Personality Prediction Using Machine Learning

Shilpa R[1],Supriya V[2], Sweta Prasad [3], Vinaya Varshini R [4],Uday Shankar SV[5]

1-4 UG Students, Department of Information Science and Engineering ,SJB Institute of Technology, Bengaluru-60.
5 Asst. Professor, Department of Information Science and Engineering ,SJB Institute of Technology, Bengaluru-60.

Abstract*: Personality is an essential component of our way of life. It influences how we live, speak, respond, and express ourselves, as well as our mental health. Personality analysis is a natural human capacity that is used on a daily basis with a variety of people and for a variety of purposes. Personality profiling, in particular, has a variety of real- world applications, including exams for screening of mental illness, human resource interview shortlisting, friend recommendations and writer recommendations on the interplay of personalities that people love reading about. This project analyses a person's written content, such as a tweet, an essay or a blog post, and finding a personality profile of the person. The type of data collected, text preparation methods, and machine learning approaches utilised to predict personality scores are all important issues in this study. The deployment of solutions has been explained, and several feature vector combinations and machine learning models have been compared. Using the methods described in this paper, accuracies of up to 65 percent were achieved.*

Keywords- data preprocessing, Datasets, Logistic Regression, MBTI .

.

1. Introduction

Personality is characterized as the trademark sets of cognitions, behaviors and passionate examples that advance from biological and ecological components. It mirrors the people as they contrast in thoughts, behavior and sentiments. Personality characteristics are constant in world as they give a sense of high and low of explicit attributes in an individual on consistent quality instead of displaying particular personality. The distinction in personality should be expected to forestall the reason for deterrents or friction in the subject of work or education [3]. In this way, in the enlistment of schooling or working environment, an individual should be distinguished.

So, in order to predict distinct personality of each individual we develop a machine learning model i.e., the Myers-Briggs Type Indicator (MBTI). This Project follows the principle of MBTI as a guideline's so that it helps to identify the personality of the user based on the following personality dimensions: Introvert (I)and Extrovert (E), Sensation(A)and Intuition(N), Thinking(T) and Feeling(F), Perceiving(P) and Judging(J). The coalescence of the above four types of personality dimensions will result in sixteen types of personality such as "INFJ" or "ENFP" etc [2]. In our model we have used algorithms like KNN, Logistic Regression, XG Boost. We have taken the dataset from the source Kaggle. In our model we first import the dataset from Kaggle. Then feed it into data

analysis to Check whether there are any missing or null values present in the dataset and the analysed data is then fed into data pre- processing to clean the data. After cleaning process, the data is sent to feature engineering and finally by comparing the algorithms with each other we choose the best algorithm to our model that can predict the personality of each individual.

