INDIRA GANDHI
DELHI TECHNICAL UNIVERSITY
FOR WOMEN

MBTI PERSONALITY PREDICTION

(BI Project Report)
(Team 5)

Submitted By
Akanksha Singh- 01304092018
Arshiya- 04604092018
Bhavana sinha-02804092018
Harshita Dwivedi- 01904092018
Sapna Rai- 03004092018
Shivani Giri- 05404092018

Under the supervision of
Mr. Rishabh Kaushal
Assistant Professor
Department of Information Technology

STUDENT UNDERTAKING

Dated: ....

This is to undertake that the work titled MBTI Personality Prediction in this Minor Project Report as part of 4rd Semester in MCA (Information Technology) with specialization in during January – May, 2020 under the guidance of Rishabh Kaushal . The report has been written by us in my own words and not copied from elsewhere. This report was submitted to plagiarism detection software on 25-05-2020 and percentage 7% similarity found was 93% unique, similarity report attached as Appendix. Anything that appears in this report which is not my original has been duly and appropriately referred / cited / acknowledged. Any academic misconduct and dishonesty found now or in future in regard to above or any other matter pertaining to this report shall be solely and entirely my responsibility. In such a situation, I understand that a strict disciplinary action can be undertaken against me by the concerned authorities of the University now or in future and I shall abide by it.

Student Signature
Student Name

Date-of-Submission, New Delhi

DEPARTMENT OF INFORMATION
TECHNOLOGY
INDIRA GANDHI DELHI TECHNICAL
UNIVERSITY FOR WOMEN
KASHMERE GATE, DELHI - 110006

Dated: ....

CERTIFICATE

This is to certify that the work titled MBTI personality prediction submitted by Team 5 in this project report as part of 4th Semester in MCA (Information Technology) with specialization in during January – May, 2020, done under my guidance and supervision. This work is her original work to the best of my knowledge and has not been submitted anywhere else for the award of any credits / degree whatsoever. The work is satisfactory for the award of Minor Project credits. Name and Signature of Faculty Advisor Designation Department of Information Technology Indira Gandhi Delhi Technical University for Women

Name and Signature of Faculty Advisor
Designation
Department of Information Technology
Indira Gandhi Delhi Technical University for Women

ACKNOWLEDGEMENT

It is high privilege for us to express our deep sense of gratitude to Assistant professor Rishabh Kaushal who helped us in the completion of the project. My special thanks to all Batch mate Seniors of Indira Gandhi Delhi Technical University, New Delhi for helping us in the completion of project work and its report submission.

Student Name

MBTI Personality Prediction

Akanksha Singh
Arshiya
Bhavana sinha
Harshita Dwivedi
Sapna Rai
Shivani Giri

May 2020

Abstract

MBTI personality types can be predicted through many ways. The most commonly used methodology has been questionnaires that are time consuming and needs the focus of participant.This project will explore the area of predicting personalities without questionnaires. People are increasingly using digital platforms like facebook, twitter ,etc. This gives us an opportunity to study if there's a way to predict their personality using these platforms.

Contents

1 Introduction

In the field of psychology, personality is studied as it speaks volumes about how people behave in their life. As the world is advancing, people are using digital platforms like social medias to express themselves. Personality can be predicted using different models. One such model is the Myers-Briggs Type Indicator (MBTI) where personalities are divided into 16 different types. The MBTI divides the traits into four classes such as: Introversion (I) or Extroversion (E), Sensing (S) or Intuition (N), Thinking (T) or Feeling (F), and Judging (J) or Prospecting (P) . Eg: INFP. In this project, we want to study the correlation between the language of individuals used on their social medias and their respective personality traits. This will help us know that to what extent can we predict personality traits from various linguistic features.

1.1 Problem Statement & Objectives

The MBTI tests use many multiple choice questions to determine the personality of an individual. But, this approach is time consuming and requires people to be focused enough to answer correctly. Thus, we think of minimizing the efforts of users and making it more efficient to predict a personality. We can thus model this as a classification problem. A successful implementation of such a classifier would demonstrate a good connection between linguistic features and potential personality in general. This won't just help users know their personality but can also be used in psychoanalysis to help find the relation between natural language and personality type.

