MBTI Personality Type Prediction

Kenny Mathew
210419104084

Under the supervision of

Premlatha T

July 2022

Acknowledgements

[To be changed by candidates as per their requirement]

We are using this opportunity to express my gratitude to everyone who supported us throughout the course of this project. I am thankful for their aspiring guidance, invaluably constructive criticism and friendly advice during the project work. I am sincerely grateful to them for sharing their truthful and illuminating views on a number of issues related to the project.

Further, we were fortunate to have Premlatha as our mentor. He / She has readily shared his immense knowledge in data analytics and guide us in a manner that the outcome resulted in enhancing our data skills.

We wish to thank all the faculties, as this project utilized knowledge gained from every course that formed the PGA program.

We certify that the work done by us for conceptualizing and completing this project is original and authentic.

Date: July 24, 2022 Name: Kenny Mathew

Place: Chennai

Certificate of Completion

[To be changed by candidates as per their requirement]

I hereby certify that the project titled "MBTI Personality Type Prediction" was undertaken and completed under my supervision by Kenny Mathew.

Mentor: Premlatha

Date: July 24, 2022

Place – Chennai

Abstract

In today's era, personality is one of the heavily re- searched and fascinating topics in psychology. What makes us who we are? Why do people act and behave in specific ways? How is our character different from the people around us? Applying MBTI, a personality type system that divides everyone into 16 distinct personalities, we can classify an individual in a particular personality type. The scope of personality computing has increased significantly. Personality recognition of users is widely used in research do- mains like recommendation systems and human-robot inter- action. Traditional recommendation systems come across problems like lack of data about the preferences of the user, free-riders problem, and the data sparsity problem. The identified user personality traits help understand users' preferences, which leads to better recommendations and tack- les the above issues. Another motive for this project is the MBTI's test-retest reliability, which hovers around a 0.5 error rate. On retest, people come out with 3-4 type preferences 75%-90% of the time. Our methodology can assist with more accuracy than currently existing tests, allowing users to rely on their outcomes. Personality classification based on digital data has proved to be an easier and more efficient alternative to traditional psychological tests. We can look into a far more significant amount of data with the help of text classification than we can with a single personality test.

Table of Contents

1. Introduction

Most people believe that there are only two types of personalities: introverts and extroverts. MBTI (Myers Briggs Type Indicator) evaluation helps us understand that personality is much more than that. With over 3.5 million assessments conducted each year, MBTI is the most widely used personality indicator globally. The Myers Briggs Type Indicator (MBTI) is a personality type system that divides everyone into 16 distinct personalities based on four dimensions, namely: Introversion (I) - Extroversion (E), Intuition (N) - Sensing (S), Thinking (T) - Feeling (F), Judging (J) - Perceiving (P). MBTI is widely used by companies, recommendation systems, and other research domains. MBTI predicted personality traits are found to retain essential properties of the traditional personality characteristics.

Researchers widely use machine learning and deep learning algorithms to predict personality and psychological traits from digital records. But as a field of research, personality prediction is at a relatively early stage. It is important to understand if the predicted personalities retain characteristics from psychological science and understand performance expectations for real-world tasks.

We're developing an MBTI personality classifier that uses machine learning models to predict a person's personality based on the recent social media posts per user as input. We find correlations between a person's MBTI personality type and writing style. The classifier also demonstrates the validity of the MBTI test. We have used a de- cent amount of mined personality annotated data from social media. Furthermore, our model would run on more data than that provided in a conventional personality test, which serves as a confirmation system and helps people rely more on the results.

2. Related Research

We referred to many research papers to develop a ma- chine learning system for the MBTI personality classification.

Sagar Patel et al. (2021) [1] did a personality analysis using social media (Twitter posts) based on MBTI. The pre- processing steps include removing hyperlinks, converting emojis to text, removing special characters, removing stop- words, and grouping different words with the same meaning and stemming. Also, the authors applied sampling methods to make the data balanced. They added new columns that divide the personalities based on four personality traits. Natural language processing techniques (NLP) for feature selection, such as N-gram, TF-IDF, Word2Vector, and glove word embedding, were used. They used K Nearest Neigh- bour (KNN), Naive Bayes, and Logistic Regression to train data. The results of these two models were found to be slightly more accurate than the SVM model. The model was run on the testing data, and accuracy, f1-score, recall was reported. Also, hyperparameter tuning was done to achieve the best results. Logistic regression performed the best using their methodology.

