Character Personality Prediction Based on Deep Learning Algorithm MBTI

Yuanshen Su
School of Sydney Smart Technology College, Northeastern University, Shenyang, China

Keywords: MBTI, Deep Learning Algorithms, Personality Prediction, Catboost Classifier.

Abstract: The Myers-Briggs Type Indicator (MBTI) is a popular and extensively utilized tool for determining personality type. With the help of the Internet, the MBTI personality test, packaged by mass media, has become popular again in recent years. The MBTI personality test is helpful for people to seek their own identity and explore their strengths and weaknesses. This article will explore the collection of users' online speech to predict user MBTI personality, so as to improve the platform's help to users and improve service quality. The research topic of this paper is character prediction based on the deep learning algorithm MBTI. In this paper, six models are established using deep learning algorithms. Examples include the CatBoost Classifier and the XGBoost Classifier model. Use test set data to make predictions and find the accuracy of individual model prediction results. Among them, the prediction completed by the CatBoost Classifier model is the most accurate among the six models, with an accuracy of 0.668588. However, to make the prediction results more applicable, the predictive model is more confident. Research requires more data for model training to have better performance in real-world use. This research can help online platforms provide different help to users with different personality traits. For example, recommend jobs that suit people's personality traits, or match and expand users' circle of friends.

1 INTRODUCTION

"Myers-Briggs Type Indicator" is the full name of the MBTI Personality Type Scale, a forced selection, self-reported personality evaluation instrument created on the basis of Jung's typology theory. Because of the different personality types, different individuals look at and understand things from their own thinking perspective, resulting in different views and views on things, thus showing different personalities and behaviors. Jung through clinical observation and psychological points Three dimensions of individual behavioral differences: (1) mental energy pointing: Extraversion—Introversion; (2) Information acquisition method: Sensing — Intuition; (3) Decision-making method: Thinking-Feeling. Based on Jung's theoretical system, Briggs and Myers are based on these three dimensions A new dimension has been added － the way it interacts with the outside world: Judging—Perceiving. Based on such a theoretical basis, personality measurement tools The Myers-Briggs Type Indicator (MBTI) was born (Qin 2011). The Myers-Briggs Type Indicator (MBTI) is a popular and extensively utilized tool for determining personality type. It uses four binary categories and sixteen total kinds to explain how individuals act and engage with their environment (Cui et al 2017). Using the bolded identifying letter for each category, an individual's MBTI personality type is defined as the sum of their four types for those four categories. The personality type ENTJ, for instance, would be possessed by someone who gets most of their energy from being around other people (E), believes in intuition and gut feelings to understand the world (N), considers decisions logically (T), and leads a planned life (J) as opposed to an impulsive one. The research will use this personality model throughout this article.

Although the MBTI test has been around for a long time, the reason for its sudden popularity in recent years is that the mass media has made it more in line with the psychology of the audience. That is, Internet media contact has satisfied the various needs of the audience. Language communication and interactive behavior in the Internet environment carry the "tagging" characteristics brought by the MBTI test, which cleverly meets the identity needs of

Su, Y.
Character Personality Prediction Based on Deep Learning Algorithm MBTI.
DOI: 10.5220/0012815900003885

In *Proceedings of the 1st International Conference on Data Analysis and Machine Learning (DAML 2023)*, pages 374-379
ISBN: 978-989-758-705-4

individuals and the social needs of groups. In terms of exploring its own characteristics and finding a sense of belonging to the group, the needs are satisfied. A cultural phenomenon that identifies a person's own style is MBTI culture. Audiences focus mainly on their own identities while engaging in cultural engagement; MBTI provides them with a new, more romantic, and creative identity, which is a major factor in their widespread participation in MBTI culture. They actively use MBTI culture to attempt and convey who they are as individuals (Sun 2022).

The MBTI test is becoming increasingly popular among young people. The popularity of searches on the Internet is also increasing. The MBTI test involves psychological theories. However, due to the limitations of online psychological testing, it tends to be a "fun game". Attract users through science and effectiveness, and then present entertaining analysis and solutions. Achieve successful business promotion and popularity. The spread of the Internet has facilitated the spread of the MBTI test, which can be completed only through a website link. At present, social mass media is hyping the topic of MBTI with its diversified content integration. In addition to the personality analysis after the test, various personality social matching, personality and professional matching, and other information will be established and promoted, so that a single personality test topic can be transformed into a form of group interaction. In the current situation, young people are under pressure to make choices in the face of uncertainties in society and future destiny. Similar to astrology and gossip in the past, the MBTI test has become one of their ways to seek answers and relax. In the face of career choices, social relationships, and other choices in life, the MBTI test can alleviate their inner confusion and pressure to a certain extent. Help young people to discover and identify with their own personality and characteristics. At the same time, it is also conducive to the expansion of social scope and the satisfaction of the group's sense of belonging. In the process of group interaction, "youth culture" sprang from this, including the creation and pursuit of fashion as well as the presentation of youth views (Tang 2022).