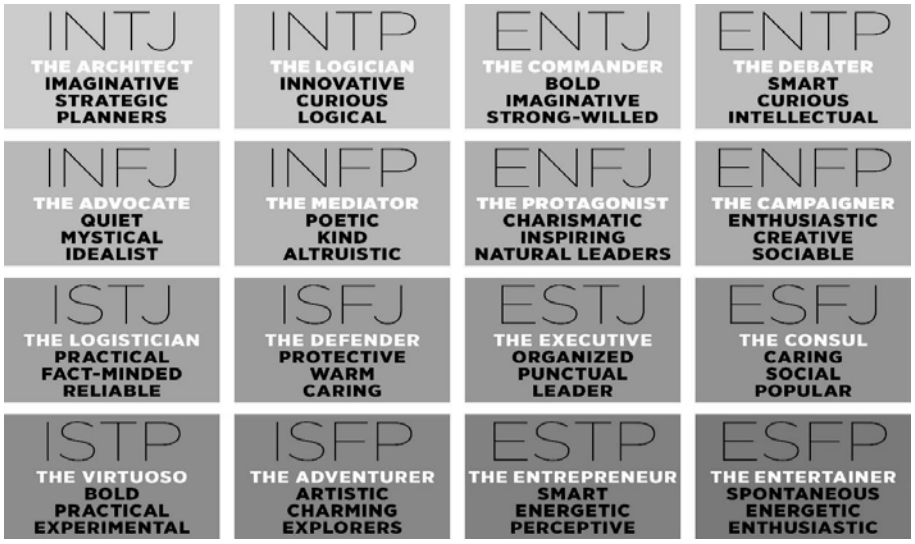

Fig. 1: The sixteen MBTI types

2. Related Work

The task of personality prediction has been contributed to by a number of researchers [5]. Few have utilized psychological exams to determine personality labels, while others have employed machine learning algorithms such as Naive Bayes to predict outcomes [7]. In it linear regression and support vector regression were used to determine Facebook user personality. This study makes use of the My Personality dataset. The study's findings demonstrate that linear regression is a better alternative for prediction. Others have used random forest, basic logistics, and J48 to find suicide-related tweets from Twitter. The dataset was collected using the Twitter streaming API, and the Martingale framework was used to forecast the outcomes. This is a unique technique. Few have Detected mental disorders using binary SVM [6]. Amazon Mechanical Turk was used to create the dataset for this. This study's future work could focus on retrieving multimedia content for prediction. Few have extracted and examined digital footprints of users. For prediction utilizing digital footprints, several meta analyses have been conducted. The Big Five personality traits were used in this study, and the results are beneficial for recommending items or services to users based on their preferences [4].

3. Dataset

The dataset imported to our project has tweets from social media of a single person as one attribute and one of the sixteen MBTI personality type of the person which is a combination of four personality dimensions representing the Mind, Energy, Nature and Tactics.

Each of the character of the four characters corresponds to each of the four MBTI classes. The dataset consists of 8675 rows and 2 columns. The 2 columns are the two attributes in the dataset namely Type, representing the MBTI personality type and Posts, representing the social media posts.

The data present in a particular row of the dataset contains posts from social media from a single person. The dataset has no null values neither in the type attribute nor in the posts attribute. The INFP, INFJ, INTP, INTJ personality types have a greater number of overall posts while the ESTJ, ESFJ, ESFP, ESTP personality types have the lesser number of overall posts in the dataset. This might be because of the fact that the ESTJ, ESFJ, ESFP and ESTP form a smaller population in the world and are one of the rarest personality types on the earth.

```
Introversion (I) /  Extroversion (E):    6676  /  1999
Intuition (N) / Sensing (S):             7478  /  1197
Thinking (T) / Feeling (F):              3981  /  4694
Judging (J) / Perceiving (P):            3434  /  5241
```

Fig.2 : distribution of MBTI dataset rows in each class

4. Methodology

This section illustrates the steps followed to develop a successful model which distinguishes personalities.

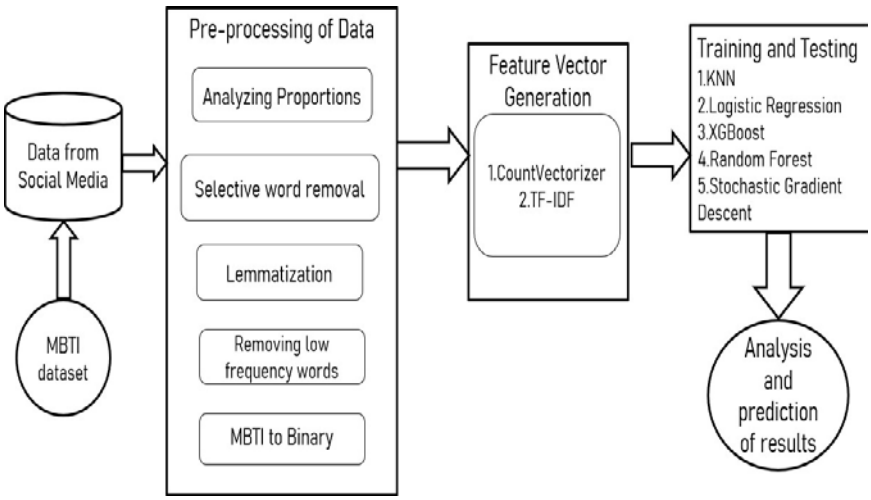

Fig.3 : The architecture of MBTI classifier

4.1 Data Analysis and Pre-Processing

Dataset that consists of attributes like Type and Posts is imported into the project. Data needs to be analysed in order to find null values if any, to understand the datatype of the values in the attributes, to know how many rows are there in the dataset and to find dependencies between attributes if any. Data Analysis is also performed in order to analyse the proportions of each personality type in the datatype.

Pre-Processing of data is done in order to remove non- words, punctuations, repeated letters and stop-words, so as to clean the data in order to increase. Pre-processing consists of selective word removal, lemmatization and conversion of MBTI 4 letter code to binary. In Selective Word Removal certain words which might cause the machine to cheat. Lemmatization brings all the words to their root form. For example, running, playing and thinking becomes run, play and think respectively. Converting the MBTI personality to binary consists of assigning each aspect with 0 or 1 so that it can be easily interpreted by the model. Here I, N, F, J are all assigned with 0, whereas E, S, T, P are all assigned with 1. So, the personality type ENFP is converted to [1,0,0,1].