- Predicting MBTI personality type using texts from social medias.
- Study the relation between natural language and MBTI personality.

1.2 Motivation

If we happen to find a correlation between natural language and MBTI personalities, it can be a contribution towards psychoanalysis.People don't always realise how do they think like and a lot of what they post on social media may have a significant relation with their true selves. This project will also help people know about themselves without solving a lot of questions that many a people find tiresome and thus don't participate into finding their ownselves and if they participate then a lot of times they don't answer correctly. Employers can find their employees using public information provided by employees if we achieve sufficient accuracy in this project.

1.3 Scope and Limitation

As we mentioned above, the project has scope in psychoanalysis however it has certain limitations.

- The project is based on texts.It does not include images, URLs etc. People don't always express their original thoughts and writing texts isn't as interesting for everyone.They prefer sharing images, blogs, etc. written by another person than writing something about the topic on their own. This makes it difficult for our code to find sufficient data. If we include other factors like the images they share, the types of blogs they share, the content of articles they shared, the connects they have or prefer, their likes or dislikes. This may add value to our accuracy and provide sufficient data.

2 Related Work

Mihai Gavrilescu [1] and Champa H N [2] used deep feed forwards neural networks for small datasets that are textual.This was proven to be successful in predicting personality.They used a 3 layered feed forward architecture on textual data which is handwritten.Even though, this model that they used had handwritten features than just text, they proved it that MBTI personalities can be predicted using deep neural architectures.

3 Methodology

3.1 Dataset Description

For this project, we used the Myers-Briggs Personality Type Dataset available on Kaggle [1].This data which is available on Kaggle was collected through the PersonalityCafe forum that provides a large number of people,their respective MBTI personality types and what they have posted.

Details	**Count**
Number of instances(posts)	8,675
Number of unique attributes	16

Table 1: Details of the dataset.

Every dataset also comprises of data attributes. Table 2 describes attributes of data. In case of supervised learning, clearly mention which attribute(s) would be considered as the *labels.*

Data Attributes	**Brief Explanation**
Personality type	Posts written by that personality type

Table 2: Details of Data Attributes.

3.2 Description of attributes

- I - Introversion
- E - Extroversion
- S - Sensing
- N - Intuitive
- F - Feeling
- T - Thinking
- P - Prospecting
- J - Judging

A combination of I or E, S or N , F or T and P or J gives us an MBTI personality.

3.3 Data Pre-processing

3.3.1 Data Proportion

As shown in Figure 1 given below, the data is clearly not in proportion.The number of posts are higher for INFP than any other type.

[1] https://www.kaggle.com/datasnaek/mbti-type

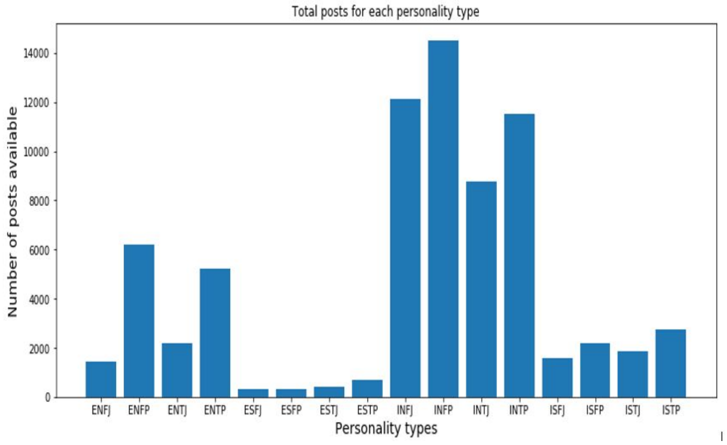


Figure 1: Graphical Representation of available data.

3.3.2 Data Cleaning

Since our project is strictly based on text, removal of URLs was necessary. We also removed all the NULL values. Next step was to remove common fillers like "or" , "a", "the", etc. This we did using python's NLTK.In order to preserve the data, we replace the null values with the hyphen symbol.

