Software Requirement Specification Document for **Personality Traits Detection Based on Facebook Using Machine learning**

Amal Khaled, Karim Ali, Mohamed Sabry, Mohamed Akram
Supervised by: Eng. Lobna Shahen, Prof. Dr. Essam Halim Houssein

January 4, 2022

Table 1: Document version history

Version	Date	Reason for Change
1.0	4-Dec-2020	SRS First version's specifications are defined.
1.1	2-Dec-2020	Add use Case ,Class Diagram ,similar systems
1.3	5-Dec-2020	Add more similar systems , System overview Diagram

GitHub: https://github.com/Amal-Safwat/Personality-Traits-Detection-Based-on-Facebook-Using-Machine-l

Contents

Abstract

The use of social media is rapidly growing. Various information's are shared widely through social media, such as Facebook. Information about users, as well as what they conveyed through status updates, is extremely valuable for behavioral learning and human psychology research. Similar studies have been undertaken in this subject, and it grows continually till now. This research aims to develop a system that can predict a person's personality based on information from Facebook accounts. We propose an approach that intends to facilitate this line of research by directly predicting the Myers Briggs Type Indicator personality analysis from user's Social Network Sites behaviors. Comparing to the conventional inventory-based psychological analysis. We demonstrate via experimental studies that users' personalities can be predicted with reasonable precision based on their online behaviors.

1 Introduction

1.1 Purpose of this document

The purpose of this software requirement document is to present a detailed description of user's personality to discover user behavior on social media using machine learning tools. The main purpose of the project is to be able to classify the traits of users so they can understand their personality well. Also, we can recommend a job that is suitable for their personality.

1.2 Scope of this document

The system is developed to predict personality of user and recommend which job is suitable for their traits. To give people better understanding for which career paths they should take specially for high school students, as it prevents time and help psychologists in understanding their patients before even they come to them.

1.3 Business Context

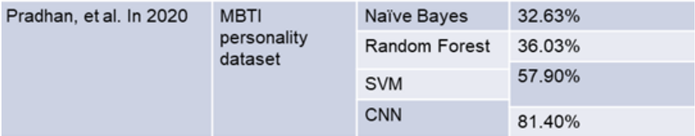
This project would be needed by anyone who works in the psychological field as it would improve their work if they know the personality of their patients before they come to them, also with some improvements it would be used by marketing companies too. It can also be used by people who needs job recommendation based on their personality and their strengths and weaknesses

2 Similar Systems

2.1 Academic

1. **Tejas Pradhan, Rashi Bhansali 2020[1]**The user 4-letter personality type is detected, with each letter corresponding to a specific personality preference by using the MBTI dataset. The user takes the test on a website by providing his interpretation of an abstract image in about 50-60 words, then a complete analysis and the percentage of each trait is previewed. A model predicts the personality type of users based on their responses to the test by using

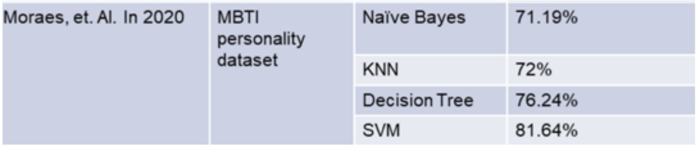
three classification algorithms which are Support Vector Machines (SVM), Naïve Bayes Classifier (NB) and Random Forest Classifier (RF). These algorithms were carried out on labelled dataset that was collected from social media with 8000 samples. The results show that SVM acquired the highest accuracy with 57.90%.



Pradhan, et al. In 2020	MBTI personality dataset	Naïve Bayes	32.63%
		Random Forest	36.03%
		SVM	57.90%
		CNN	81.40%

Figure 1: Model accuracy

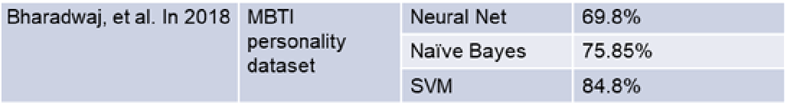
2. **Roshal Moraes, Larissa Lancelot Pinto 2020[2]**The system aims to automate the personality assessment process by analyzing the text extracted from the candidates' tweets. Their system architecture is clarified in showen figure, They used different ML algorithms which are Decision Tree (DT), Naïve Bayes (NB), K-nearest Neighbors (KNN) and Support Vector Machine (SVM). They used the MBTI personality type dataset. Their results show that SVM acquired the highest accuracy with 81.64%.



Moraes, et. Al. In 2020	MBTI personality dataset	Naïve Bayes	71.19%
		KNN	72%
		Decision Tree	76.24%
		SVM	81.64%

Figure 2: Model accuracy

3. **Srilakshmi Bharadwaj, Srinidhi Sridhar 2018[3]**The objective of the system is to analyze Social Media posts and create a personality profile of the person, by scoring him/her on the MBTI scale using Computational Psychology. They predicted personalities using different Machine Learning algorithms which are Naïve Bayes (NB), Neural Net, and support vector machine (SVM). These algorithms are used with Myers-Briggs Type Indicator (MBTI) their results observed that SVM gave the highest accuracy up to 88%.

