
***Human Personality Prediction from text using Myers
Briggs Type Indicator***

By

Momin Ali	ID:17182103011
Shuva Chakraborty	ID:17182103014
Md.Arifur Rahman	ID:17182103038
Shakil Ahmed	ID:17182103010
Arifur Rahman Kawser	ID:17182103012

Submitted in partial fulfillment of the requirements of the degree of

**Bachelor of Science in
Computer Science and Engineering**



DEPARTMENT of COMPUTER SCIENCE AND ENGINEERING
BANGLADESH UNIVERSITY of BUSINESS AND TECHNOLOGY

JUNE 2022

Declaration

We declare that this work has been presented on our own and has been propagated by us as the result of our original research.

We confirm that:

- This work is done mainly while in candidacy for a research degree at this university.
- This report has not been submitted previously for any degree at this university or any other educational institution.
- We have lifted the work from others; the source is always given. With the exclusion of such quotations, this thesis is entirely our work.

Momin Ali

ID: 17182103011

Signature

Shuva Chakraborty

ID: 17182103014

Signature

Md. Arifur Rahman

ID: 17182103038

Signature

Shakil Ahmed

ID: 17182103010

Signature

Arifur Rahman Kawser

ID: 17182103012

Signature

Approval

This report “**Human Personality Prediction from text using Myers Briggs Type Indicator**” submitted by **Momin Ali, Shuva Chakraborti, MD. Arifur Rahman, Shakil Ahmed, and Arifur Rahman Kawser ID no: 17182103011, 17182103014, 17182103038, 17182103010, and 17182103012** Department of Computer Science and Engineering (CSE), Bangladesh University of Business and Technology (BUBT) under the supervision of **T.M. Amir-Ul-Hoque Bhuiyan, Assistant Professor**, Department of Computer Science and Engineering (CSE) has been accepted as appeasement for the partial fruition of the requirement for the degree of Bachelor of Science (B.SC.) in Computer Science and Engineering and endorsed as to its contents.

Supervisor:

T.M. Amir-Ul-Hoque Bhuiyan

Assistant Professor

Department of Computer Science and Engineering (CSE)

Bangladesh University of Business and Technology (BUBT)

Mirpur-2, Dhaka-1216.

Md. Saifur Rahman

Assistant Professor & Chairman

Department of Computer Science and Engineering (CSE)

Bangladesh University of Business and Technology (BUBT)

Mirpur-2, Dhaka-1216.

Dedication

Dedicated to our parents and teachers for all their love and inspiration.

Acknowledgment

We would like to express our heartfelt gratitude to the almighty Allah who offered upon our family and us kind care throughout this journey until the fulfillment of this research. Also, we express our sincere respect and gratitude to our supervisor, TM Amir-Ul-Haque Bhuiyan, Assistant Professor Department of Computer Science and Engineering, Bangladesh University of Business and Technology (BUBT). Without his guidance, this research work would not exist. We are grateful to him for his excellent supervision and for putting his utmost effort into developing this project. We owe him a lot for his assistance, encouragement, and guidance, which has shaped our mentality as a researcher. Finally, we are grateful to all our faculty members, office department, BUBT, for making us compatible to complete this research work with the proper guidance and support throughout the last four years.

Abstract

This paper provides a review of personality predictions using the Myers Briggs Type Indicator(MBTI) .The use of social media increases day by day. People share their emotions and information via social media posts. Those posts are important assets for research in the field of behavior learning of human personality. In this paper, we attempt to build a system that can predicat a person's personality using their social media posts using Myers Briggs Type Indicator(MBTI).while other previous research used various types of old machine learning models, this research we used XGBoost classifiers for batter accuracy The results succeed to perform the accuracy of IE:78.90%, NS:86.17%, FT:74.05%, and JP:65.81%.

Table of Contents

1	Introduction	1
1.1	Introduction	1
1.2	Problem Statement	4
1.3	Problem Background	4
1.4	Research Objectives	4
1.5	Motivation	5
1.6	Flow of the Research	5
1.7	Significance of the Research	6
1.8	Research Contribution	7
1.9	Thesis Organization	7
1.10	Summary	8
2	Background	9
2.1	Introduction	9
2.2	Literature Review	9
2.3	Problem Analysis	15
2.4	Summary	15
3	Proposed Model	16
3.1	Introduction	16
3.2	Feasibility Analysis	16
3.3	Requirement Analysis	16
3.4	Research Methodology	16
3.5	Design, Implementation, and Simulation	18
3.6	Summary	18
4	Implementation, Result And Discussion	19
4.1	System Requirement	19
4.1.1	Google Colab	19
4.1.2	XGBoost Classifier	20
4.1.3	Python Libraries	21

4.2 Testing Process	23
4.3 Results and Discussion	23
4.3.1 Model Accuracy	23
4.3.2 Result	24
4.4 Summary	24
5 Standards, Impacts, Ethics, and Challenges	25
5.1 Sustainability	25
5.2 Impacts on Society	25
5.3 Ethics	26
5.4 Challenges	26
5.5 Summary	27
6 Standards, Impacts, Ethics, and Challenges	28
6.1 Design Constraints	28
6.2 Component Constraints	28
6.3 Budget Constraints	28
6.4 Summary	29
7 Schedules, Tasks, and Milestones	30
7.1 Timeline	30
7.2 Gantt Chart	30
8 Conclusion	32
8.1 Conclusion	32
8.2 Future Works and Limitations	32
References	33

List of Figures

1.1 The figure illustrates the flow of the thesis work.	6
3.1 Before Pre-Processing of Data	17
3.2 After Pre-Processing of Data.	17
4.1 Testing Process.	23
4.2 Model Accuracy	23
4.3 Result	24
7.1 Gantt Chart	31

Introduction

1.1 Introduction

Human Personality Prediction is an age-old topic. Personality is the combination of an individual's behavior, emotion, motivation, and characteristics of their thought patterns. Different human personality predictions have been done from Facebook data and handwritten data [1]. The daily behavior of the people exposes their personality traits. With the manifestation of the social platforms, some perspectives of this behavior are being recorded in their social accounts. Those give necessary input to develop algorithms that can predict the personality traits of people. We're developing an MBTI personality classifier that uses XGBoost models to predict a person's personality based on the 50 recent social media posts per user as input. We find correlations between a person's MBTI personality types using their social media posts. We have used a decent 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.

There are four major types of Human Personality traits Those are:

Extraversion (e) – Introversion (i): Extraversion is a measure of how energetic, sociable, and friendly a person is. Extraverts are commonly understood as being a 'people's person' drawing energy from being around others directing their energies toward people and the outside world.

Introversion is a personality trait characterized by a focus on internal feelings rather than on external sources of stimulation.

Sensing (S) – Intuition (N): Sensing (S) and Intuition (N) are how you process information. Someone who is strong in sensing lives in the now and enjoys facts. While being Intuitive means you try and find the deeper meaning in things.

Thinking (T) – Feeling (F): How do people make decisions based on the information that they gathered from their sensing or intuition functions. People who prefer thinking place a greater emphasis on facts and objective data.

They tend to be consistent, logical, and impersonal when weighing a decision. Those who prefer feeling are more likely to consider people and emotions when arriving at a conclusion.

Judging (J) – Perceiving (P): People with a Judging preference want things to be neat, orderly, and established. The Perceiving preference wants things to be flexible and spontaneous. Judgers want things settled, Perceivers want things open-ended.