The research of this paper aims to achieve MBTI personality prediction of users by collecting their online speech on Internet platforms. After completing the advanced prediction of the user's MBTI personality, both social media platforms and job search software can better serve users according to their personality characteristics. On the other hand, in order to analyze the personality of college students who wish to start a business, it is crucial to use effective tools such as MBTI, which can also help people understand themselves better (Han and Zhan 2019). At the same time, corresponding method training can also be carried out according to personality characteristics to achieve the goal, which requires different methods for different personalities (Li 2021). Correspondingly, people can also choose people suitable for development according to the needs of the job and specific functions. For example, positive personality traits can not only scientifically guide the growth of managers' psychological capital, but also contribute significantly to the continuous enhancement and shaping of managers' core competitiveness (Lyu 2019). At present, the practical application of this aspect is still blank. Therefore, it is very useful and necessary to conduct this research.

The research topic of this paper is character prediction based on the deep learning algorithm MBTI. After completing the data collection and cleaning, the researcher split the dataset into training and testing sets. And through deep learning algorithms to establish a variety of models. Today's various deep learning algorithm models have their own limitations and imperfections. Therefore, this paper will select and build a variety of models for subsequent research. After training each model, the accuracy test is carried out, and finally, the model with the highest accuracy and reasonable accuracy is selected to predict the user's MBTI personality and achieve the research purpose.

2 METHOD

The Myers-Briggs Personality Type (MBTI) Dataset was sourced from Kaggle for this article. Founded in Melbourne in 2010, Kaggle provides a platform for developers and data scientists to host machine learning competitions, host databases, and write and share code. The researcher can get the exact data set that needs it. This data then needs to be processed. First change uppercase letters to lowercase, and then initialize the form. Replace some special text such as URL links, numbers, dates, and emojis with corresponding escape tags, and save the processed results in a new CSV file. Then extract 5,000 words with the highest frequency of occurrence. The text content of the dataset is then converted into a feature vector. And encode the classification label as a branch. Finally, the processed data is written back to the database. The specific operation is as follows:

1. Hierarchical splitting to ensure even distribution of data, and divide the dataset and training set.
2. Clean data
3. Tokenizing words

Let's view the complete dataset first. It should be noted that the sample sizes of the 16 MBTl kinds vary. INFP was the most prevalent type in the sample, making up 21.1%, while ESTJ was the least prevalent kind, making up only 0.447%. Furthermore, there is an unequal sample size for two categories within the same characteristic; for instance, there are more people in the "introvert" category ("I") than in the "extrovert" category ("E"). Due to their modest size, the MBTl datasets currently in use might not be sufficiently trained for all characteristics. This research project requires larger and more diverse datasets, including data like photos, in order to increase forecast accuracy.

But at this time, for factors such as model training accuracy and dataset integrity, this paper only considers the top 5000 feature words with the highest word frequency. Exclude common words that have no actual meaning in feature word selection. Use the Lemmatizer object as a separator to restore the word to its original form. Input the training dataset into the TfidfVectorizer object to extract and vectorize the training data. This builds a lexical inverse document frequency matrix that represents the importance of feature words in each document. First, get the list of feature words extracted by TfidfVectorizer, and then use these feature words to generate a word cloud that shows words between the 500th and 3500th of the feature word list.

First, call the transform() method to input the training dataset train_data.posts to the transform() method of the TfidfVectorizer object, and use the array() method to convert the returned sparse matrix into a dense array. This generates a two-dimensional array representing the feature vectors of each document in the training dataset and assigns them to the train_post variable.

Then, similarly, the same transformation is performed on the test data set test_data. posts, converted to a feature vector representation, and assigned to the test_post variable. In this way, train_post and test_post represent the feature vector representations of each document in the training and test datasets, respectively, and can be used for training and testing machine learning models.

Encode classification labels as numeric representations to facilitate the training and testing of machine learning models.

First, define a LabelEncoder object that encodes the target variable. Next, call the fit_transform() method to input the type column from the training dataset as a parameter to the target_encoder object. This will label the column according to the type of target variable in the training dataset, and return an encoded one-dimensional array representing the category to which each sample belongs. In this binary classification problem, the result may get an output labeled 0 or 1. Finally, similarly, encode the target variable column in the test data set and assign the encoded result to the test_target variable. In this way, train_target and test_target represent the target variable encoding results in the training dataset and the test dataset, respectively, and can be used for training and testing machine learning models. Figure 1 illustrates the percentage of each personality type in the dataset.