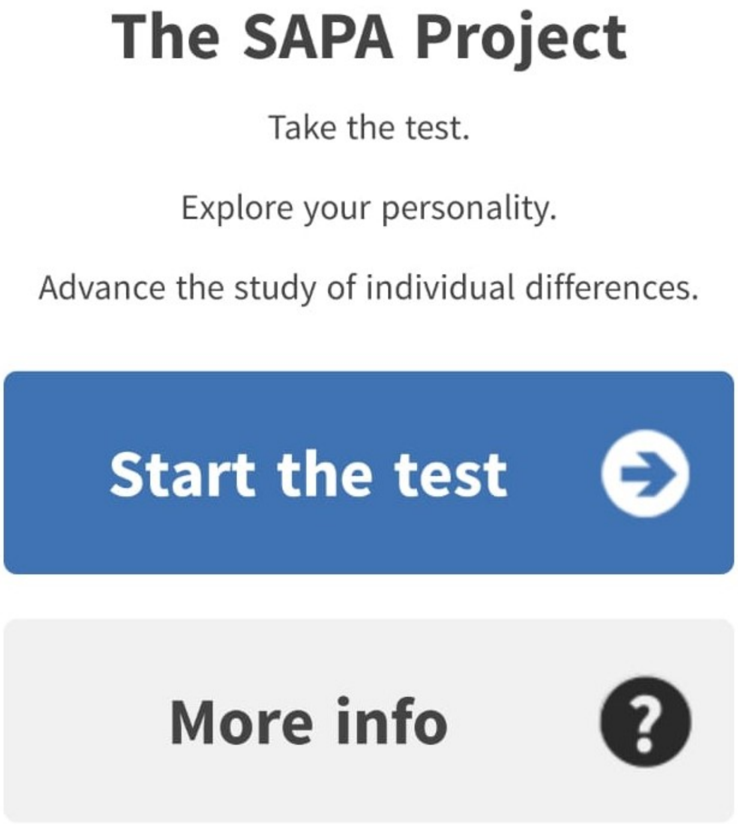


Bharadwaj, et al. In 2018	MBTI personality dataset	Neural Net	69.8%
		Naïve Bayes	75.85%
		SVM	84.8%

Figure 3: Model accuracy

2.2 Business Applications

Sapa-project website offers an enneagram personality test which divide the human personalities into 9 different personalities, each one has percentages of each one Through a questionnaire we can find our percentage of each personality and to each one of them different traits, weaknesses and strength points[4].



The SAPA Project

Take the test.

Explore your personality.

Advance the study of individual differences.

Figure 4: Sapa-Website

3 System Description

3.1 Problem Statement

The main problem that faces the community is that the human errors are huge in the process of hiring people, we always have no time, patience and of course not enough knowledge to be able to judge which candidates are the best for the job. Knowing someone's personality, preferences and what career they would be best at would minimize these human errors. In the interviews people spend most of the time trying to identify the candidate personality, which is a problem that costs a lot of time. Through knowing the candidate's personality through their social media this time would be minimized which would decrease the cost to the minimum and make the interview job easier for everyone. As the candidate may not be sure about which career will be the best fit for his/her personality and which environment will be more suitable for his/her traits. So, he/she may need job recommendation that matches his/her strengths and weaknesses.

3.2 System Overview

The concept of the system in this study is shown in Figure[5].

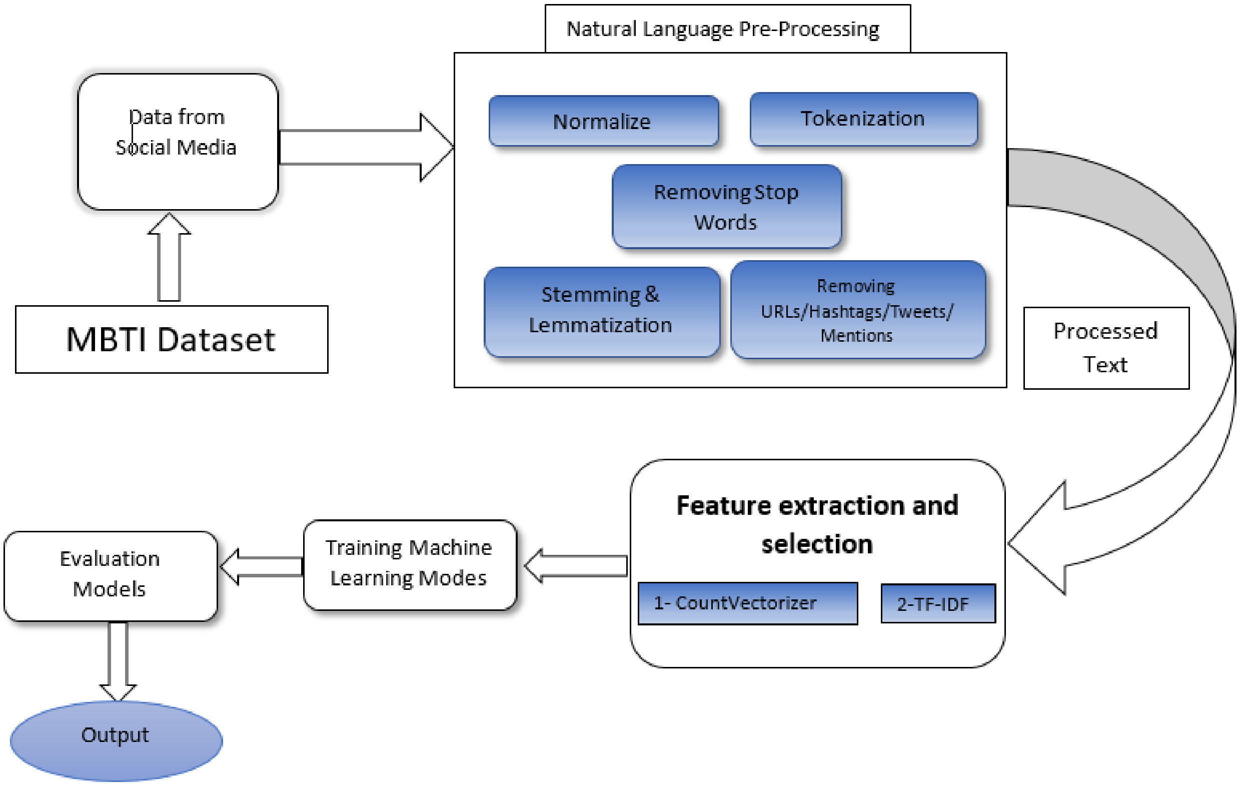


Figure 5: Project Overview.