The MBTI Types

Each type is then listed by its four-letter code:

- **ISTJ:** (introversion, sensing, thinking, judgment) is a four-letter code representing one of the 16 personality types found on the Myers-Briggs Type Indicator (MBTI). People with an ISTJ personality type tend to be reserved, practical and quiet.
- **ISTP :** (introverted, sensing, thinking, perceiving) is one of the 16 personality types identified by the Myers-Briggs Type Indicator (MBTI). People with ISTP personalities enjoy having time to think alone and are fiercely independent.
- **ISFJ :**(Introverted, Sensing, Feeling, Judging) ISFJ is one of the 16 personality types identified on the Myers-Briggs Type Indicator (MBTI), the personality test developed by Isabel Myers and her mother Katherine Briggs based on the theories of psychoanalyst Carl Jung. People who have ISFJ personalities are known for being warm-hearted, responsible, and reserved.
- **ISFP :** (Introverted, Sensing, Feeling, Perceiving) ISFP is a four-letter code representing one of the 16 personality types identified by the Myers-Briggs Type Indicator. People with an ISFP personality are frequently described as quiet, easy-going, and peaceful.
- **INFJ:** (Introverted, Intuitive, Feeling, Judging) INFJ is one of the 16 personality types identified by the Myers-Briggs Type Indicator (MBTI). Scoring as an INFJ means your personality type is best described as Introverted, Intuitive, Feeling, and Judging.
- **INFP:** (Introverted, Intuitive, Feeling, Perceiving) INFP is a four-letter abbreviation for one of the 16 personality types identified by the Myers-Briggs Type Indicator. The INFP personality type is often described as an "idealist" or "mediator" personality. People with this kind of personality tend to be introverted, idealistic, creative, and driven by high values.
- **INTJ:**(introverted, intuitive, thinking, and judging) is one of the 16 personality types identified by a personality assessment called the Myers-Briggs Type Indicator (MBTI).

Sometimes referred to as the "Architect" or the "Strategist," people with INTJ personalities are highly analytical, creative, and logical.

- **INTP** :(introverted, intuitive, thinking, perceiving) is one of the 16 personality types described by the Myers-Briggs Type Indicator (MBTI). People who score as INTP are often described as quiet and analytical.
- **ESTP** : (Extraverted, Sensing, Thinking, Perceiving) ESTP is one of the 16 personality types identified by the Myers-Briggs Type Indicator (MBTI). People with this personality type are frequently described as outgoing, action-oriented, and dramatic.
- **ESTJ**:(Extraverted, Sensing, Thinking, Judging) ESTJ is one of the 16 personality types identified by the Myers-Briggs Type Indicator (MBTI). ESTJs are often described as logical, take-charge kinds of people.
- **ESFP** : (extraverted, sensing, feeling, perceiving) is one of the 16 personality types identified by the Myers-Briggs Type Indicator.¹ People with ESFP personality types are often described as spontaneous, resourceful, and outgoing.
- **ESFJ**: (Extroverted, Sensing, Feeling, Judging)ESFJ, also known as "The Caregiver" or "The Consul," is one of the 16 personality types identified by the Myers-Briggs Type Indicator. People with an ESFJ personality type tend to be outgoing, loyal, organized, and tender-hearted.
- **ENFP**: (Extraverted, Intuitive, Feeling, Perceiving)The ENFP personality type is one of the 16 different types identified by the Myers-Briggs Type Indicator (MBTI). This acronym stands for Extraverted, Intuitive, Feeling, and Perceiving.
- **ENFJ**: (Extraverted, Intuitive, Feeling, Judging)ENFJ, also known as the giver or protagonist personality, is one of the 16 different personality types identified by the Myers-Briggs Type Indicator. Some other types are known by the acronyms ESFJ, ENFP, INFP, ISFJ, and INTP. People with ENFJ personality types are often described as warm, outgoing, loyal, and sensitive.
- **ENTP** :(Extroverted, Intuitive, Thinking, Perceiving)ENTP is one of the 16 different personality types identified by the Myers-Briggs Type Indicator. People with this personality type are often described as innovative, clever, and expressive. ENTPs are also known for being idea-oriented, which is why this personality type has been described as "the innovator," "the visionary," and "the debater."

- **ENTJ:** (Extraverted, Intuitive, Thinking, Judging)The acronym ENTJ represents one of the 16 personality types that are identified by the Myers-Briggs Type Indicator. This popular personality assessment was developed by Isabel Myers and her mother Katherine Briggs. The assessment tool is based on Carl Jung's theory of personality types. Other people often describe people with this type of personality as assertive, confident, and outspoken.

1.2 Problem Statement

We are going to analyze all kinds of social media data to predict personality. It can be social media text. The analysis becomes more committed because of its human behavior analytical power. There are some problems with the text analyzing human personality. In this project, we are going to apply Myers Briggs Type Indicator with XGboost classifier for better accuracy than in previous works. For human personality prediction. We are going to use a large dataset from Kaggle where we can work with quality data.

1.3 Problem Background

Personality is the combination of an individual's behavior, emotion, motivation, and characteristics of their thought patterns. Identifying human personality is a real-world computer vision problem. Recognizing human personality from text is very challenging to identify who is Extraversion (e) – Introversion (i), Sensing (S) – Intuition (N), Thinking (T) – Feeling (F), Judging (J) – Perceiving (P). Most importantly there is no accurate data on a person. Lacking a proper dataset is the critical challenge of our research. We are going to use a large dataset where we can ensure quality data.

1.4 Research Objectives

The objectives of our research work are as follows:

- In the development stage, the project required different types of resources. Including Material (text), Information (text of person), People, Developers, and Technologies (ML platforms).

- Using different approaches in our model to detect personality interacting with different types of data including text .
- We will work on how it can be improved.
- By detecting human personality, it can be predicting any upcoming precursor, able to predict anyone's personality, identify if anyone is in depression, which will help our society to avoid different undesired events
- Our project is related to all the work that has been done on human personality. One by one they are aggregating their ideas.

1.5 Motivations

With the rapid growth of social media, users are getting involved in virtual socialism, generating a huge volume of textual and image content. Considering the contents such as status updates/tweets and shared posts/retweets, liking other posts is reflecting the online behavior of the users. Predicting the personality of a user from these digital footprints has become a computationally challenging problem. To reduce this problem we will apply the XGBoost model. For Human Personality Prediction we should take texts; We are basically focused on the accuracy of our model. We will work on a method that can do personality prediction from people's social media posts.

1.6 Flow of the Research

The research work is being developed in several steps. Firstly we analyzed the research topic and then studied the basic theory of the XGBoost model and Myers Briggs Type Indicator concepts. Then we investigated the application of the XGBoost model on Myers Briggs Type Indicator concepts. We investigated the lack of accuracy in present architectures and motivated them to increase the accuracy of results. Figure 1.1 illustrates the overall steps of the research procedure in the following diagram.

