An Analysis of Different Machine Learning Algorithms for Personality Type Predictions Based on Social Media Posts

Ruiwen Yu
University High School, Irvine, California, 92603, U.S.A.

Keywords: Personality Type Predictions, MBTI, Machine Learning.

Abstract: In the digital era, people have increasingly used social media as a platform to communicate with each other and express feelings. Researchers have vast data to predict users' personalities, who play various practical roles in real life. This paper focuses on MBTI personality trait prediction based on users' social media posts. Comparisons and Analyses are made between the three most common methods of personality prediction, including Naive Bayes (NB), Support Vector Machine (SVM), and Extreme Gradient Boosting (XGBoost), to find out the most effective one. Results show no evident correlation between the accuracy of the models and the training method; each model's capability to predict varies. Models trained on XGBoost have the highest accuracy on average, although all three outcomes are incredibly close. One model trained on XGBoost with well-tuned hyperparameters obtained the best performance. Further analyses show other factors. For example, data distribution also influences the model's performance in some ways.

1 INTRODUCTION

Personality is the combination of characteristics or qualities that form an individual's distinctive character. It originated from the Latin word persona, which means mask or a character played by the actor. Right now, a person’s personality represents the combination of characteristics or qualities that form an individual's distinctive character. It can be the way they perceive the world, the way they make decisions, their values, beliefs, and goals of life.

Researchers increasingly gain interest in predicting users’ personality types. Personality recognition is very helpful and important in the age when social media has become a major place for people to communicate and share their thoughts. It allows the recommendation system to work more effectively, recommending the appropriate content to targeted groups of people who find that interest them, therefore improving the user experience (G. Ryan et al., 2023). In addition, it enables both sides to know more about each other’s thoughts and preferences in some scenarios that require communication. This mutual understanding is helpful to avoid misunderstandings and make the communication more productive, therefore achieving the goals with minimum efforts. Moreover, knowing one’s personality type can help users learn more about themselves online, and increase self-awareness. Most importantly, further machine learning research on personality prediction has the potential to help advance analytic psychology and personality science (W. Bleidorn et al., 2018).

This research compares and analyzes the performance of different machine-learning methods. Comparisons are made on multiple angles. The goal is to find the most effective and accurate algorithm for personality prediction.

2 RELATED WORKS

There are three main aspects of machine learning in personality prediction (C. Stachl et al., 2020). The first one is customized recommendation systems in social media, matching content to people. It makes sure the contents recommended are tailored to each individual’s preferences. The second would be personal evaluations that answer methodological questions, including predictions of life outcomes, task performance, etc. The third one predicts the actual personality traits of people. It can be based on various means, like facial features, handwriting, signatures, and more (K. Ilmini and T. G. I. Fernando, 2016)-(K.

Yu, R.
An Analysis of Different Machine Learning Algorithms for Personality Type Predictions Based on Social Media Posts.
DOI: 10.5220/0012816400003885

In *Proceedings of the 1st International Conference on Data Analysis and Machine Learning (DAML 2023)*, pages 386-391
ISBN: 978-989-758-705-4

Ilmini and T. G. I. Fernando, 2020). However, researchers commonly use digital footprints like posts on social media to train models that predict user traits.

There are two popular personality identifiers: Big Five and MBTI. Multiple researches have been conducted on both of them. Big Five refers to Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. This study will mainly focus on MBTI, one of the most well-known personality type classification systems. It separates one's personality into four dimensions: Extraversion (E) or Introversion (I), Sensing (S) or Intuition (N), Thinking (T) or Feeling (F), and Judging (J) or Perceiving (P). The MBTI test classifies individuals by their scores on the four dimensions. In general, extraversion means focusing their attention.

3 CASE ANALYSIS

Table 1 shows different methods and techniques on personality predictions conducted by other researchers and the accuracy of the best-performing model if there are multiple methods used.

Table 1: Different methods and techniques for personality predictions (B. Cui and C. Qi, 2017)-(Z., S. Ashraf, and N. Sabahat, 2020).

