INTERIM PROJECT REPORT ON

Personality Prediction through Social Media Posts

SUBMITTED BY

AJINKYA PINGALE (403061)
JAY NAIK (403046)
MANOJ NANDHA (403047)
AKASH MISAL (403043)

UNDER THE GUIDANCE OF
Prof.(Mrs.) Mamta Bhamare

Department Of Computer Engineering
MAEER's MAHARASHTRA INSTITUTE OF TECHNOLOGY
Kothrud, Pune 411 038
2019-2020

MAHARASHTRA ACADEMY OF ENGINEERING AND EDUCATIONAL RESEARCHES

MAHARASHTRA INSTITUTE OF TECHNOLOGY PUNE

DEPARTMENT OF COMPUTER ENGINEERING

C E R T I F I C A T E

This is to certify that

AJINKYA PINGALE (403061)

JAY NAIK (403046)

MANOJ NANDHA (403047)

AKASH MISAL (403043)

of B. E. Computer successfully completed project report in

Personality Prediction through Social Media Posts

to my satisfaction and submitted the same during the academic year 2019-2020 towards the partial fulfillment of degree of Bachelor of Engineering in Computer Engineering of Pune University under the Department of Computer Engineering , Maharashtra Institute of Technology, Pune.

Prof.(Mrs.) Mamta Bhamare
(Project Guide)

Dr.(Mrs.) V. Y. Kulkarni
(HOD Computer Department)

Place: Pune

Date:

ACKNOWLEDGEMENT

I take this opportunity to express my sincere appreciation for the cooperation given by Dr. (Mrs.) V. Y. Kulkarni, HOD (Department of Computer Engineering) and need a special mention for all the motivation and support.

I am deeply indebted to my guide Prof.(Mrs.) Mamta Bhamare for completion of this project report for which she has guided and helped me going out of the way.

For all efforts behind the project report, I would also like to express my sincere appreciation to staff of department of Computer Engineering, Maharashtra Institute of Technology Pune, for their extended help and suggestions at every stage.

Ajinkya Pingale
Jay Naik
Manoj Nandha
Akash Misal

Contents

List of Figures

Abstract

Our focus for this project is using machine learning to build a classifier ca-pable of sorting people into their Myers-Briggs Type Index (MBTI) personality type based on text samples from their social media posts. The motivations for building such a classifier are twofold. First, the pervasiveness of social media means that such a classifier would have ample data on which to run personality assessments, allowing more people to gain access to their MBTI personality type, and perhaps far more reliably and more quickly. There is significant interest in this area within the academic realm of psychology as well as the private sector. For example, many employers wish to know more about the personality of potential hires, so as to better manage the culture of their firm. Our second motivation centers on the potential for our classifier to be more accurate than currently available tests as evinced by the fact that retest error rates for personality tests administered by trained psychologists currently hover around 0.5. That is, the there is a probability of about half that taking the test twice in two different contexts will yield different classifications. Thus, our classifier could serve as a verification system for these initial tests as a means of allowing people to have more confidence in their results. Indeed, a text-based classifier would be able to operate on a far larger amount of data than that given in a single personality test.

Keywords:
Deep Learning, Machine Learning , Natural Language Processing(NLP)

Chapter 1

Introduction

1.1 Overview

This is a project deals with detecting personality of a person through his tweets . This model detects personality using Mayers Briggs Type Indicators (MBTI) and Big Five Model . We Used Natural Language Processing (NLP) to proccess our labelled dataset and clean it for our use .
We then use a specifically Designed Recurrent Neural Network (RNN) so as to accomadate it in our model as it provides better results than traditional baseline models. We then calculated accuraccy for every category of MBTI Model for ex. (I Vs E) , (N Vs S) , (F Vs T) , (P Vs J) and for Big Five Model we calculated accuracy for Each Five traits of the Model
We Then Implemented these models to be used in an Android App , so that it would be more easier for peoples to access it and use this model . What this does is that it just uses a Twitter ID and then evaluate their tweets and provide graphical statisics for every traits and provide explanation for each trait
Thus this would really enable everyone to use this model, and we used an intutative Graphical UI that uses card view to provide results and thus feels more organised and is comfortable to interact.

1.2 Motivation

The motivations for building such a classifier are two fold -¿
First, the pervasiveness of social media means that such a classifier would have ample data on which to run personality assessments, allowing more people to gain access to their MBTI personality type, and perhaps far more re-liably and more quickly. There is significant interest in this area within the academic realm of psychology as well as the private sector. For example, many employers wish to know more about the

personality of potential hires, so as to better manage the culture of their firm.
Our second motivation centers on the potential for our classifier to be more accurate than currently available tests as evinced by the fact that retest error rates for personality tests administered by trained psychologists currently hover around 0.5. That is, the there is a probability of about half that taking the test twice in two different contexts will yield different classifications. Thus, our classifier could serve as a verification system for these initial tests as a means of allowing people to have more confidence in their results.

1.3 Problem Defination and Objectives

1.3.1 Problem Defination

This Project is mainly divided into two parts . The First Part involves of use of Natural Language Processing(NLP) to firstly analyze the post and materials posted on social media. Various Prepocessing techniques are used at this step firstly to make the data ready for the processing and second main part is that of analysis of this part of data , so as to predict the behaviour of person based on personality traits model i.e Big Five Model, MDTI Model or DISC Model.

1.3.2 Objective

Following are the Objectives of this Project -

- To Succesfully Implement Personality Prediction through tweets
- Use New Innovative LSTM Approach to get better Results
- Make the Model Yeild Better Results than the Traditional Baseline Methods
- Get an High Accuracy Model that can work for MBTI as well as Big Five Model
- Make an User Freindly Application Implementing this Model and Make it free and available for everyone to use

Chapter 2

Literature Survey

Sr no.	Paper Title, authors, year of publication	Concepts described in the paper	Gaps in the paper
978-1-538	Persona Identification Traits based on MBTI - A Text Classification Approach By Srilaxmi Bharadwaj , Srinidhi Sridhar , Rahul Choudhary and Ram Srinath	MBTI , CountVectorisation ,LIWC,Lemmetization TF-IDF	lack of implementation of deep learning methods , only baseline methods
978-1-4673	Predicting Student Personality Based on a Data - Driven Model from Student Behaviour on LMS and Social Networks by Mohamed Soliman Halawa, Mohamed Elemam Shehab and Essam M. Ramzy Hamed	JRIP,KNN IBK,MBTI,LMS OneR,Random Forest J48	Designed Specifically for Student Behaviour ignoring some of 16 MBTI traits
1556-4681	Emerging Trends in Personality Identification Using Online Social Networks by Vishal Kaushal and Manasi Patwardhan in ACM Transactions on Knowledge Discovery from Data, Vol. 12, No. 2, Article 15.	Models Used , NLP Techniques , Various Machine Learning Classifiers Models of Personality	No detailed analysis of impact of various ML Classifiers that are being used here

Sr no.	Paper Title, authors, year of publication	Concepts described in the paper	Gaps in the paper
2169-3536	DI Xue, Zheng Hong and Shize Guo, "Personality Recognition on Social Media With Label Distribution Learning," in IEEE Conference, vol. 29, pp. 265-276, 2019.	Pearson Correlation, Label Distribution, PT-SVM Classifiers, Big Five Model	Here the prediction takes place for very local chinese market using Textmind a chinese language specific psyo-analysis tool which is not easily applicable to other language dataset
2169-3538	Personality Predictions Based on User Behaviour on the Facebook Social Media Platform by Micheal M. Tadesse, Bo Xu , Liang Yang at IEEE Conference Supported by Natural Science Foundation of China (No. 61632011)	Big Five Model SNA, Splice, LIWC XGBoost, Gradient Boost OpenNLP	Model is specifically based on very basic sample dataset so accuracy may or may not hold true.
1541-1672/17	Navonil Majumder, Soujanya Poria, Alexander Gelbukh, "Deep Learning-Based Document Modeling for Personality Detection from Text,"2017	Network Architechture, Feature Extraction, Pre-processeing Convolution NN Neural Network, Word Vector	Near Perfect Model with visibally no literature gaps

Chapter 3

Software Requirement Specification

3.1 Assumption and Dependencies

3.1.1 Assumption

- One Of the Assumption is that tweets are written by the user himself rather than by some representative or agencies. As this would not give exct personality of the person
- We are taking into Assumption that the Application user has stable internet connection so that we can contact server to get latest tweets
- We are Assuming the user to have decent processing power so that processing time doesn't go out of hand

3.1.2 Dependencies

- We are Dependant on Twitter Servers to extract Tweets and in case of any failures in twitter servers , our application also wouldnt be able to function properly
- Our RNN Model is dependant on various Python Machine Learning and Keras , Sci-kit Learn Libraries and it wont run without these libraries

3.2 Functional Requirements

In software engineering, a functional requirement defines a system or its component. It describes the functions a software must perform. A function is nothing but inputs, its behavior, and outputs. It can be a calculation, data manipulation, business process, user interaction, or any other specific functionality which defines what function a system is likely to perform.

3.2.1 System Feature 1

Calculate Traits through the Tweets
Model Should First Calculate Exact Traits for every Trait of MBTI Model and Big Five Model

3.2.2 System Feature 2

Data Acquisition
Models Extract the Data From two Dataset for training for MBTI Model it extract Kaggle Dataset named MBTI_1 while for Big Five Model it uses MyPersonality Dataset

3.2.3 System Feature 3

Tweets Extraction
We Use Tweepy Library to Extract 200 latest Tweets from the entered username So the System Should Extract the Top Tweets in an array seperated by single quotes

3.2.4 System Feature 4

Data Preprocessing
The Twitter Data needs to be preprocessed before feeding into the RNN Model for personality extraction

3.3 Non Functional Requirements

3.3.1 Model Accuracy should be High

The RNN Model Should have a minimum acccuracy of 70% so that we do not give false results for any username

3.3.2 Sprcific Trait Points

In Big Five Calculation we should get exact points for every trait like OCEAN so that we get great results

3.3.3 Big Five Trait Point shouldn't exceed 5 points

We give each Trait in Big Five Model points out of Five (5) . Points for each trait shouldn't exceed 5

3.3.4 Seccurity

Security requirements ensure that the software is protected from unauthorized access to the system and its stored data. It considers different levels of au-thorization and authentication across different user roles. For instance, data privacy is a security characteristic that describes who can create, see, copy, change, or delete information. Security also includes protection against viruses and malware attacks.

3.3.5 Reliability

Reliability defines how likely it is for the software to work without failure for a given period of time. Reliability decreases because of bugs in the code, hardware failures, or problems with other system components. To measure software relia-bility, you can count the percentage of operations that are completed correctly or track the average period of time the system runs before failing.

3.3.6 Performances

Performance is a quality attribute that describes the responsiveness of the sys-tem to various user interactions with it. Poor performance leads to negative user experience. It also jeopardizes system safety when it's is overloaded.

3.4 System Requirements

3.4.1 Hardware Requirements

- NVIDIA Graphicc Card (For Faster Processing)
- 4GB DDR4 Ram or Above
- i5 - 5th Generation or above

3.4.2 Software Requirements

- Ubuntu 18.04 or Above
- Anaconda Enviornment
- Tensorflow , NLTK , Keras and ML Libraries
- Recommended Python Version Above 3.5

3.4.3 Application Requirements

- Android OS with Lollipop or Above
- Good Midrange Processor Snapdragon 625 or above
- 3GB LPDDR4x RAM

3.5 Software Devlopment Life Cycle

Software development life cycle is essentially a series of steps, or phases, that provide a model for the development and life cycle management of an application or piece of software.It is well defined, structured sequence of stages in software engineering to develop the intended software product or module.

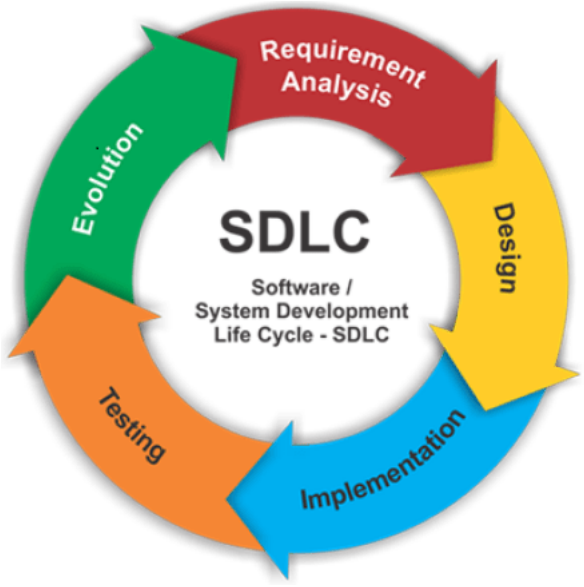


Figure 3.1: Software Devlopment Life Cycle

3.5.1 Requirement Analysis

Requirement analysis is the most important and fundamental stage in SDLC. It is performed by the senior members of the team with inputs from the customer, the sales department, market surveys and domain experts in the industry. This information is then used to plan the basic project approach and to conduct product feasibility study in the economical, operational and technical areas.
We Firstly Gathered All the Requirements and then we analyzed what we needed to do exatly as a project . So We Dedicated Majority Of our time in Gathering Requirements and Planning the Project.