The system steps of this study are as follows:

1-Data Extraction: This phase involves identifying and collecting data sets. Facebook has been the major source for social media data.

2-Pre-processing Steps:

- **Natural Language Pre-processing:** Text Pre-processing is traditionally an important step for natural language processing (NLP) tasks. It transforms text into a more digestible form so that machine learning algorithms can perform better[6].

1. Normalize: Changing all the capital letters into lowercase.
2. Tokenization: A common task in Natural Language Processing (NLP). It is a way of separating a piece of text into smaller units called tokens.
3. Filtering: Remove words that are meaningless which are stop words.
4. Stemming & Lemmatization: The objective of stemming is to reduce related words to the same stem even if the stem is not a dictionary word. for this phase we used WordNet lemmatizer as in this phase we translate language to English.
5. Removing URLs / Hashtags / Mentions.

3-Feature Extraction and selection: Methods of feature extraction are:

- countVectorizer: We used to transform a given text into a vector on the basis of the frequency (count) of each word that occurs in the entire text.

- TF-IDF: Means Term Frequency - Inverse Document Frequency. This is a statistic that is based on the frequency of a word in the corpus but it also provides a numerical representation of how important a word is for statistical analysis.

4-Training Machine Learning Models In this phase we will apply some machine learning algorithms to choose the one that gives best accuracy:

- Naïve Bayes

- Support vector machine(SVM)

- Recurrent neural network (RNN)

- k-nearest neighbor (kNN)

- XGBoost

5- Evaluate models: Evaluating results from personality model.

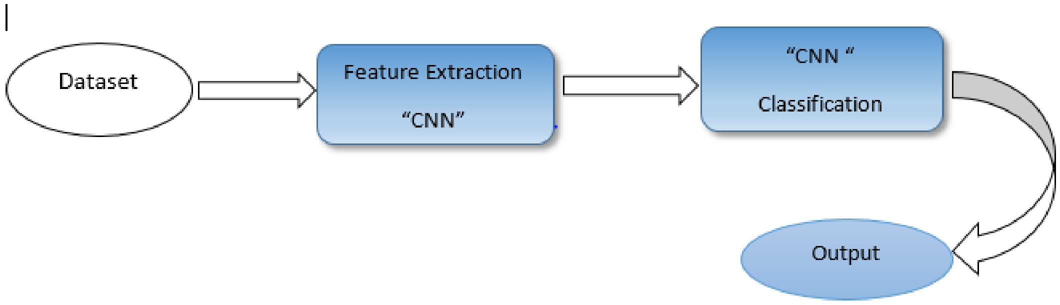


Figure 6: Project Overview.

The steps of this study are as follows[7][8]:

1. Input:Pre-processed data.

2. Feature Extraction using CNN.

3. Classification Using CNN.

4. Output Result

3.3 System Scope

1. Understand human behaviour through knowing their personality types.
2. To help psychologists in understanding their patients before even they come to them.
3. To give people better understanding for which career paths they should take specially for high school students.
4. Detect facial expression from images

3.4 System Context

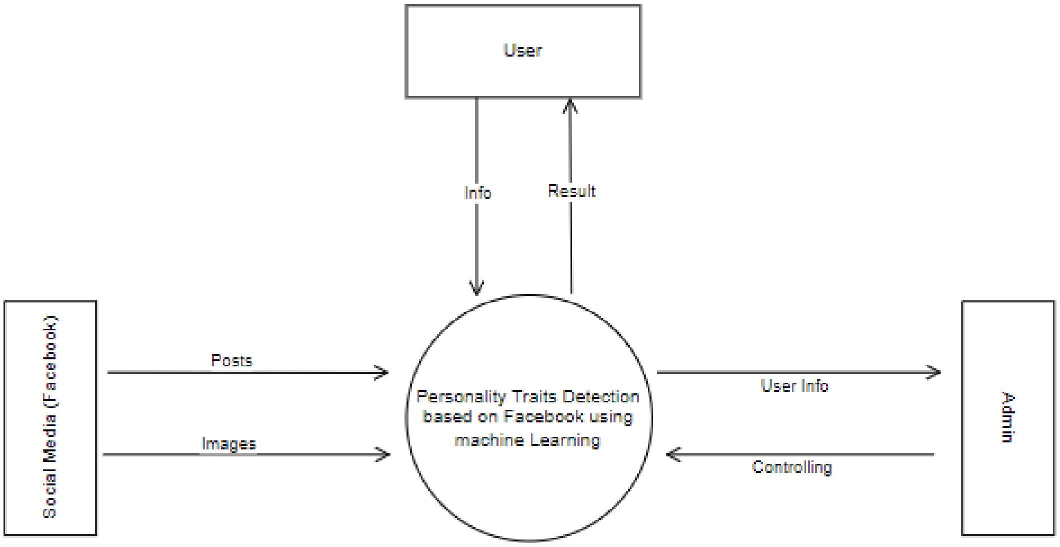


Figure 7: System context diagram

3.5 Objectives

The project has two main objectives the first one, aims to build a personality profile detection using text written by the user, we intend to use machine learning and deep learning models to build a classifier that will take in Posts as input and produce as output a prediction of the MBTI personality type.

The second object aims to create a model where the user will be provided with career guidance which matches his personality. For example: if a user has the ability to speak well and is able to convince the opposite person. So, this user will be good in the marketing field.

We tried more than machine learning algorithm to choose the best model to classify the personality and build the personality profile. At the end the selected machine learning model is Support Vector Machine (SVM) and the selected deep learning model is Convolution Neural Network (CNN)

3.6 User Characteristics

The user here of this project has to have some psychological knowledge, Facebook user, and wants to be more self aware and know his career path or a psychologist who wants to know his patients personalities or an interviewer who wants to know the career recommendations for his interview.