Title	**Method(s)**	**Best Performing Result**
(B. Cui et al., 2017)	Softmax, Naive Bayes, Regularized Support Vector Machine, Deep Learning	Deep Learning, 38%
(K. A., U. Kulsum et al., 2021)	Naive Bayes, Support Vector Machine, Extreme Gradient Boosting	Extreme Gradient Boosting, 52%
(T. M, 2021)	Stochastic Gradient Descent, Logistic Regression, Random Forest, K - Nearest Neighbor, Extreme Gradient Boosting, Support Vector Machine	Support Vector Machine, 32%
(S. Ontoum and J. H. Chan, 2022)	Naive Bayes, Support Vector Machine, Recurrent neural network	Recurrent neural network, 50%
(M. H. Amirhosseini and H. B. Kazemian, 2020)	Extreme Gradient Boosting	33%
(Z., S. Ashraf et al., 2020)	K-means clustering + Extreme Gradient Boosting	54%

Most of these methods were implemented on the split dataset based on four categories (I/E, N/S, T/F, J/P). Some researchers used the average of four categories instead of products in their paper. Table 1 shows the multiplication of each category as the performing result for better comparisons. Some commonly used methods are Naive Bayes (NB), Support Vector Machine (SVM), and Extreme Gradient Boosting (XGBoost).

3.1 Data Collection

This publicly accessible dataset posted in Kaggle has 8675 rows and 2 columns (MBTI, 2017). The first column is the personality types. The second column contains each user's posts from personalitycafe.com, and each person has up to 50 posts. However, through visualization in Fig. 1., the data is skewed. INFP, INFJ, INTP, and INTJ are the top four personality types regarding the frequency, with INFP having 3832 out of 8675 rows, while ESTJ, ESFJ, ESFP, and ESTP are the personality types that have the lowest frequency, with ESTJ merely having 39 rows (MBTI, 2017).

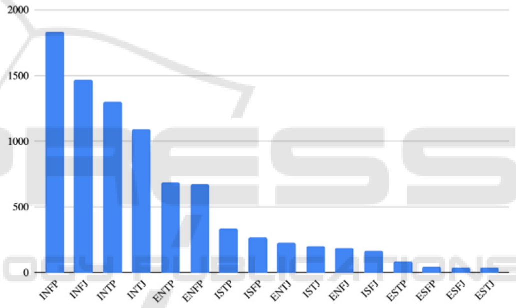

Figure 1: The frequency of each individual personality type.

3.2 Data Preprocessing

Most researchers cleaned and reformatted the dataset before putting it in training. Data preprocessing can improve the efficiency of the training process, and most likely improve the performance as well. Some common steps for data cleaning and preprocessing are listed below (shown in figure 2) (MBTI, 2017).

- Remove all URLs.
- Convert uppercase to lowercase.
- Remove punctuation marks.
- Remove numbers and dates.
- Remove emojis and special characters.
- Remove non-English characters.
- Remove stop words like the, to, and, of, etc.
- Tokenization (split text into words).
- Stemming (reduce words to their base form).
- Split the Data into 4 categories.
- Split the Data into Training and Testing Sets.

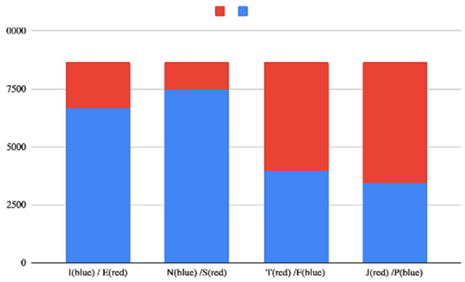

Figure 2: The marginal distribution of each category in the dataset.

3.3 Models

The most commonly chosen ones are naive Bayes, Support Vector Machine, and Extreme Gradient Boosting.