3.5.2 Designing

A design approach clearly defines all the architectural modules of the product along with its communication and data flow representation with the external and third party modules (if any). The internal design of all the modules of the proposed architecture should be clearly defined with the minutest of the details in DDS.
SRS is the reference for product architects to come out with the best architecture for the product to be developed. Based on the requirements specified in SRS, usually more than one design approach for the product architecture is proposed and documented in a DDS - Design Document Specification.
We Talked with our mentors and using their inputs and our researh regarding the Subject , we made a clear product Architecture that would be robust and then would work with every model

3.5.3 Project Implementation

In this stage of SDLC the actual development starts and the product is built. The programming code is generated as per DDS during this stage. If the design is performed in a detailed and organized manner, code generation can be accomplished without much hassle.
Developers must follow the coding guidelines defined by their organization and programming tools like compilers, interpreters, debuggers, etc. are used to generate the code. Different high level programming languages such as C, C++, Pascal, Java and PHP are used for coding. The programming language is chosen with respect to the type of software being developed.
We Used Top to Bottom Approach to build the Project , We First Completed Preproccessing Module , then Data Extraction Module and finally move to RNN Model Building

3.5.4 Testing

This stage is usually a subset of all the stages as in the modern SDLC models, the testing activities are mostly involved in all the stages of SDLC. However, this stage refers to the testing only stage of the product where product defects are reported, tracked, fixed and retested, until the product reaches the quality standards defined in the SRS.
We Divided The Testing Phase into Functional , Non Functional and Maintainence Phase . So that we can concentrate on each and every aspect of testing and do not miss out on any bugs.

3.5.5 Evolution

The last Phase Adds some important Tweaks to make the product experience better. This Tweaks adds so that end user feel more comfortable while using the application and we do not feel out of place

We Concentrated more on User Interface of Our Android App making the UI Standard throughout , using the card layout , and using gradient coloring just to name a few

Chapter 4

Feasibility Study

4.1 Technical Feasibility

The project requires the basic knowledge of NLP and ML classifiers. It makes use of some basic preprocessing techniques and use word vector methods to be used in pyschological models such as BIG FIVE which help in the processing of data.

4.2 Hardware Feasibility

Adequate GPU and processing power is required so that the model can be trained in an efficient manner. Enough storage space is required as the training files take a lot of space.

4.3 Software Feasibility

Pyhton, basic libraries for machine learning classification models and for NLP techniques like N-grown algorithm, Anaconda, tensorflow etc.

4.4 Economic Feasibility

The project is economic as the only expenditure that is required is the cost of setting up a machine with good computing and software capabilities and no additional hardware is required.

4.5 Schedule Feasibility

The project can be completed in a timely manner as the only time taking part will be training the model in an accurate manner. Everything else depends on the technical

aspect.

4.6 Operational Feasibility

Provided the training model has high accuracy, the model will be helpful in predicting personality using the given dataset.

Chapter 5

System Design

5.1 System Architechture

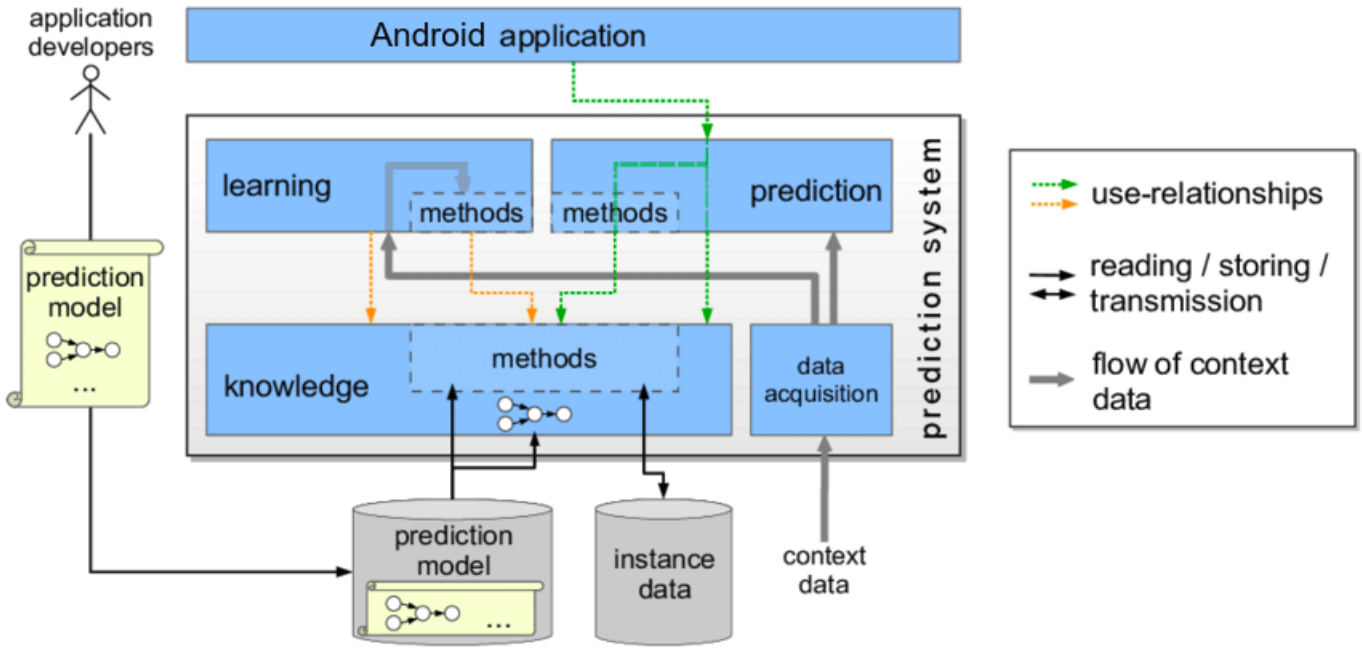


Figure 5.1: System Architecture

5.2 Mathematical Model

For overall system:
Let S be the system such that,
S = I, O, FN, SC, FC
Where,

- Input

I = Set of all possible inputs
I1 :User Tweets/Posts(Based on their accounts)

- Output

O = Set of all possible outputs
O1 : Personality Traits (INFP,INFJ,INTP,INTJ,ENTP,ENFP ,ISTP,ISFP,ENTJ,ISTJ,ENFJ,ISFJ,ESTP,ESFP,ESFJ,ESTJ)

- Function

FN1 : Preprocess()
FN2 : wordembedding()
FN3 : model()

- Success Conditions

Sc = Set of success cases
SC1 : Exact trait to be recognised

- Failure Conditions

Fc = Set of failure cases
FC1 : Wrong Trait Prediction

To Calculate RNN Output for this we use the formulae for RNN Learning

$$Ht = \sigma(U * Xt + W * Ht\text{-}1)$$
$$yt = \text{Softmax}(V * Ht)$$
$$J^t(\theta) = -\sum_{j=1}^{|M|} y_{t,j} \log \bar{y}_{t,j}$$
$$J(\theta) = -\frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{|M|} y_{t,j} \log \bar{y}_{t,j}$$

