Personality Type Prediction based on the Myers–Briggs Type Indicator

Monika Sharma Nupur Sudhakar Snigdha Gupta
Netaji Subhas University of Technology

Abstract

With the development of social networks, a large variety of approaches have been developed to define users' personalities based on their social activities and language use habits. Particular approaches differ with regard to different machine learning algorithms, data sources and feature sets. The goal of this paper is to investigate the predictability of the personality traits of social media users based on different features. The results for the prediction accuracy show that even if tested under the same data set, the personality prediction system built on the XGBoost classifier outperforms the average baseline for all the feature sets, with a highest prediction accuracy of 74.2%. The best prediction performance was reached for the extraversion trait which achieved a higher personality prediction accuracy of 78.6%.

The Myers–Briggs Type Indicator (MBTI) is currently considered as one of the most popular and reliable methods. In this study, a new machine learning method has been developed for personality type prediction based on the MBTI. The results of this study can assist psychologists in regards to identification of personality types and associated cognitive processes.

Introduction

Today's world is witnessing a great increase in the use of social media. People use them as a platform to share their feelings, emotions and experiences along with a lot of personal information. They express themselves on certain issues related to their lives and family well beings, psychology, financial issues, interaction with societies and environment, as well as politics. In some cases, these expressions can be used to characterize the individual behaviour and personality. All such information could be used in advantageous ways to help increase the business and understand the user's needs. Personality prediction has gained a lot of focus nowadays. It studies behaviour of users and reflects the thinking, feelings etc.

Termed personality prediction, the process involves extracting the digital content into features and mapping it according to a personality model.

The Myers Briggs Type Indicator (or MBTI for short) is a personality type system that divides everyone into 16 distinct personality types across 4 axes:

- Introversion (I) – Extroversion (E)
- Intuition (N) – Sensing (S)
- Thinking (T) – Feeling (F)
- Judging (J) – Perceiving (P)

MBTI has been used in this study in order to predict the personality type on individuals.

The most popular personality types will be identified and current organizational culture and task allocation can be modified based on this information.

We aim to produce a machine learning model that can attempt to determine a person's personality type based on some text they have written.

Figure 1 shows these 16 personality types that result from the interactions among the preferences of an individual.

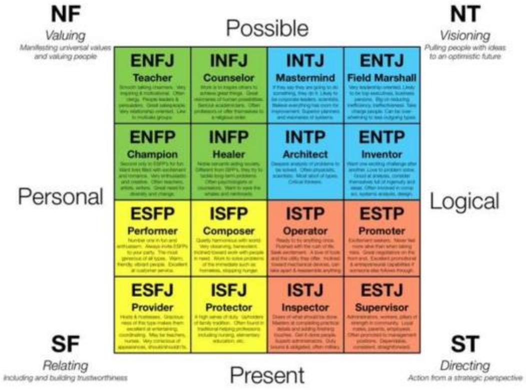


Figure 1: Personality types in the Myers Briggs Type Indicator

Each keyword represents a specific personality type.

Figure 2 describes the cognitive functions of each MBTI personality type.

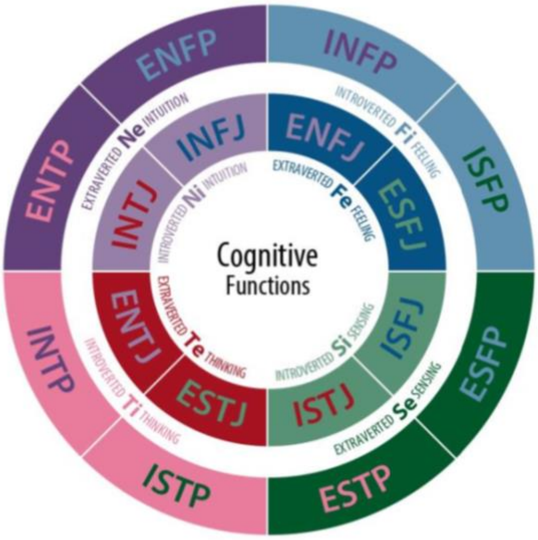


Figure 2: Cognitive functions of each personality type

The background colour of each type represents its dominant function and the colour of the text represents its auxiliary function.

Literature Survey

There is significant growing interest in automated personality prediction using social media among researchers in both the Natural Language Processing and Social Science fields. So far, the application of traditional personality tests has mostly been limited to clinical psychology, counselling and human resource management. However, automated personality prediction from social media has a wider application, such as social media marketing or dating applications and websites.