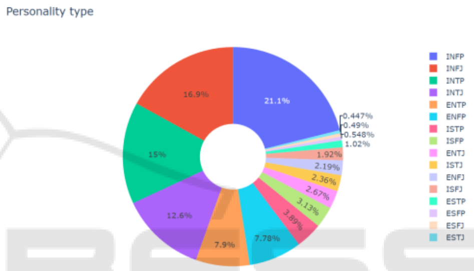

Figure 1: Personality type ratio chart (Picture credit: Original).

In terms of model selection, this research established a variety of models, experimented, and calculated their accuracy. Examples include the CatBoost Classifier and the XGBoost Classifier model.

3 RESULT

Finally, based on the results obtained after data cleaning, set the size of the plot, remove the axes, and display the resulting word cloud. The results of the study are shown in Figure 2, every post's word count approximated the normal distribution, albeit a little bit to the left. In the end, this research divided the dataset into an 8:2 ratio across the training and testing sets.

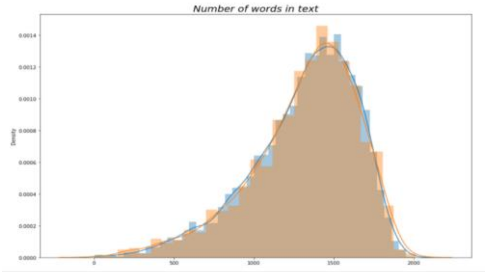

Figure 2: Word cloud diagram (Picture credit: Original).

In the two models that this paper mainly chooses, CatBoost (Categorical Boosting): CatBoost is a gradient boosting algorithm developed by Yandex and open-sourced in 2017. It has some unique features that make it particularly well-suited for datasets with categorical features (Hancock and Khoshgoftaar 2020). CatBoost Classifier is Yandex's open-source machine learning algorithm. It can be easily integrated with deep learning frameworks. CatBoost is a GBDT framework based on oblivious trees-based learning with fewer parameters, supports categorical variables and high accuracy, and the main pain point solved is efficient and reasonable processing of categorical features, which can be seen from its name, CatBoost is composed of Categorical and Boosting. In addition, CatBoost solves the problems of Gradient Bias and Prediction shift, thereby reducing the occurrence of overfitting and improving the accuracy and generalization ability of the algorithm. It automatically handles categorical features in a special way. First, do some statistics on categorical features, calculate the frequency of occurrence of a certain category feature, and then add hyperparameters to generate new numerical features. Catboost can help automate the handling of categorical features. Additionally, Catboost incorporates combinatorial class features, which dramatically enhance the feature dimension by utilizing links between features. In contrast to the conventional boosting approach, which computes the average, Catboost's basic model uses symmetric trees, and the leaf-value computation method prevents the model from overfitting by optimizing other procedures.

XGBoost Classifier: XGBoost is a scalable machine learning system developed in 2016 by Professor Tianqi Chen of the University of Washington. XGBoost is not just a model, but a set of tools that make it simple for users to resolve classification, regression, or ranking issues. Internally, a gradient boosting tree (GBDT) model is implemented, and the algorithm in the model has undergone numerous optimizations to provide great accuracy while operating at breakneck speed. At the same time, XGBoost considers both system optimization and machine learning principles. The scalability, portability, and accuracy offered by XGBoost pushed the limits of machine learning computing, running on a single machine more than ten times faster than popular solutions at the time, and even processing a billion amounts of data in distributed systems. Compared to other machine learning libraries, XGBoost has the advantages of being straightforward and simple to use, allowing users to get good results; both scalable and effective. Processing large-scale datasets requires little hardware resources, like as memory, and may be done quickly and efficiently; with strong resilience. In contrast to the deep learning model, it may get similar results without requiring precise parameter adjustments; XGBoost has an inbuilt boosted tree model to automatically manage missing data. All things considered, the tree-boosting classifier is the Extremely Gradient Boosted Decision Tree or XGBoost classifier. Using the decision tree approach, this ensemble classifier divided the targets by dividing the data into smaller pieces. This classifier helps achieve good accuracy for many NLP applications and machine learning models. In addition, it is quicker than other well-known classifiers and scalable in a variety of settings. When compared to other classifiers, the XGBoost ensemble classifier's scalability properties contribute to an improvement in accuracy (Ghosal and Jain 2023).

