Prediction of MBTI personality leveraging machine learning algorithms

Ze Li

Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong, 999077, China

1155157157@link.cuhk.edu.hk

Abstract. In this study, the author attempted to implement a machine learning approach to determine users' corresponding MBTI personality types by relying only on the content of their online forum postings. Models based on different algorithms are built and trained, and the natural language of the collected data set is converted into machine language for machine learning and used in subsequent tests to determine the correctness of the predicting results. The data set is collected from the forum and divided into two parts, the training set is leveraged to train the model and the test data set is leveraged to make personality predictions and compare with the training data set to measure the correctness of the predicting outcomes. The results show that logistic regression algorithm and vectorized representation of text with TfidfVectorizer can best accomplish the prediction task. This study completed a preliminary comparison of algorithms for personality prediction from text, which became the basis for subsequent personality model predictions using other media.

Keywords: machine learning, logistic regression, MBTI personality.

1. Introduction

Personality is one of the hot topics right now, with the Myers-Briggs type indicator (MBTI) having the highest level of discussion. According to Funder, personality traits could be constructed by the way of thinking, feeling, and acting, this means that personality affects every aspect of a person's life [1,2]. Personality theories were developed from different psychological theories, and among them, MBTI is a theoretical model of personality types developed by Briggs Myers and her mother based on the 8 psychological types classified by Carl Jung [3,4]. This theory attempts to divide a personality into four dimensions, as shown in Table 1, each with two opposite types: Introversion or Extroversion, Sensing, or Intuition, Thinking or Feeling, Judging or Perceiving. Each person will have only one of the two types for each dimension above. Combining the four dimensions gives the final personality type, total of 16 kinds.

Table 1. Types of MBTI personality.

Dimensions	Type 1	Type 2
Mind	Introverted (I)	Extraverted (E)
Energy	Sensing (S)	Intuitive (N)
Nature	Feeling (F)	Thinking (T)
Tactics	Perceiving (P)	Judging (J)

There are many ways to predict personality, such as self-testing by filling out a questionnaire or using the test function that comes with social media [5]. The Internet provides a stage for everyone to express themselves, so is it possible to determine a person's personality type from the comments, emojis or other such electronic traces left by each person on the Internet? Advances in “big data” analytics offers the possibility of this question, and machine learning with classification algorithms provide the affirmative answer [6].

This works attempts to predict personality leveraging different machine learning algorithms. By comparing the effectiveness of these methods, this paper demonstrates that machine learning is a promising solution towards personality prediction.

2. Method

The flowchart of the code implementation is demonstrated in Figure 1. In this section, the information of the dataset, the data preprocessing, and the modelling process will be elaborated sequentially.

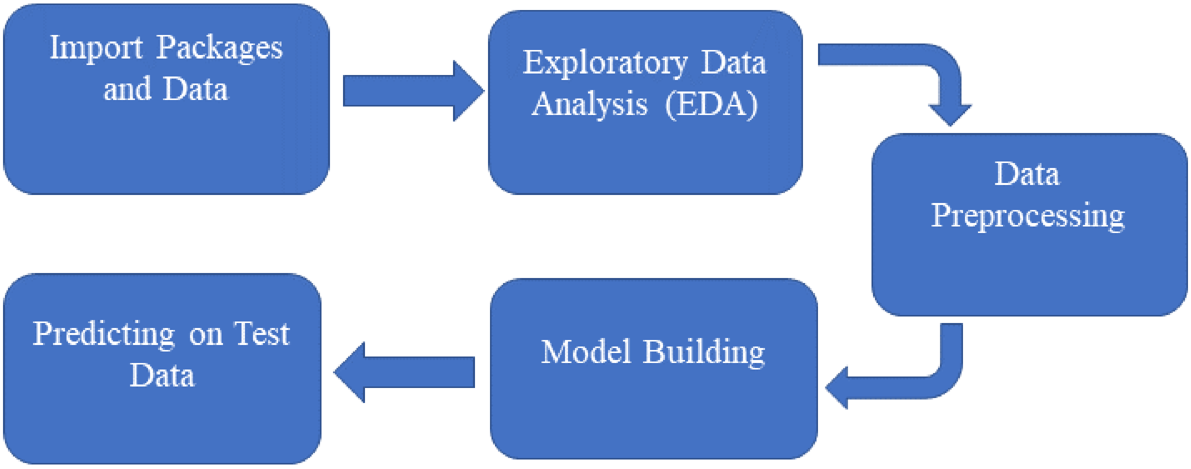


Figure 1. Flowchart of the implementation.

