Predicting Myers-Briggs Personality From Posts in Social Media

Shaochang Tan

October 25, 2023

Abstract

Myers-Briggs personality type is a well-established classification system that assigns people to one of 16 distinct personality types, each characterized by preferences in four key dimensions (extraversion vs. introversion, sensing vs. intuition, thinking vs. feeling, and judging vs. perceiving). In this study we tried to predict an individual's Myers-Briggs personality type based on their social media posts. We used a dataset of over 8,000 individuals, their respective MBTI personality types, and the textual content they have authored. We used several machine learning models to predict the four personality dimensions, and evaluate the performance of these models using standard evaluation metrics for binary classification tasks. By comparing the performance of these models, we gained insights into their strengths and weaknesses, and identified the best model for predicting MBTI personality types. The code for this project can be found at https://github.com/petertheprocess/MBTI-Predict-NLP

1 Introduction

Myers-Briggs personality type is a well-established classification system that assigns people to one of 16 distinct personality types, each characterized by preferences in four key dimensions (extraversion vs. introversion, sensing vs. intuition, thinking vs. feeling, and judging vs. perceiving). all based on their social media posts. I plan to work on predicting an individual's Myers-Briggs personality type based on their social media posts.

2 Dataset

In the pursuit of predicting Myers-Briggs Personality Types (MBTI) from social media posts, it is imperative to prepare the dataset to ensure its suitability for analysis. The chosen dataset, namely the 'MBTI Myers-Briggs Personality Type' dataset[1], was collected from the Personality Cafe forum. It provides a substantial collection of individuals, their respective MBTI personality types, and the textual content they have authored. The dataset comprises over 8675 rows of data, where each row includes the following elements:

(1) Type: This refers to the individual's four-letter MBTI code or type.

(2) Posts: This section contains the text of the last 50 things the individual has posted. Each entry within this section is separated by the "|||" delimiter, which consists of three pipe characters.

Type	Posts
ENTP	'I'm finding the lack of me in these posts ver...'
INTJ	'Dear INTP, I enjoyed our conversation the o...'
ENTJ	'You're fired.———That's another silly misconce...'

Table 1: dataset preview

2.1 Distribution

This dataset comprises 8675 data points, as visually demonstrated in Figure 1. It is evident from this representation that the dataset is afflicted by a substantial class imbalance, a common challenge in the realm of machine learning, particularly in the context of classification tasks. This imbalance poses a significant hurdle to the effectiveness of our classification models.

Breaking the classes down and analyze the type in 4 different dimensions, It becomes evident that two primary dimensions, namely I/E (Introversion vs. Extraversion) and N/S (Intuition vs. Sensing), bear the brunt of this imbalance. In the case of the I/E dimension, individuals classified as introverted significantly outnumber their extraverted counterparts by more than a threefold margin. The N/S dimension faces an even more pronounced imbalance, with the ratio of intuition-oriented individuals to sensing-oriented individuals exceeding seven to one. Conversely, the remaining two personality dimensions do not exhibit such stark imbalances, illustrating the heterogeneity of the dataset.

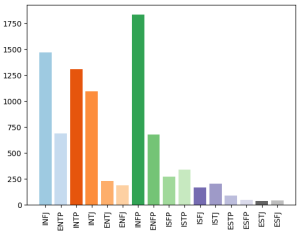

Figure 1: distribution for all 16 MBTI types

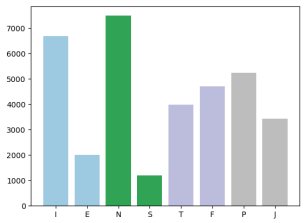

Figure 2: distribution for each MBTI dimensions

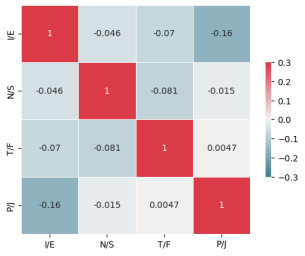

Figure 3: Correlation coefficients of 4 personality dimensions

2.2 Correlation

In this section, we conducted an in-depth examination of the interrelationships between four fundamental personality dimensions. Our objective was to understand how these dimensions interact and potentially influence with each other. Upon analyzing the dataset, we observed that the absolute value of correlation coefficients for the majority of these dimensions were relatively low, all falling below the threshold of 0.1. This suggests that, there is limited association or connection between the studied personality traits. However, the I/E and P/J dimensions displayed a more noteworthy correlation with a coefficient of -0.14. This higher coefficient indicates a relatively stronger relationship between these two aspects of personality. In practical terms, this could mean that a person's tendency toward Introversion or Extraversion is somehow related to their preference for Judging or Perceiving when making decisions or approaching tasks.