3.3.3 Lemmatization

We used imported WordNetLemmatizer from nltk.stem to lemmatize the text which means that infected forms of the same word are treated as one form of the root word(e.g. "running", "ran", "run" all become "run").

3.3.4 Tokenization

Tokenization is necessary. Here, we split the available text into words using python's Natural Language ToolKit (NLTK).We tokenized further find the useless words.To apply this, we needed bag of words. We have defined a set of useless words with nltk.stopwords to tokenize correct posts.

3.3.5 Bag of words

We built bag of words by removing all the stopwords and punctuation marks in order to have only necessary data on which we can apply our machine learning algorithm.

3.3.6 Splitting

Since each number of personality type has different number of posts,they must split accordingly. We have split the data into 2 parts. 80 percent is for training and 20 percent is for testing.

3.4 Proposed Methodology

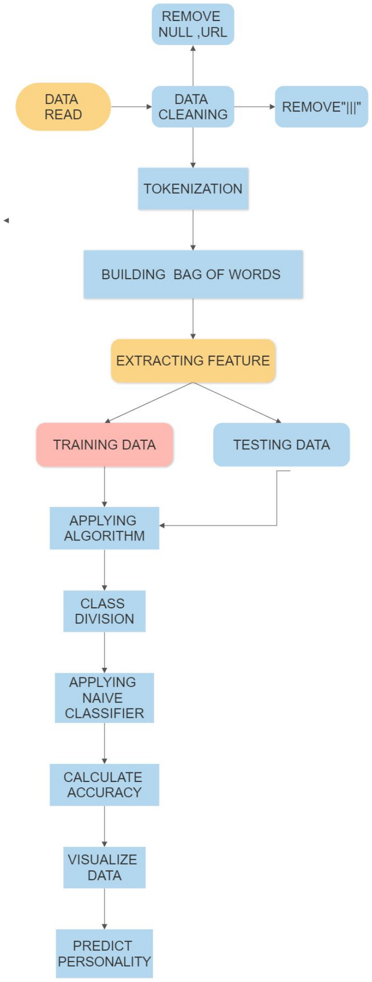


Figure 2: Flowchart

- Domain : Predictive analytics
  - Task of classification and prediction is the key

- - Predicting/filling the information that is unknown
  - In our project, we are predicting the personality trait of a person

- Machine learning task : Supervised learning
  - Given data and associated target response, model is trained and then is used to predict the correct response for a new data.
    In our project,
    - * Data : post of user(post column in dataset)
    - * Target response : personality type of user (type column in dataset)
    - * New data : New user whose personality we want to predict

- Type of problem under supervised learning : Classification(multi-class classification)
  - Class : 16 personality type , a user is assigned one out of these class

3.4.1 Algorithm

The algorithm that we use here is **Naive Bayes Classifier Algorithm**.This is a classification technique which is based on Bayes Theorem with an assumption of independence among the given data values/set. In our project , text given in several posts on social media platforms are the data values which forms our data set . This works well on the text/categorical data instead of numeric data. A classifier under the supervised learning based on probabilistic logic(bayes theorem). In lay-man language, we can say that a existence of the number of posts of a particular person/ individual is unrelated / independent from the existence of number of posts of a another person /individual and that's the assumption in naives bayes classifier. For each attribute from each class set ,it uses probability to make predictions.

$$\{X1, X2, ..., Xn\} - -- \geq \{C1, ..., Ck\} \tag{1}$$

In our project,
X1 denotes the no of posts of a particular person/ individual.
X2 denotes the no of posts of another person/ individual.
Xn denotes the no of posts of nth person/individual .
C1is the probability of the number of posts describing a particular trait personality trait.
Ck is the probability of the number of posts describing a kth – personality trait . The data model which is yielded is called as Predictive model with probabilistic problems at foundation.
Bayes theorem gives us a way to calculate posterior probability by the given equation:-

Posterior probability= Likelihood Class Prior Probability /Predictor Prior Probability*

Input=50 posts

Trained Data	**58.354826823876195**
Test Data	**10.4463235294117645**

Output:- The language used in the number of posts (twitter posts) reflecting/describing each 16 types/ traits of personality and predicting the personality trait people possess.

