of the model, which contains all the relevant information for the prediction. These features can be raw data, such as pixel values, text characters, or sensor readings, or they can be data that has been pre-processed or feature-engineered.

w is the weight vector that contains the model parameters used to map the input feature x to the output h(x). During training, the model learns these weights in order to minimize prediction errors.

The size and symbol of the weights reflect the degree and orientation of the influence of the corresponding features on the prediction results.

b is the offset term, which is a constant value used to adjust the baseline of the model output. The offset term can help the model fit the data better because it allows the output to shift away from the origin without any input.

e is a natural constant, also known as the Euler number, which is approximately equal to 2.71828. It has a wide range of applications in mathematics, especially in calculus and complex analysis.

The training process of the logistic regression model is mainly to solve the weight w and the bias term b by the maximum likelihood estimation method [10]. In the training process, it is necessary to compute the loss function (e.g., the cross-entropy loss function) in order to assess the model's predictive accuracy. The parameters are then updated continuously through the use of optimization algorithms (like the gradient descent method) aiming to minimize the loss function.

3. Experimental result

The dataset consists of a table with two columns, with the dataset is a table containing two columns, with 8,675 rows. Notably, this dataset is characterized by the absence of any null values, which is an advantage in data analysis as it eliminates the need to deal with missing data issues [4]. However, a potential issue arises from the fact that all values within the dataset are in text format. This means that prior to utilizing this data, there is a need to transform it into a format that is appropriate for model processing, such as numerical or categorical data.

By utilizing the command to conduct descriptive statistical analysis on the dataset, some statistical information pertaining to one of the columns, presumed to be the "Type" column, was obtained. The results showed that this column has 8,675 non-null values, which matches the total number of rows, indicating that there are no missing values in this column. Additionally, there are 16 unique personality type indicators in this column, demonstrating the diversity of the data.

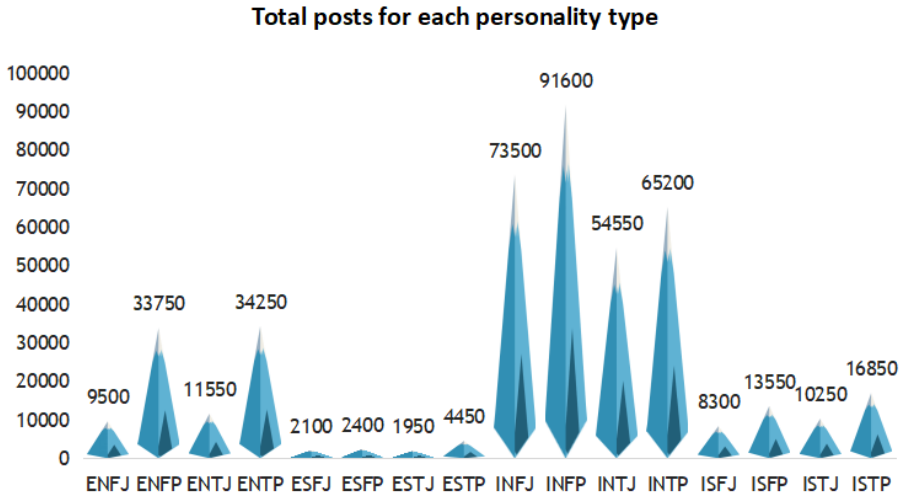

Fig. 2 Personality types

As shown in Fig. 2, among these 16 unique personality types, the INFP type appears most frequently. This suggests that INFP is the most common personality type in our dataset and may have a substantial influence on the analytical outcomes. However, the conclusion drawn from the analysis of user comments suggests that individuals who frequently engage in commenting on social media platforms may exhibit a greater propensity towards introversion, sensing, and emotional characteristics.

Finally, the posts in the dataset are unique, with no duplicate posts. This means that each post provides unique information, enhancing the value of the dataset and the reliability of the analysis.

4. Conclusion

In this paper, two machine learning classifiers were examined for the purpose of categorizing personality types. The central aim of this study was to investigate new algorithms and enhance the precision of the baseline algorithm, logistic regression. Performance metrics such as precision, recall, accuracy, and F1 score were employed to assess the efficacy of all classifiers under consideration. Notably, the cumulative results indicated that logistic regression continued to surpass the performance of all newly introduced algorithms. Furthermore, personality prediction constitutes an abstract modeling approach, suggesting that it provides a conceptual description of phenomena. In contrast, concrete models in machine learning yield direct, analogous outcomes. As a consequence, abstract models may not achieve extremely high accuracy rates when juxtaposed with concrete models.

In future research, techniques for hyperparameter tuning such as grid search and Bayesian optimization could be utilized to identify the most optimal model configurations. Additionally, ensemble learning methods like bagging and boosting could be utilized to enhance the stability and generalization capabilities of the models.

References

[1] Pooja K, Rakshitha G H, Spoorthi S, et al. Personality Prediction Using Machine Learning. 2023 International Conference on Computational Intelligence for Information, Security and Communication Applications (CIISCA). IEEE, 2023: 267-272.

[2] Sönmezöz K, Uğur Ö, Diri B. MBTI personality prediction with machine learning. 2020 28th Signal Processing and Communications Applications Conference (SIU). IEEE, 2020: 1-4.

[3] Zhang H. Mbti personality prediction based on bert classification. Highlights in Science, Engineering and Technology, 2023, 34: 368-374.

[4] Ashraf N, Ahmad R S, Bano S, et al. Enhancing MBTI Personality Prediction from Text Data with Advance Word Embedding Technique. VFAST Transactions on Software Engineering, 2024, 12(3): 35-43.

[5] Mushtaq Z, Ashraf S, Sabahat N. Predicting MBTI personality type with K-means clustering and gradient boosting. 2020 IEEE 23rd International Multitopic Conference (INMIC). IEEE, 2020: 1-5.

[6] Nisha K A, Kulsum U, Rahman S, et al. A comparative analysis of machine learning approaches in personality prediction using MBTI. Computational Intelligence in Pattern Recognition: Proceedings of CIPR 2021. Springer Singapore, 2022: 13-23.

[7] Justindhas Y, Mohanraj S M, Shivani R. A synoptic survey on personality prediction system using mbti. 2022 3rd International Conference on Electronics and Sustainable Communication Systems (ICESC). IEEE, 2022: 950-955.

[8] Abidin N H Z, Remli M A, Ali N M, et al. Improving intelligent personality prediction using Myers-Brigg's type indicator and random forest classifier. International Journal of Advanced Computer Science and Applications, 2020, 11(11): 232285846.

[9] Kukreja S, Choudhary H, Abbas A. Personality Prediction: Ensemble Learning of Logistic Regression and SVM with LightGBM for Automated Personality Classification. 2023 5th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N). IEEE, 2023: 30-35.

[10] Arya S, Joy A M, D'Souza J N. Prediction of MBTI with textual data using different pre-trained transformer models. 2023 Fourth International Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE). IEEE, 2023: 1-7.