3.3.1 Naive Bayes

Naive Bayes (NB) is a simple machine learning algorithm for classification and text analysis tasks. It's based on Bayes' fundamental probability theorem that calculates conditional probabilities. Naive Bayes is "naive" because it assumes that all words used in the classification are probabilistically independent. The research uses the Naive Bayes method. The accuracy of their outcomes is 26%, 37%, and 41%, respectively, and the average is 35% approximately (B. Cui and C. Qi, 2017)-(S. Ontoum and J. H. Chan, 2022). The research model is overfitted because their training accuracy is 32% and the testing accuracy is 26% (B. Cui et al., 2017). The research uses the default parameters for the model and obtained 37% accuracy surprisingly (K. A., U. Kulsum et al., 2021). There is not much information provided in research that achieved 41% accuracy (S. Ontoum and J. H. Chan, 2022). The reasons for such a big difference in the accuracy of each model remain unclear. It could be the inherent randomness in the training process. Looking closely at each sub-category shown in Table 2, N/S appeared to be the most accurately predicted one (85% accurate on average). Three articles maintain relatively consistent performance on the prediction for I/E and falls behind on the prediction for T/F and J/P (B. Cui et al., 2017).

Table 2: The resulting accuracy for each category from models trained on Naive Bayes (B. Cui and C. Qi, 2017)-(S. Ontoum and J. H. Chan, 2022).

	I/E	N/S	T/F	J/P	Overall
(B. Cui et al., 2017)	0.750	0.845	0.624	0.603	0.2586
(K. A., U. Kulsum et al., 2021)	0.76	0.85	0.80	0.73	0.37
(S. Ontoum and J. H. Chan, 2022)	0.7828	0.8695	0.8063	0.7478	0.4104

3.3.2 Support Vector Machine

Support Vector Machine (SVM) is a supervised machine learning algorithm for classification and regression tasks. SVMs are particularly good at binary classification problems, which is what personality prediction is. Finding a hyperplane that best divides the data points of different categories while maximizing the margin between them is the basic goal of SVM. The distance between the hyperplane and the nearest data points from each class is referred to as the margin in SVM. The nearest data points are the support vectors. The hyperplane that maximizes this margin is considered the optimal decision boundary. Nisha, Amirhosseini, and Mushtaq M research use SVM, and their test accuracy is 33%, 39%, 32%, and 42%, as shown in Table 3 (B. Cui and C. Qi, 2017)-(S. Ontoum and J. H. Chan, 2022). The average is about 36%, similar to the NB method. There is also insufficient information shown in each research study to determine the factors that cause the difference. The pattern for each sub-group continues: N/S is the most well-predicted group with high accuracy between 86%-87%, and the I/E group falls in second place with 81% accuracy. It is more apparent in this chart that T/F has a higher accuracy than the J/P group.

Table 3: The Resulting Accuracy for Each Category from Models Trained on Support Vector Machine (B. Cui and C. Qi, 2017)-(S. Ontoum and J. H. Chan, 2022).

	I/E	N/S	T/F	J/P	Overall
(B. Cui et al., 2017)	-	-	-	-	0.33
(K. A., U. Kulsum et al., 2021)	0.80	0.86	0.80	0.72	0.39
(T. M, 2021)	0.7756	0.8621	0.7303	0.6550	0.3198
(S. Ontoum and J. H. Chan, 2022)	0.8515	0.8732	0.8049	0.7270	0.4197

3.3.3 Extreme Gradient Boosting

Treme Gradient Boosting (XGBoost) is an upgraded version of Gradient Boosting that is more effective and performs better. Gradient Boosting is a technique for ensemble learning that integrates the predictions from multiple weak learners to produce a strong

Table 4: The Resulting Accuracy for Each Category From Models Trained on Extreme Gradient Boosting in (K. A., U. Kulsum et al., 2021)-(M. H. Amirhosseini and H. B. Kazemian, 2020).