To increase the accuracy of our model, we took 4 classifiers of personality traits to classify the individual's personality(using MBTI)

Input=50 posts

Data	**Introvert-Extrovert**	**Intuition-Sensing**	**Thinking-Feeling**	**Prospecting-Judging**
Trained	**57.401437**	**67.658438**	79.504422	73.613670
Test	**49.705882**	**73.566176**	53.198529	48.768382

Output:-
The language used in the number of posts (twitter posts) reflecting/describing each(4)classifiers of personality and predicting the MBTI Trait using the classifiers.

Algorithm 1

```
 split ← []
for i in range 16 do
    split += [len(features[i])*0.6]
    split ← $np.array(split, dtype = int)$end for

        train ← []
        for i in range 16 do
            train +=features[i][:split[i]

        end for
        sentiment$_c lassifier = NaiveBayesClassifier.train(train)$
        nltk.classify.util.accuracy(sentiment$_c lassifier, train) * 100$
        test ← []
        for i in range 16 do
            test +=feautures[i][split[i]:]

        end for
        nltk.classify.util.accuracy(sentiment$_c lassifier, test) * 100$
```

4 Experiment Setup & Results

We are splitting our data set into training and test data. We are calculating accuracy in 4 trials (50:50, 60:40, 70:30 ,80:20) . This splitting is done on complete dataset where we have 16 classes each class representing a personality type. The accuracy through this method turned out to be 10% approximately as shown in Figure 3.

Hence, instead of selecting all 16 personalities as a unique feature, we decided to simplify the dataset. The MBTI personality type divides everyone into 16 personality types across 4 axis.

1. Introversion(I) or Extroversion(E)
2. Intuition(N) or Sensing(S)
3. Thinking (T) or Feeling (F)
4. Prospecting (P) or Judging (J)

Now we have 4 classes, we create 4 classifiers (Naive Bayes Classifier to classify the person into a particular personality) . We got approximately 53% accuracy after classifying the personality types into 4 classes rather than 16 types, as shown in Figure 4.

In Figure 5, the graph shows which trait has higher percentage than the other and thus chooses the higher trait to predict the personality type.

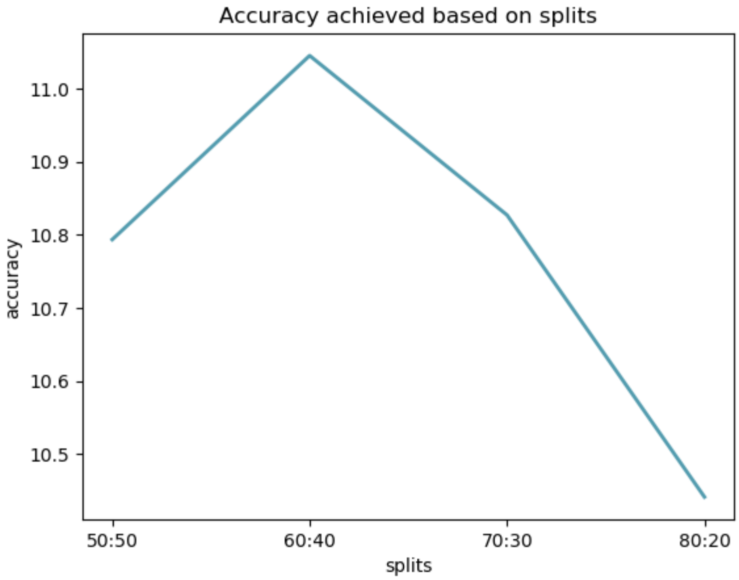


Figure 3: Splits vs Accuracy

In Figure 6,we tried predicting the personality of Barack Obama based on his tweets and we got INFJ as shown in Figure 6 which is different from his original personality which is ENFJ.

