Project Report

Text Mining

BAN 675-01

Fall 2018

Predicting Ones MBTI (personality) Type Using Text Data

Submitted By-

Anh Nguyen (xc8374)

Saif Ur Rahman (ct8692)

Surbhi Bagdi (yp7669)

Thao T. Dinh (pe3293)

Tong Xu (im7348)

1. INTRODUCTION

MBTI, short for Myers- Briggs Type Indicator, is a personality metric developed by a mother-daughter duo, Katharine Cook Briggs and Isabel Briggs Myers. This test is based on Carl Jung's theory of Psychological types. The MBTI test is largely used by organizations and individuals to optimize organizational dynamics and to know oneself better.

An individual is tested in four different areas as can be illustrated by **Figure 1**

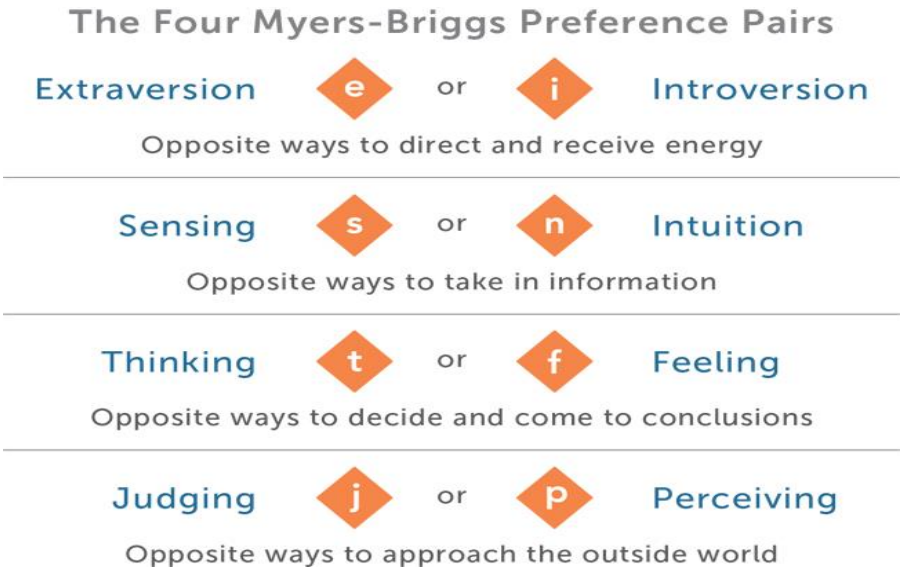

Figure 1 – The Four Myers-Briggs Preference Pairs **Source: Internet**

The identification of preferences of each of the four dichotomies implicit in Carl Jung's theory-

- **Favorite world:** Do you prefer to focus on the outer world or on your own inner world? *This is called Extraversion (E) or Introversion (I).*
- **Information:** Do you prefer to focus on the basic information you take in or do you prefer to interpret and add meaning? *This is called Sensing (S) or Intuition (N)*.
- **Decisions:** When making decisions, do you prefer to first look at logic and consistency or first look at the people and special circumstances? *This is called Thinking (T) or Feeling (F).*
- **Structure:** In dealing with the outside world, do you prefer to get things decided or do you prefer to stay open to new information and options? *This is called Judging (J) or Perceiving (P).*

User's Personality Type: When you decide on your preference in each category, you have your own personality type, which can be expressed as a code with four letters.

An individual is tagged with 4 letters according to MBTI. For instance, if a person is type ISFJ, then that person is an introvert, sensing, feeling, and judging.

The project shall attempt to help a person in finding his/her MBTI. So, rather than having an individual to answer a series of questions only to find in the end his/her personality type, the model needs to pick-up the existing messages produced by a user and predict his/her personality type.

We set out to predict each users' MBTI personality type from the 50 social media posts. Our algorithm takes in an excerpt of text as input and outputs the predicted MBTI personality label (e.g. ENTJ). We apply 2 methods of classical Supervised Learning for this task. Then we perform comparisons and analysis on their resulting error and accuracy to find the method that is most effective for this problem.

2. RESEARCH QUESTION

- Which is the most common MBTI personality type according to the dataset?
- What are the patterns that can be detected in specific types and their style of writing?
- What are the words used most frequently in specific types?
- Is there a correlation amongst the personality type? (i.e. whether each letter type is independent or dependent of/on one another)?
- Is data classification imbalanced? If yes, what is the imbalanced data classification problem?
- Define target variable for the dataset
- Learn and build a typical type of Machine Learning model to predict MBTI personality categories and evaluation metrics.

3. IMPLICATION OF RESEARCH

For time immemorial, man has remained mystery to himself as well as others- he longs to get a bit of insight as to who he really is like.

Furthermore, as our communication happens to be online these days, we are interested in finding out if there is a strong relationship between one's use of language online and his/her real personality.

There are two main implications of this research:

- **Personality tests for self-**

Results from personality tests can make us feel as if we have a sense of agency over our lives.