4 Functional Requirements

4.1 System Functions

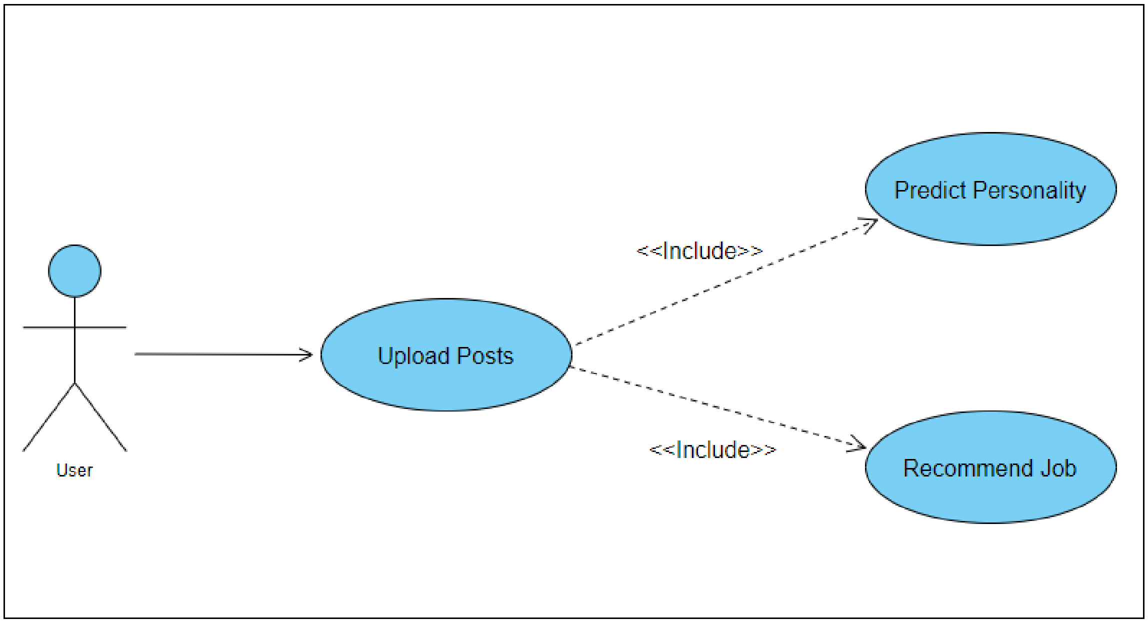


Figure 8: Use case diagram

- Upload Posts: The user uses this function to upload his last 50 posts.
- Predict Personality: This abstract use case is not initialized by the user and initialized by Upload posts use case and its output is the user's personality according to MBTI [9].
- Recommend Job: This abstract use case is not initialized by the user and initialized by Upload posts use case and its output is the suitable job for the user based on his traits.

4.2 Detailed Functional Specification

Show the details of all functions shown in section 4.1.Describe each function in the following structure.

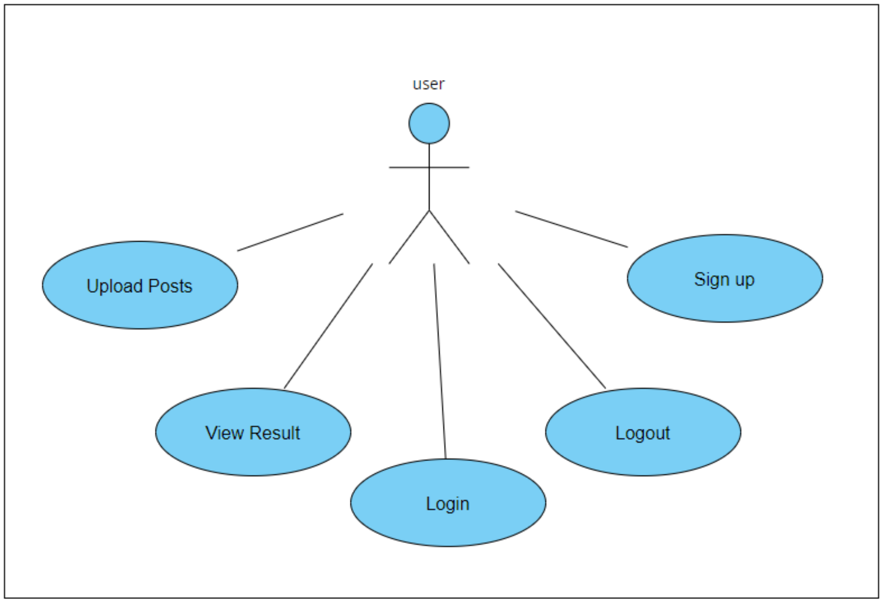


Figure 9: Use case diagram

Table 1: Sign Up Function Description

Name	User Register
Code	FR01
Priority	Extreme
Critical	The user has to sign up to save his info in the database.
Description	This function takes the information from the user and insert it in the database
Input	Username , E-mail , Phone number, Password.
Output	Boolean (true for success, false otherwise)
Pre-condition	None
Post-condition	User information saved in the database.
Dependency	None
Risk	The user may not enter the needed information, or enter it wrongly, so it must be validated.

Table 2: User log in Function Description

Name	User Login
Code	FR02
Priority	Extreme
Critical	The user has to login to view his results.
Description	This function takes the information from the user and check if it is in the database
Input	E-mail , Password.
Output	Boolean (true for success, false otherwise)
Pre-condition	He has to sign up frist
Post-condition	User access his profile .
Dependency	FR01
Risk	The user may login without signing up or enter the info wrongly ,so it must be validated.

Table 3:Uploading Data Function Description

Name	Uploading Data
Code	FR03
Priority	Extreme
Critical	The user has to upload the last 50 posts.
Description	This function let the user upload the last 50 posts (csv file) and the system starts processing.
Input	CSV file.
Output	If successful the system starts processing the data,if not the system asks the user to re-check the input.
Pre-condition	The user must be logged in.
Post-condition	The posts will be uploaded and the system starts processing .
Dependency	FR02
Risk	The user may upload the wrong file.