M = vocabulary , J(θ) = Cost function

5.3 UML Diagrams

5.3.1 Use Case Diagram

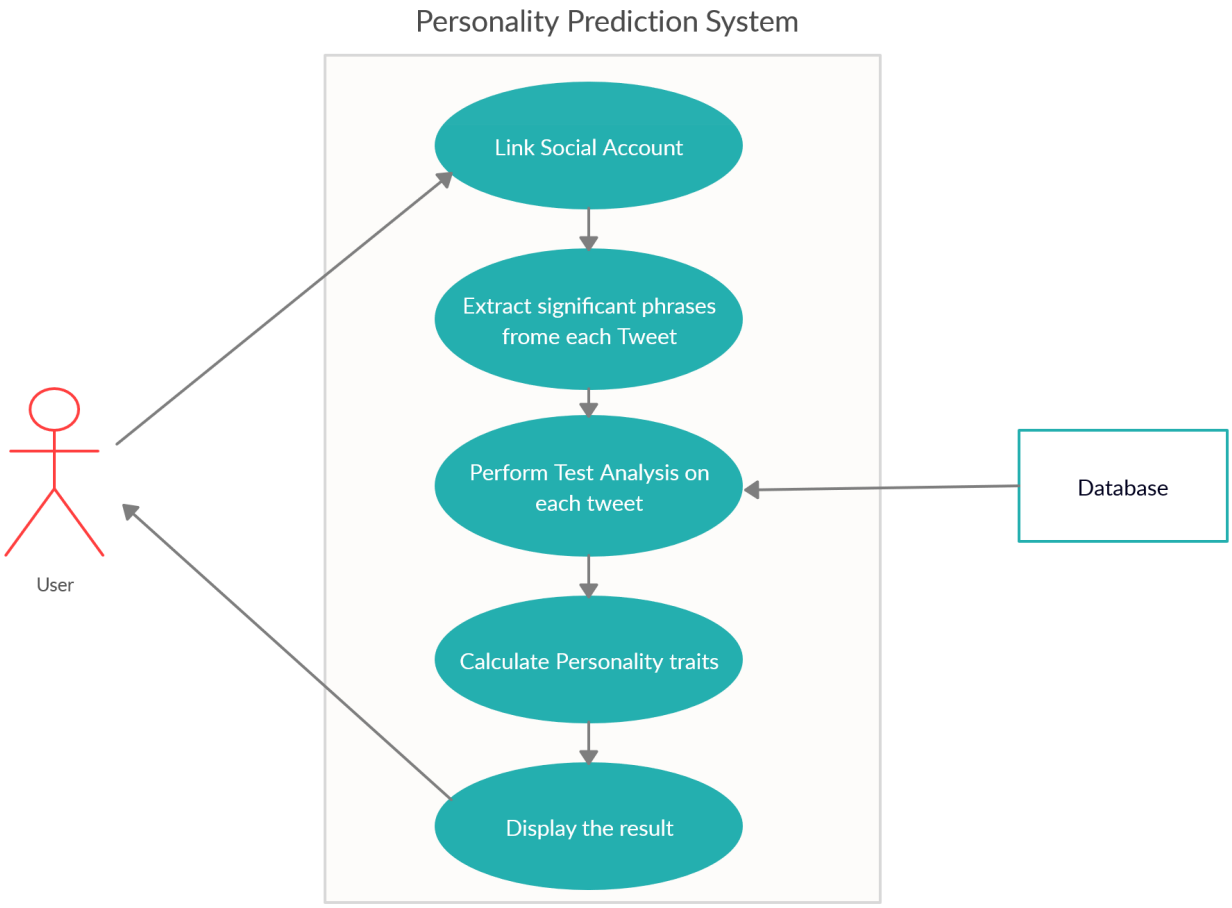


Figure 5.2: Use Case Diagram

5.3.2 Class Diagram

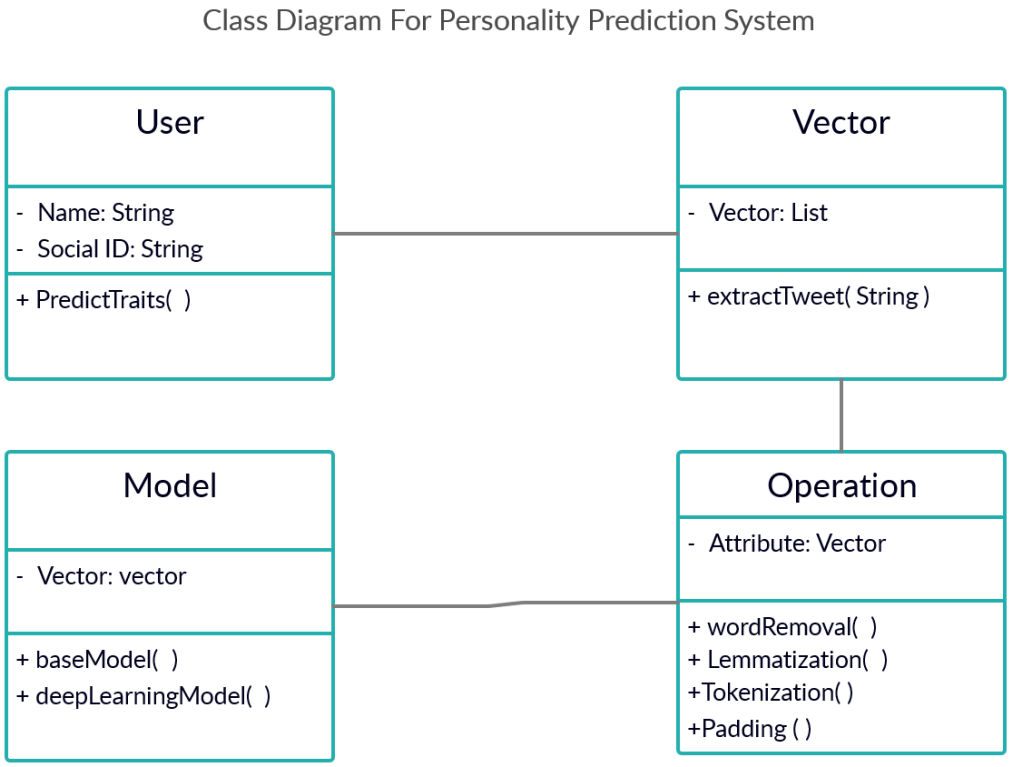


Figure 5.3: Class Diagram

5.3.3 Activity Diagram

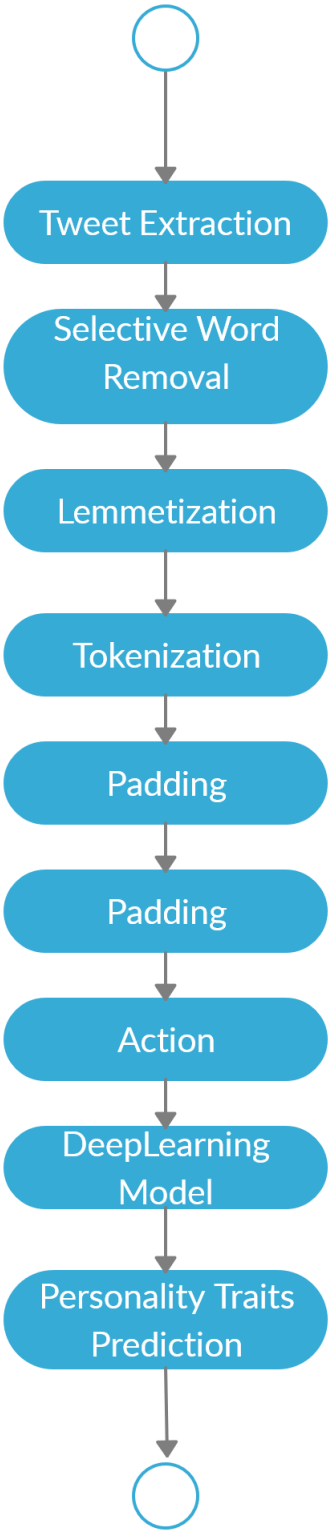


Figure 5.4: Activity Diagram

5.3.4 Sequence Diagram

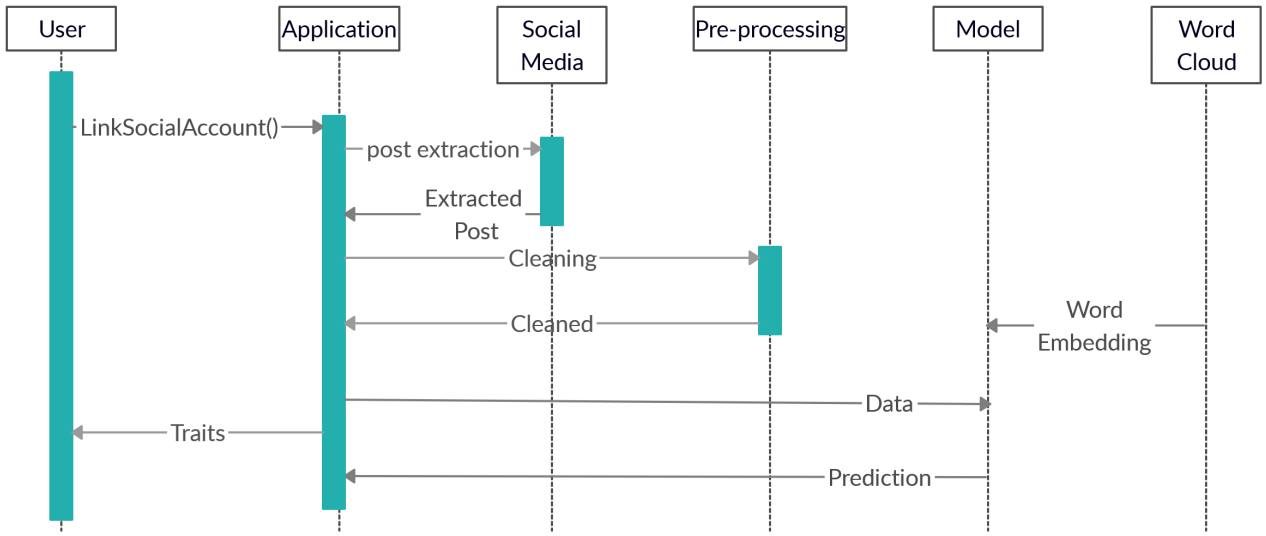


Figure 5.5: Sequence Diagram

Chapter 6

Project Plan

6.1 Project Estimate

The estimation is a process of giving value of things approximately which can be used for further transactions. This often can be considered as a rough calculation for a transaction.
The project estimation is a rough idea of how much input a project is going to need. It gives the developer and the client/customer a idea about the value needed to build a project. This estimates can be used for considering whether a project is feasible in real time.
The project estimation has three parts :-

1. Effort Estimation :-

The Effort Estimate includes estimation about the efforts that are to be put in a project . The Efforts in this Project are as follows -
1. Finding Suitable Dataset
2. Cleaning Dataset
3. Building a Suitable Classfiction Model
4. Getting Proper Accuracy through prediction model

2. Cost Estimation :-

The Cost Estimate contains the cost that incurr during project devlopment. There are no costs that we incurr during the project devlopment . We Used Chaquopy Python Library for Android that is free for one time use but would require us to buy licence key for further usage also if we wish to upload this app on to PlayStore it would additionally require around 2000 Rs . So there is no cost involved as such in the project making.

3. Resource Estimation :-

The Resource Estimation includes the resources that you need to devlop a specific project . To devlop this we didnt need any special resource except a Desktop with Python Enabled and approx 4GB RAM for processing . As a user , you need just an Android Smartphone with decent ram and Android 5.0 Lollipop .

6.1.1 Reconciled Estimates

A reconciliation is an independent cost estimate that the end user can compare against the contractor's cost estimate, mitigating budget shortfalls and correcting identified deficiencies. Reconciliations can help ensure that differences between the two estimates are appropriate and reasonably expected. There are no Reconciled Estimate in our project as the estimated cost of our project and actual cost matches exactly and there are no additional or hidden cost incurred.

6.1.2 Project Resources

These can be divided into two parts -

1. Hardware Resource :- Hardware Resource we need are a laptop with deent enough specification

2. Software Resource :- Software Resources we need are Python libraries like nltk , tensorflow etc and Ubuntu or Windows Operating Systems

6.2 Risk Management

6.2.1 Risk Identification

Risk identification is the process of determining risks that could potentially prevent the program, enterprise, or investment from achieving its objectives. It includes documenting and communicating the concern. Following were the risk attached to our project -

1.RNN Model Risk :-

The Major Risk was to build a RNN Model that would incorporate our dataset and provide desired results as desired because we didnt wanted to use traditional baseline methods and thus we dedicated most of our time to make the RNN best and precise. The Risk was big because whole of our project was based on this RNN Method and we also had to optimise the Algorithm to use MBTI and Big Five Model

2. DataSet Selection Risk :-

Another Major Risk was the selection of a Suitable Dataset that would be labelled and thus we could learn more from the data , initially we were not finding a suitable dataset but then we found this dataset and found it suitable for all of our MBTI and Big Five Model.

3. Opportunity Risk :-

This Risk says that time spent and effort developing something when another solution would have been far more successful. So we had to devlop a solution that produe a viable solution and give better results than any other solution currently available in the market.

4. Output Risk :-

This Risk deals with the priniple that the risk was we could not achieve the results we expeted to produce . This ould mean the failure of whole of project and would also have lead to ccomplete abondening of the project

6.2.2 Risk Analysis

Risk analysis is the process of identifying and analyzing potential issues that could negatively impact key business initiatives or projects. This process is done in order to help organizations avoid or mitigate those risks. Risk analysis can help an organization improve its security in a number of ways. Depending on the type and extent of the risk analysis
The Above Risks were too big so we deided to deal with this Risks and calulated the severity of each of the Risk . The First two risk were the most dangerous risk therfor we decided to dedicate majority of time solving these risk and we made sure during each step to see whether these risk are being handled succesfully or not
The Other Two Risks are the direct outcome of the failure to make peace with first two risk , if we fail to avoid the first two risks then we are sure to fail in other two risks as well . Therefor , concentrating on these two risk is not necessary as we have to dedicaate majority of our resources to counter first two risks

6.2.3 Risk Management , Mitigation , Monitoring

he goal of the risk mitigation, monitoring and management plan is to identify as many potential risks as possible.

Risk Mitigation is a problem avoidance activity, Risk Monitoring is a project tracking activity,Risk Management includes contingency plans that risk will occur.The following actions will be taken to reduce risks in this project:

- Dedicate all the major resources to Building the RNN Model
- Find and try to make the dataset labelled so that more can be learnt through it
- Regularly make backups of project report and important documents, if an unplanned day off occurs make up for work lost.
- Immediately storing the data and applying processing on it
- Github to backup code and prevent loss, use own initiative to keep on top of project plan

6.3 Project Schedule

Project scheduling is a mechanism to communicate what tasks need to get done and which organizational resources will be allocated to complete those tasks in what timeframe.A project is made up of many tasks, and each task is given a start and end (or due date), so it can be completed on time.
Project scheduling occurs during the planning phase of the project.

6.3.1 Project Task Set

The primary tasks included during the project stages are:

1. Problem Statement Formulation
2. Concept Scoping and Planning
3. Risk Assessment
4. Proof of Concept
5. Implementation of each module
6. Integration of modules with the main system
7. Testing and Deployment

6.3.2 Task Network

We Combined Some Tasks and Distribute them among various group members so that we could achieve our tasks . We combine tasks to form networks and thenthrough co-operation of every group member we tried to perform each task network and thus we could perform task faster and could perform tasks faster

6.3.3 Timeline Chart

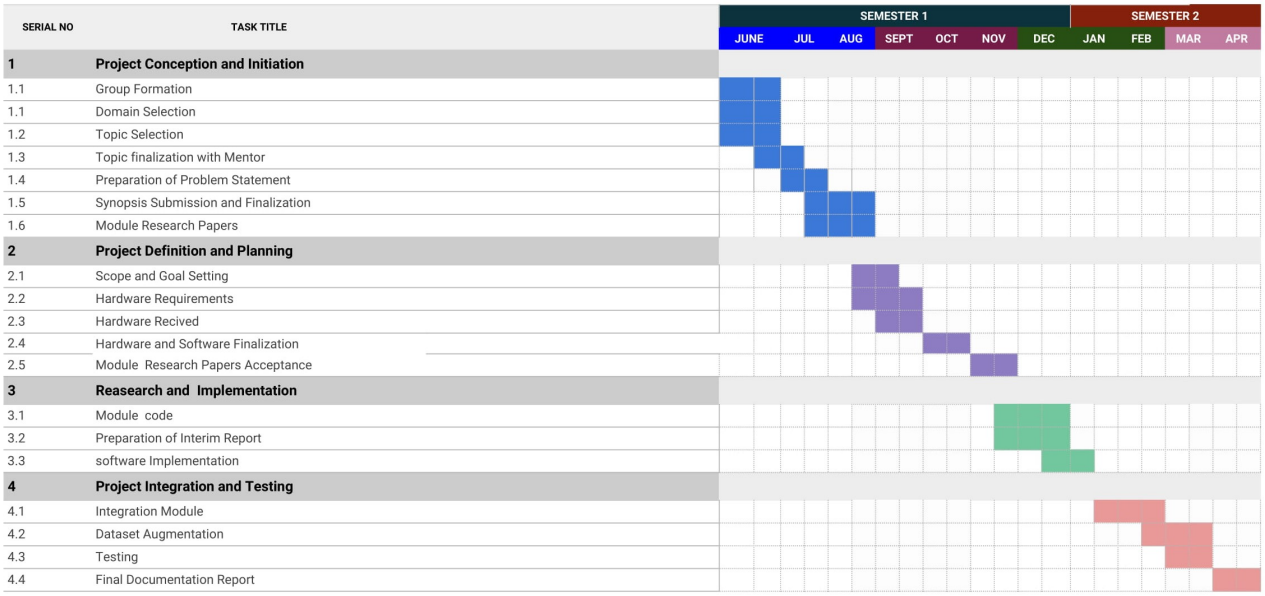


Figure 6.1: Timeline Diagram

6.4 Team Organisation

Our project team consisted of the following team members, all from the Computer Department:

1. Ajinkya Pingale
2. Jay Naik
3. Manoj Nandha
4. Akash Misal

Along with these team members, our internal guide Prof. Mamta Bhamre Mam was the most important and indispensable part of the team who provided us with

sound guidance, motivation, support and deadlines without whom our project would not have been timely managed.