2.1. Dataset

In this research, it is necessary to test the efficiency and prediction accuracy of an algorithm, so two sets of data are used in the research, the training, and the testing data. They are all collected from the Personality Cafe website forums randomly [7]. The training data consists of 6505 rows, with 2 columns and there are 2169 rows with 2 columns in test data set.

The reason for collecting data in this forum is that the posts in this forum are more likely to reflect the personality type of the author than posts in other forums. Therefore, these data are easier to discern and process. And the randomly selected data can reflect the real personality type distribution of the post authors to a certain extent, thus making the data more valid. But data that do not reflect the personality type is discarded.

2.2. Data preprocessing and modeling

The baseline for this research is building and training a model that is capable of predicting labels for each of the four MBTI dimensions. That is to say, the four dimensions of personality type were predicted separately, and then the four results were combined to obtain that person's personality type.

Although invalid data is discarded at the time of collection, the data used for the research still need to be preprocessed. Filtering data based on two criteria: ease of use and usability. Delimeters, URLs, punctuation and numbers were removed. Stopwords were removed by vectorizer in the model building section, as CountVectorizer and TfidfVectorizer have the ability of removing stopwords. Words were converted to lowercase and they were lemmatized as well. These pre-processes were designed to avoid having multiple copies of the same words and remove noise, which may negatively affects the accuracy of the model [8]. In terms of results, the preprocessing of the data also reduces the number of training samples, thus making the algorithm more efficient. Both the train and test data were preprocessed.

Next, the training data is used to perform exploratory data analysis. The data analysis performances are shown in Table 2. Through data analysis it could be achieved that how many posts each personality type has made and how many total words each personality type has written.

Table 2. The numbers of samples of different type of personality.

Type	posts	word_count
INFP	1386	1766459
INFJ	1100	1433173
INTP	960	1182968
INTJ	830	1021534
ENTP	530	657644
ENFP	496	640173
ISTP	255	309575
ISFP	198	228704
ENTJ	167	211562
ISTJ	145	181368
ENFJ	143	187741
ISFJ	124	157912
ESTP	71	85690
ESFP	36	38738
ESFJ	35	45280
ESTJ	30	37560

The table shows that “INFP” personalities that have posted the most posts and tend to write the most words. And the least number of posts is made by “ESTJ” personalities. “ESFP” personalities posted the least number of words.

Based on the training data, a hypothesis about the final results of the research could be made, which is that the distribution of personality types should be similar to the distribution obtained from the data analysis. The “INXX” personality type would occupy a huge proportion and the “ESXX” personality type would be much less.

Model building requires various classification machine learning techniques. The classification techniques applied were Logistic Regression, Multinomial Naive Bayes, Support Vector Classifier (SVC) and Random Forrest Classifier. As mentioned before, CountVectorizer and TfidfVectorizer were also used in this section. The words were vectored with TfidfVectorizer. Then multiple iterations of the parameters for each feature were performed. “E” or “I”, “N” or “S”, “T” or “F”, “J” or “P”, correspond to the four classification "Mind", "Energy", "Nature" and "Tactic". They were modeled and parameterized separately.

After this model was fitted, the testing set is leveraged to make predictions of personality type from four dimensions. Combining the results of the four variables gave the personality type corresponding to this piece of data. The results obtained are shown in the next section.