Table 4:View results Function Description

Name	View Result
Code	FR04
Priority	Extreme
Critical	The user sees the result of personality trait.
Description	The system takes the sample and starts processing the data then shows the result to the user
Input	None
Output	The user's personality
Pre-condition	The samples have to be uploaded to the system.
Post-condition	None.
Dependency	FR03
Risk	None

5 Design Constraints

5.1 Standards Compliance

All we need is to have an active social media account that the user post in it and give us permission to access the account to extract a personality trait and recommend a suitable job for the user.

5.2 Hardware Limitations

A PC or Mobile phone that can access social media and supports internet connection.

6 Non-functional Requirements

6.1 Security

Security is very important for the system so no one has the access to the user account unless he allowed to access any data.

6.2 Reliability

The system is reliable and ready to handle any issues. The time needed to analysis posts on the system has an average speed.

6.3 Maintainability

The code is easy and simple and can be maintained in anytime.

6.4 Portability

The system will be available in any windows on pc since the code is written by python.

6.5 Extensibility

We are looking forward to add a job recommendation for the user to recommend for him the best job that suits his/her personality.

6.6 Usbility

The system is simple so the user doesn't need time to learn how to use it.

7 Data Design

7.1 Database Design Description

The database system contains four entities Admin, Primary_User, Secondary_User,and Report, with their relationship. Admin entity has attributes such as AdminId, Name, Email, and Password. Report entity has attributes such as ReportId, UserId. Primary_User has attributes such as UserId, Name, Email ,Password , and Gender. Secondary_User has attributes such as UserId, Name, Email ,Password , and Gender.

We describe the structure of a database with the help of this diagram:

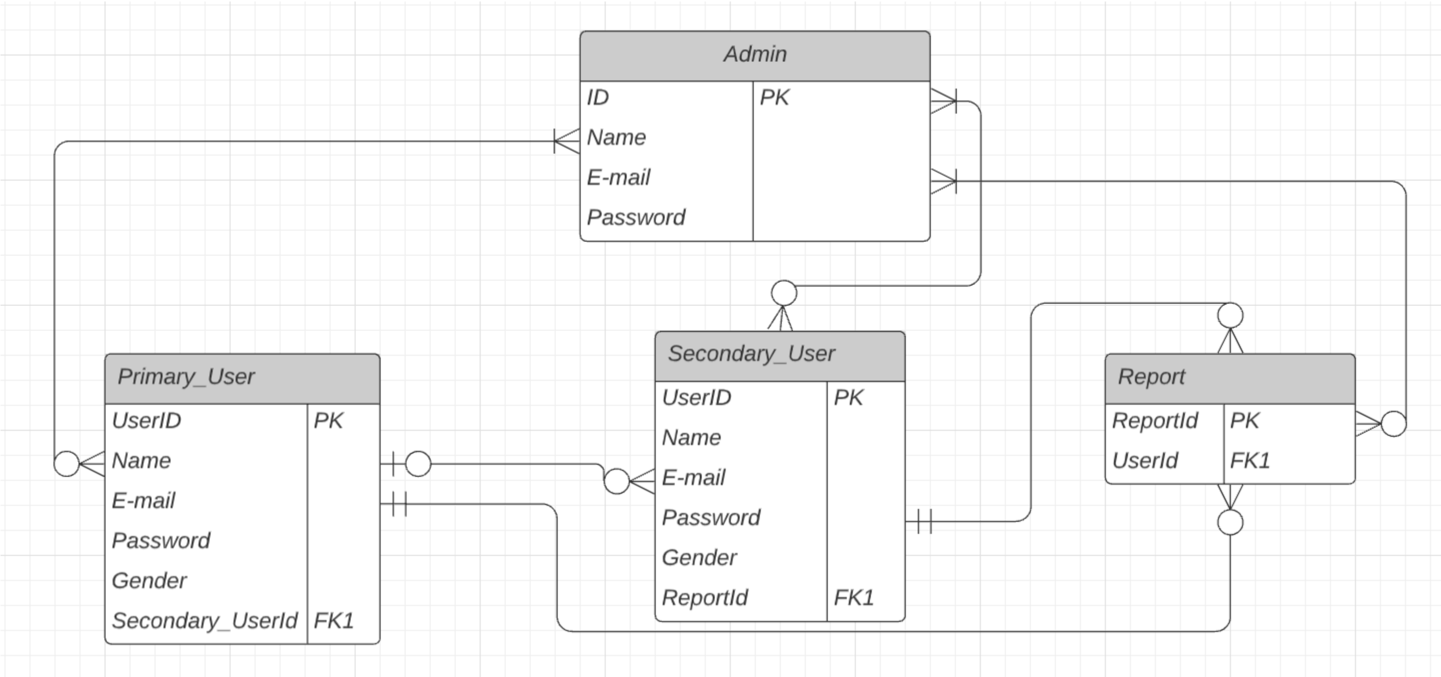


Figure 10: Database diagram

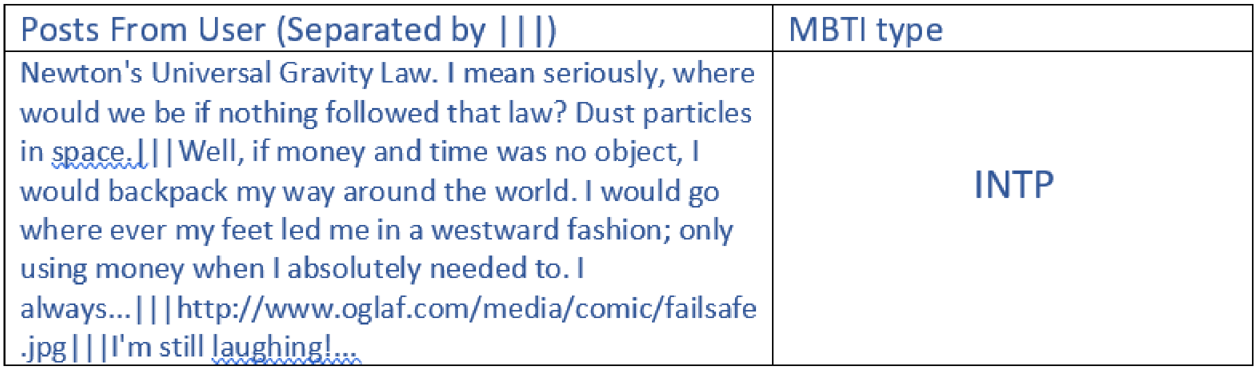
7.2 Dataset Description

The dataset we used contains posts that makes one of the 16 MBTI types as shown in Figure 11 The tags are a combination of four different characters. Each on is corresponding to the one of the trait classes in the four MBTI test. This dataset contains about 8650 rows. The distribution of MBTI traits (number of rows out of 8600) in each class is as follows[10][9]:

- Introverted (I): 6664; Extroverted(E): 1996
- Sense(S): 7466; Intuition (N): 1194
- Thinker (T): 4685; Feeler (F): 3975

- Judge (J): 5231; Perceive (P): 3429

Each row represents a different user.



Posts From User (Separated by |||)	MBTI type
Newton's Universal Gravity Law. I mean seriously, where would we be if nothing followed that law? Dust particles in space.|||Well, if money and time was no object, I would backpack my way around the world. I would go where ever my feet led me in a westward fashion; only using money when I absolutely needed to. I always...|||http://www.oglaf.com/media/comic/failsafe.jpg|||I'm still laughing!...	INTP

Figure 11: Dataset Sample

8 Preliminary Object-Oriented Domain Analysis

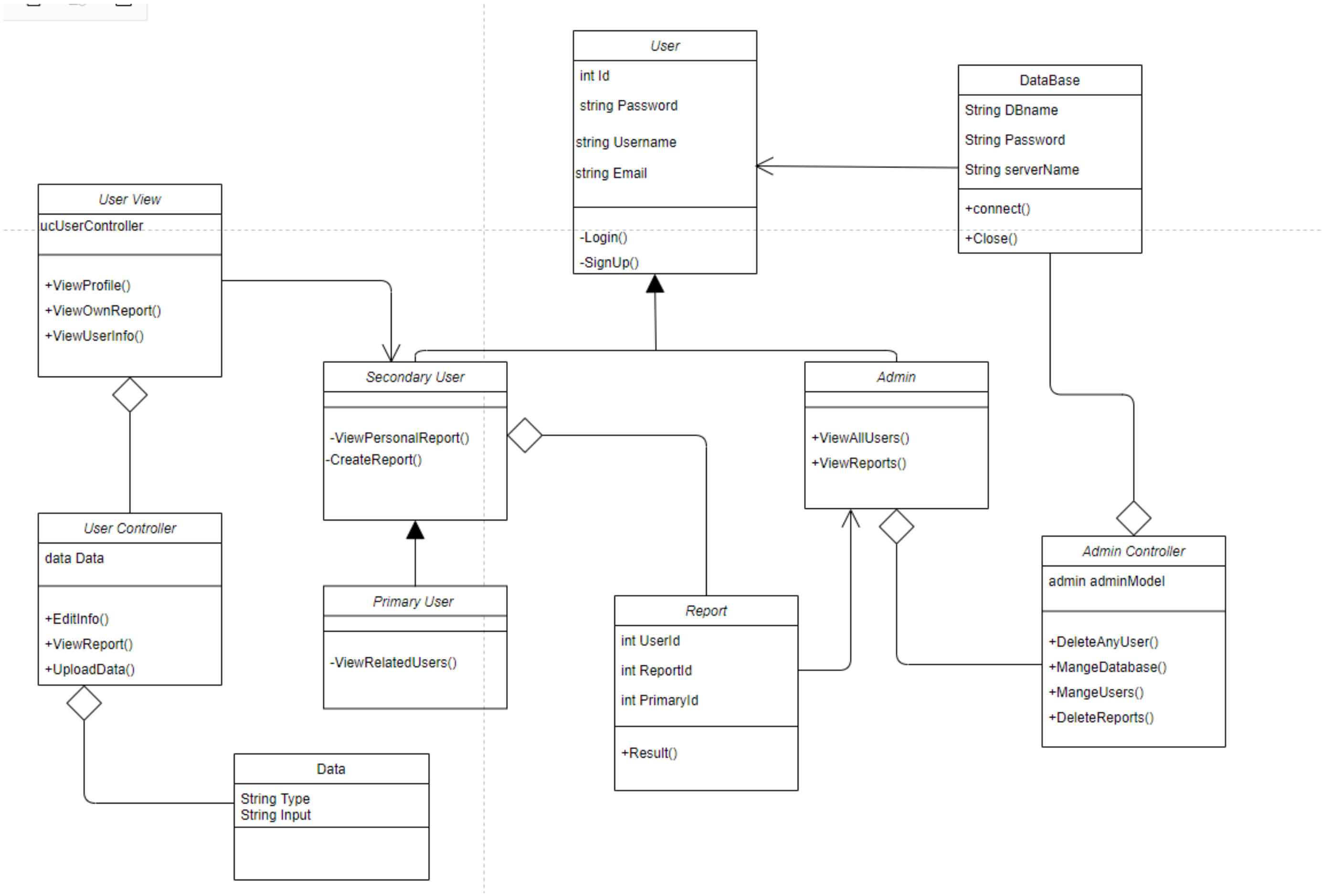


Figure 12: Class diagram

8.1 Class Description

Primary user : is a user that supervises some of the secondary users reports by their permission. Example : Psychologist, Interviewer, etc.