6.4.1 Team Struture

Our Team has all the members from the Computer Department and we have tried to make the group as diverse as possible so that we can collabrate on more distinct ideas and distincts approaches

6.4.2 Management Reporting and Communication

The management of the timely completion of tasks was seen to by our internal guide. We, the group members of the project managed the entire project by splitting it into multiple unit tasks and by assigning deadlines to them. The reporting of the project status to our guide was done as and when the unit tasks were accomplished. Similarly, the reporting was timely and up to date. The communication between the group members was through college meetings, discussions, emails and text message groups. The distribution of tasks, queries and integration was resolved through communication. Similarly,the communication of the members with Prof. Mamta Bhamre Mam was through reviews, personal meetings and emails.

Chapter 7

Project Implementation

7.1 Project Modules

7.1.1 Dataset Description

We are using Two Different Dataset for Training Our Models . One for training model to predict MBTI Trait and other for predicting Big Five Trait . We Searched Various Websites to get to the best Dataset for our use as it would have been not an impressive model if it hadn't had enough training Material in it . So we Decided to choose Two Dataset so that we could learn more through each and Information Doesnt get repeated , We also Decided to select large dataset so that model is the most acccurate with enough examples of each trait .These Two Datasets are from different sources and here are its description -

DataSet for MBTI

We used a kaggle Dataset for training MBTI model this Dataset is a labelled dataset and has more than 8600+ Rows and It has two columns One Showing a speccific Trait and other showing Tweets of that Type . This Dataset isnt a Clean one and there is lot of irrelevant things that we would have to remove in preprocessing Step.Each Tweet in a row are seperated by 3 pipe pattern and this tweets also has some additional charectars like @ "" etc. It has more than 433000 Tweets
This data was collected through the PersonalityCafe forum and then uploaded to kaggle, as it provides a large selection of people and their MBTI personality type, as well as what they have written.
It Contains Majority of Posts for INFP (21%) and INFJ (17%)

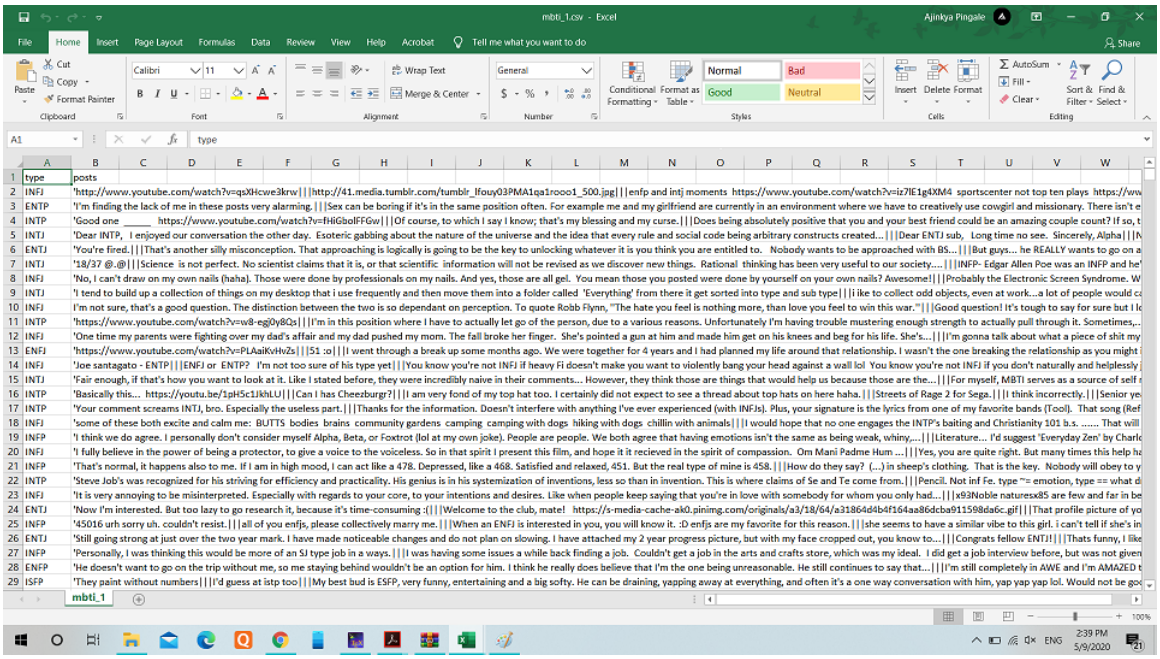


Figure 7.1: MBTI Dataset

DataSet for Big Five

We used a MyPersonality Dataset for training Big Five model this Dataset is a labelled dataset and has more than 9918 Rows and It has 21 columns Each One Showing a speccific Tweet and Out of Five points on each of five traits . This Dataset isnt a Clean one and there is lot of irrelevant things that we would have to remove in preprocessing Step.It has more than 9918 Tweets as it contains only one tweet per person or AuthorId

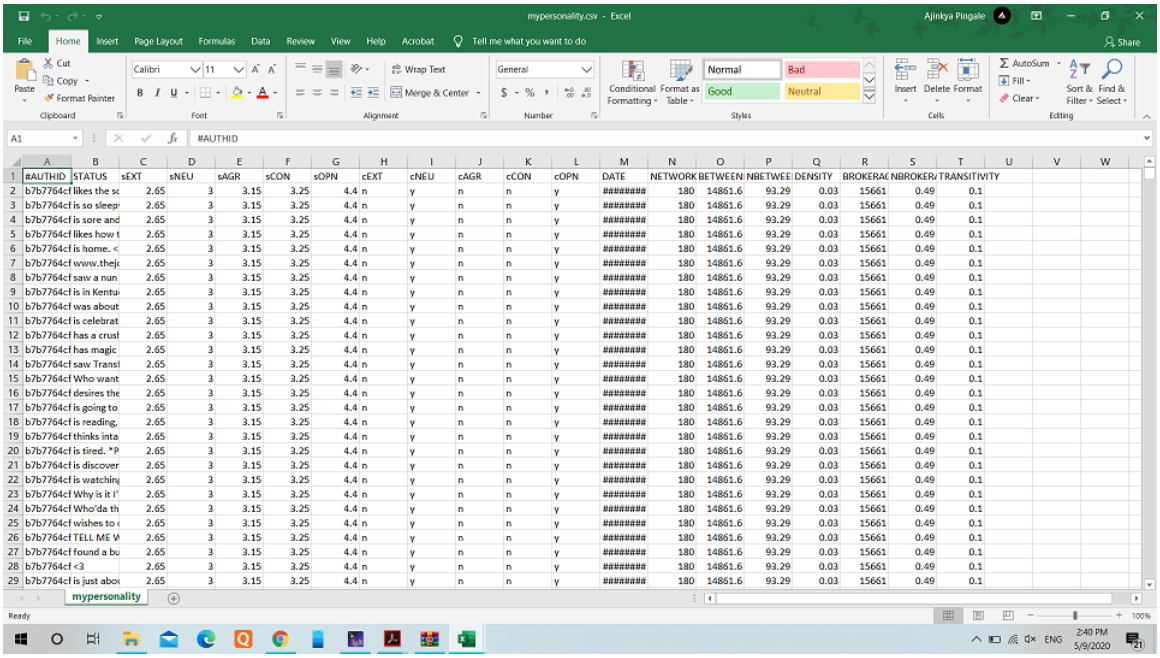


Figure 7.2: Big Five Dataset

7.1.2 Preprocessing

To Process these Dataset we Used Various Predefined Python Libraries and used them to manipulate Database so That we can feed it to the model

- We First Seperated Each Tweet From the Column and removed links , quotations etc
- We Lemmatized The Word To Reach The Root Word
- Using StopsWords Libraires From NLTK We Removed Unneccessary StopWords
- And At last we Removed MBTI Keyword from each Tweet
- Same Process was repeated for each and every tweet and data was divided into cclean and unclean data

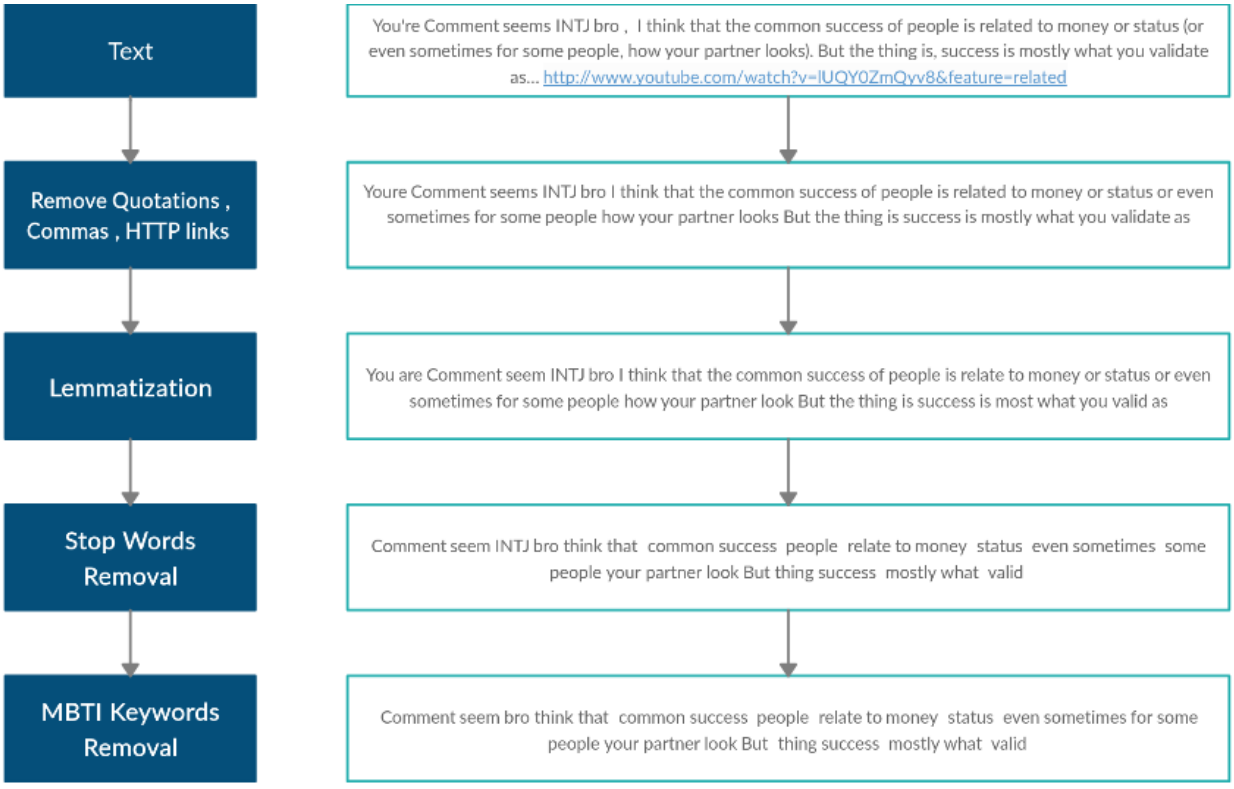


Figure 7.3: Example of Data Preprocessing

7.1.3 MBTI Model

The Myers–Briggs Type Indicator (MBTI) is an introspective self-report questionnaire indicating differing psychological preferences in how people perceive the world and make decisions.
he MBTI is based on the conceptual theory proposed by Swiss psychiatrist Carl Jung, who had speculated that people experience the world using four principal psychological functions – sensation, intuition, feeling, and thinking – and that one of these four functions is dominant for a person most of the time. The four categories are Introversion/Extraversion, Sensing/Intuition, Thinking/Feeling, Judging/Perception. Each person is said to have one preferred quality from each category, producing 16 unique types. Most of the research supporting the MBTI's validity has been produced by the Centre for Applications of Psychological Type, an organization run by the Myers-Briggs Foundation, and published in the Centre's own journal, the Journal of Psychological Type, raising questions of independence, bias, and conflict of interest

Extraversion – Introversion

It signifies the source and direction of a person's energy expression. An extravert's source and direction of energy expression is mainly in the external world, while an introvert has a source of energy mainly in their own internal world.