Table 1: CatBoost Classifier Training set.

	precision	recall	F1-score	support
ENFJ	0.87	0.54	0.67	152
ENFP	0.81	0.75	0.78	540
ENTJ	0.89	0.59	0.71	185
ENTP	0.78	0.78	0.78	548
ESFJ	0.93	0.41	0.57	34
ESFP	1.00	0.24	0.38	38
ESTJ	1.00	0.35	0.52	31
ESTP	0.93	0.56	0.70	71
INFJ	0.79	0.82	0.81	1176
INFP	0.77	0.88	0.82	1465
INTJ	0.80	0.81	0.81	873
INTP	0.75	0.86	0.80	1043
ISFJ	0.88	0.63	0.73	133
ISFP	0.80	0.64	0.71	217
ISTJ	0.87	0.66	0.75	164
ISTP	0.88	0.75	0.81	270
accuracy			0.79	6940
macro avg	0.86	0.64	0.71	6940
weighted avg	0.80	0.79	0.79	6940

After cleaning the data, this research gets the training set and the test set, and gets the training and testing process and results of the two models.

For the CatBoost Classifier, Table 1 shows Training set data.

Table 2 shows Test set data.

Table 2: CatBoost Classifier Test set.

	precision	recall	F1-score	support
ENFJ	0.64	0.37	0.47	38
ENFP	0.74	0.64	0.69	135
ENTJ	0.79	0.41	0.54	46
ENTP	0.61	0.57	0.59	137
ESFJ	1.00	0.25	0.40	8
ESFP	0.00	0.00	0.00	10
ESTJ	1.00	0.12	0.22	8
ESTP	0.73	0.44	0.55	18
INFJ	0.69	0.74	0.71	294
INFP	0.66	0.80	0.72	367
INTJ	0.67	0.63	0.65	218
INTP	0.64	0.79	0.71	261
ISFJ	0.75	0.45	0.57	33
ISFP	0.56	0.43	0.48	54
ISTJ	0.75	0.37	0.49	41
ISTP	0.68	0.64	0.66	67
accuracy			0.67	1735
macro avg	0.68	0.48	0.53	1735
weighted avg	0.67	0.67	0.66	1735

For the XGBoost Classifier, Table 3 shows Training set data.

Table 3: XGBoost Classifier Training set.

	precision	recall	F1-score	support
ENFJ	0.99	0.92	0.95	152
ENFP	0.94	0.91	0.92	540
ENTJ	0.99	0.90	0.94	185
ENTP	0.94	0.91	0.92	548
ESFJ	1.00	0.91	0.95	34
ESFP	1.00	0.92	0.96	38
ESTJ	1.00	0.84	0.91	31
ESTP	1.00	0.94	0.97	71
INFJ	0.91	0.90	0.91	1176
INFP	0.89	0.95	0.92	1465
INTJ	0.92	0.92	0.92	873
INTP	0.90	0.93	0.91	1043
ISFJ	1.00	0.96	0.98	133
ISFP	0.99	0.92	0.95	217
ISTJ	0.99	0.92	0.95	164
ISTP	0.97	0.96	0.96	270
accuracy			0.92	6940
macro avg	0.96	0.92	0.94	6940
weighted avg	0.92	0.92	0.92	6940

Table 4 shows Test set data.

Table 4: XGBoost Classifier Test set

	precision	recall	F1-score	support
ENFJ	0.62	0.34	0.44	38
ENFP	0.69	0.61	0.65	135
ENTJ	0.75	0.39	0.51	46
ENTP	0.59	0.57	0.58	137
ESFJ	1.00	0.12	0.22	8
ESFP	1.00	0.10	0.18	10
ESTJ	1.00	0.12	0.22	8
ESTP	0.38	0.17	0.23	18
INFJ	0.67	0.74	0.71	294
INFP	0.67	0.81	0.73	367
INTJ	0.68	0.66	0.67	218
INTP	0.67	0.77	0.72	261
ISFJ	0.68	0.45	0.55	33
ISFP	0.69	0.46	0.56	54
ISTJ	0.65	0.37	0.47	41
ISTP	0.66	0.63	0.64	67
accuracy			0.67	1735
macro avg	0.91	0.46	0.50	1735
weighted avg	0.67	0.67	0.66	1735

For the same operation, this research trains six models, which are the CatBoost Classifier, XGBoost Classifier, Linear Support Vector Classifier, Support Vector classifier, logistic regression, and Decision Tree Classifier.

Then, the test set data is substituted into the model for prediction, and the prediction results are compared with the MBTI personality distribution of the original data to calculate the prediction accuracy of different models. Finally, the prediction accuracy of each model is shown in Table 5.