5 Conclusion and Future Work

There is a slight difference between the personality predicted by the model and the personality predicted by 16personalities.com.This might be because:

1. We have not scraped the profile but have copied few posts of the user into the test file.
2. We are using Naive Bayes classifier, the accuracy of which is 50%,so according to the accuracy of the model,we are getting a good result.
3. We didn't proportionalise the data and thus it's more likely that our code predicts INFP or traits related to INFP as it has the highest number of posts. Our data is very imbalanced.

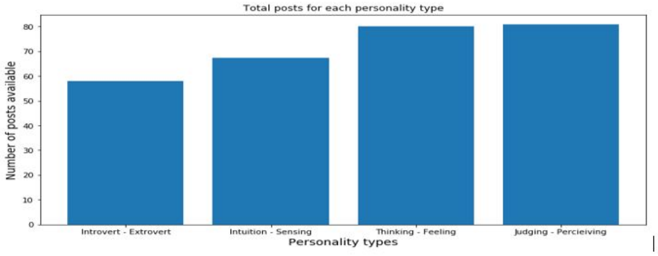


Figure 4: After classification into 4 classes

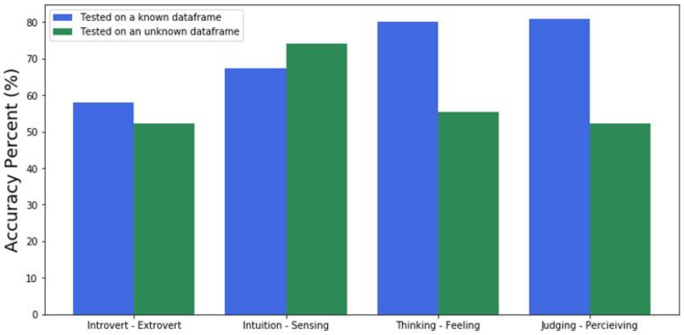


Figure 5: Model Classifying trait

References

[1] Mihai Gavrilesku. Study on determining the myers briggs personality type based on individual's handwriting. *The fifth IEEE International Conference on E-Health and bioengineering*, 11,2015.

[2] Champa H N and Dr. K R Anandakumar. Artificial neural network for human behaviour prediction through handwriting analysis. *International Journal of Computer Applications*, 2010.

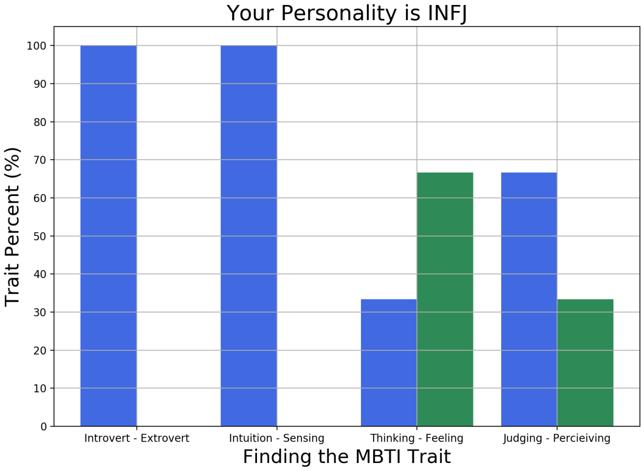
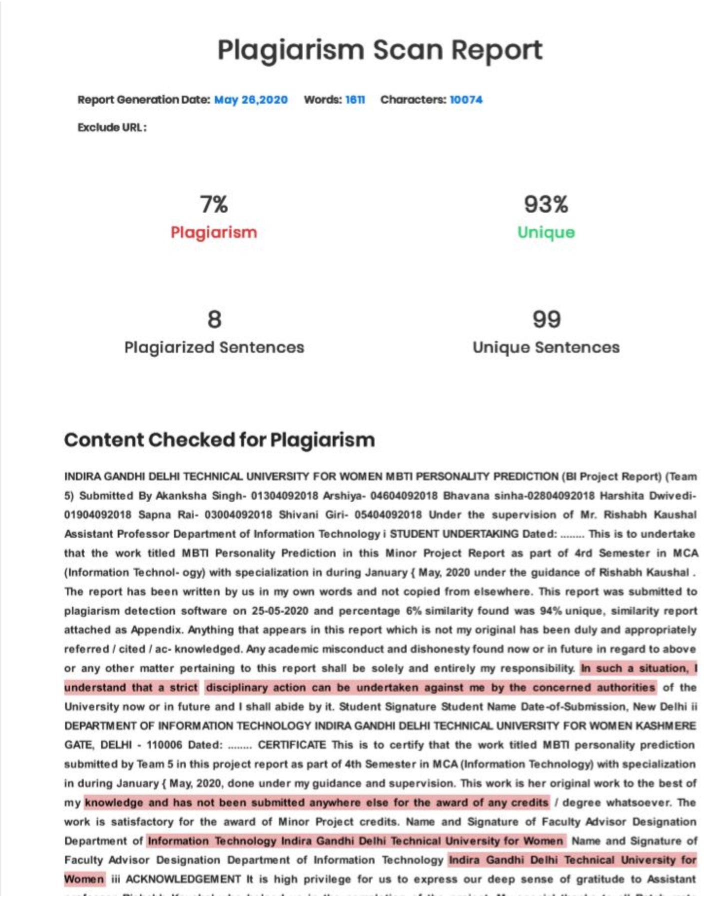