Given these findings, we can reasonably assume that, at least within the context of our dataset and analysis, the various personality dimensions we investigated are largely independent and do not exhibit strong connections or associations with one another. These findings provide strong theoretical support and assurance for our approach to categorizing and analyzing the four personality dimensions separately. The remarkably low correlation coefficients observed in our analysis suggest that each dimension operates independently within the context of our research, strengthening our confidence in treating them as distinct categories.

3 Preprocessing

The dataset was preprocessed by removing URLs, user mentions, and non-alphabetic characters, and converting all text to lowercase. Stop words were removed using the NLTK library. In addition, we used the gensim[2] library to train a word2vec model on the text data, which was used to generate word embeddings for each word in the text. We also extracted other features such as post length, average word length and average emotion using sentiment analysis tool from TextBlob[4]. Finally, the dataset was split into training and test sets, and the minority classes were oversampled by SMOTE algorithms[5].

4 Model training

For our MBTI personality prediction task, we trained several machine learning models to predict the four personality dimensions. Specifically, we trained the following models: **Logistic Regression**: We chose logistic regression because it is a simple and interpretable model that is well-suited for binary classification tasks like ours.*Support Vector Machine* (SVM): We chose SVM because it is a powerful model that can handle both linear and nonlinear classification tasks, and can be tuned using different kernel functions. **Naive Bayes**: We chose Naive Bayes because it is a simple and efficient model that can handle high-dimensional data and missing values. **XGBoost**: We chose XGBoost because it is a state-of-the-art ensemble learning model that can handle high-dimensional and nonlinear data, and can be tuned using different hyperparameters.

We implemented these models using scikit-learn[6], a popular machine learning library in Python. To evaluate the performance of these models, we used k-fold cross-validation with k=5, and measured the F1 score as our primary evaluation metric due to the imbalanced nature of the dataset. By comparing the performance of these models, we can gain insights into their strengths and weaknesses, and identify the best model for predicting MBTI personality types.

5 Evaluation

5.1 Evaluation metrics

To evaluate the performance of our MBTI personality prediction model, we used several standard evaluation metrics for binary classification tasks. Specifically, we used the following metrics: **Accuracy:** The proportion of correctly classified instances out of the total number of instances. **Precision:** The proportion of true positive predictions out of all positive predictions. **Recall:** The proportion of true positive predictions out of all actual positive instances. **F1 score:** The harmonic mean of precision and recall, which provides a balanced measure of both metrics. Since our dataset is imbalanced, we used the weighted score for each metric, which is calculated as follows:

$$\text{Weighted score S} = \frac{\sum_{i=1}^{n} w_i \cdot S_i}{\sum_{i=1}^{n} w_i}$$

where n is the number of classes, S_i is the score (can be accuracy, precision, recall and F1 score) for class i, and w_i is the weight for class i, which is proportional to the number of instances in class i in the test set. The weighted score provides a more accurate measure of the model's performance on the test set, taking into account the class imbalance. we focus on the F1 score as our primary evaluation metric. The F1 score provides a balanced measure of precision and recall, which is important in imbalanced datasets where

Best Model		Precision	Recall	F1 score	Support
XGBoost (lambda=10)	I	0.82	0.81	0.82	1350
	E	0.37	0.39	0.38	385
	W-AVG	0.72	0.72	0.72	1735
XGBoost (lambda=10)	N	0.86	0.88	0.87	1477
	S	0.23	0.20	0.21	258
	W-AVG	0.78	0.78	0.78	1735
SVM (C=500)	T	0.71	0.74	0.72	804
	F	0.77	0.74	0.75	931
	W-AVG	0.74	0.74	0.74	1735
XGBoost (lambda=10)	P	0.67	0.68	0.67	1058
	J	0.48	0.47	0.48	677
	W-AVG	0.60	0.60	0.60	1735

Table 2: Model performance

the number of positive instances is much smaller than the number of negative instances.

5.2 Results

In this experiment, we trained and evaluated four machine learning models for predicting the four personality dimensions of the Myers-Briggs Type Indicator (MBTI). Specifically, we trained logistic regression, support vector machine (SVM), naive Bayes, and XGBoost models using scikit-learn, a popular machine learning library in Python.