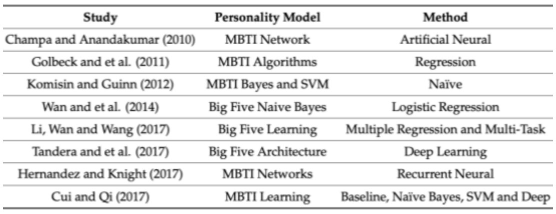
Research on personality type prediction from textual data is scarce. However, important steps have been taken in this endeavour through machine learning. Classic machine learning techniques and neural networks have been used successfully for predicting MBTI personality types.

Most studies on personality prediction have focused on the Big Five or MBTI personality models, which are the two most used personality models in the world.

A personality trait is a characteristic pattern of thinking, feeling, or behaving that tends to be consistent over time and across relevant situations. Based on this explanation, the Big Five personality model can be defined as a set of five broad trait dimensions, namely, (1) extroversion, (2) agreeableness, (3) conscientiousness, (4) neuroticism and (5) openness. In fact, the Big Five personality model uses descriptors of common language and suggests five broad dimensions commonly used to describe the human personality. Research proposes that considering controversy about the reliability and validity of these two models, the MBTI model has more applications, especially in industry and for self-discovery of personality types.

In this study, it was found that classification techniques such as logistic regression, Naïve Bayes, Random Forest, K Nearest Neighbour (KNN), Linear Discriminant Analysis (LDA) and Support Vector Machine (SVM) have all been used for personality type prediction based on the MBTI or Big Five personality type models.

Table 1: Research on personality type prediction and personality models used



Study	Personality Model	Method
Champa and Anandakumar (2010)	MBTI Network	Artificial Neural
Golbeck and et al. (2011)	MBTI Algorithms	Regression
Komisin and Guinn (2012)	MBTI Bayes and SVM	Naïve
Wan and et al. (2014)	Big Five Naive Bayes	Logistic Regression
Li, Wan and Wang (2017)	Big Five Learning	Multiple Regression and Multi-Task
Tandera and et al. (2017)	Big Five Architecture	Deep Learning
Hernandez and Knight (2017)	MBTI Networks	Recurrent Neural
Cui and Qi (2017)	MBTI Learning	Baseline, Naïve Bayes, SVM and Deep

According to the literature, the MBTI model has been more popular among researchers and, considering controversy about reliability and validity of these two models, the MBTI model has more applications in different disciplines. That is why the MBTI personality model was used in this study.

Implementation

a) <u>Dataset</u>

The dataset has been taken from Kaggle website which is a .csv file. It has 2 columns and 8675 rows. The first column represents the personality type and the second column represents the last 50 things that have been posted on the social media. Each row represents the user's personality type and posts.

b) <u>Data Visualisation</u>

Using libraries like seaborn, matplotlib, word cloud, we have visualised and interpreted the data.

Figure 3 shows a plot of number of posts versus personality types.

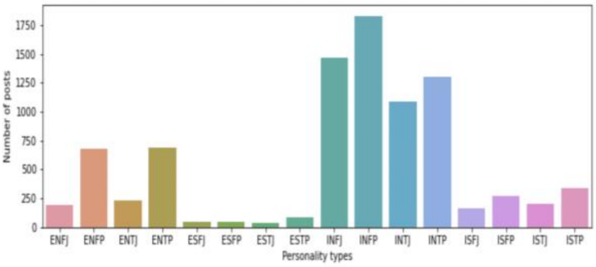


Figure 3: Number of posts versus personality types

INFP personality type has the largest number of posts whereas ESTJ has the smallest number of posts.

Figure 4 shows a plot of number of words in posts versus personality types.

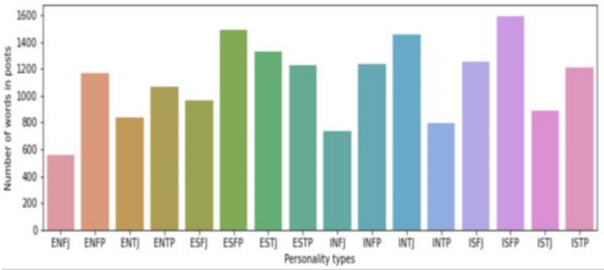


Figure 4: Number of words in posts versus personality types

Figure 5 shows a word cloud which is used for representing text data in which the size of each word indicates its frequency or importance.