3. Result

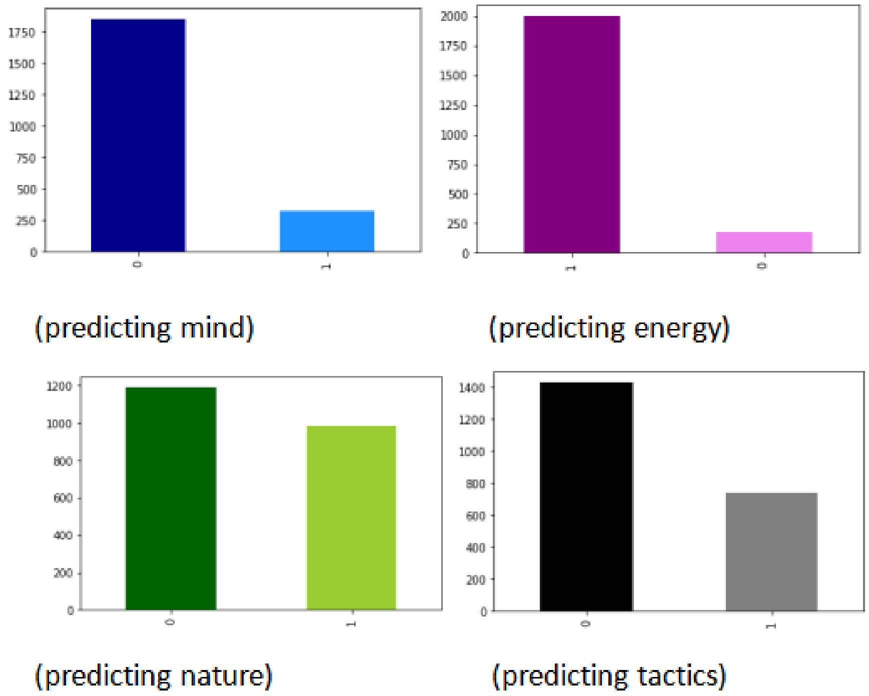
In the model building session, four algorithms are used to make predictions for each of the four dimension models. Therefore, the statistical data will be compared with the predicted data after processing to calculate the log loss between the predicted data and the actual data (the training data are considered as the actual data). Thus, the advantages and disadvantages of the four algorithms are judged.

Table 3. Performance of different algorithms for personality prediction.

Model	Mind Log Loss	Energy Log Loss	Nature Log Loss	Tactics Log Loss	Average Log Loss
Logistic Regression	4.87	3.59	4.89	6.56	4.98
SVC	4.87	4.28	5.52	6.56	5.31
Random Forest	6.39	4.46	9.30	10.65	7.70
Multinomial Naive Bayes	7.59	4.54	6.67	9.73	7.13

From the table of logarithmic losses, it seems that Logistic Regression and SVC algorithms perform better, however, Random Forest and Multinomial Naive Bayes algorithms yield results with relatively larger errors. It is worth mentioning that the best results were obtained by Logistic Regression and vectoring the words with TfidfVectorizer. While SVC with CountVectorizer gets a close second. Anyway, in the next interpretation of the results, the model of Logistic Regression with TfidfVectorizer will be used.

The prediction results for the four characteristic models are shown in Figure 2.



(predicting mind)

(predicting energy)

(predicting nature)

(predicting tactics)

Figure 2. Prediction of the four characteristic models.

The personality distributions of the training dataset, the prediction and the global personality are respectively demonstrated in Figure 3, Figure 4 and Figure 5.