4.2 Feature Engineering

Feature engineering is the process where we convert raw data from the pre-processing stage into features that can make it easy for the model to understand the underlying problem so that we can get an improved model accuracy for the given data. After pre-processing some words like “is”, “the”, “but” (in English) will be predominantly present. These words will carry very little meaning and information about what the data is trying to convey. If we fed this data directly to the classifier, since the importance given to these words would be higher solely because they appear a greater number of times, then we’ll end up over-shadowing the frequencies of terms which are rarer but far more

important.

Tf-idf transform is used to re-weight the features from the count vectorizer into floating point values so that it can be effectively used by the classifier. It is used to analyze how much a word is relevant to a corpus in a collection of corpuses.

Tf-idf represents Term Frequency and Inverse Document Frequency. It is the product of the before mentioned terms Term Frequency and Inverse Document Frequency. Term frequency represents the frequency by which a word appears in a document. Inverse Document Frequency applies for a set of documents and represents the importance of a word in the documents and is decided by how rare the word is present in the document.

4.3 Training and Testing

Text Classification is a task of higher importance in Supervised Machine Learning. The algorithms taken into consideration for text classification for this project are Random Forest, Stochastic Gradient Descent, K-Nearest Neighbour, Logistic Regression and XG-Boost.

Random Forest has its roots from ensemble learning concept. Ensemble learning process integrates multiple classifiers with each solving smaller problem to solve a large problem so as to upgrade performance of the classifiers. Stochastic gradient descent has its applications in finding model attributes that correlate with the actual and predicted outputs. KNN does not learn immediately from the given dataset, it will stockpile the dataset and while doing the task of classification, it takes action on the stored dataset. If the target variable is a discrete variable only then can the algorithm of logistic regression be used. Logistic regression constructs a regression model whose work is to give the probability that a given input belongs to a category named as ‘1’ as the result. This algorithm uses sigmoid function to model the data. XGBoost is an algorithm which is built for the sole purpose of improving the computational speed and classifier model performance and it is based on the concept of gradient boosted decision trees.

Now the pre-processed and vectorized text is fed to the models in order to train them. After the models have been trained, the testing part of split dataset is loaded to the pre-trained models.

In our project the train test split ratio of 70:30 was decided and approved after comparing the accuracies for different train test split ratios. Now after feeding the models with the testing part of the dataset, the accuracies achieved by the different models are compared. We have considered two metrics based on which the model is chosen. These two metrics are Accuracy and Performance. Based on these factors, Logistic Regression was chosen to be the primary model for this project, as depicted in Fig 3.

4.4 Web Application

In order to provide our model with a User Interface so that the user can interact freely and easily with the Personality prediction system and find out their personality, a web application is built. The input given by the user can be a long text content or the twitter handle of the certain person whose personality is to be known. The Paragraph entered by the user is sent to the pre-trained models (back end). The text is pre-processed and vectorized and given to the model which predicts the personality type. This result is sent to the front end. The result is displayed to the user along with certain description about the specific personality type that the user got back as a result.

5. Results

The main result will fetch the predicted value of the personality type of individuals. The major concern was to enhance the accuracy, so we have taken different algorithm in concern to check the accuracy and choose the most accurate algorithm which gives the maximum accurate results.

The Logistic Regression is created for binary classification and do not support classification tasks with more than two classes by default. Splitting the multi-class classification dataset into numerous binary classification datasets and fitting a binary classification algorithm to each of them is one method for applying binary classification algorithms for multi-classification situations. Instead of using multiple classifier we are using one classifier to classify the personality trait which will be repeated on loop for 4 times to yield the final output.

The testing and training are done using test_train_spilt with different values to look for best variable to divide the dataset into two parts.

Table 1. accuracy obtained for each of the four personality dimension

Algorithms	Introversion (I) / Extroversion (E)	Intuition (N) / Sensing (S)	Feeling (F) / Thinking (T)	Judging (J) / Perceiving (P)
Logistic Regression	77.54%	86.06%	86.06%	64.51%
KNN model	75.76%	75.76%	53.82%	44.99%
XGBoost model	76.18%	85.37%	67.73%	62.84%

.