Sensing – Intuition

It represents the method by which someone perceives information. Sensing means that a person mainly believes information he or she receives directly from the external world. Intuition means that a person believes mainly information he or she receives from the internal or imaginative world.

Thinking – Feeling

It represents how a person processes information. Thinking means that a person makes a decision mainly through logic. Feeling means that, as a rule, he or she makes a decision based on emotion, i.e. based on what they feel they should do.

Judging – Perceiving

It reflects how a person implements the information he or she has processed. Judging means that a person organizes all of his life events and, as a rule, sticks to his plans. Perceiving means that he or she is inclined to improvise and explore alternative options.

7.1.4 Big Five Model

The "big five" are broad categories of personality traits. While there is a significant body of literature supporting this five-factor model of personality, researchers don't always agree on the exact labels for each dimension.
It is important to note that each of the five personality factors represents a range between two extremes. For example, extraversion represents a continuum between extreme extraversion and extreme introversion.
These Five Traits are -

Openness

This trait features characteristics such as imagination and insight.People who are high in this trait also tend to have a broad range of interests. They are curious about the world and other people and eager to learn new things and enjoy new experiences.

Conscientiousness

Standard features of this dimension include high levels of thoughtfulness, good impulse control, and goal-directed behaviors.Highly conscientious people tend to be organized and mindful of details. They plan ahead, think about how their behavior affects others, and are mindful of deadlines.

Extraversion

Extraversion (or extroversion) is characterized by excitability, sociability, talkativeness, assertiveness, and high amounts of emotional expressiveness.People who are high in extraversion are outgoing and tend to gain energy in social situations. Being around other people helps them feel energized and excited.

Agreeableness

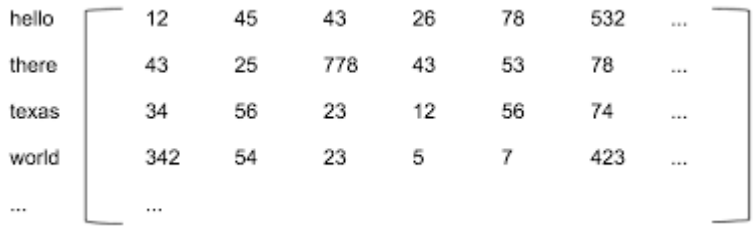
This personality dimension includes attributes such as trust, altruism, kindness, affection, and other prosocial behaviors.People who are high in agreeableness tend to be more cooperative while those low in this trait tend to be more competitive and sometimes even manipulative.

Neuroticism

Neuroticism is a trait characterized by sadness, moodiness, and emotional instability. Individuals who are high in this trait tend to experience mood swings, anxiety, irritability, and sadness. Those low in this trait tend to be more stable and emotionally resilient.

7.1.5 Embedding Matrix

The embedding matrix is randomly initialized and set as parameters to this context-guessing model. A cost can be calculated by seeing how closely the model guessed the context embedding, then the whole model can be trained using gradient descent Embedding Matrix -



7.1.6 Glove

Global Vectors for Word Representation, or GloVe, is an "unsupervised learning algorithm for obtaining vector representations for words." Simply put, GloVe allows us to take a corpus of text, and intuitively transform each word in that corpus into a position in a high-dimensional space. This means that similar words will be placed together.

GloVe is an unsupervised learning algorithm for obtaining vector representations for words. Training is performed on aggregated global word-word co-occurrence statistics from a corpus, and the resulting representations showcase interesting linear substructures of the word vector space. The pretained model is available in size of 50d,100d,200d and 300d. We Used Glove pretrained vectors for embedding layer

7.1.7 MBTI Architecture Module

Based on the twitter dataset which was available the average post length was aroung 40. So we set the max length of the posts as 40. Those 40 words are given as input to our lstm model. This model consists of an input layer of length 40 followed by the embedding layer. In the embedding layer, each word in post is converted into a vector 0f length 50 by using the glove 6b.50D matrix., resulting into matrix of 40*50 matrix input.

This matrix is used in the lstm layer which has 50 layers. Activation function is sigmoid with a dropout of 0.1. Finally at the last layer we added a dense layer with sigmoid activation function and used loss function of binary cross entropy. This model predicts two classes. We trained 4 models to predict 8 traits. Each model predciting between two classes

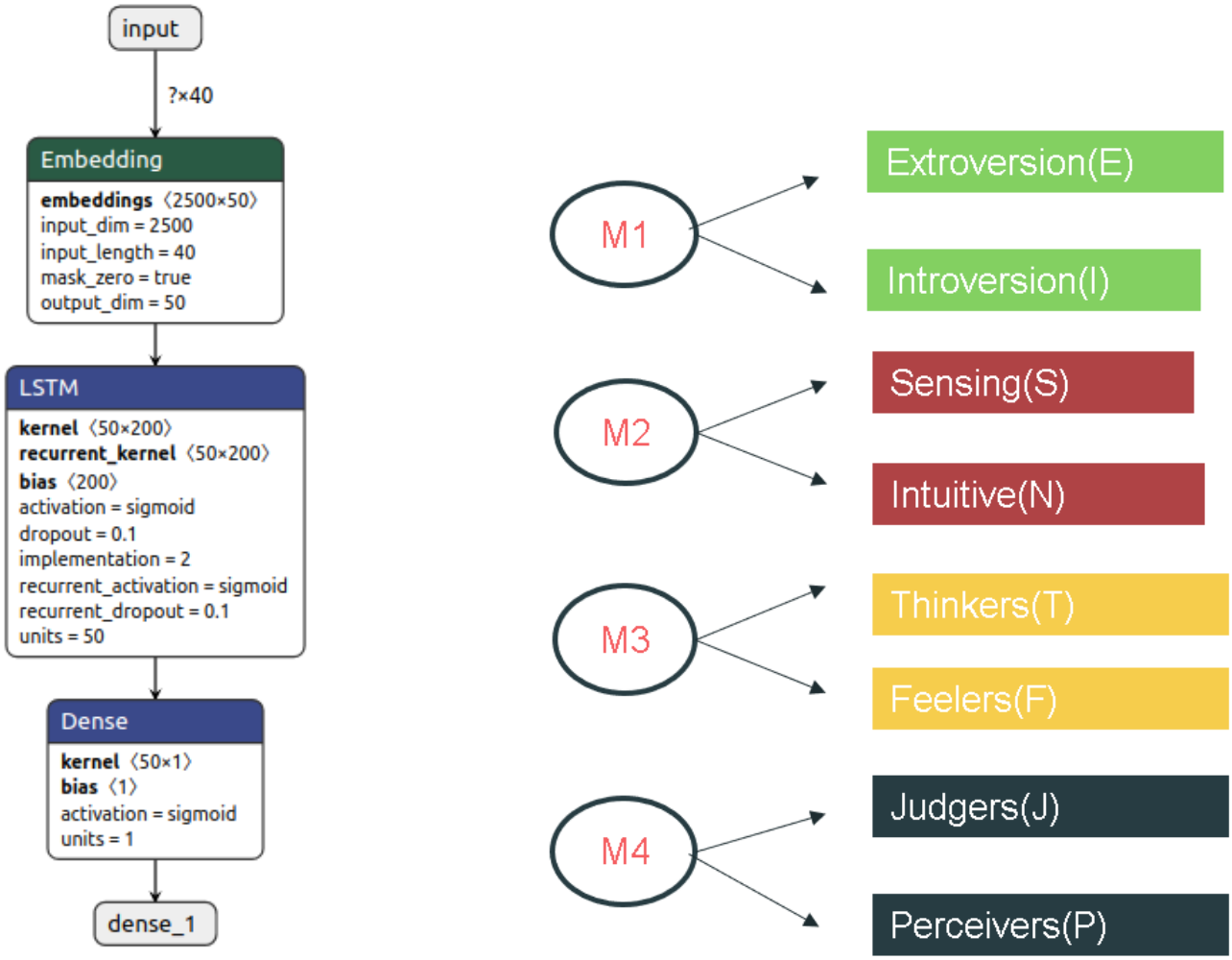


Figure 7.4: MBTI Module Architecture

7.1.8 Big Five Architecture Module

Based on my personality dataset, the average length of the post was found to be around 15. Therefore we set max length as 15 in this case. It's almost similar to the MBTI model, except we used the size of the embedding matrix to 300. Therefore after the embedding layer, The resultant matrix is of size 15*300. We added different dropouts to every layer in this model. LSTM layer size is 10 . We added a dense layer having activation function of relu. The last layer's activation function is set to mean suqre error function. This layer predicts value ranging from 0 to 5. We trained 5 models for 5 traits of Big 5.

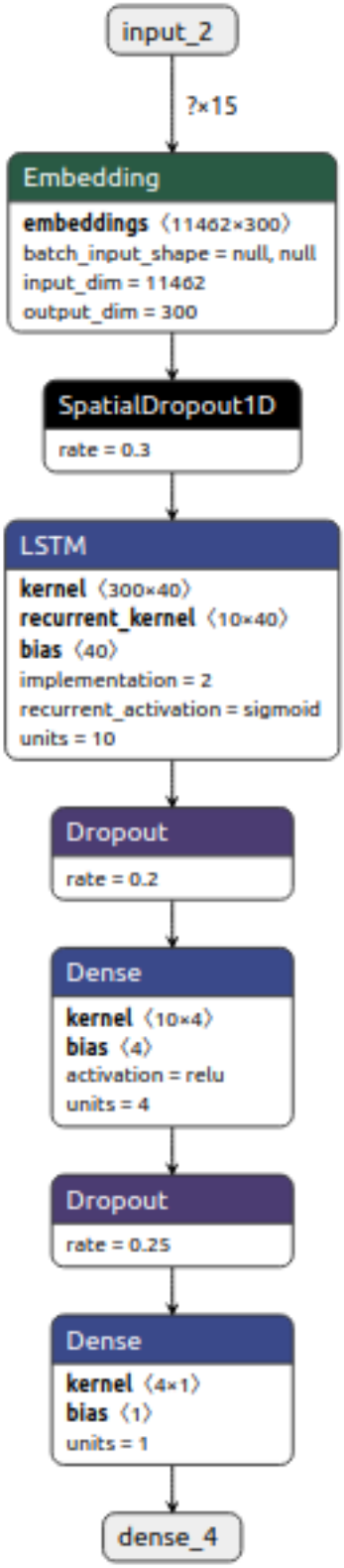


Figure 7.5: Big Five Architecture Module

7.2 Tools and Technologies Used

7.2.1 Jupyter - Notebook

The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text. Uses include: data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more.
We used Jupyter-Notebook along with Anaconda Python Enviornment to run and devlop our Python Files

7.2.2 Android Studio

Android Studio is the official integrated development environment (IDE) for Google's Android operating system, built on JetBrains' IntelliJ IDEA software and designed specifically for Android development. It is available for download on Windows, macOS and Linux based operating systems.It is a replacement for the Eclipse Android Development Tools (ADT) as the primary IDE for native Android application development.

Android Studio was announced on May 16, 2013 at the Google I/O conference. It was in early access preview stage starting from version 0.1 in May 2013, then entered beta stage starting from version 0.8 which was released in June 2014.The first stable build was released in December 2014, starting from version 1.0.

7.2.3 Python Libraries - Keras , Sklearn , tensorflow etc

Keras is an open-source neural-network library written in Python. It is capable of running on top of TensorFlow, Microsoft Cognitive Toolkit, R, Theano, or PlaidML. Designed to enable fast experimentation with deep neural networks, it focuses on being user-friendly, modular, and extensible. It was developed as part of the research effort of project ONEIROS (Open-ended Neuro-Electronic Intelligent Robot Operating System), and its primary author and maintainer is François Chollet, a Google engineer. Chollet also is the author of the XCeption deep neural network model

Scikit-learn (formerly scikits.learn and also known as sklearn) is a free software

machine learning library for the Python programming language. It features various classification, regression and clustering algorithms including support vector machines, random forests, gradient boosting, k-means and DBSCAN, and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy.

TensorFlow is a free and open-source software library for dataflow and differen-

tiable programming across a range of tasks. It is a symbolic math library, and is also used for machine learning applications such as neural networks.It is used for both research and production at Google. TensorFlow was developed by the Google Brain team for internal Google use.

7.3 Algorithms

7.3.1 Baseline Approach

A baseline is a method that uses heuristics, simple summary statistics, randomness, or machine learning to create predictions for a dataset. You can use these predictions to measure the baseline's performance (e.g., accuracy)– this metric will then become what you compare any other machine learning algorithm against.
Instead of building a model, predict the mean of the output variable. That means you take all observations with your output value, find the mean of all observations and use that as the prediction for every observation,
Build a linear regression model, since it is one of the simplest models. Then you just take the predictions and score them, and then that score is the score to beat with other machine learning models (e.g. like neural networks or gradient boosting).