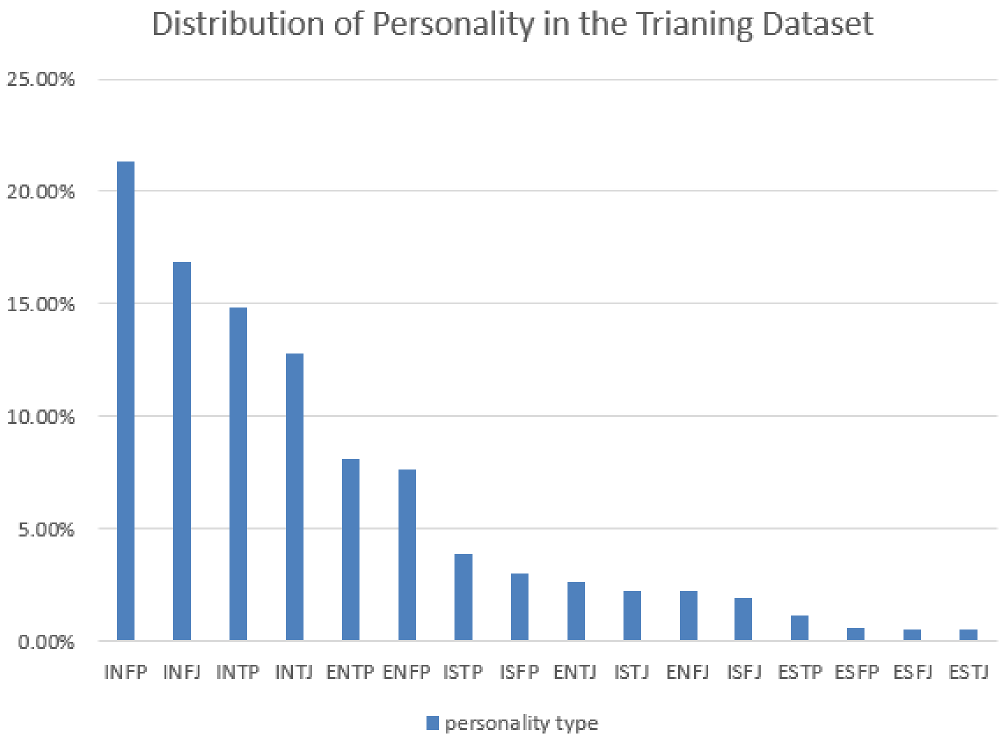


Figure 3. Distribution of personality in the training dataset.

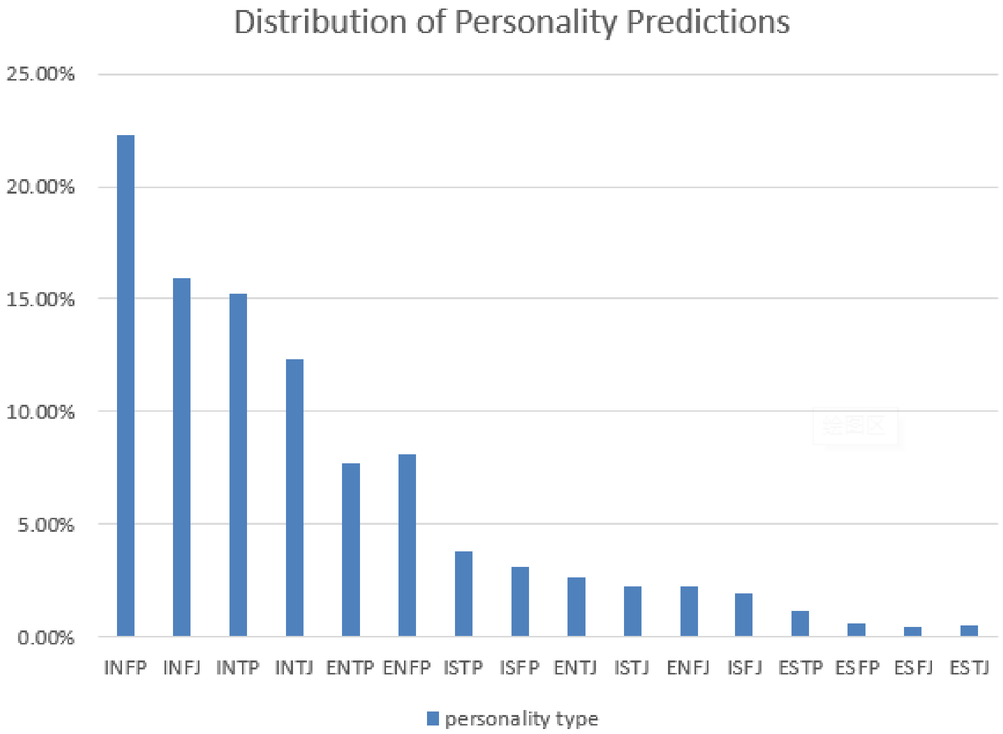


Figure 4. Distribution of personality in the prediction results.