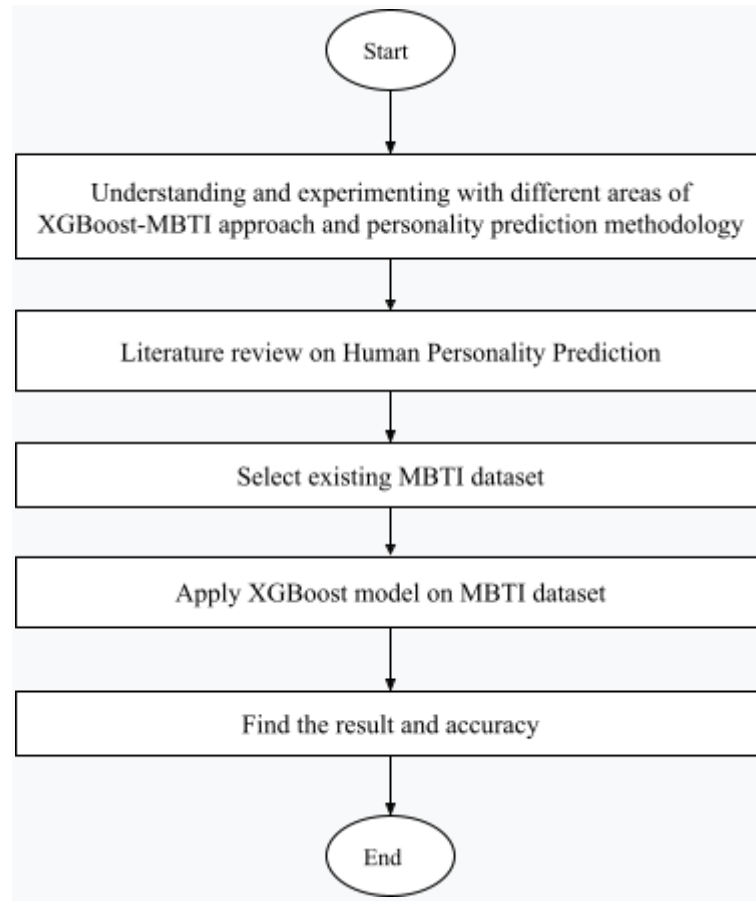


Figure 1.1: The figure illustrates the flow of the thesis work.

1.7 Significance of the Research

We observe that most of the Personality Prediction system datasets on the web are built on the Myers-Briggs Type Indicator (MBTI) traits of humans. These are Extraversion (E) – Introversion (i), Sensing (S) – Intuition (N), Thinking (T) – Feeling (F) Judging (J) – Perceiving (P). Therefore we use **Myers-Briggs Type Indicator (MBTI)** approach to identify individuals' personalities. Nowadays, Depression is a serious problem in society. People are suffering from depression and are on the verge of death. In this issue, human personality prediction systems can help to detect depressed people. It can help parents to identify their children's mentality to guide them. Human personality prediction systems also can identify abnormal people. It will warn us before the accident happens to those abnormal people. This human personality classification problem becomes a state-of-the-art

architecture to estimate human personality prediction and significantly impact society and the country.

1.8 Research contribution

The overall contribution of the research work are :

- The project requires the study of research on the Myers-Briggs personality type indicator and database learning.
- Data collection and information.
- Engineering design and development.
- Knowledge of software engineering and image processing.
- Accurate text will be used.
- We have worked on an XGBoost model
- When we started working on the project we didn't know which model would be good or bad. Which is why we had to work on XGBoost model.
- We will work hard on the human personality by focusing on the best of the model accuracy in our proposal.
- We want to decompose our model's small work in different sub-models

1.9 Thesis Organization

The thesis work is organized as follows.

Chapter 2 highlights the background and literature review on the field of the Human Personality Prediction system. Chapter 3 contains the human personality prediction systems proposed architecture and a detailed walk-through of the overall procedures. Chapter 4 includes the details of the tests and evaluations performed to evaluate our proposed architecture. Chapter 5 explains the Standards, Impacts, Ethics, Challenges, Constraints, Timeline, and Gantt Chart. Finally, Chapter 6

contains the overall conclusion of our thesis work. Chapter 7 illustrates the time schedules that we managed while the thesis work was attended. Finally, Chapter 8 contains the allover conclusion of our thesis work.

1.10 Summary

This chapter includes a broad overview of the problem that we aimed explicitly at our research work's objectives, the background, and the research work's motivation. This chapter also illustrates the overall steps on which we carried out our research work.

Background

2.1 Introduction

Nowadays in psychology, personality is one of the heavily researched and entrancing topics. The concept of predicting anyone's personality has come to the notice of many researchers during the past few years. The majority of existing approaches focus on the Big Five Traits of personality prediction and a few used Myers Briggs Type Indicator (MBTI). However, in the past, no researchers used MBTI with XGBoost properly. In this research work, we focused on better accuracy of human personality prediction by using MBTI along with the XGBoost model.

In this chapter, We have presented a short brief of some of the work by the researchers.

2.2 Literature Review

Human Personality Prediction and MBTI have inducted significant attention in computer vision researchers during the past few years. François Mairesse and Marilyn Walker[14] train the generation of the big five personality style through data-driven parameter estimation using NB,J48,NN ADA, SVM.b Its a generation technique that can target multiple stylistic effects simultaneously andover a continuous scale controlling stylistic dimensions that are commonly understood and thus meaningful to users and application developers. Uses a base generator to produce multiple utterances by randomly varying its parameter.

Menasha Thilakaratne, Ruvan Weerasingheand Sujana Perera [15] leveraging social media data to predict personality traits. They used SVM (Support Vector Machine) Learning Algorithm and my Personality 2007 dataset.The state-of-the-art algorithms design for predicting personality traits exploited only the linguistic features of the social media posts.Accuracies are Openness .3748 Conscientiousness .3386 Extraversion .3524 Agreeableness.2959 neuroticism 0.3955. This model can link up open data cloud DBpedia.

Xiaoli he [16] predicted personality using lexicon based analysis from social media using randomized Logistic Regression and Randomized Lasso method. It uses kaggle , twitter_100g, twitter_500g, twitter_2000g datasets. Main features are using all three types of n grams(unigrams,1-2grams,1-2-3 grams) , extracting binary n grams for each sample with better accuracy. This model compares MBTI and Big five traits from datasets.

Bruce Ferwerda and Marko Tkalcic [17] predict personality using instagram pictures with 44 item BFI personality questionnaires. For each picture it uses color-based and content-based features. Color space used for color based features. Google vision used for content features. In this work the visual features as well as the content features consist of information for personality prediction that

attain similar results.

Jennifer Golbeck, Cristina Robles and Karen Turner [19] predict personality with social media data (facebook). They use two algorithm m5sap rules and a gaussian process. They collect all the data from facebook using 45-question of the Big Five Personality Inventory Traits. It considers two structural features: number of friends and network density.

Hong-Wei Ng and Stefan Winkler [20] formulate the problem of identifying the faces to be removed as a quadratic programming problem, naive method and they use FaceScrub dataset and imdb. And the features: automatically remove outliers from a set of faces and Face Recognition. The accuracy is 0.728 makes it the best for removing outliers among the different methods that they implemented. And the limitation is the false positives term is not used, because the unsupervised One-Class SVM probably misclassified many outliers as nonfaces.

Lauge Sorensen; Mads Nielsen; Pechin Lo; Haseem Ashraf; Jesper H. Pedersen; Marleen de Bruijne [21] were used regions of interest (ROIs) , k nearest neighbor (kNN) classifiers . Experiments on two subsets of the DLCST database that are denoted data A set and B. Training a texture-based classifier to recognize COPD in pulmonary CT images using supervised learning techniques in a fully automatic model are the features. Measures RA and PD, with an AUC of 0.713 compared to 0.596 and 0.598, respectively. And the limitations are less sensitive to inspiration level—a major source of variability in computerized quantitative CT measures.