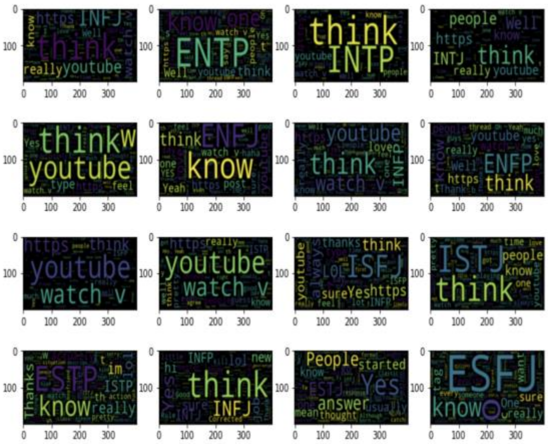


Figure 5: Word Cloud of the words that appear in the posts

We can infer that we would also have to remove the personality type words like ENTP, EFSJ, INTP, etc.

c) <u>Data Cleaning</u>

We have cleaned the posts column. Words with less than 3 characters and more than 30 characters, links, special characters and personality type words have been removed.

We have added a number of words column where we have analysed how many words has each user posted. Four more columns (IE, NS, TF, JP) which are the axes across which personality is divided have also been added. The value 1 for these columns indicate the first personality type like I, N, T, J and 0 indicates the second personality type like E, S, F, P.

Figure 6 shows a pie chart distribution of people with different personality types.

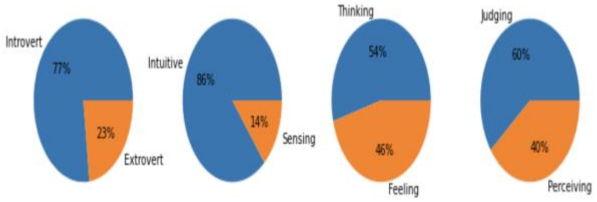


Figure 6: Distribution of different personality types

This shows that among all the users :-

- 77% are introverts and 23% are extroverts.
- 86% are intuitive and 14% are sensing.
- 54% are thinking and 46% are feeling.
- 60% are judging and 40% are perceiving.

Next, we use CountVectorizer to transform text of posts into a vector on the basis of the frequency (count) of each word that occurs in the entire text. English stopwords imported from the Natural Language Toolkit library have been removed through CountVectorizer.

The columns have been split into X and Y columns and further the data has been split into 80% training set and 20% test set.

d) Model

Logistic Regression, Multinomial Naïve Bayes, Support Vector Machine, Random Forest Classifier and XGB Classifier have been used to implement the model.

We are training these five models separately for all four personality axes.

Performance Analysis

The accuracy of the five models for each personality axis is as follows –

- For IE :-
  - SVC: 0.7648414985590778,
  - Logistic Regression: 0.7469740634005764
  - Multinomial Naïve Bayes: 0.7596541786743516
  - XGB Classifier: 0.7648414985590778
  - Random Forest Classifier: 0.7648414985590778
- For NS :-
  - SVC: 0.8518731988472622
  - Logistic Regression: 0.8403458213256484
  - Multinomial Naïve Bayes: 0.8484149855907781
  - XGB Classifier: 0.8570605187319885
  - Random Forest Classifier: 0.8524495677233429
- For TF :-
  - SVC: 0.7717579250720461
  - Logistic Regression: 0.7654178674351585
  - Multinomial Naïve Bayes: 0.7613832853025937
  - XGB Classifier: 0.7406340057636888
  - Random Forest Classifier: 0.7152737752161383
- For JP :-
  - SVC: 0.6368876080691642
  - Logistic Regression: 0.6161383285302594
  - Multinomial Naïve Bayes: 0.6593659942363113
  - XGB Classifier: 0.631700288184438
  - Random Forest Classifier: 0.6184438040345821

Conclusion

In this paper, we provide an outline of insights for research on social networks and personality psychology. The study investigates the literature on the uses of social media frame-work as behavioural feature study by exploring the relationship between users' personalities and their behaviours in social networks. To predict a user's personality, we conducted a comparative study of best behavioural indicators for social media usage of the same set of features to capture the ways the users socialize, communicate and connect with each other.