Pavel Novikov et al. (2021) [2] reviewed 220 research papers and articles to check if predicted personality estimates retain the characteristics of the original traits. Digitally available data is widely being used instead of traditional psychological tests for personality analysis. The authors stated that the automatic assessment should predict traits that are consistent with time (future behavior). They found that many predicted personality traits do not retain the characteristics of traditional personality traits. Most of the research papers reviewed used a Big-5 dataset where personality traits are distributed in a five-dimensional space. These traits remain consistent with time and include general characteristics shown by humans. The authors found that for most studies, the correlation between predicted and reported personality was below 0.5. The studies using the Big-5 dataset have the same correlation above 0.6. More work on analyzing personality prediction using psychometric validation instruments is required.

Nur Haziqah Zainal Abidin et al. (2020)[3] aimed to im- prove MBTI personality prediction using random forest classifiers. Dataset used here is the same as [1]. Exploratory analysis done on the posts included visualizing the number of words per post. More features like words per comment, ellipsis per comment, links per comment, music per comment, question marks per comment, images per comment, and exclamation marks per comment, were included. Also, the authors discussed a correlation matrix between all additional features for each personality. Same as [1], they have added four columns that divide characters into four dimensions. The authors used the Random Forest, Linear Regression, KNN, and SVM models of sklearn. They found that personality prediction using textual information was most accurate for the Random Forest machine learning model (almost 100%) using this methodology.

3. Aim and Objective

The primary aim of this research is to make model that is able to predict MBTI personality type and create an API so that the user can get a model predicted value for the given input.

The primary objective of this project is to bring together diverse, novel and impactful work on personality prediction in one place, thereby accelerating research in this field.

The objective of this project is as follows;

i. To train model with a dataset consisting of various posts which are taken from social media posts.
ii. To use the model to generate the personality type based on the textual input provided to the model.
iii. To use the model and create an API using streamlit.
iv. To deploy the API so that user can interact with the web application

4. Significance and Scope of Study

Personality traits are generally referred to as relatively stable patterns of thoughts, feelings, and behaviors that have been associated with a wide range of important life outcomes and choices. Specifically, personality traits have repeatedly been related to the individual (e.g., well-being, psychopathology), interpersonal (e.g., relationship satisfaction), and social-institutional outcomes (e.g., occupational choices, job success). Hence, in the recent years, there has been a massive increase in the interest to develop models which use online data on human behavior and preferences (i.e., digital footprints) to automatically predict an individuals' personality traits.

Social media gives users the opportunity to build an online persona through posting of content such as text or through interaction with others. The way in which users present themselves is a type of behavior usually determined by differences in demographic or psychological traits. The behavioral residue harvested from websites and online social media platforms is also another valuable source of data on behavior linked to personality traits. Hence, automated personality prediction has important practical applications in diverse areas ranging from recommendation systems, computational advertising, marketing science, job screening to aiding in psychological counselling, intervention and therapy, enhanced human-computer interaction, etc.

5. Research methodology

The methods used to acquire the MBTI personality ac- curacy scores are described in this section:

5.1. Pre-Processing

For better feature extraction, some preprocessing is per- formed on the textual data in column ’posts’. The process reduces data inconsistencies, outliers, or duplicates, which can otherwise negatively affect a model’s accuracy. The be- low approaches minimize the data’s complexity and make it satisfactory for machine learning model training.

1) To lower case: The textual data is converted into lowercase using str.lower() function. As a result, two identical words written in different letter cases can be interpreted as similar.

2)Removal of URL/links: The web URLs do not give us any direct text information regarding a person’s personality. They are incompetent in the classification of personality. These links are removed using the regular expression for URLs.

3) Removal of special characters and numbers: The special characters such as ’.’, ’,’, ’———” etc., are primarily outliers and noise. Also, numbers rarely give some helpful information about someone’s personality. Thus, they are removed as well using a regular expression.

4) Removal of extra space: Extra space gives meaning- less information. So, they are removed using regular ex- pressions.