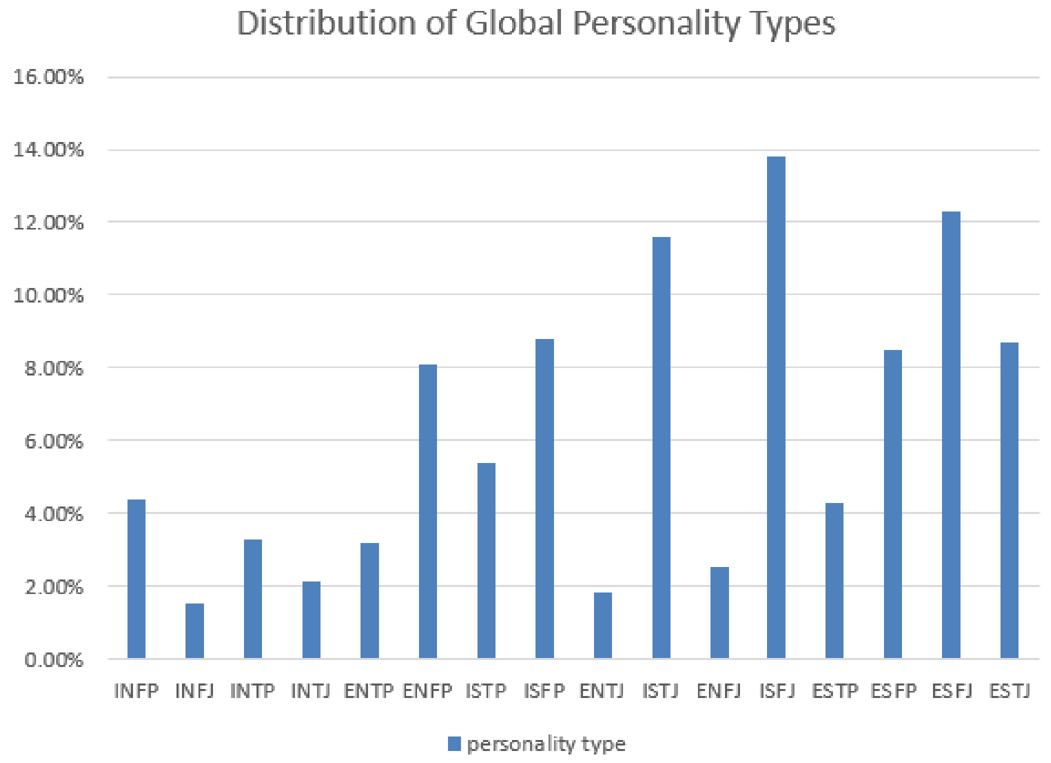


Figure 5. Distribution of the global personality types.

Via comparing the training outputs with the predicted results, the predictions are basically consistent with the results of the training set, which is in line with the expected desired situation. Thus, it seems that the logistic regression algorithm has successfully completed the prediction of personality.

However, the predicted data results differed significantly from the global personality distribution results [9]. The reason for this is easy to understand, because in the data collection of this prediction model, the personality data are taken from online forums, which are not able to reflect the real personality distribution. But the data collected matches reality to test the accuracy of the model. According to the previously predicted model, “INXX” personalities tend to post more on online platforms like forums, while “ESXX” personalities are less likely to post, so counting the number of people of different types on platforms like forums also tends to fit this pattern. However, this does not represent the distribution of the number of different personalities in reality. So, comparing the realistic model, it is reasonable to predict that the model will show a larger proportion of IN-type personalities and a smaller proportion of ES-type personalities.

In summary, although the data used to train and test the model are not representative of the worldwide personality distribution, the predicted models compared with the collected training data indicate that the prediction models have been able to perform the prediction task relatively successfully.