In the paper “Is an irritable ADHD profile traceable using personality dimensions? Replicability, stability, and predictive value over time of data-driven profiles”, [22] the Recruitment procedure, Diagnostic procedure and exclusion process, Assessment of personality and clinical outcomes methods they used. They used a Big Five questionnaire for children (BFQ-C) as a dataset. And the features are: examined whether an irritable ADHD profile could be identified in a new sample based on personality dimensions with the accuracy Community detection identified three communities at baseline of 38(21%), 73 (41%) and 67 (38%) children. And the limitations are replication studies using both temperament and personality questionnaires will be needed to clarify whether these divergent findings are due to differences in sample characteristics. measurement instruments.

In the paper “Stronger Together: Personality, Intelligence and the Assessment of Career Potential”, [23] they use machine learning algorithms. They collect data from people’s digital data, such as their social media footprint. Both intelligence and personality consistently predict job performance, making them valuable metrics for organizations. Importantly, they also offer a theoretical framework and explanation for individual potential. Personality traits were classified correctly in 40–63% of cases, speech clips achieved accuracies of 70–80% and speech signals such as rate, energy, pitch, silent intervals successfully distinguished between high and low extraverts in 86% of cases. And the limitation is faking and acceptability to applicants.

Leqi Liu, Daniel Preotiuc-Pietro, Zahra Riahi Samani, Mohsen E. Moghaddam, Lyle Ungar [24] Analyzing Personality through Social Media Profile Picture Choice. State-of-the-art, Text-based prediction, Big Five, Face++², EmoVu3 methods have been used here. They used TwitterSurvey, TwitterText as a dataset. Their system provides robust accuracy. Twitter Text dataset Feat=85,Ope=.162,Con=.189,Ext=.180,Agr=.150,Neu=.145 prove their good accuracy. One of the limitations is Analyze Images using interpretable features on Data set orders of magnitude.

C. Sutherland, Xizi Liu, Lingshan Zhang, Yingtung Chu, Julian A. Oldmeadow³, and Andrew W. Young [25] developed the first data-driven non-Western (Chinese) model of facial impressions.

Here three methods have been created through three studies and them. They use “Ambient Images”; Santos & Young, 2005; see Jenkins, White, Van Montfort, & Burton, 2011, as dataset. In their paper the main feature extract is "Facial judgments likely function similarly across culture to judge the opportunities or threats afforded by others". There have some drawback like Somewhere Yorkshire or in Guangdong, observers form very similar first impressions of a stranger, simply from seeing their face.

Daniel Preotiuc-Pietro, Jordan Carpenter, Salvatore Giorgi, Lyle Ungar [26] Studying the Dark Triad of Personality through Twitter Behavior. They have presented a data-driven, multi-modal exploration of the expression of the dark triad in social media. They also use LIWC, Face++ API methods in their work. Here the ‘Dirty Dozen’ dataset of Twitter has been used. This system will automatically extract features that capture multiple aspects of online behavior. In result, All features are significant at $p < .01$, two tailed t-test. Here The dark triad of personality can be studied through more complex questionnaires.

Tomoki Tokuda, Junichiro Yoshimoto, Yu Shimizu, Go Okada, Masahiro Takamura, Yasumasa Okamoto, Shigeto Yamawaki & Kenji Doya [27] develop Identification of depression subtypes and relevant brain regions system using a data-driven approach. Data Driven, Clustering Method have been used here. They use FC data in resting state fMRI, clinical questionnaires, and gene expression data as datasets. Their main goal is to allow for flexible cluster structures with possible overlaps of patients and controls. It provides better accuracy with High proportion of depression-related features(60%), including 39 numerical features and 19 categorical features. They did not include the cerebellum because the image on this brain region was not reliably obtained. Small size of study is also a drawback of their research work.

Joanne Hinds and Adam Joinson [28] predict Human and Computer Personality From Digital Footprints. They used Bayesian inference, Teach methodology in their work. Social Media platforms like Facebook, Twitter are used as dataset here. There are some feature extract as Experience-sampling methods could provide a useful approach for examining personality over time, in different contexts, or across multiple platforms and devices simultaneously. In result we see that Human Openness=0.19, Computer Openness=0.39 which is more accurate. But Both humans and computers are limited by their traditions in practice, methods, and theoretical applications.

Nizar Omheni, Omar Mazhoud, Anis Kalboussi and Ahmed HachKacem [29] Predict Human Personality Traits From Annotation Activities. They used the Five Factor Model here. They used Group Of 120 Volunteers as a dataset. They Show that a user's personality traits can be predicted from their annotation practices. Result of Openness= $R^2=0.03$, F test= 0.57, P-value =0.76, Conscientiousness= $R^2=0.12$, F test= 2.52, P-value =0.03 shows their better accuracy. There's a limitation which is The sample size as they expect more significant results with a larger sample.

Leqi Liu , Daniel Preotiuc-Pietro, Zahra Riahi Samani, Mohsen E. Moghaddam, Lyle Ungar [32] Analyzed Personality through Social Media Profile Picture Choice. To predict personality there was used twitter profile picture and used two APIs based on deep learning methods – Face++2 and EmoVu3 – for facial feature extraction. Using the TwitterText data set, observed an overall correlation of $r>.145$ across all traits with conscientiousness the most predictive at $r = .189$ and using TwitterSurvey data set observed an overall correlation of $r>.046$ across all traits with conscientiousness the most predictive at $r = .190$. Limitations of this study, we consider this represents a necessary experiment in analyzing social media profile images using interpretable features on a data set orders of magnitude larger than previously.

[35] In December 2020 Tanay Gottigundala Predicted Personality Type From Writing Style In this

work, specifically focused on the predictive strength of an individual's Myers-Briggs Type Indicator (MBTI) type based on their posts and comments from an online personality forum. The dataset used in this research was scraped from an online personality forum on personalitycafe.com and it is freely available on Kaggle, an online data science community. Designing software that is customized by a user's personality type could be a paradigm shift in the field of user experience, and the implementation of an automated personality type predictor could be instrumental in facilitating it. The three classification methods used in this experiment performed somewhat similarly when their parameters were optimized, Random Forest=45.35%, Extra Trees=45.29%, Gradient Boosting=44.88%. While the prediction power of the classifier trained in this study is weak outside of writing from the personality forum, it is likely that a more complete dataset with increased variance could expand the results of this experiment to a much larger scope.

In this study, [36] Assem Talasbek, Meirambek Zhaparov, Seong Moo-Yoo, Yong Kab Kim has done Personality Classification Experiment by Applying k-Means Clustering. The method consists of three parts: data collection, data preparation, and hyper-parameter tuning. To develop a prototype of a classification system, we created a Google form. Here, questions were based on four dichotomies. For each dichotomy there are allocated five questions resulting in an overall number of 40 questions. Participants of our survey had to answer each question "Yes" or "No". Using Google's script, we automated the process of personality calculation and made some simulations using k-means clustering. The algorithm continues by switching back and forth between two stages. According to the results of our research above 28 percent (30 survey participants) belong to cluster 00. Clusters 07 and 14 placed the farthest from all and have the smallest amount which is equal to 0.94 percent of total. Since we have only one person in each cluster like 07 and 14, the distance means of these clusters are equal to 0. Limitation that in this work they do not have some target variable for their training data, they had to use unsupervised learning.