It was observed test_size=0.33 and the random_state=7 which giving good result and set fixed for executing different algorithms to look for maximum accuracy.

We have taken K-Nearest Neighbours (KNN) algorithm, Logistic Regression, XG BOOST algorithm, Random Forest, Stochastic Gradient Descent and found the logistic regression performing exceptionally better than others with the accuracy of 65%.

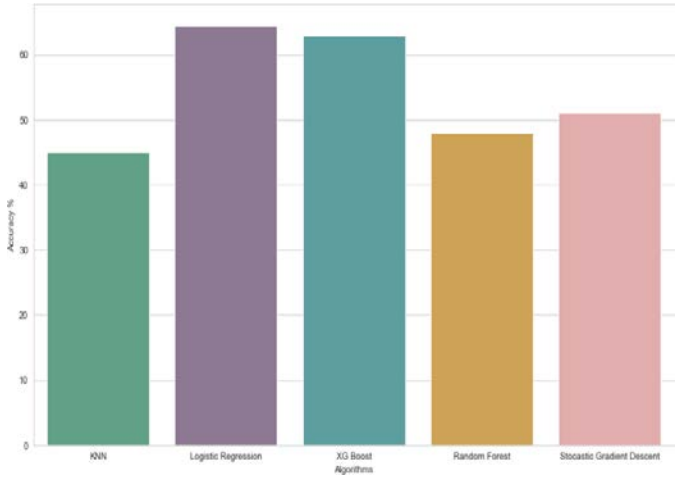

Fig.2 Overall accuracy of different algorithm to predict the personality type correctly

6. Conclusion

The major objective of the project is to use the user's content to generate a personality profile. The early analysis included looking for trends in attitudes as well as analysing distribution. In the data collected from social media sites, there are a lot of raw emotions which in turn become beneficial to predict the personality. After that, for data cleansing, the text was pre-processed by removing hyperlinks, numerals, punctuation, and context- sensitive words. Various machine learning models in various combinations was studied. Nowadays, many organizations have also started shortlisting candidates based on personality as it increases work efficiency as the person works on what is good than what they need to do. Our model includes algorithms like KNN, XG Boost, Random forest, Stochastic Gradient Descent and Logistic regression. These algorithms were compared. It was discovered that the best outcomes were obtained through the use of logistic regression. However, accuracy is almost same after using XG Boost. Our model helps the users to easily identify their personality from this model.

7. References

[1] S. Bharadwaj, S. Sridhar, R. Choudhary and R. Srinath, "Persona Traits Identification based on Myers-Briggs Type Indicator(MBTI) - A Text Classification Approach," 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2018, pp. 1076-1082, doi: 10.1109/ICACCI.2018.8554828.

[2] C. Y. Yaakub, N. Sulaiman and C. W. Kim, "A study on personality identification using game-based theory," 2010 2nd International Conference on Computer Technology and Development, 2010, pp. 732-734, doi: 10.1109/ICCTD.2010.5646417.

[3]A.Ariyanto, E. C. Djamal and R. Ilyas, "Personality Identification of Palmprint Using Convolutional Neural Networks," 2018 International Symposium on Advanced Intelligent Informatics (SAIN), 2018, pp. 90-95, doi: 10.1109/SAIN.2018.8673353.

[4] Azucar, Danny & Marengo, Davide &Settanni, Michele. (2018). Predicting the Big 5 personality traits from digital footprints on social media: A meta- analysis. Personality and Individual Differences. 124. 150-159. 10.1016/j.paid.2017.12.018.

[5]Aditi V. Kunte, Suja Panicker. "Using textual data for Personality Prediction:A Machine Learning Approach" , 2019 4th International Conference on Information Systems and Computer Networks (ISCON), 2019

[6]Predicting Mental health disorders using Machine Learning for employees in technical and non-technical companies,2020 IEEE International Conference on Advances and Developments in Electrical and Electronics Engineering (ICADEE),Date of Conference: 10-11Dec.2020,DateAddedtoIEEE Xplore: 08March021,I NSPECNumber: 20532983,DOI: 10.1109/ICADEE51157.2020.9368923,Publisher: IEEE,Conference Location: Coimbatore, India

[7]B. Y. Pratama and R. Sarno, "Personality classification based on Twitter text using Naive Bayes, KNN and SVM," 2015 International Conference on Data and Software Engineering (ICoDSE), 2015, pp. 170-174, doi: 10.1109/ICODSE.2015.7436992.