Table 5: Prediction accuracy.

Models	Test accuracy
CatBoost Classifier	0.668588
XGBoost Classifier	0.666282
Linear Support Vector classifier	0.661671
Support Vector classifier	0.648991
Logistic regression	0.628242
Decision Tree classifier	0.509510

According to the prediction accuracy data obtained in Figure 3, it can be seen directly, and the CatBoost Classifier model should be selected for MBTI personality prediction. This is because the CatBoost Classifier model has the highest predictive accuracy compared to other models.

4 DISCUSSION

In the process of processing data, although the larger the amount of data in theory, the more widely applicable and reasonable the prediction results can be reflected. However, due to various limitations, this research had to reduce the number of data sets processed. For example, it is too difficult to process a large number of data sets; Some machine learning algorithm models will greatly reduce their prediction accuracy when faced with massive data sets. For example, one of the disadvantages of the XGBoost Classifier is that when a research project has massive training data and can find a suitable deep learning model, the accuracy of deep learning can be far ahead of XGBoost. Therefore, this research must fix a certain number of data sets for research and prediction. Deep learning algorithms and model selection are most important according to the characteristics of the dataset and the target. Some models are difficult and error-prone to train, and the final predictions are not accurate. However, models like the CatBoost Classifier are built smoothly and the results are highly accurate. In summary, the CatBoost Classifier model has the highest accuracy among the six model predictions, reaching approximately 0.668588. Finally, the CatBoost Classifier model should be used to predict the user's MBTI personality. However, due to the limited data set, and even the scarcity of experimental data on some platforms, it is not enough to support the model training of traditional deep learning methods (Liu et al 2021). Therefore, the predictive model obtained in this study has certain limitations. Within a reasonable error range, the prediction results of the CatBoost Classifier model have certain reference values and effects.

5 CONCLUSION

In this paper, six models are established to predict the MBTI personality of users based on online speeches through deep learning algorithms. In this study, after cleaning and dividing the dataset into two parts, six models were trained with the training set data. After the model is built, the test set data is put into all models for prediction. Based on the prediction results, the prediction accuracy calculated shows that the CatBoost Classifier is the most accurate prediction, reaching 0.668588. Therefore, the CatBoost Classifier is more suitable for this prediction problem in terms of accuracy. Because the research method itself will bring some errors, such as the instability of the MBTI test results themselves. The prediction accuracy obtained in this study can be preliminarily used for MBTI personality prediction under error conditions. MBTI personality prediction can help online platforms better serve users. It can also help people recognize their own personality status, which is of great help to young people looking for jobs or making friends. However, due to the limitation of the size of the experimental data. The usefulness of the model is not entirely satisfactory, and future studies must use larger datasets for training and prediction to enhance the broad fit of the model. A wider variety of data could be collected for future studies. MBTI prediction itself involves a variety of information such as text, pictures, and sounds. Future research could try to fuse these different kinds of data, not just online reviews. This can improve the effectiveness of MBTI personality prediction.

REFERENCES

J Qin. (2011). Overview of the MBTI Personality Type Scale. Intellect (06), 298-299.

B. Cui..,DfaBDHF & C. Qi, (2017). Survey analysis of machine learning methods for natural language processing for MBTI Personality Type Prediction. Google Scholar Google Scholar Reference, 1.

X. Sun. (2022). Identity Confessions and Style Markers in MBTI Culture. Creative Review Tan (06), 50-53.

X. Tang. (2022). Analysis of the popular phenomenon of MBTI test in the era of social media: A case study of user video of station B. New Media Studies(17),83-87.

Z. Han & L. Zhan. (2019). Research on Innovation and Entrepreneurship Guidance for College Students in the New Era Based on MBTI Theory. Rural Economy and Science and Technology(14),296+298.

S. Li. (2021). Methods to reduce fraud among accounting practitioners through MBTI personality type scale analysis. Jiangsu Business Review(02),101-103.

Y. Lyu. (2019). Research on the influence mechanism of MBTI personality and psychological capital on the growth of managers of high-tech enterprises. Journal of Hunan Industrial Vocational and Technical College (06),38-41+63.

J. T. Hancock, & T. M. Khoshgoftaar. (2020). CatBoost for big data: an interdisciplinary review. Journal of Big Data, 7(1).

S. Ghosal & A. Jain..(2023).Depression and Suicide Risk Detection on Social Media using fastText Embedding and XGBoost Classifier. Procedia Computer Science.

W. Liu, X. Li, Z. Zhao, Q. Guo, X. Tang & W. Zhou. (2021). Small sample classification evaluation method based on improved relationship network. Experimental Technology and Management(11),194-199.