[11] In this paper, they develop an alternative approach to the identification of personality types, which they apply to four large data sets comprising more than 1.5 million participants. They find robust evidence for at least four distinct personality types, extending and refining previously suggested typologies. They show that these types appear as a small subset of a much more numerous set of spurious solutions in typical clustering approaches, highlighting principal limitations in the blind application of unsupervised machine learning methods to the analysis of big data. In their analysis, they use four different data sets from web-based questionnaires that measure the personality traits (so-called domains) of the FFM3 using different scales: The Johnson-300 data set (145,338 respondents), the Johnson-120 data set (410,376 respondents), theirPersonality-100 data set (575,380 respondents) the and BBC-44 data set (386,375 respondents). For all data sets, they only consider respondents who gave responses to all of the items. For theirPersonality-100 data set, they considered only one set of responses of each individual (indicated by the variable 'best protocol') in case the same person took the test several times. In addition, they obtained the gender and age of each participant. Although it is clear from these demographics that their data sets are not representative of the general population, it has been well established that data from web-based questionnaires are more diverse and are of at least as good quality as data obtained through more traditional approaches. For the Personality-100 and the BBC-44 data sets, gender or age is not available for 350,263 (60.9%) and 4 (0.001%) respondents, respectively. In the analysis involving these demographics, they only considered respondents for which both variables are available.

[12] In this paper there has been effort given to create a data model to identify both the student personality type and the dominant preference based on the Myers-Briggs Type Indicator (MBTI) theory. Based on the Jungian's psychological types, Myers-Briggs evaluates personality types and preferences through four aspects of personality:

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

The proposed model utilizes data from student engagement with the learning management system (Moodle) and the social network, Facebook. With this knowledge, educators will be more capable of matching students with their respective learning styles. To achieve this result, methods and techniques from several scientific fields and application areas are used. The most well-known techniques are: Data Mining, Knowledge Discovery, User Modeling, Web Mining, User Profiling, Artificial Intelligence and Agent Technologies, etc. These techniques utilize data generated from student engagement on all forms of web-based learning systems to provide a high degree of personalized education. The data generated from the student's engagement with these tools was so vast and enriching. The data can help them create a model that predicts and classifies the student's personality and preferences using data mining tools, in which researchers can test their model with different classification methods. The first phase is to predict student PT; they defined and initialized their counter of 8 preferences measured in the MBTI model. Then each student behavior measures one of the four preferences scales of the MBTI (i.e. Extraversion/Introversion, Sensing/Intuition, Thinking/Feeling, and Judging/Perceiving). In the first seven student behaviors (i.e. visited pages, time spent, comments, likes, shares, posts, and chat sessions). If the number of a certain behavior is greater than the average number (Avg) of this behavior, the value of the corresponding preference counter will be incremented. The last two attributes (i.e. Number of Early Assignments Submissions, Number of Late Assignments Submissions) will be compared and the highest will increment the value of the corresponding preference counter. After all student behaviors have been measured, they will compare the counters to find the greater of each of the four Jungian psychological types preferences. The greater preference counter of each of the four Jungian psychological type's preferences will be represented by one letter in the student PT. The second phase is to predict student DP which is based on the student PT; every student personality type has one of 4 dominant

preferences. So, the dominant preference is assigned according to the predicted student PT. The experiment used 10 classification algorithms in order to determine the highest classification accuracy. The algorithms were NaiveBayes, Bayes Net, Kstar, Random forest, J48, OneR, JRIP, KNN/IBK, Random Tree, and Decision Table. In order to measure the model performance, the dataset was split into two groups: the training set (20%) which was used to train the model, and the test set (80%) used to test the model in order to measure its accuracy, precision, recall, F-value, true positives (TP) rate, and false positives (FP) rate, which are defined by equations by means of a confusion matrix. Predicting students' learning styles and learning preferences are important and difficult tasks in e-learning, to make the learning process more personalized to the students and to overcome the “one-size-fits-all” learning model problem. To resolve this problem, we used data mining to help predict student personality and dominant preference using classification algorithms on a new model dataset from student engagement with the LMS (Moodle) and social network (Facebook groups) using MBTI theory. This new model proved its success by achieving high accuracy in different classification algorithms. The results show that the best classification algorithm was OneR algorithm, which had the highest accuracy of 97.40%, followed by Random forest at 93.23%, and J48 at 92.19%. The new model shows a potential and a promising method to predict student personality and dominant preference. This model is useful in following up with student personality prediction and dominant preference through each college year. Initial data can be gathered from a primary assessment done at college entry. This model used a retrograde dataset through a complete college year, which allowed us to have a more accurate prediction based on the actual activity of the student. This methodology provides a more accurate evaluation and prediction of student personalities. With a larger dataset, instructors will be able to modify the e-learning program according to each student's personality and dominant preference. A wide-scale study is recommended to ensure the validity of this model and to compare it with the current models so as to provide them with the capability of providing more efficient and personalized e-learning systems for students around the globe.

H. Andrew Schwartz and Lyle H. Ungar[[13](#)] work on automated methods using data-driven content analysis of social media. They used Facebook data to Insight, gaining an understanding of possible psychological and behavioral factors with the accuracy of Lexica .264 topics .307 controls .30. But the main problem is that social media has a lot of biased and noisy data which are irrelevant to their project.

However, all these Personality Prediction techniques with the MBTI concept cant provide better accuracy than what we are actually looking for. In this paper, we have discussed how the MBTI-XGBoost combined approach to solve this issue.

2.3 Problem Analysis

Over the past few years, Human Personality Prediction Analysis has emerged as trending research work for researchers. But in previous research work related to human personality prediction, they mainly focused on Big Five Personality Traits and a few other concepts. A very few used the MBTI-XGBoost model combined approach. In this thesis work, we will apply the XGBoost model along with the MBTI concept, and our main focus on providing better accuracy in personality prediction systems.

2.4 Summary

In this chapter, we reviewed the latest and previous technologies of the Human Personality Prediction system including the limitations. Our thesis target is to solve the accuracy issue of the MBTI-XGBoost model combined approach of the Human Personality Prediction System.

Proposed Model

3.1 Introduction

In this section of the thesis work, we analyze the feasibility of our project by analyzing the dataset and using the model of our project. Finally, this chapter illustrates the model's overall architecture, which is given by a detailed explanation.

3.2 Feasibility Analysis

This thesis work required twelve months of study by five researchers under one supervisor. We used the existing MBTI dataset for this project and this made our research convenient and gave a boost to our work procedure. As our research is based on MBTI so we go for an existing dataset rather than create a dataset for our project. It's quite challenging to train that kind of huge dataset which contains almost 8675 data.

3.3 Requirement Analysis

This Human Personality Prediction System required someone's social media post to predict their personality. This could be their posts on FaceBook, Twitter, LinkedIn or any kind of social media platform like this. After providing a person's social media post, this prediction system gives Myers Briggs Type Indicator results of the given person. And this will be done within a few seconds.

3.4 Research Methodology

The methods we used to acquire the MBTI personality accuracy scores are Pre-Processing and Features Extraction

In Pre-Processing firstly we convert all uppercase letters to lowercase and then we Remove URL/links from the dataset. After that we remove special characters and numbers from the dataset.