Figure 6: Personality prediction of Barack Obama

6 Appendix: Similarity Report



Plagiarism Scan Report

Report Generation Date: May 26,2020 Words: 1611 Characters: 10074

Exclude URL:

7%	93%
Plagiarism	Unique
8	99
Plagiarized Sentences	Unique Sentences

Content Checked for Plagiarism

INDIRA GANDHI DELHI TECHNICAL UNIVERSITY FOR WOMEN MBTI PERSONALITY PREDICTION (BI Project Report) (Team 5) Submitted By Akanksha Singh- 01304092018 Arshiya- 04604092018 Bhavana sinha-02804092018 Harshita Dwivedi-01904092018 Sapna Rai- 03004092018 Shivani Giri- 05404092018 Under the supervision of Mr. Rishabh Kaushal Assistant Professor Department of Information Technology i STUDENT UNDERTAKING Dated: .... This is to undertake that the work titled MBTI Personality Prediction in this Minor Project Report as part of 4rd Semester in MCA (Information Technol- ogy) with specialization in during January { May, 2020 under the guidance of Rishabh Kaushal . The report has been written by us in my own words and not copied from elsewhere. This report was submitted to plagiarism detection software on 25-05-2020 and percentage 6% similarity found was 94% unique, similarity report attached as Appendix. Anything that appears in this report which is not my original has been duly and appropriately referred / cited / ac- knowledged. Any academic misconduct and dishonesty found now or in future in regard to above or any other matter pertaining to this report shall be solely and entirely my responsibility. In such a situation, I understand that a strict disciplinary action can be undertaken against me by the concerned authorities of the University now or in future and I shall abide by it. Student Signature Student Name Date-of-Submission, New Delhi ii DEPARTMENT OF INFORMATION TECHNOLOGY INDIRA GANDHI DELHI TECHNICAL UNIVERSITY FOR WOMEN KASHMERE GATE, DELHI - 110006 Dated: .... CERTIFICATE This is to certify that the work titled MBTI personality prediction submitted by Team 5 in this project report as part of 4th Semester in MCA (Information Technology) with specialization in during January { May, 2020, done under my guidance and supervision. This work is her original work to the best of my knowledge and has not been submitted anywhere else for the award of any credits / degree whatsoever. The work is satisfactory for the award of Minor Project credits. Name and Signature of Faculty Advisor Designation Department of Information Technology Indira Gandhi Delhi Technical University for Women Name and Signature of Faculty Advisor Designation Department of Information Technology Indira Gandhi Delhi Technical University for Women iii ACKNOWLEDGEMENT It is high privilege for us to express our deep sense of gratitude to Assistant

Figure 7: Plagiarism report