Secondary user : might be a normal user or someone who is supervised by one of the primary users. Example : Patient, Interviewee, etc

- User
  1. Class name:User
  2. Super classes:N/A
  3. Sub classes:Secondary User ,Admin
  4. Purpose:To represent the common attributes and operations for all user’s classes.
  5. Collaborations:Database class
  6. Attributes:id, password, username, email
  7. Operations:Login, SignUp

- Admin
    1. Class name: Admin
    2. Super classes:User
    3. Sub classes:N/A
    4. Purpose: This class holds all functionalities for Admin
    5. Collaborations:Report, Admin Controller
    6. Attributes:N/A
    7. Operations:ViewAllUsers,ViewReports
- Secondary User
    1. Class name:Secondary User
    2. Super classes: User
    3. Sub classes:Primary User
    4. Purpose:This class holds all functionalities for Users
    5. Collaborations:Report,UserView
    6. Attributes:N/A
    7. Operations:ViewPersonalReport,CreateReport
- Primary User
    1. Class name:Primary User
    2. Super classes:Secondary User
    3. Sub classes:N/A
    4. Purpose:This class holds all functionalities for primary users
    5. Collaborations:N/A
    6. Attributes:N/A
    7. Operations:ViewRelatedUsers
- Admin Controller
    1. Class name:Admin Controller
    2. Super classes:N/A
    3. Sub classes:N/A
    4. Purpose:This class gives admin control over database
    5. Collaborations:Database, Admin
    6. Attributes: adminModel
    7. Operations:DeleteAnyUser,MangeDatabase,MangeUser,DeleteReports

- Database

  1. Class name:Database
  2. Super classes:N/A
  3. Sub classes:N/A
  4. Purpose:Holds all data
  5. Collaborations:User, Admin Controller
  6. Attributes:DBname, Password ,ServerName
  7. Operations:Connect,close

- Report

  1. Class name:Report
  2. Super classes:N/A
  3. Sub classes:N/A
  4. Purpose: Gives the user the result
  5. Collaborations:Admin, Secondary User
  6. Attributes:UserId, ReportId, PrimaryId
  7. Operations:Result

- User View

  1. Class name:User view
  2. Super classes:N/A
  3. Sub classes:N/A
  4. Purpose:This class allows users to view their own data
  5. Collaborations:User Control, Secondary User
  6. Attributes:User Control
  7. Operations:ViewProfile, ViewOwnReport, ViewUserInfo

- User Controller

  1. Class name: User Controller
  2. Super classes:N/A
  3. Sub classes:N/A
  4. Purpose:This class allows users to control their data
  5. Collaborations:User View, Data
  6. Attributes:Data
  7. Operations:EditInfo, ViewRports, UploadData

- Data

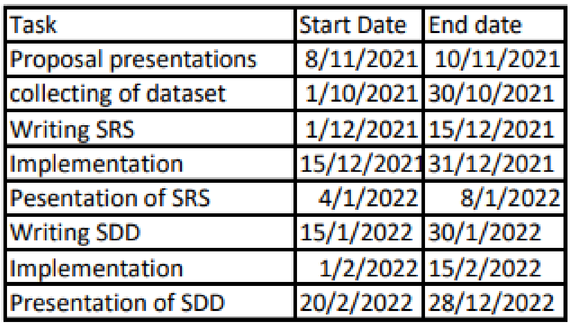
  1. Class name:Data
  2. Super classes:N/A
  3. Sub classes:N/A
  4. Purpose:This class holds data uploaded by the user
  5. Collaborations:User Controller
  6. Attributes:Type, Input
  7. Operations:N/A

9 Operational Scenarios

1. System Processes scenario, The system will wait for the user of social media to agree to grant him access to his posts and media on social media before beginning to collect and analyse them in order to provide him with an accurate personality type and suitable employment based on the images and posts.

2. User Processes scenario, The user must first apply for an account, then go through the process of analysing his personality traits using posts and photographs from his social media accounts, and agree to give us access.

10 Project Plan



Task	Start Date	End date
Proposal presentations	8/11/2021	10/11/2021
collecting of dataset	1/10/2021	30/10/2021
Writing SRS	1/12/2021	15/12/2021
Implementation	15/12/2021	31/12/2021
Pesentation of SRS	4/1/2022	8/1/2022
Writing SDD	15/1/2022	30/1/2022
Implementation	1/2/2022	15/2/2022
Presentation of SDD	20/2/2022	28/12/2022