And then remove extra space and stopwords. Removal of MBTI personality names occurs then. That means MBTI personality names such as 'INFJ', 'ISTP' used by people in their posts can wrongly influence the results. We also remove those kinds of data. And then we did Lemmatization word removal. Lemmatization means words having the same meaning should be taken as a single feature. Lemmatizer is used to group words with the same purpose together for example gone, going, went to go, etc.

```
{x} [13] MB['posts'] = [soup(text).get_text() for text in MB['posts']]

MB['posts']

0      'http://www.youtube.com/watch?v=gsXHcwe3krw|]|...
1      'I'm finding the lack of me in these posts ver...
2      'Good one _____ https://www.youtube.com/wat...
3      'Dear INTP, I enjoyed our conversation the o...
4      'You're fired.|||That's another silly misconce...
...
8670   'https://www.youtube.com/watch?v=t8edHR_h908|]|...
8671   'So...if this thread already exists someplace ...
8672   'So many questions when i do these things. I ...
8673   'I am very conflicted right now when it comes ...
8674   'It has been too long since I have been on per...
Name: posts, Length: 8675, dtype: object
```

Figure 3.1: Before Pre-Processing of Data

```
[19] list_posts[0]

moment sportscenter top ten play prank life changing experience life repeat today may perc experience immerse last thing friend posted faceb
ook committing suicide next day rest peace hello sorry hear distress natural relationship perfection time every moment existence try figure hard
time time growth welcome stuff game set match prozac wellbrutin least thirty minute moving leg mean moving sitting desk chair weed moderation may
be try edible healthier alternative basically come three item determined type whichever type want would likely use given type cognitive function
whatnot left thing moderation sims indeed video game good one note good one somewhat subjective completely promoting death given sim dear favori
te video game growing current favorite video game cool appears late sad someone everyone wait thought confidence good thing cherish time solitude
b c revel within inner world whereas time workin enjoy time worry people always around yo lady complimentary personality well he...
```

Figure 3.2: After Pre-Processing of Data

After Pre-Processing the raw or annotated text is converted 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). If These High-frequency data are fed into the classifier, the model overshadows the less in a number of data. TF-IDF and count vectorizer is used to convert text into features, providing a more focused text view. First, vectorize the data using count Vectorizer and convert the post into the matrix of ken 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 word 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. $Tf(t) = (\text{No. of times term appears in document}) / (\text{Total no. of terms in the document})$ Idf tells us the importance of words in the dataset, and it is decided by how rare the word is in the datasets. $Idf(t) = \log_{10}(\text{Total number of documents} / \text{Number of documents with term } t \text{ in it})$. Therefore, $Tf-idf = Tf * Idf$ After examining the importance of words in the datasets, all the relevant terms are removed, making the model less complex by reducing the input.

3.5 Design, Implementation, and Simulation

In the beginning period of our research work, we decided to work on the MBTI-XGBoost model combined approach. But when we see lower accuracy there, we focus on accuracy and fix our goal to provide better accuracy in this approach. All of our proposed architecture depends on the MBTI concept and XGBoost model. We used some python libraries which are pandas, NumPy, matplotlib, seaborn, etc. Those were implemented in Google Colab. There's a huge dataset (MBTI based dataset) we choose for our project from Kaggle. After that, we train and test this dataset and use it for simulation.

3.6 Summary

In this chapter, we reviewed the architecture of our thesis work (the Human Personality Prediction system) including the research methodology, data pre-processing and features extraction. And lastly we discuss about working procedure of our MBTI-XGBoost combined approach.

Implementation, Testing, and Result Analysis

4.1 System Requirement

1. Google Colab
2. XGBoost Classifier
3. Python libraries

4.1.1 Google Colab

Google Colab is a free Jupyter Notebook environment. It is free cloud-based assistance by Google which implies you don't need to pay anything. Probably the best thing about Colab is that you don't have to introduce anything in advance. Truth be told, a significant number of the Data Science and Machine Learning libraries like Pandas, NumPy, Tensorflow, Keras, and OpenCV come pre-introduced with Colab. The journals you make are saved on your Google Drive. So Colab likewise uses the joint effort highlights of Google Docs, where you can impart your notepad to numerous individuals effectively and every one of you can chip away at a similar scratchpad simultaneously with next to no issue. Google additionally gives the utilization of a free NVIDIA Tesla K80 GPU. Assuming you interface Colab to Google Drive, that will surrender you to 15 GB of circle space for putting away your datasets. You can run the meeting in an intuitive Colab Notebook for 12 hours, which is enough for an amateur. Google has its independent custom chips called TPUs. Another thing to remember is that the dataset you transfer in the Colab notepad gets erased once the meeting is finished.

4.1.2. XGBoost Classifier

XGBoost is an algorithm that has recently been dominating applied machine learning and Kaggle competitions for structured or tabular data. XGBoost is an implementation of gradient boosted decision trees designed for speed and performance.

The implementation of the model supports the features of the scikit-learn with new additions like regularization. Three main forms of gradient boosting are supported:

- Gradient Boosting algorithm also called gradient boosting machine including the learning rate.
- Stochastic Gradient Boosting with sub-sampling at the row, column and column per split levels.
- Regularized Gradient Boosting with both L1 and L2 regularization.

The library provides a system for use in a range of computing environments, not least:

- Parallelization of tree construction using all of your CPU cores during training.
- Distributed Computing for training very large models using a cluster of machines.
- Out-of-Core Computing for very large datasets that don't fit into memory.
- Cache Optimization of data structures and algorithm to make best use of hardware.

The two reasons to use XGBoost are also the two goals of the project:

- Execution Speed.
- Model Performance.

4.1.3 Python Libraries

We installed many python libraries like pandas, NumPy, matplotlib, seaborn, etc that are defined in the code to run the project.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
sns.set_style("whitegrid");
plt.rcParams['figure.dpi'] = 360
from wordcloud import WordCloud
import spacy
from spacy.lang.en import English
nlp = English()
from collections import Counter
import shutil
from glob import glob
import os
from sklearn.model_selection import train_test_split
import sys
from sklearn.linear_model import LinearRegression
from sklearn import metrics
import category_encoders as ce
from sklearn.impute import SimpleImputer
from sklearn.linear_model import LogisticRegression
from sklearn.pipeline import make_pipeline
```

```
from sklearn.preprocessing import StandardScaler
import time

from nltk.corpus import stopwords

from nltk import word_tokenize

from nltk.stem import PorterStemmer, WordNetLemmatizer

import string

import re

import nltk

from nltk.corpus import stopwords

from nltk.stem.porter import PorterStemmer

from sklearn.feature_extraction.text import CountVectorizer

from bs4 import BeautifulSoup as soup

nltk.download('stopwords')

from nltk.corpus import stopwords

from nltk import word_tokenize

from nltk.stem import PorterStemmer, WordNetLemmatizer


from nltk.tokenize import word_tokenize

!pip install num2words

import num2words

from nltk import ne_chunk
```

4.2 Testing Process



```
test_set = """ I feel lonely today because my girl friend left me. I haven't done anything wrong with her.
I don't even know why she left me. It's one kind of trait. I will take revenge soon for this.
"""

# The type is just a dummy so that the data prep fucntion can be reused
mydata = pd.DataFrame(data={'type': ['ENFJ'], 'posts': [test_set]})

test_set, dummy = pre_process_data(mydata, remove_stop_words=True)

my_X_CV = CV.transform(test_set)
my_X_TF = TF.transform(my_X_CV).toarray()
```

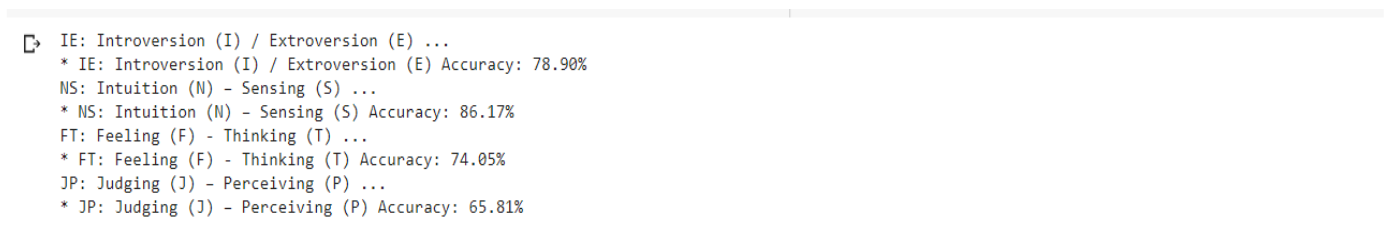
1 of 1 rows

Figure 4.1: Testing Process

4.3 Results and Discussion

Our model provides Accuracy of 78.90% for IE (Introversion (I) - Extraversion (E)). 86.17% for NS (Intuition (N) - Sensing (S)). 74.05% in FT (Feeling (F) - Thinking (T)) and 65.81% for JP (Judging (J) - Perceiving (P))

4.3.1 Model Accuracy