7.3.2 LSTM

LSTM is a recurrent neural network (RNN) architecture that REMEMBERS values over arbitrary intervals. LSTM is well-suited to classify, process and predict time series given time lags of unknown duration. Relative insensitivity to gap length gives an advantage to LSTM over alternative RNNs, hidden Markov models and other sequence learning methods.
The structure of RNN is very similar to hidden Markov model. However, the main difference is with how parameters are calculated and constructed. One of the advantage with LSTM is insensitivity to gap length. RNN and HMM rely on the hidden state before emission / sequence. If we want to predict the sequence after 1,000 intervals instead of 10, the model forgot the starting point by then. LSTM REMEMBERS.

7.3.3 Recurrent Neural Network

Recurrent Neural Network is a generalization of feedforward neural network that has an internal memory. RNN is recurrent in nature as it performs the same function for every input of data while the output of the current input depends on the past one computation. After producing the output, it is copied and sent back into the recurrent network. For making a decision, it considers the current input and the output that it

has learned from the previous input.

Unlike feedforward neural networks, RNNs can use their internal state (memory) to process sequences of inputs. This makes them applicable to tasks such as unsegmented, connected handwriting recognition or speech recognition. In other neural networks, all the inputs are independent of each other. But in RNN, all the inputs are related to each other.

First, it takes the X(0) from the sequence of input and then it outputs h(0) which together with X(1) is the input for the next step. So, the h(0) and X(1) is the input for the next step. Similarly, h(1) from the next is the input with X(2) for the next step and so on. This way, it keeps remembering the context while training.

Chapter 8

Software Testing

SOFTWARE TESTING is defined as an activity to check whether the actual results match the expected results and to ensure that the software system is Defect free. It involves execution of a software component or system component to evaluate one or more properties of interest.Software testing also helps to identify errors, gaps or missing requirements in contrary to the actual requirements. It can be either done manually or using automated tools. Some prefer saying Software testing as a White Box and Black Box Testing.

8.1 Types of Testing

Software Testing that we have Implemented in our Project can be Classified into three Types -

8.1.1 Functional Testing

FUNCTIONAL TESTING is a type of software testing whereby the system is tested against the functional requirements/specifications. Functions (or features) are tested by feeding them input and examining the output. Functional testing ensures that the requirements are properly satisfied by the application.

8.1.2 Non Functional Testing

NON-FUNCTIONAL TESTING is defined as a type of Software testing to check non-functional aspects (performance, usability, reliability, etc) of a software application. It is designed to test the readiness of a system as per nonfunctional parameters which are never addressed by functional testing.

8.1.3 Maintainence Testing

Maintenance Testing is done on the already deployed software. The deployed software needs to be enhanced, changed or migrated to other hardware. The Testing done during this enhancement, change and migration cycle is known as maintenance testing.

8.2 Test Cases and Results

We Analyzed the Personalities of Following Popular Personalities whose MBTI Trait was already known

8.2.1 Donald J Trump

The President Of United State Of America tweets using the official twitter id of 'potus'. Using Our Model We Calculated the personality to be ISTP with I calculated at 54% whereas Studies suggest it to be ESTP with I calculated at 48% which is quite close

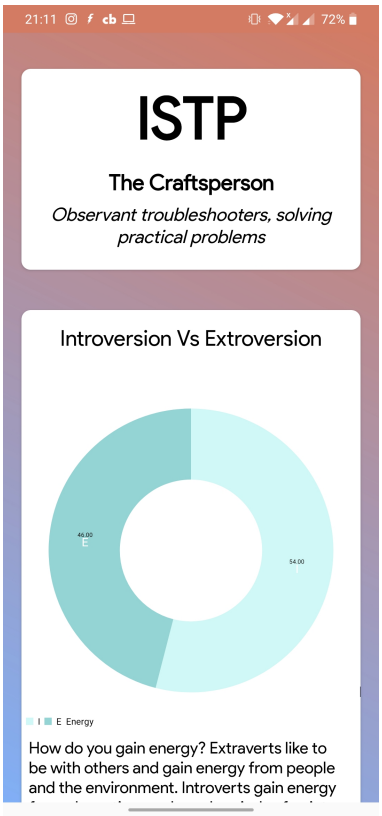


Figure 8.1: Trump Personality Prediction

8.2.2 Leonardo Di Caprio

The Famous Oscar Winning Actor Leonardo Di Caprio who uses twitter if of 'LeoDiCaprio'. Our Model Predicted the personality to be ESTP where as studies suggest that result to be ESFP

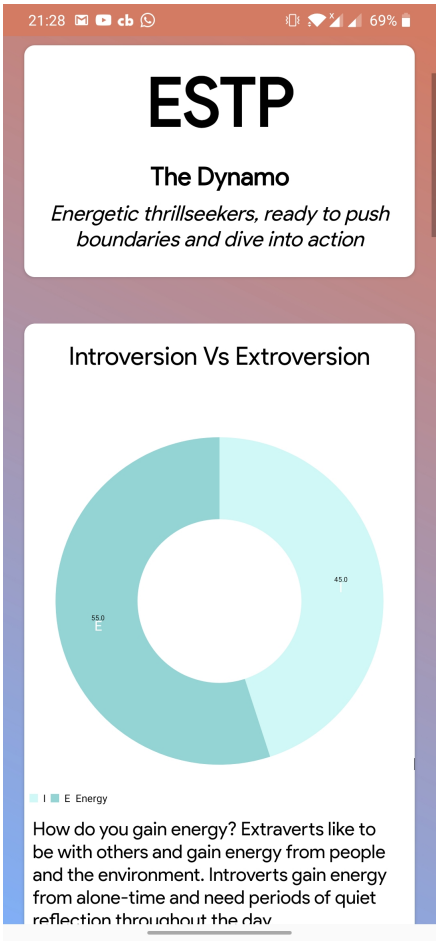


Figure 8.2: Leonardo Personality Prediction

Chapter 9

Results

9.1 Outcomes

We were able to achieve great results and accuracy through this models . Accuracies of these traits under some specific algorithms are below . We will first Discuss the Accuracy of MBTI Model and then of Big Five Model