	I/E	N/S	T/F	J/P	Overall
(K. A., U. Kulsum et al., 2021)	0.86	0.90	0.84	0.80	0.52
(T. M, 2021)	0.7553	0.8594	0.6715	0.6215	0.2709
(M. H. Amirhosseini and H. B. Kazemian, 2020)	0.7901	0.8596	0.7419	0.6542	0.3296

model. It works by sequentially training these weak learners to fix faults produced by the earlier ones. A distinct decision tree that calculates the residual values from the previous ones is built after each iteration. These individual models are then combined to make the final prediction. Techniques like regularization and early stopping can be used to prevent overfitting. XGBoost incorporates several optimizations and features, such as regularization terms, parallel processing, and handling of missing data, to make the training process faster and the resulting models more accurate. It can handle complex tasks effectively with high stability. As demonstrated in Table 4, the accuracy of outcomes is 52%, 27%, and 33%, respectively. The average accuracy is 37%, similar to the other two methods. The accuracy trend for each subcategory remains the same.

This study looks closely at Mushtaq's research with a higher accuracy of 54% (Z., S. Ashraf et al., 2020). It also has a relatively more complex data analysis process than other models. After data cleaning, the authors created a Term Frequency-Inverse Document Frequency (TF-IDF) Matrix (often used in text classification and information retrieval) that measures how frequently a term appears in a document and the rarity or uniqueness of a term across a collection of documents. Principal component analysis (PCA) is applied to reduce the dimension and visualize the dataset. PCA transforms a dataset with many potentially correlated variables into a new dataset with a smaller number of variables called principal components. PCA aims to retain as much of the original data's variance as possible in principal components while reducing the dimensionality of the dataset. Then, K-means clustering is implemented. K-means clustering is an unsupervised machine learning algorithm for dividing a dataset into non-overlapping groups or clusters. K-means clustering aims to group similar data points together while minimizing the variation within each cluster. Next, the authors used default parameters from the XGBoost classifier to train the model, and the initial results were between 65-75% accuracy on average. Adjusting hyperparameters was the critical step that raised their accuracy to 85–90%, making a vast improvement.

3.4 Discussion

By comparing each method, no single method that has an overall better outcome stands out. The overall accuracy of each method analyzed ranges from 35%-37%. This might not seem very well, but it is a lot better than random guesses that will have 6.25% accuracy. Not to mention that a few well-trained models have above 50% accurate overall outcomes. Regarding subgroups, N/S appeared to be the most accurately predicted one, and I/E, T/F, and J/P followed in descending order. In the best model from Nisha's research, the accuracy for the N/S group gets as high as 90%. And the worst sub-group outcome at J/P has 60% accuracy (K. A., U. Kulsum et al., 2021). These models are capable of making fair personality predictions. As Mushtaq showed in an actual vs. predicted values chart, although the results are not always exactly the personality type, they appeared promising (Z., S. Ashraf et al., 2020). When analyzing, Nisha's and Ontoum's research shows a consistently higher accuracy and high values for best-performing results in Table 1 (K. A., U. Kulsum et al., 2021), (S. Ontoum and J. H. Chan, 2022). However, there is nothing explicitly different shown in the paper that could be a reasonable explanation. One noticeable variation is in the data preprocessing, despite all researchers using the same dataset. In the articles, Cui, Nisha, Amirhosseini, and Mushtaq mentioned hyperparameter tuning, while M and Ontoum did not (B. Cui and C. Qi, 2017)-(Z., S. Ashraf, and N. Sabahat, 2020). Differences in the hyperparameter could be another factor. Mushtaq's research demonstrated that well-tuned hyperparameters can yield much better results (Z., S. Ashraf et al., 2020).

3.5 Limitations

However, it is still worth noting that there are many limitations. The size and quality of the dataset play an important role in the effectiveness and accuracy of models.

3.5.1 Lack of Resources

For example, various researches have been conducted on personality prediction and machine learning based on the MBTI scheme. They all use the Kaggle dataset, which reveals the problem of a lack of available data and resources. Many researchers use the Big Five personality traits instead of the MBTI because more datasets are available. As Halevy's research discussed, large amounts of data can sometimes compensate for the shortcomings of simple models, and inadequate amounts of data can negatively influence the model (A. Halevy et al., 2009). However, limiting to a specific dataset is also good. Knowing that one particular dataset is used in all papers helped eliminate potential variation in the result that might be due to the differentiation in the dataset. It is more transparent and straightforward to compare each model's results from different literature.