```
IE: Introversion (I) / Extroversion (E) ...
* IE: Introversion (I) / Extroversion (E) Accuracy: 78.90%
NS: Intuition (N) - Sensing (S) ...
* NS: Intuition (N) - Sensing (S) Accuracy: 86.17%
FT: Feeling (F) - Thinking (T) ...
* FT: Feeling (F) - Thinking (T) Accuracy: 74.05%
JP: Judging (J) - Perceiving (P) ...
* JP: Judging (J) - Perceiving (P) Accuracy: 65.81%
```

Figure 4.2: Model Accuracy

4.3.2 Result

```
[ ] print("The result is: ", translate_back(result))
```

The result is: INTP

Figure 4.3: Result

4.4 Summary

XGBoost, which stands for Extreme Gradient Boosting, is a scalable, distributed gradient-boosted decision tree (GBDT) machine learning library. It provides a parallel tree boosting and is the leading machine learning library for regression, classification, and ranking problems. In this chapter, we are using the XGBoost Classifier for Execution Speed and Model Performance. We got accuracy about IE:78.90% , NS:86.17%, FT:74.05% and JP:65.81%.

Standards, Impacts, Ethics, and Challenges

5.1 Sustainability

The project is based on identifying the personality of an individual using machine learning algorithms - XGBoost. We have used mainly four types of Human Personality traits to measure progress. These are :

- 1. Extraversion (e) – Introversion (i)**
- 2. Sensing (S) – Intuition (N)**
- 3. Thinking (T) – Feeling (F)**
- 4. Judging (J) – Perceiving (P)**

If we want to present a clear, concise way that is easy for stakeholders to understand then we will make it clear to them that our project will give more accuracy in less time. Compared to the previous project, we have tried our best to bring out more accurate results and fill up the targets of personality prediction.

5.2 Impacts on Society

Personality is the sum of the ideas, attitudes, and values of a person which determine his role in society and form an integral part of his character. Personality is acquired by the individual as a result of his participation in group life. Behaviors and actions: Personality not only influences how we move and respond to our environment but also causes us to act in certain ways. Multiple expressions: Personality is displayed in more than just behavior. It can also be seen in our thoughts, feelings, close relationships, and other social interactions. Personality is important to society because Personality affects academic and job performance, social and political attitudes, the quality and stability of social relationships, physical health and mortality, and risk for mental disorder Society affects our personality make-up. Society provides patterns and platforms for realization, activity, and

socialization. We react and develop traits based on what we face in interaction on the social plane

5.3 Ethics

Ethics is about making the best possible decisions concerning people, resources, and the environment. Ethical choices diminish risk, advance positive results, increase trust, determine long-term success and build reputations. Not only do ethics allow us to act in a way consistent with our beliefs, but they're also a key to executing projects successfully. This is because ethics lead to trust, which leads to leadership, which in turn leads to project success.

In our project, we have followed the ethical issues. We have promised not to disclose the data of the users. We have not copied algorithms from others. We have not stolen any dataset.

5.4 Challenges

The main challenge in our project is to find relevant datasets from multiple sources. It's quite challenging to pre-process the huge dataset which contains almost nine thousand pieces of data. Firstly we convert all uppercase letters to lowercase and then we Remove URL/links from the dataset. After that, we remove special characters and numbers from the dataset. And then removing extra space and stopwords is also challenging for us. After that, we face some problems in the testing phase. It's quite time-consuming and we have to be patient when the testing phase comes.

5.5 Summary

Having sustainability be relevant in all project areas will ensure that the environmental damage is minimized. Sustainability plays an important role in a project. Our project is about personality prediction. That's why it affects society and we have carefully looked after that matter. And obviously, we have followed the ethical issues. We have taken on challenges to gather the datasets and train them.

Constraints and Alternatives

6.1 Design Constraints

Design limitations are based on the results of all designs and it takes patience and intelligence to apply the limitations of personality prediction results to one. The languages that were used to understand the algorithms used at work were really subtle, memorizing the names of the algorithms, remembering the steps of their work, and designing a result using them was really difficult. It takes a considerable amount of time to spend on Dataset, Dataset Train. Later results were found to be a little more promising

6.2 Component Constraints

The component requirements of the proposed architecture include,

- Processor Requirement: Intel i3 (3rd Gen, 3GHz)
- Memory Requirement: 4GB (DDR3, 1600 bus)

6.3 Budget Constraints

The estimated budget is to be calculated by the current market price of the component requirements.

6.4 Summary

Design limitations are based on the results of all designs and it takes patience and intelligence to apply the limitations of personality prediction results to one . And In components constraints, we discuss the lowest requirement possible. In this chapter, we discuss our Design Constraint limitations, Component Constraints, and Budget Constraints.

Schedules, Tasks, and Milestones

7.1 Timeline

Based on the timeline, we have divided the whole thesis work into 3 parts.

The first semester of the three-semester work process, in collaboration with our supervisor, includes work plans and reviews related to the thesis work. The second-semester work process includes collaborative work of prototype designing and prototype analysis. In the third semester, we report on the overall architecture and overall workflow of implementation and testing with results.

7.2 Gantt Chart