Figure 13: Project Plan

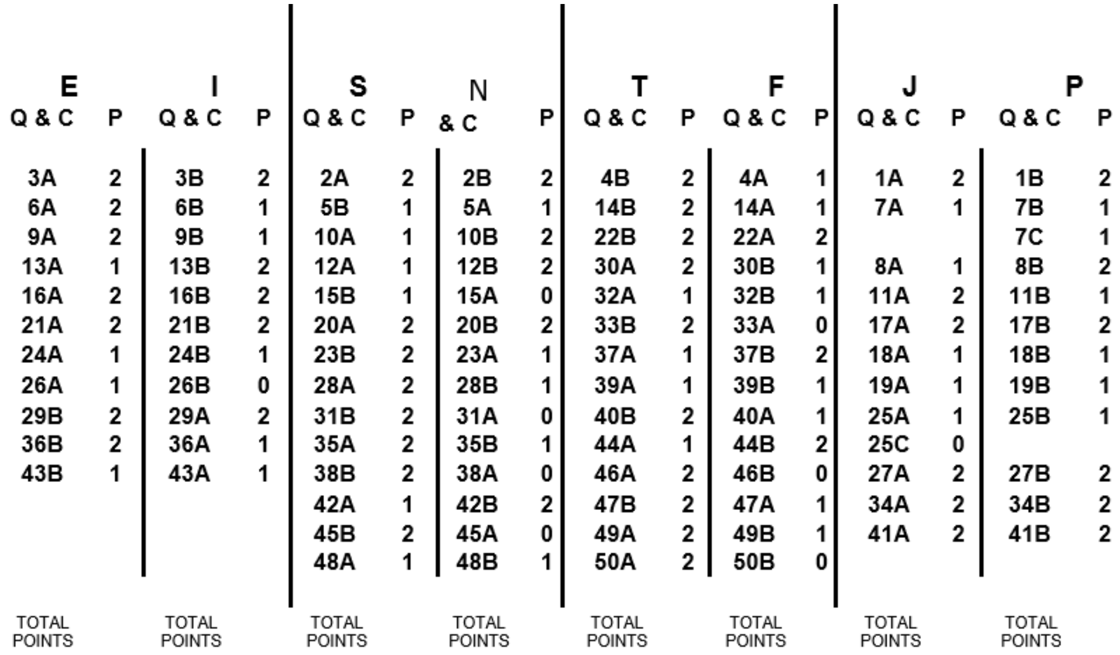
11 Appendices

11.1 Abbreviations

MBTI	Myers Briggs Type Indicator
SNSs	Social Network SiteS
NLP	Natural language processing
PP	Personality Prediction
RNN	Recurrent neural network
SVM	Support Vector Machine
TF-IDF	Term frequency inverse document frequency
XGBoost	Extreme Gradient Boosting
CNN	Convolutional neural network

11.2 Survey

For the first 26 questions, we have a scale from 1(strongly agree) to 5(strongly disagree), if 1 is chosen then it's considered A, if 5 it is considered B, if 2 then it is half of the value of A, if 4 then half the value of B, 3 then it doesn't matter. For questions 7 and 25 A, B and C are clear. For the last questions it is clear too.



E		I		S		N		T		F		J		P	
Q & C	P	Q & C	P	Q & C	P	& C	P	Q & C	P	Q & C	P	Q & C	P	Q & C	P
3A	2	3B	2	2A	2	2B	2	4B	2	4A	1	1A	2	1B	2
6A	2	6B	1	5B	1	5A	1	14B	2	14A	1	7A	1	7B	1
9A	2	9B	1	10A	1	10B	2	22B	2	22A	2			7C	1
13A	1	13B	2	12A	1	12B	2	30A	2	30B	1	8A	1	8B	2
16A	2	16B	2	15B	1	15A	0	32A	1	32B	1	11A	2	11B	1
21A	2	21B	2	20A	2	20B	2	33B	2	33A	0	17A	2	17B	2
24A	1	24B	1	23B	2	23A	1	37A	1	37B	2	18A	1	18B	1
26A	1	26B	0	28A	2	28B	1	39A	1	39B	1	19A	1	19B	1
29B	2	29A	2	31B	2	31A	0	40B	2	40A	1	25A	1	25B	1
36B	2	36A	1	35A	2	35B	1	44A	1	44B	2	25C	0		
43B	1	43A	1	38B	2	38A	0	46A	2	46B	0	27A	2	27B	2
				42A	1	42B	2	47B	2	47A	1	34A	2	34B	2
				45B	2	45A	0	49A	2	49B	1	41A	2	41B	2
				48A	1	48B	1	50A	2	50B	0				
TOTAL POINTS		TOTAL POINTS		TOTAL POINTS		TOTAL POINTS		TOTAL POINTS		TOTAL POINTS		TOTAL POINTS		TOTAL POINTS	

Figure 14: Survey Explanation

In Case Of TIE:

1. between E I, select I
2. between S N, select N
3. between T F, male will select 'T' females 'F'
4. between J P, select P

References

[1] Tejas Pradhan, Rashi Bhansali, Dimple Chandnani, et al. "Analysis of Personality Traits using Natural Language Processing and Deep Learning". In: *2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA)*. IEEE. 2020, pp. 457–461.

[2] Roshal Moraes, Larissa Lancelot Pinto, Mrunal Pilankar, et al. "Personality Assessment Using Social Media for Hiring Candidates". In: *2020 3rd International Conference on Communication System, Computing and IT Applications (CSCITA)*. IEEE. 2020, pp. 192–197.

[3] Srilakshmi Bharadwaj, Srinidhi Sridhar, Rahul Choudhary, et al. "persona traits identification based on myers-briggs type indicator (MBTI)-a text classification approach". In: *2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*. IEEE. 2018, pp. 1076–1082.

[4] David Condon and William Revelle. "Selected personality data from the SAPA-Project: On the structure of phrased self-report items". In: *Journal of Open Psychology Data* 3.1 (2015).

[5] Jennifer Golbeck, Cristina Robles, and Karen Turner. "Predicting personality with social media". In: *CHI'11 extended abstracts on human factors in computing systems*. 2011, pp. 253–262.

[6] Mariam Hassanein, Wedad Hussein, Sherine Rady, et al. "Predicting personality traits from social media using text semantics". In: *2018 13th International Conference on Computer Engineering and Systems (ICCES)*. IEEE. 2018, pp. 184–189.

[7] Hartayuni Sain and Santi Wulan Purnami. "Combine sampling support vector machine for imbalanced data classification". In: *Procedia Computer Science* 72 (2015), pp. 59–66.

[8] Navonil Majumder, Soujanya Poria, Alexander Gelbukh, et al. "Deep learning-based document modeling for personality detection from text". In: *IEEE Intelligent Systems* 32.2 (2017), pp. 74–79.

[9] Isabel Briggs Myers. *The 16 MBTI® Type*. URL: `https://www.myersbriggs.org/my-mbti-personality-type/mbti-basics/the-16-mbti-types.htm`.

[10] Siti Khotijah. "Big Five Personality Test + Clustering". In: (). URL: `https://www.kaggle.com/khotijahs1/big-five-personality-test-clustering/data`.