3.5.2 Skewed Dataset

This dataset itself is also not perfect due to the nature of each type of personality, introverted and intuitive people tend to spend more time on the internet, which means that they tend to post more texts and data. This tendency is clearly shown in Figure 3 and Figure 4 (MBTI, 2017).

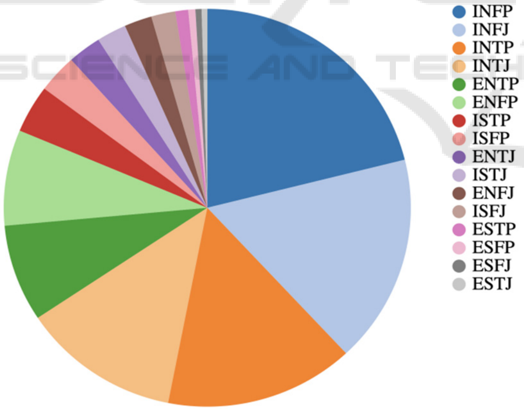

Figure 3: A pie chart showing each personality's distribution in the dataset (MBTI, 2017).

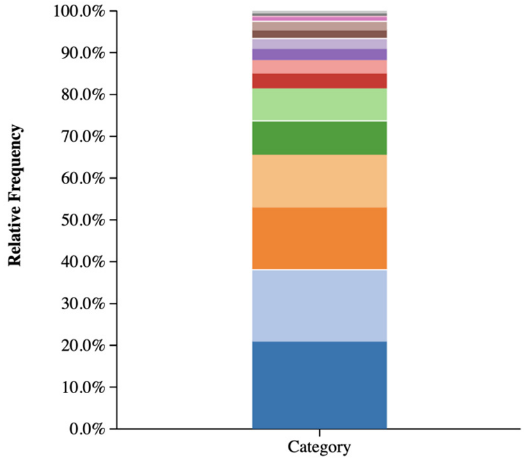

Figure 4: A segmented bar graph showing each personality's distribution in the dataset (MBTI, 2017).

The highly unbalanced amount of data might also be a source of error. Mushtaq's research talked about the unbalanced data problem (Z., S. Ashraf et al., 2020). Their model fails to predict the I/E trait, and they believe their results would have been different if they had a better data set. Another research, on the contrary, used the synthetic minority oversampling technique (SMOTE), attempting to solve the problem by generating texts based on the existing data. Improvement in accuracy showed that their approach was effective (G. Ryan et al., 2023).

3.5.3 Unrepresentative Data

1) As previously mentioned, the MBTI personality test results are not always accurate. Since most people determine their personality types by taking online tests, chances are that some people have mistaken their type. Moreover, the test can only reflect the temporary status of the person. As time progresses, their type could change, and reports show that the MBTI type of one person has about a 50% chance to change every few months (B. Cui et al., 2017).

2) Bias and Noise: Collected through the PersonalityCafe forum, there is an inherent bias in the sample group compared to the entire population. The sample group is limited. In addition, people could intentionally or unintentionally provide faulty information, distracting a model's accuracy. This led to a rise of suspicion about the accuracy and quality of the dataset (Y. Mehta et al., 2019). Furthermore, the texts people post on Twitter are highly dependent on moods and can be easily influenced by many other factors. The fact that some personality types have many similar traits (for example, INTP and ENTP) also needed to be considered. The data is noisy.