9.1.1 MBTI Model

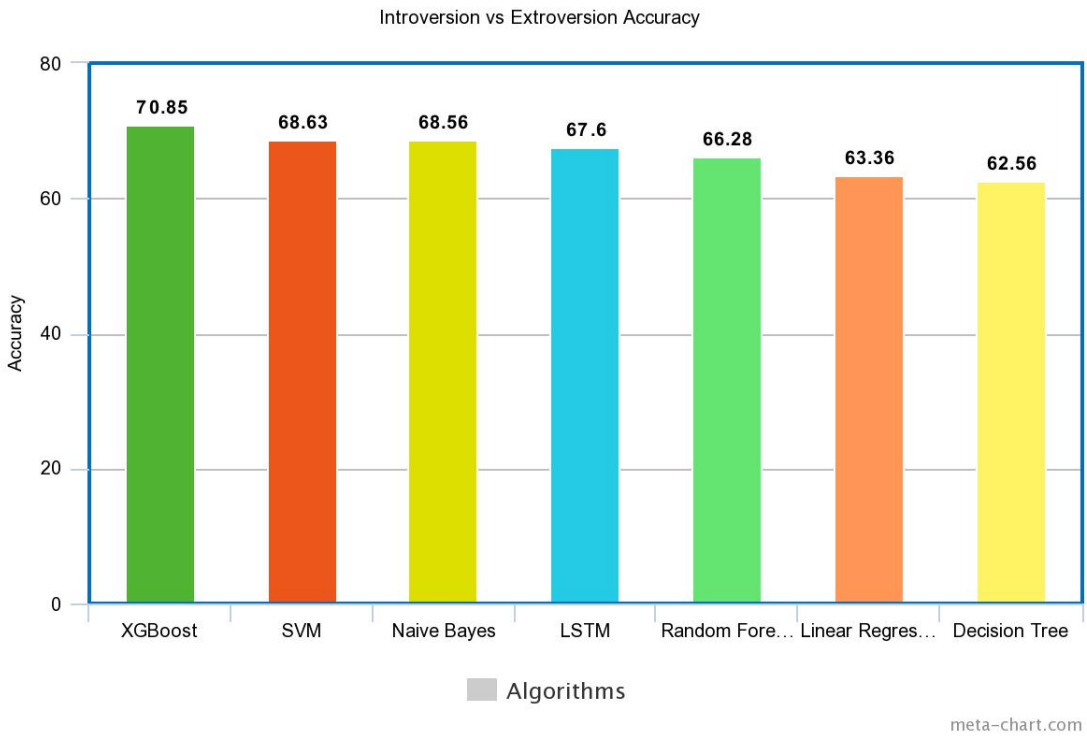


Figure 9.1: Introversion Vs Extroversion

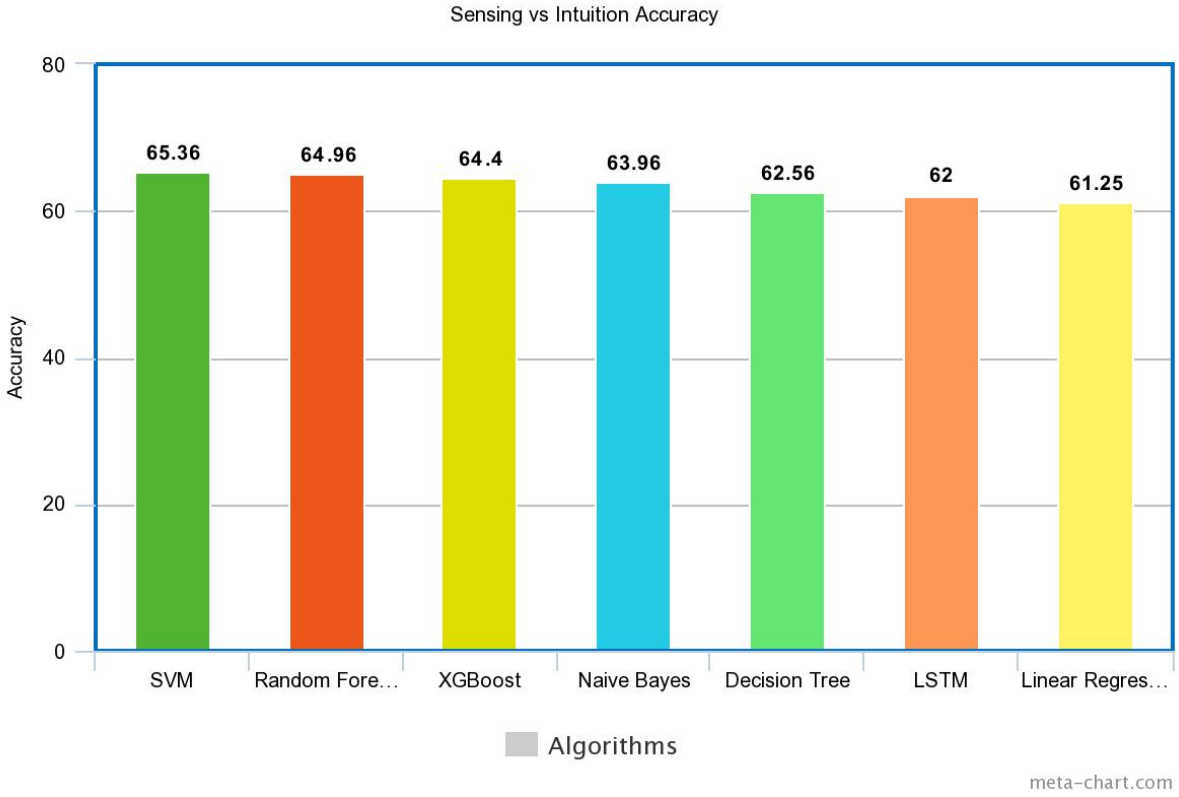


Figure 9.2: Intution Vs Sensing

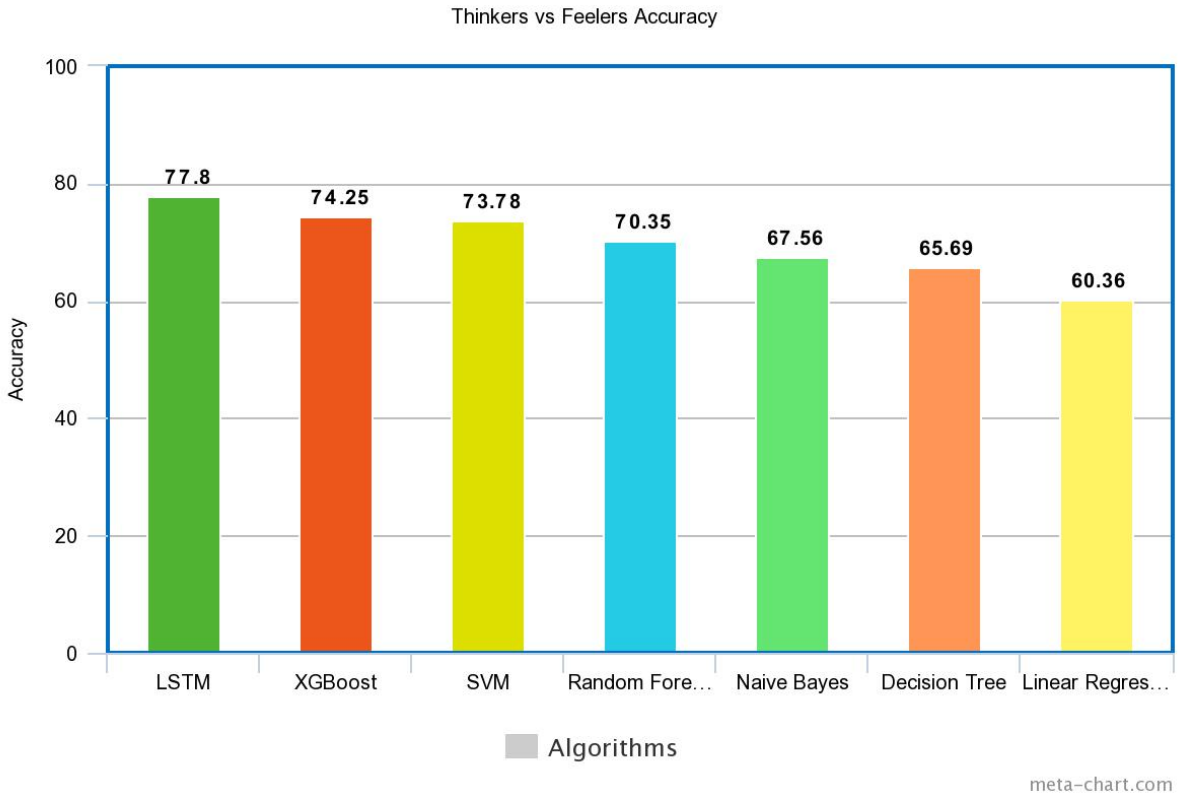


Figure 9.3: Thinking Vs Feeling

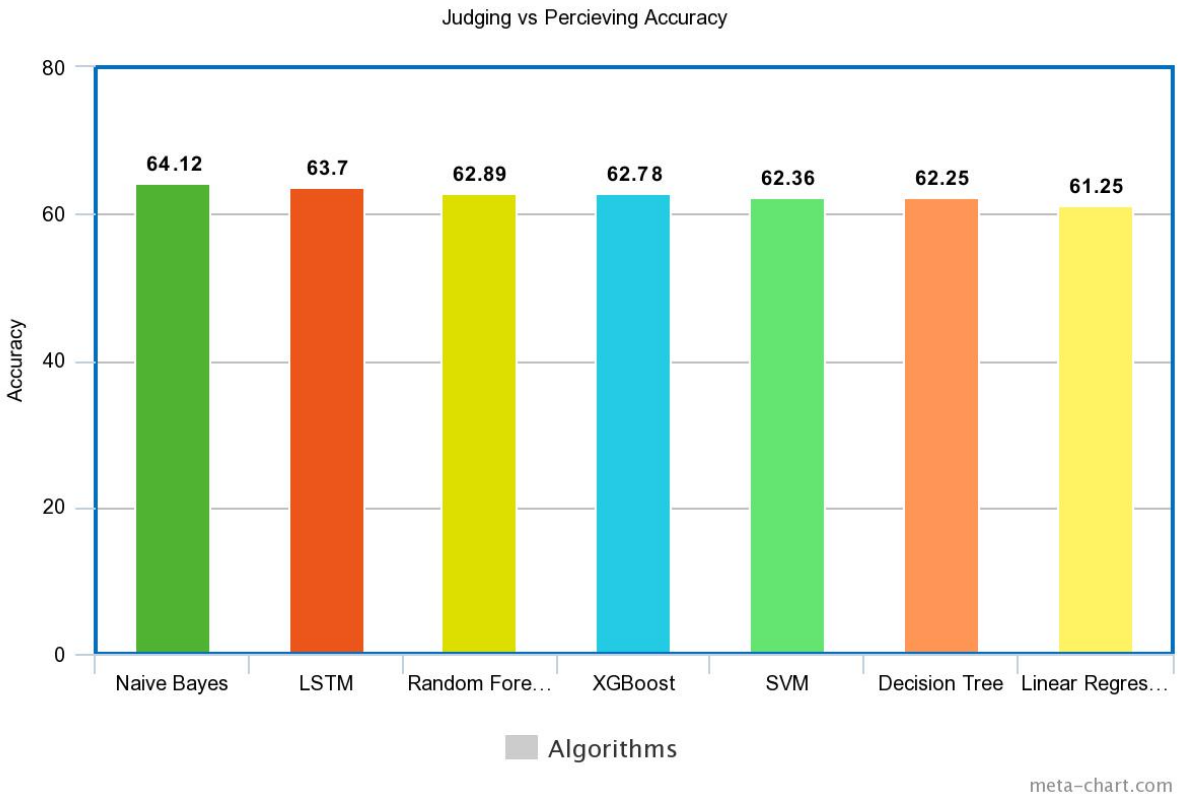


Figure 9.4: Percieving Vs Judging

9.1.2 Big Five Model

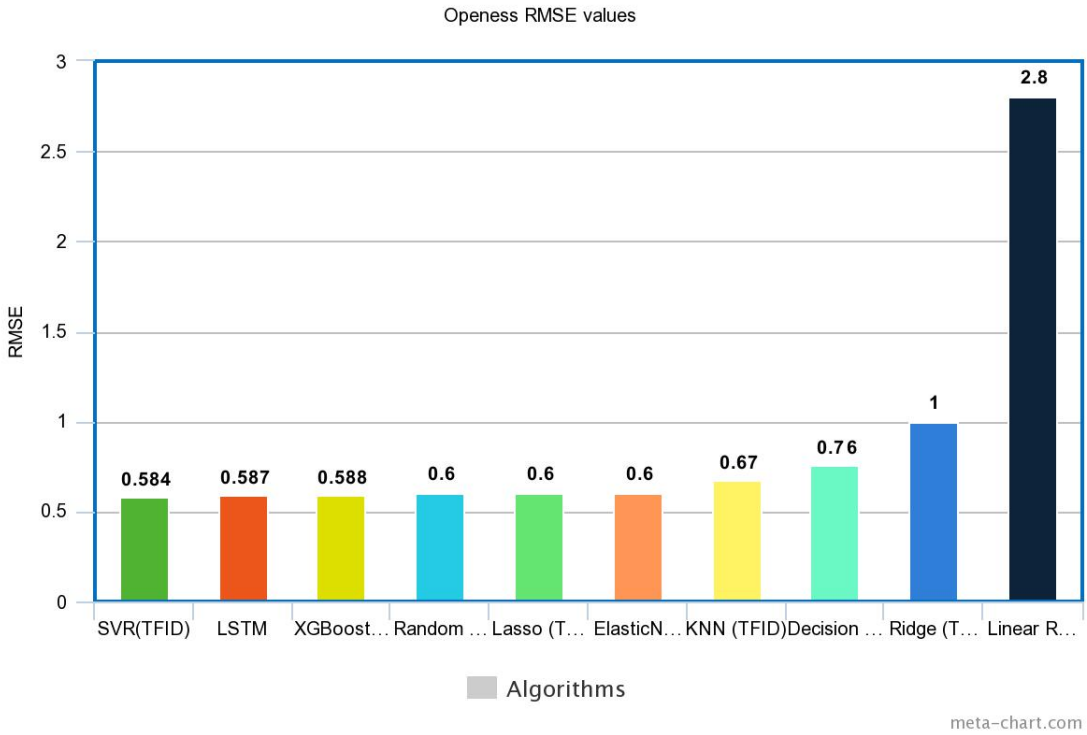


Figure 9.5: Openness