To find the best hyperparameters and best model, we used GridSearchCV with 5-fold cross-validation and F1 score as the evaluation metric. We searched the parameter space for each model by the means of PipelineHelper[7] and selected the model with hyperparameters that resulted in the highest F1 score.

Figure 4 shows the grid search results on each personality dimension. We can see that the XGBoost model with L2-penalty factor 500 outperformed the other models on the I/E, N/S and P/J dimensions, while the SVM model outperformed the other models on the T/F dimensions. These results suggest that machine learning models can be effective for predicting MBTI personality types, and that XGBoost is a promising model for this task.

Table 2 shows performance of best models for each dimension on the test set. We can see that our models have some degree of generalization ability, but the score of minority classes are consistently low, indicating that the issue of class imbalance in the dataset has not been fully resolved.

6 Future work

Despite using SMOTE to address the issue of class imbalance, the problem still persists, with low precision and recall for the minority classes. Future work could involve collecting more data for the minority classes to address the continued issue of class imbalance in the

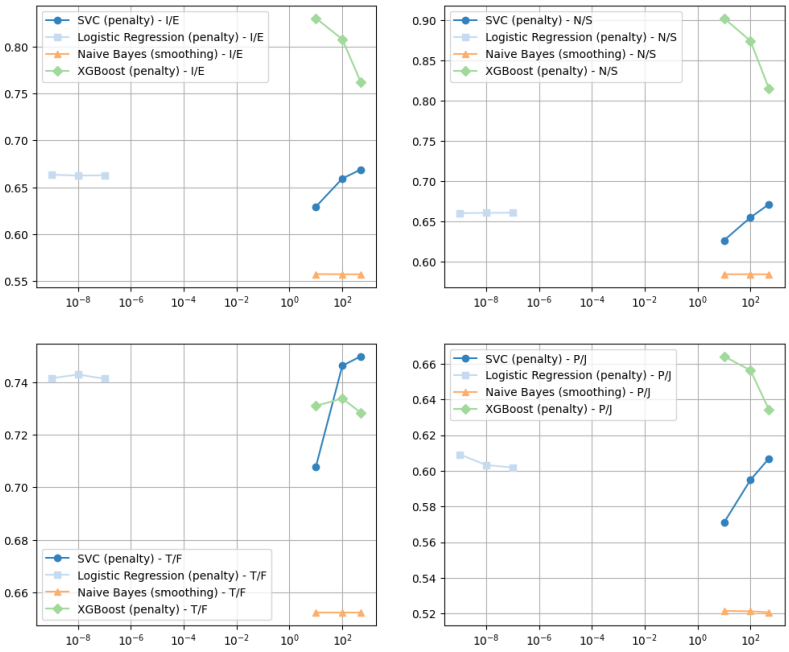

Figure 4: Grid Search Results

dataset. Besides, pre-trained models such as BERT and GPT could be used to improve the performance of the model.

References

[1] J, Mitchell. "(MBTI) Myers-Briggs Personality Type Dataset." Kaggle, 22 Sept. 2017, www.kaggle.com/datasets/datasnaek/mbti-type.

[2] "Gensim: Topic Modelling for Humans." Gensim, radimrehurek.com/gensim/. Accessed 25 Oct. 2023.

[3] Mikolov, Tomas, et al. "Efficient Estimation of Word Representations in Vector Space." arXiv.Org, 7 Sept. 2013, arxiv.org/abs/1301.3781.

[4] "Simplified Text Processing." TextBlob, textblob.readthedocs.io/en/dev/index.html. Accessed 29 Oct. 2023.

[5] Chawla, N. V., et al. "Smote: Synthetic minority over-sampling technique." Journal of Artificial Intelligence Research, vol. 16, 2002, pp. 321–357, https://doi.org/10.1613/jair.953.

[6] Pedregosa, Fabian, et al. "Scikit-Learn: Machine Learning in Python." Journal of Machine Learning Research, vol. 12, 2011, pp. 2825–2830, https://doi.org/10.5555/1953048.2078195.

[7] Bmurauer. "Bmurauer/Pipelinehelper: Scikit-Helper to Hot-Swap Pipeline Elements." GitHub, github.com/bmurauer/pipelinehelper. Accessed 25 Oct. 2023.