4 CONCLUSION

Personality prediction is a valuable tool in hiring and developing the recommendation system. It can also help users obtain a better mutual understanding in conversations, improve self-awareness, and achieve personal growth. This paper discusses implementing different machine learning methods in personality prediction based on MBTI. Among the three analyzed methods, Naive Bayes, Support Vector Machine, and Extreme Gradient Boosting, Extreme Gradient Boosting obtains the highest accuracy on average. However, the average accuracy for the three methods is not different. This research cannot get into other methods that are uncommon but show outstanding performance, like Deep Learning and Recurrent Neural Networks. Through detailed analysis of each sub-group, this research also finds that the model's capability to predict each sub-group generally falls in this sequence: N/S > I/E > T/F > J/P. As discussed in 3.5, this dataset is highly skewed, unbalanced, and noisy. However, surprisingly, the N/S category is the most unbalanced, as shown in Fig. 2. The sequence above is an exact sequence of how "unbalanced" the data is, ranging from high to low. The possible conclusion that can be drawn from this observation is that the "equal amount of data on each category" does not matter as much. More data is most helpful, and it can train the model to grasp the underlying pattern for each category better. Unfortunately, due to the lack of resources, this conclusion cannot be testified – there is no alternative dataset. The nature of each category or the inherent bias from the dataset may also be factors for the difference in accuracy rate. A richer and greater variety of datasets in the future can lead to more robust conclusions. Another thing that remains unclear is why models in some research have greater accuracy when the model is trained on the same method and dataset. Moreover, these are places where future research can be conducted.

REFERENCES

G. Ryan, P. Katarina, and D. Suhartono, "MBTI Personality Prediction Using Machine Learning and SMOTE for Balancing Data Based on Statement Sentences," Information, vol. 14, no. 4, pp. 217, Apr. 2023.

W. Bleidorn and C. J. Hopwood, "Using Machine Learning to Advance Personality Assessment and Theory," Personality and Social Psychology Review, vol. 23, no. 2, pp. 190–203, May 2018.

C. Stachl et al., "Personality Research and Assessment in the Era of Machine Learning," European Journal of Personality, vol. 34, no. 5, pp. 613–631, Sep. 2020.

K. Ilmini and T. G. I. Fernando, "Persons' Personality Traits Recognition Using Machine Learning Algorithms and Image Processing Techniques," Advances in Computer Science: An International Journal, vol. 5, no. 1, pp. 40–44, Jan. 2016.

N. Lemos, K. Shah, R. Rade, and D. Shah, "Personality Prediction Based on Handwriting Using Machine Learning," 2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS), Dec. 2018.

I. Maliki and M. A. B. Sidik, "Personality Prediction System Based on Signatures Using Machine Learning," IOP Conference Series, vol. 879, no. 1, pp. 012068, Jul. 2020.

B. Cui and C. Qi, "Survey Analysis of Machine Learning Methods for Natural Language Processing for MBTI Personality Type Prediction," 2017.

K. A., U. Kulsum, S. Rahman, Md. F. Hossain, P. Chakraborty, and T. Choudhury, "A Comparative Analysis of Machine Learning Approaches in Personality Prediction Using MBTI," in Advances in intelligent systems and computing, 2021, pp. 13–23.

T. M, "Comparative Study of Personality Prediction From Social Media by Using Machine Learning and Deep Learning Method," IJERT, Jun. 2021.

S. Ontoum and J. H. Chan, "Personality Type Based on Myers-Briggs Type Indicator With Text Posting Style by Using Traditional and Deep Learning," arXiv (Cornell University), Jan. 2022.

M. H. Amirhosseini and H. B. Kazemian, "Machine Learning Approach to Personality Type Prediction Based on the Myers–Briggs Type Indicator®," Multimodal Technologies and Interaction, vol. 4, no. 1, pp. 9, Mar. 2020.

Z., S. Ashraf, and N. Sabahat, "Predicting MBTI Personality type with K-means Clustering and Gradient Boosting," 2020 IEEE 23rd International Multitopic Conference (INMIC), Nov. 2020.

"(MBTI) Myers-Briggs Personality Type Dataset," Kaggle, Sep. 22, 2017. https://www.kaggle.com/datasets/datasnaek/mbti-type

A. Halevy, P. Norvig, and F. Pereira, "The unreasonable effectiveness of data," IEEE Intelligent Systems, vol. 24, no. 2, pp. 8–12, Mar. 2009.

Y. Mehta, N. Majumder, A. Gelbukh, and Z. Wang, "Recent Trends in Deep Learning Based Personality Detection," Artificial Intelligence Review, vol. 53, no. 4, pp. 2313–2339, Oct. 2019.