4. Discussion

As analyzed in the previous results, the use of logistic regression algorithms can build a predictive model with a high degree of accuracy, thus obtaining results that are relatively in line with expectations. If the algorithm of multiple classification is used, that is, based on the input data, the prediction type is directly determined as a specific one of the sixteen personalities, the accuracy of the resulting results will be greatly reduced. With binary classification, the machine only needs to judge and output 0 or 1, and combine the results obtained after 4 judgments. Because in the MBTI personality model, each personality consists of four dimensions, namely mind, energy, nature, and tactics, and these four dimensions do not affect each other, and each trait is taken from one of the two opposing types, so the use of binary classification does not negatively affect the accuracy of the prediction model. Instead, it is more efficient and accurate to simply pick the output of the one that fits better among the opposite types.

According to another study of personality prediction systems based on EEG signals, the authors concluded after comparative experiments that the DeepLSTM model achieved the best classification accuracy among the many algorithms [10]. It follows that the classification algorithm is undoubtedly the optimal algorithm in terms of prediction models for personality models, and the choice of classification algorithm varies according to the situation. And in this study, only four algorithms were compared, which is one of the places where the experiment can be improved.

5. Conclusion

During this study, postings from two groups of Cafe website forum were collected to evaluate the personality prediction model. The machine is trained using the training dataset and is used as a reference target for testing accuracy. The test dataset is then used to make personality predictions for different algorithmic models and the results are analyzed. We choose log loss as an assessment criterion for the prediction accuracy of the model, and from the results, Logistic Regression and vectoring the words with TfidfVectorizer achieved the optimal performances.

The prediction model is consistent with that of the collected data, but differs significantly from the realistic personality distribution. The reason for this is that the data were collected in online forums, and different personalities do not have the same preferences for forum posting. Presumably, IN-type personalities are more likely to post in such forums and to post more content, while ES-type personalities are the opposite. So, it is understandable that such prediction results are presented.

In this study, we only analyze personality predictions based on textual forms (data taken from forum postings). And after that, more forms of personality prediction can be carried out based on this research, such as speech into text and then personality prediction, or using neural networks to make direct personality prediction for people in image videos. As to whether logistic regression and binary classification can still achieve excellent results in such a case requires more research.

References

[1] Funder, D. C. (2012). Accurate personality judgment. Current Directions in Psychological Science, 21(3), 177-182.

[2] Tandera, T., Suhartono, D., Wongso, R., & Prasetio, Y. L. (2017). Personality prediction system from facebook users. Procedia computer science, 116, 604-611.

[3] Gjurković, M., & Šnajder, J. (2018). Reddit: A gold mine for personality prediction. In Proceedings of the second workshop on computational modeling of people's opinions, personality, and emotions in social media, 87-97.

[4] Nisha, K. A., Kulsum, U., Rahman, S., Hossain, M. F., Chakraborty, P., & Choudhury, T. (2022). A comparative analysis of machine learning approaches in personality prediction using MBTI. In Computational Intelligence in Pattern Recognition: Proceedings of CIPR 2021, 13-23.

[5] Bhatti, S. K., Muneer, A., Lali, M. I., Gull, M., & Din, S. M. U. (2017). Personality analysis of the USA public using Twitter profile pictures. In 2017 International Conference on Information and Communication Technologies, 165-172.

[6] Hinds, J., & Joinson, A. (2019). Human and computer personality prediction from digital footprints. Current Directions in Psychological Science, 28(2), 204-211.

[7] Yang, Q., Farseev, A., & Filchenkov, A. (2021). Two-Faced Humans on Twitter and Facebook: Harvesting Social Multimedia for Human Personality Profiling. In Proceedings of the 2021 Workshop on Intelligent Cross-Data Analysis and Retrieval, 39-47.

[8] Anand, V., Bochkay, K., Chychyla, R., & Leone, A. (2020). Using Python for text analysis in accounting research. Foundations and Trends® in Accounting, 14(3–4), 128-359.

[9] NERIS Analytics Limited (2023) Global data source:16Personalities. URL: https://www.16personalities.com/country-profiles/global/world#global

[10] Bhardwaj, H., Tomar, P., Sakalle, A., & Ibrahim, W. (2021). Eeg-based personality prediction using fast fourier transform and deeplstm model. Computational Intelligence and Neuroscience, 2021, 1-10.