The Gantt chart describes the work execution process of the thesis work.

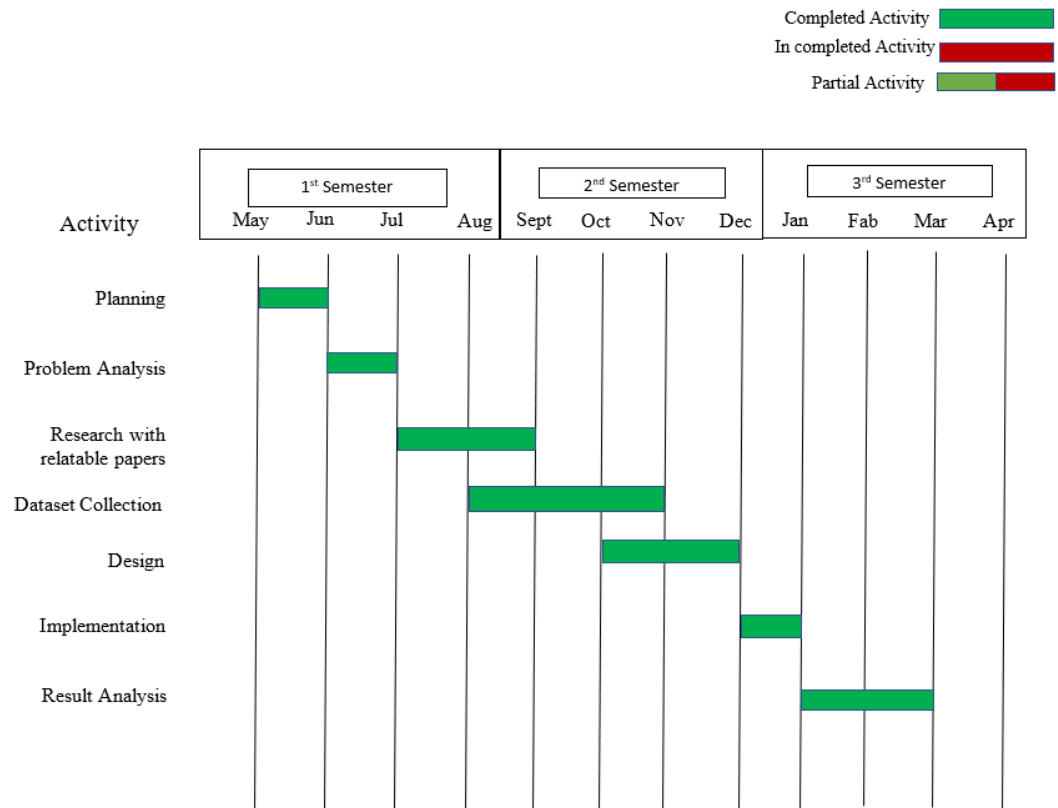


Figure 7.1: Gantt Chart

Conclusion

8.1 Conclusion

The ever-increasing social media users have dramatically contributed to significant growth as far as the volume of online information is concerned. Often, the content that these users put on social media can give valuable insights into their personalities (For example, in terms of predicting job satisfaction, specific preferences, as well as the success of professional and romantic relationships) and getting it without the hassle of taking a formal personality test. Termed personality prediction, the process involves extracting the digital content into features and mapping it according to a personality model. Owing to its simplicity and proven capability, a well-known personality model, called the MBTI traits. To date, there are many algorithms that can be used to extract embedded contextualized words from textual data for personality prediction systems; some of them are based on an ensemble model and deep learning. Although useful, existing algorithms such as RNN and LSTM suffers from few limitations. In this paper, we work on the XGboost model which is used before but we especially focused on the accuracy of human personality prediction and how it increases than the previous. More precisely, our results achieve maximum accuracy of IE:78.90%, NS:86.17%, FT:74.05%, and JP:65.81%.

8.2 Future Works and Limitations

Our training method is time-consuming. When we test personality from a random person's social media posts, sometimes if we give any irrelevant sentence or irrelevant words it gives results that are unacceptable. In the near future, we will focus on time-saving methods when we train data. And if we

give any irrelevant sentences or irrelevant words it gives us an alarming message “Your sentences or words are Incorrect . Please give valid words or sentences”.

References

- [1] François Mairesse and Marilyn Walker Trainable Generation of Big-Five Personality Style Through Data-driven Parameter Estimation Proceedings ACL-08: HLT, pages 165–173,
- [2] Menasha Thilakaratne, Ruwan Weerasinghe and Sujana Perera Knowledge-driven Approach to Predict Personality Traits by Leveraging Social Media Data 016 IEEE/WIC/ACM International Conference on Web Intelligence
- [3] Xiaoli he DATA-DRIVEN DEVELOPMENT OF PERSONALITY PREDICTIVE LEXICA FROM SOCIAL MEDIA New Brunswick, New Jersey May, 2020
- [4] Bruce Ferwerda and Marko Tkalcic Predicting Users Personality from Instagram Pictures UMAP'18,157-160 July 8–11, 2018, Singapore
- [5] Jennifer Golbeck, Cristina Robles and Karen Turner Predicting Personality with Social Media CHI 2011,253-261 May 7–12, 2011, Vancouver, BC, Canada.
19
- [6] Hong-Wei Ng and Stefan Winkler.A DATA-DRIVEN APPROACH TO CLEANING LARGE FACE DATASETS.Pages 343-347,2014.
- [7] Lauge Sørensen*, Mads Nielsen, Pechin Lo, Haseem Ashraf, Jesper H. Pedersen, and Marleen de Bruijne.Texture-Based Analysis of COPD:A Data-Driven Approach.Pages 70-78,2012
- [8] Tessa F. Blanken¹ · Ophélie Courbet² · Nathalie Franc³ · Ariadna Albajara Sáenz⁴ · Eus J.W. Van Someren¹ ·Philippe Peigneux⁴ · Thomas Villemonteix². Is an irritable ADHD profile traceable using personality dimensions? Replicability, stability, and predictive value over time of data-driven profiles.Pages 633-645,2020.
- [9] Franziska Leutner * and Tomas Chamorro-Premuzic.Stronger Together: Personality, Intelligence and the Assessment of Career Potential.Pages 1-10,2018.

[10] Leqi Liu, Daniel Preotiuc-Pietro, Zahra Riahi Samani, Mohsen E. Moghaddam, Lyle Ungar. Analyzing Personality through Social Media Profile Picture Choice. Proceedings of the Tenth International AAAI Conference on Web and Social Media 211-220 (ICWSM 2016)

[11] Clare A. M. Sutherland, Xizi Liu, Lingshan Zhang, Yingtung Chu, Julian A. Oldmeadow, and Andrew W. Young. Facial First Impressions Across Culture: Data-Driven Modeling of Chinese and British Perceivers' Unconstrained Facial Impressions. *Personality and Social Psychology Bulletin* 1–17 © 2017 by the Society for Personality and Social Psychology, Inc

[12] Daniel Preotiuc-Pietro, Jordan Carpenter, Salvatore Giorgi, Lyle Ungar. Studying the Dark Triad of Personality through Twitter Behavior. Proceedings of the 25th ACM International Conference on Information and Knowledge Management. 2016 Pages 761–770

[13] Tomoki Tokuda, Junichiro Yoshimoto, Yu Shimizu¹, Go Okada, Masahiro Takamura, Yasumasa Okamoto, Shigeto Yamawaki & Kenji Doya. Identification of depression subtypes and relevant brain regions using a data-driven approach, 2018.

[14] Joanne Hinds and Adam Joinson. Human and Computer Personality Prediction From Digital Footprints. *Current Directions in Psychological Science* 1–8 © The Author(s) 2019

[15] Nizar Omheni, Omar Mazhoud¹, Anis Kalboussi and Ahmed HadjKacem. Prediction of Human Personality Traits From Annotation Activities. Proceedings of the 10th International Conference on Web Information Systems and Technologies (WEBIST-2014), pages 263-269

[16] Leqi Liu , Daniel Preotiuc-Pietro, Zahra Riahi Samani, Mohsen E. Moghaddam, Lyle Ungar Analyzed Personality through Social Media Profile Picture Choice. Proceedings of the Tenth International AAAI Conference on Web and Social Media (ICWSM 2016).

[17] In December 2020 Tanay Gottigundala Predicted Personality Type From Writing Style, Tanay Gottigundala December 2020 .

[18] Assem Talasbek, Meirambek Zhaparov, Seong Moo-Yoo, Yong Kab Kim has done Personality Classification Experiment by Applying k-Means Clustering, International Journal of Emerging Technologies in Learning (iJET) 2020.

[19] Martin Gerlach, Beatrice Farb, William Revelle & Luís A. Nunes Amaral. A robust data-driven approach identifies four personality types across four large data sets. Pages 735-742, 2018.

[20] Mohamed Soliman Halawa, Mohamed Elemam Shehab, Essam M. Ramzy Hamed. Predicting Student Personality Based on a DataDriven Model from Student Behavior on LMS and Social Networks. Pages 294-299, 2015.

[21] H. Andrew Schwartz and Lyle H. Ungar Data-Driven Content Analysis of Social Media ANNALS, AAPSS, 659, May 2015 79-91