5) Removal of stop words: In English, stop words include

words such as ’for’, ’them’, ’you’ etc. These kinds of words are essential to make sense for a language, but they are meaningless for feature extraction and training of models. These words are accessed from the nltk library in python.

6) Removal of MBTI personality names: MBTI personality names such as ’INFJ’, ’ISTP’ used by people in their posts can wrongly influence the results.

Consequently, they were also deleted.

7) Lemmatization: Words having the same meaning should be taken as a single feature. Lemmatizer is used to group words with the same purpose together (gone, going, went to go).

5.2. Feature Extraction

After preprocessing, the raw or annotated text is con- verted into features, providing a simpler, more focused view of the text to the machine learning model and enhancing performance. The technique applied for this step is- Term Frequency and Inverse Document Frequency (TF-IDF): Our dataset is unbalanced for a few personality types, implying some words appear more often and carry little meaning and information about the data. If these high-frequency data are fed into the classifier, the model overshadows the less in number data. TF-IDF and count vectorizer is used to con- vert text into features, providing a more focused text view. First, vectorize the data using countVectorizer and convert the post into the matrix of token counts for the model. Then TF-IDF normalization is used to scale the feature from the count vectorizer into floating-point values. TF-IDF analyzes how much a word is relevant to a corpus in a corpus collection and provides the importance of words in data. After vectorizing, the dataset had 1500 features for each user post.

The term frequency represents the frequency of each of the words present in the dataset.

Tf(t) = (No. of times term t appears in a document)/(Total no. of terms in the document)

Idf tells us the importance of words in the dataset, and it is decided by how rare the word is in the datasets.

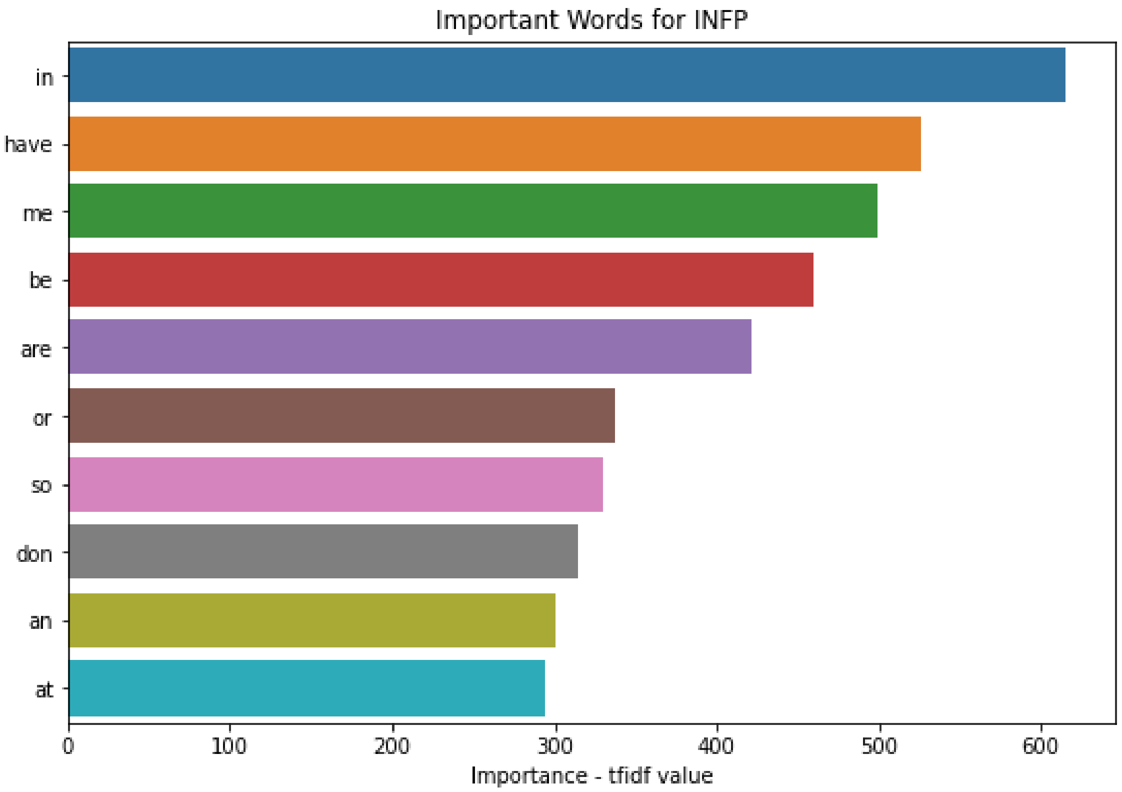
Idf(t) = log10(Total number of documents / Number of documents with term t in it). Therefore, Tf-idf = Tf*Idf

After examining the importance of words in the datasets (Figure4), all the relevant terms are removed, making the model less complex by reducing the input.