-Chris Madden

A single glance at something that promises to define the “ME” and one has an urge to chomp on the clickbait as if it were a tantalizing meal. One could take what these tests say too seriously or find it fun to hear tidbits of feedback about oneself that resonate as true. It’s always fascinating to learn about oneself.

Moreover, the research could allow us to know about the possibility of user’s ‘online persona’ as distinct from their in-person one, which suggests that people have a likelihood of behaving in a completely different way online.

- **Personality tests at work-**

This algorithm could support HR Department, managers and employers from companies and organizations in hiring and training process. The MBTI test is the most common and widely used personality metric by the fortune 100 companies, universities, military, hospitals to sort-of-summarize who you really are and what would you bring to the table.

This study could also be beneficial for psychiatrists in their professional work. Social media messages, being a method of communication with its own quirks and styles of language use distinct from prose or speech, contain a certain amount of representational power and reflect the personality of the author.

This project would help its audience learn more about personality tests, like MBTI in this case. And would save them substantial amount of time that they would otherwise waste in taking overwhelmingly lengthy tests.

Furthermore, this project can help its audience know its personality type, and hence, its strengths/weaknesses.

4. DATA DESCRIPTION

We obtained our data from the (MBTI) Myers-Briggs Personality Type Dataset from Kaggle (URL:kaggle.com/datasnaek/mbti-type). The dataset contains more than 8600 observations (people), with two columns: the Myers-Briggs where each observation gives a person's:

- Myers-Briggs personality type (as a 4-letter code)
- An excerpt containing the last 50 posts (each entry separated by "|||")

The posts are drawn from the PersonalityCafe online forum, a platform for all kinds of conversations and discussions, and they obtain the labels by allowing the user to input MBTI type as account info. The personality type of a user and 50+ user posts are stored as strings. Below is an abbreviated example of the first 5 records of the dataset:

	type	posts
0	INFJ	'http://www.youtube.com/watch?v=qsXHcwe3krw|||...
1	ENTP	'I'm finding the lack of me in these posts ver...
2	INTP	'Good one _____ https://www.youtube.com/wat...
3	INTJ	'Dear INTP, I enjoyed our conversation the o...
4	ENTJ	'You're fired.|||That's another silly misconce...

Figure 2

Data Set Excerpt

5. SUMMARY STATISTICS

As mentioned, the dataset comes with just two columns: The MBTI type itself and 50 posts made by the person of the said MBTI type. The dataset has no null values, which means there is no need for cleaning the data, and with the shape = (8675, 2).

The dataset is quite skewed and is not uniformly distributed among the 16 personality types. For example, the most common label, INFP, occurs 1832 times whereas the least frequent, ESTJ, only occurs 39 times (a ratio of 97.9%).

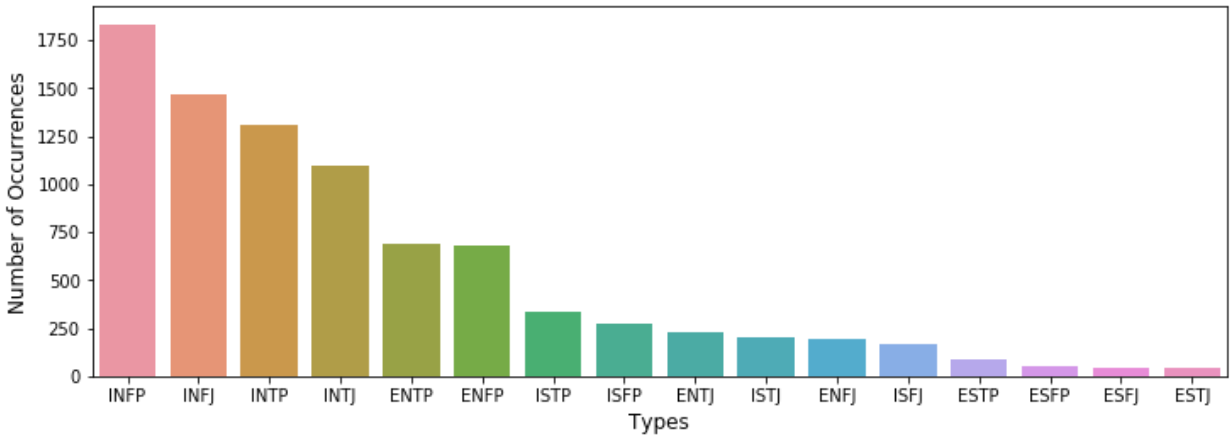

Figure 3

Bar Plot showing each personality type

For smooth work on the dataset, an assumption made is that each letter type is independent of other types i.e. A person's introversion/extraversion is totally not related to their judgement/perception. Nevertheless, we want to still test them and below is the result:

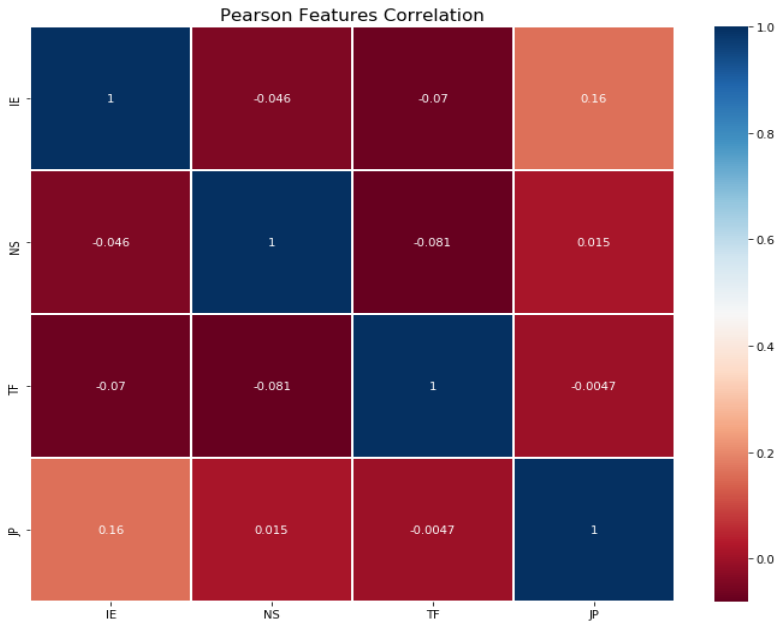

Figure 4

Pearson Features Correlation Map

The correlation for all 4 categories is very close to 0, which indicates that type letters have little to no correlation.

The next idea is counting how many words per comment (we take mean values) in each group. It could be even more helpful to count the number of weblink per comment or question mark per comment for each of personality group. As we hope that we can gain some helpful insights from these findings later. *The table below is the result that we've got for first 5 types of personality groups.*

	type	posts	words_per_comment	variance_of_word_counts	http_per_comment	qmark_per_comment
0	INFJ	'http://www.youtube.com/watch?v=qsXHcwe3krw|||...	11.12	135.2900	0.48	0.36
1	ENTP	'I'm finding the lack of me in these posts ver...	23.40	187.4756	0.20	0.10
2	INTP	'Good one _____ https://www.youtube.com/wat...	16.72	180.6900	0.10	0.24
3	INTJ	'Dear INTP, I enjoyed our conversation the o...	21.28	181.8324	0.04	0.22
4	ENTJ	'You're fired.|||That's another silly misconce...	19.34	196.4576	0.12	0.20

Figure 5

Word Count Excerpt in each group

We also tried to explore the dataset in different ways such as joint plots, pair plots and heat maps to explore relationship between data. However, we still believe that word clouds would be the most effective method to visualize text data like this dataset. We produced 16 Word Clouds for 16 groups of personality. These word clouds are generated such that the size of each word is proportional to its appearance frequency in the top posts. We consider these word clouds to be illustrative of some of the unique ways that different MBTIs use language.

Figure 6

Word Cloud of 16 MBTI types

6. METHODOLOGIES EMPLOYED TO CONDUCT ANALYSIS

When it comes to performing machine learning, trying to distinguish between two categories is much easier than distinguishing between 16 categories. Dividing the data in 2 small groups will perhaps be more useful when it comes to accuracy, so we tried to re-classify the dataset into 2 groups "I-E" standing for Introverts vs Extroverts. We took 90% of the records as train dataset and the rest is for testing.

We performed 2 different classification techniques: **Naïve-Bayes and Logistic Regression** algorithm to the data created in the previous phase to classify and predict the user's personality accordingly. After the 2 models are run, we will compare the accuracy rates of the 2 to find out the final one which has the higher predictive accuracy rate.

7. ANALYSIS RESULT AND EXPLANATION

After doing some research about algorithms for MBTI type classification, we know that Logistic Regression is the one of the best method with high accuracy. Within the limitation of the research, we adopted two common approaches which are Logistic Regression and Naïve Bayes to illustrate the results for MBTI type classification. Based on the accuracy evaluation on the train data, we are led to the conclusion that Logistic Regression have the highest accuracy. In other word, this method is a better algorithm for our dataset.

Algorithm	Accuracy
Logistic Regression	77%
Naive Bayes	74%

Table 1.

Accuracy Evaluation of each algorithm

Topics most frequently spoken by each personality type on a website can be known and content can be suited accordingly to increase traffic.

Pre - Employment assessment:

MBTI personality test helps companies understand what a potential employee will need from them as they fulfill their role.

- If taken before an interview, it helps in narrowing the selection of candidates and helps decides which respondents should be invited for the interview.
- The test can also aid in helping managers decide which questions they should ask the candidate when they do come for an interview.
- With only having a limited amount of time in an interview, candidate skills and abilities can often be overlooked on a CV or face-to-face, however a personality test offers a deeper insight into how they might fit into a company work culture.
- There is often a lower turnover of staff within a company if applicants have traits like the rest of the workforce.
- Personality tests usually detect interpersonal characteristics that may be required for some jobs.

References:

1. https://www.kaggle.com/datasnaek/mbti-type
2. Mining Text data by Charu C. Aggarwal and Cheng Xiang Zhai
3. Natural Language Processing with Python by Steven Bird, Ewan Klein and Edward Loper