5.3. Training

Text classification is a supervised machine learning task. So, the training dataset is trained on various supervised ma- chine learning models. After preprocessing and feature ex- traction, the resulting dataset is split into training and testing datasets in the ratio 80:20. Since the data was found to be slightly unbalanced for IE and NS dimensions, we used Stratified KFold cross-validation using GridSearchCV to get more accurate and stable results. The models used for

The figure above shows the Word cloud depicting the importance of different words for INFP type, i.e., each word’s size is proportionate to how often it appears in the top posts.



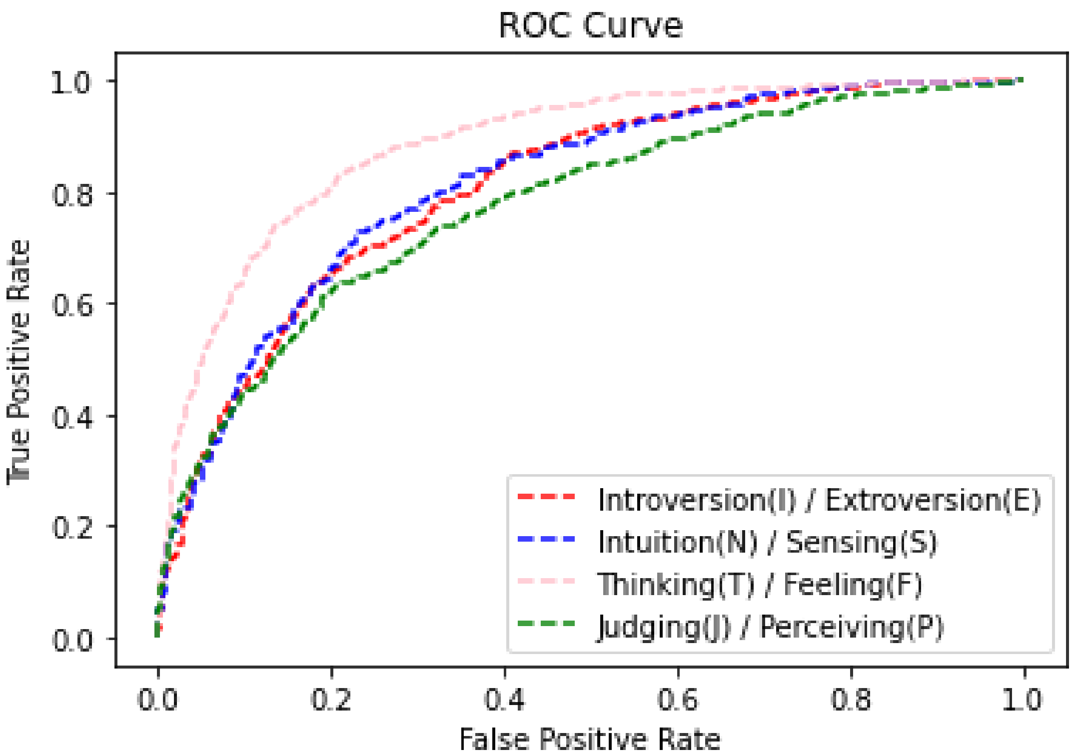
The figure above shows the importance of different words for INFP type using the TF-IDF value.

MBTI personality classification in our project are Logistic Regression, Naive Bayes and Random Forest Classifier, K- Nearest Neighbor (KNN), SGDClassifier and Support Vector Machines (SVM). To build all these models, we have used sklearn, NumPy and pandas libraries. Gaussian Naive Bayes, one of the simplest machine learning models, gives a low accuracy on the testing dataset. Multivariable logistic regression (applies regularization as default) showed excel- lent results. Random forest classifier and KNN was less accurate than logistic regression. Results of the SGD classifier and SVM were similar to logistic regression.

5.4. Testing and Validation

After training the dataset on six supervised machine learning models, we analyzed their results and performance. Prediction on the test dataset using each model is used to analyze the performance of each model. The performance evaluation is not merely based on accuracy scores. The overall performance evaluation includes analysis of accuracy scores, confusion matrix, AUC-ROC curve, precision and recall-score. The confusion matrix shows if the model is overfitting or underfitting the test data. The AUC-ROC curve shows the performance at all classification thresholds. For a model to be efficient, the roc curve should be closer to the upper left corner. The best model is the one that has the best overall performance. To avoid overfitting and underfitting of the test dataset, we performed hyperparameter tuning for all machine learning models and analyzed performance with various values of parameters. The tuned parameters were used to obtain the most efficient model. In Random Forest Classifier, we have performed hyperparameter tuning on max depth and min samples split using Grid-SearchCV. To get the best value for max depth of the tree, we analyzed accuracy vs depth graphs for both train and test datasets. In the KNN machine learning model, we decided the best value of K by observing its performance on a range of values of K. To get the best value of K; we analyzed accuracy vs K graph or both train and test datasets. For the logistic regression model, we used penalty='l2' and max iter as 500. SGDClassifier with loss='log' was used for performance analysis.

Model	I/E	N/S	F/T	J/P
LR	0.81	0.86	0.80	0.72
SGD	0.81	0.86	0.79	0.71
SVC	0.80	0.86	0.80	0.72



Logistic Regression, Support vector classifier and Stochastic gradient descent provide the best accuracy of around 80% for I/E, N/S, F/T, and approximately 72 for J/P. The precision and recall values of these algorithms are also good. Random forest, k-nearest neighbour relatively has lower performance. Still, they are giving an accuracy of around 75%. The accuracy, precision and recall results of different models are shown in the respective table. The ROC curve for logistic regression also supports the results of the analysis. We conclude that the Logistic Regression model performs the best for personality classification based on The Myers Briggs Personality Model.

6. Required Resources

The research will need below hardware and software resources throughout the implementation.

6.1. Software Requirements

- Package Manager: Anaconda Navigator 1.9.12
- Presentation Layer: Jupyter lab 0.35.4
- Language: Python 3.6.X
- Python Libraries for machine learning: Pandas and NumPy for data processing, scikit-learn for the models, seaborn and matplotlib for data visualization.

6.2. Hardware Requirements

A laptop with below configuration will be used.

- Operating System. Windows: 64-bit
- Processor: Intel® Core™ i7 8th gen 8650U Processor
- Memory: 8 GB
- GPU: NVIDIA GTX 1050 or above

7. References

1. Patel, S., Nimje, M., Shetty, A. and Kulkarni (n.d.). Personality Analysis using social media. [online] Available at:ijert.org/[Accessed 25 Nov. 2021].
2. Novikov, P., Mararitsa, L. and Nozdrachev, V. (2021). Inferred vs traditional personality assessment: are we predicting the same thing? [online] Available at: arxiv.org/[Accessed 25 Nov. 2021].
3. Haziqah, N., Abidin, Z., Remli, M., Ali, N., Nincar- ean, D., Phon, E., Yusoff, N., Adli, H. and Busalim, A. (2020). Improving Intelligent Personality Prediction using Myers-Briggs Type Indicator and Random Forest Classifier. IJACSA) International Journal of Advanced Computer Science and Applications, [online] 11(11). Available at:thesai.org/[Accessed 25 Nov.2021].
4. “6839354.pdf.” Accessed: Oct. 12, 2021. [Online]. Available:https://web.stanford.edu/
5. N. H. Z. Abidin et al., “Improving Intelligent Personality Prediction using Myers-Briggs Type Indicator and Random Forest Classifier,” IJACSA, vol. 11, no. 11, 2020, doi: 10.14569/IJACSA.2020.0111125.
6. S. Patel, M. Nimje, A. Shetty, and S. Kulkarni, “Personality Analysis using Social Media,” International Journal of Engineering Research, vol. 9, no. 3, p. 4, 202
7. arxiv.org. (n.d.). Staff Photo. [online] Available at: arxiv.org/[Accessed 25 Nov. 2021].