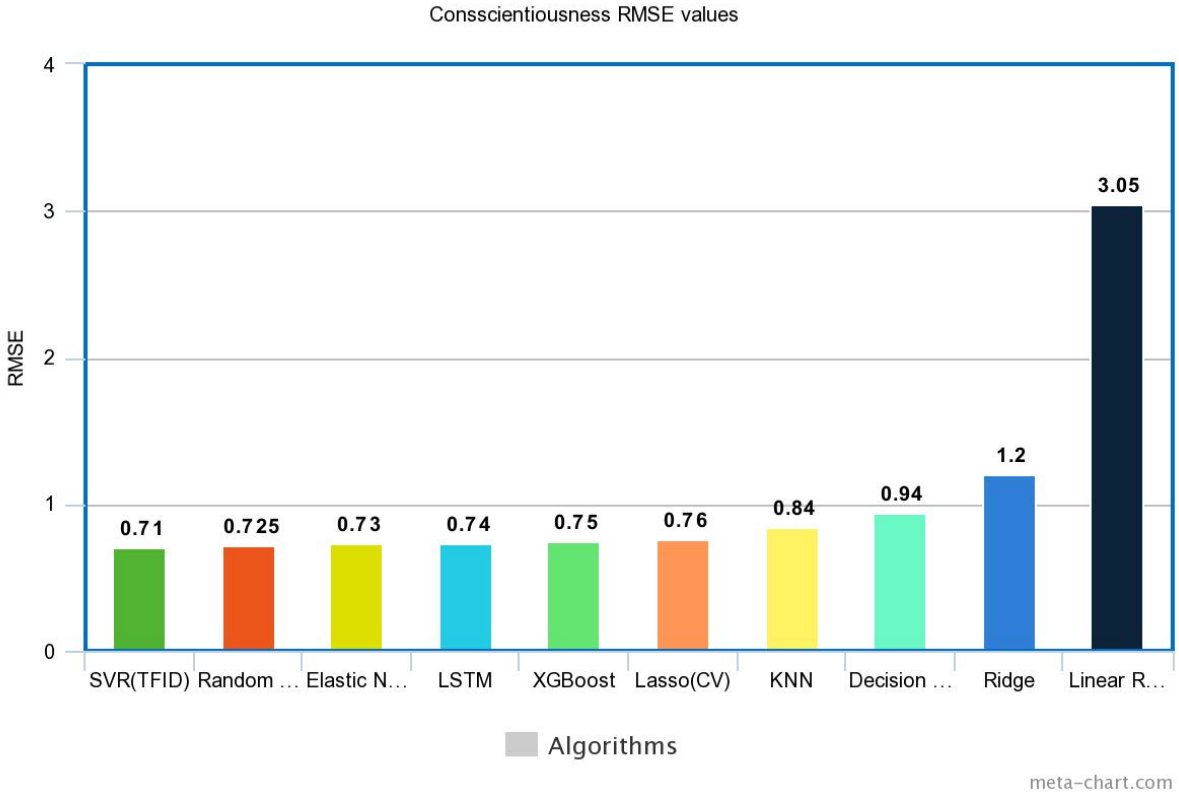


Figure 9.6: Conscientiousness

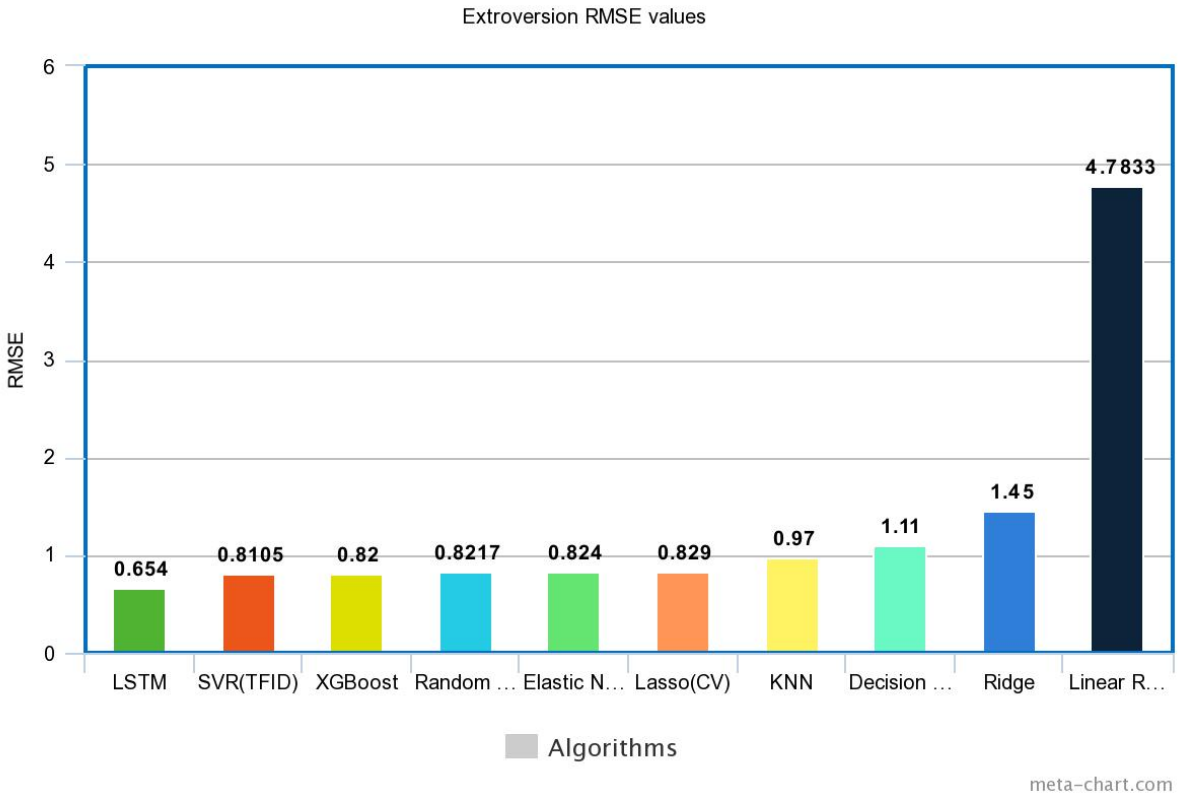


Figure 9.7: Extraversion

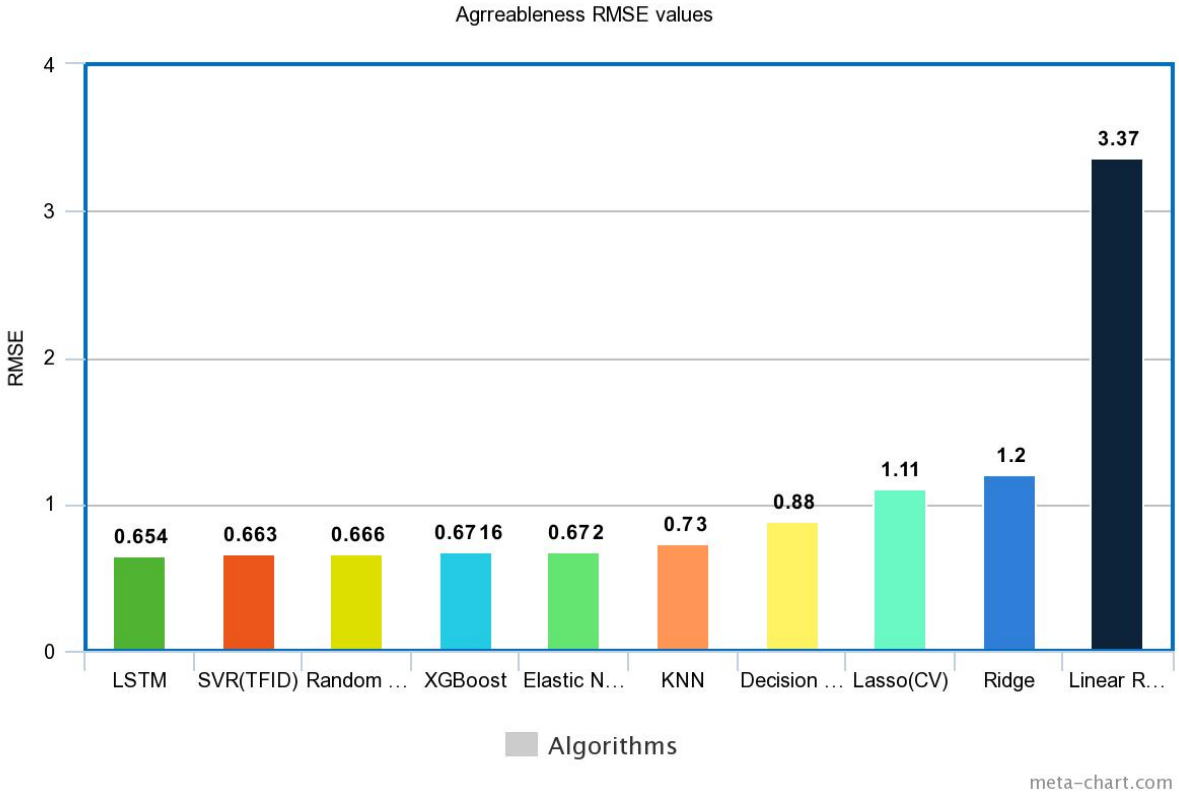


Figure 9.8: Agreeableness

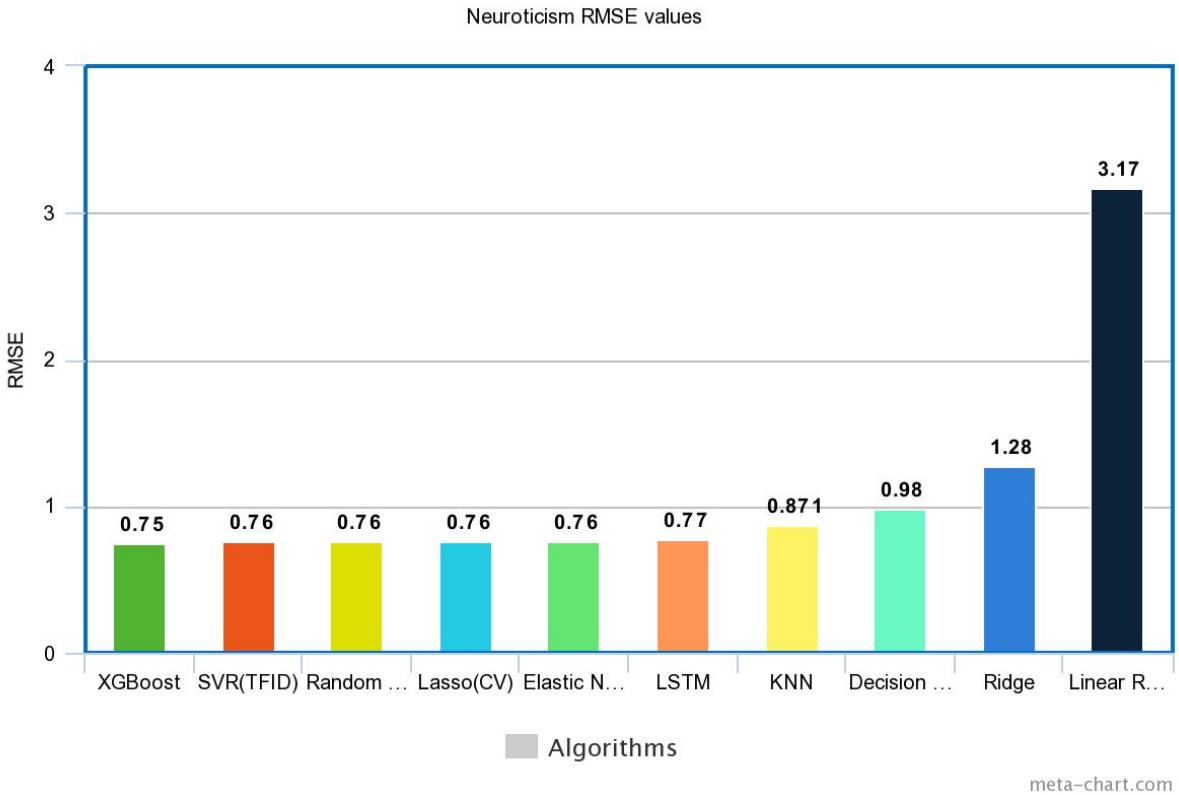


Figure 9.9: Neuroticism

9.2 Screenshots

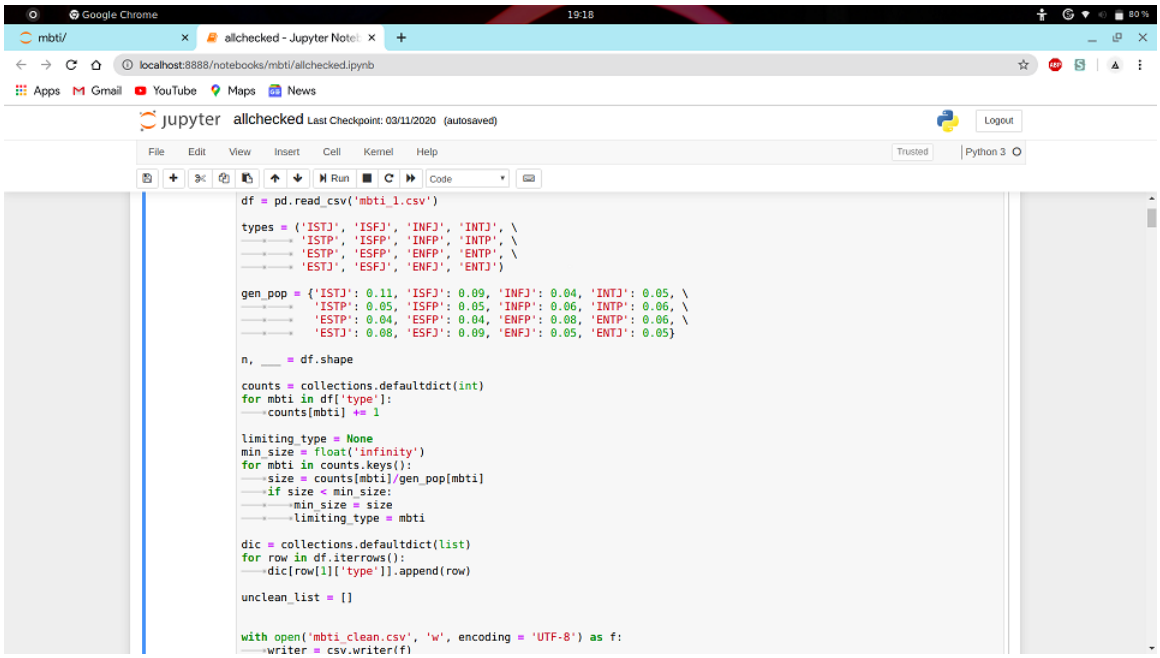


Figure 9.10: Code Snippet 1

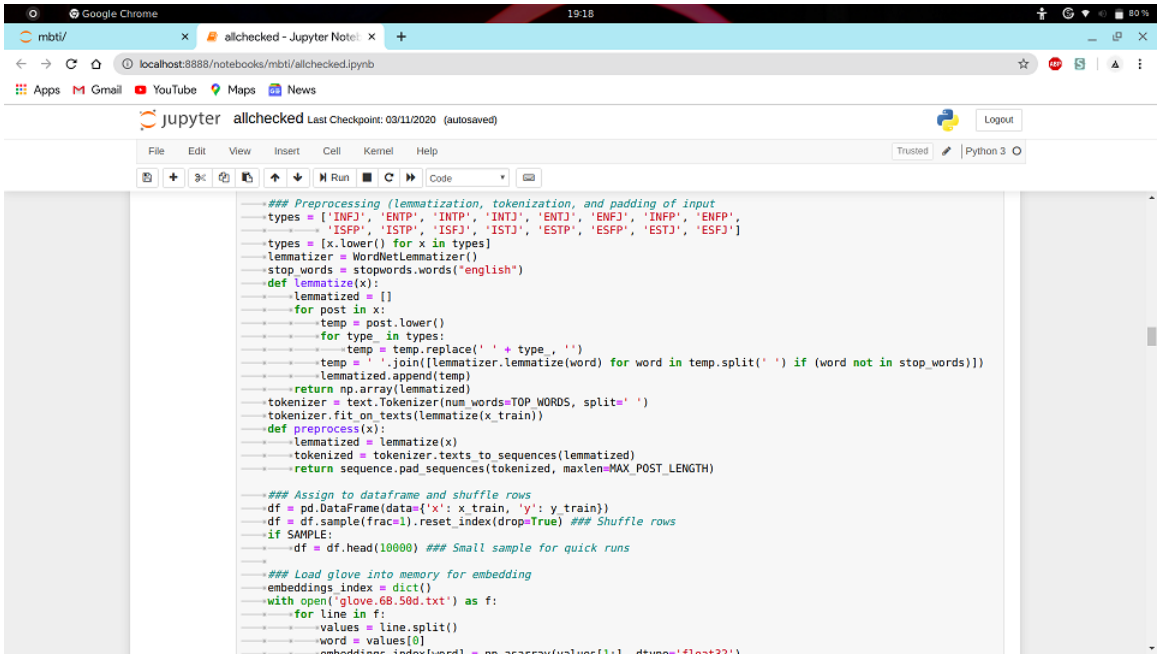


Figure 9.11: Code Snippet 2

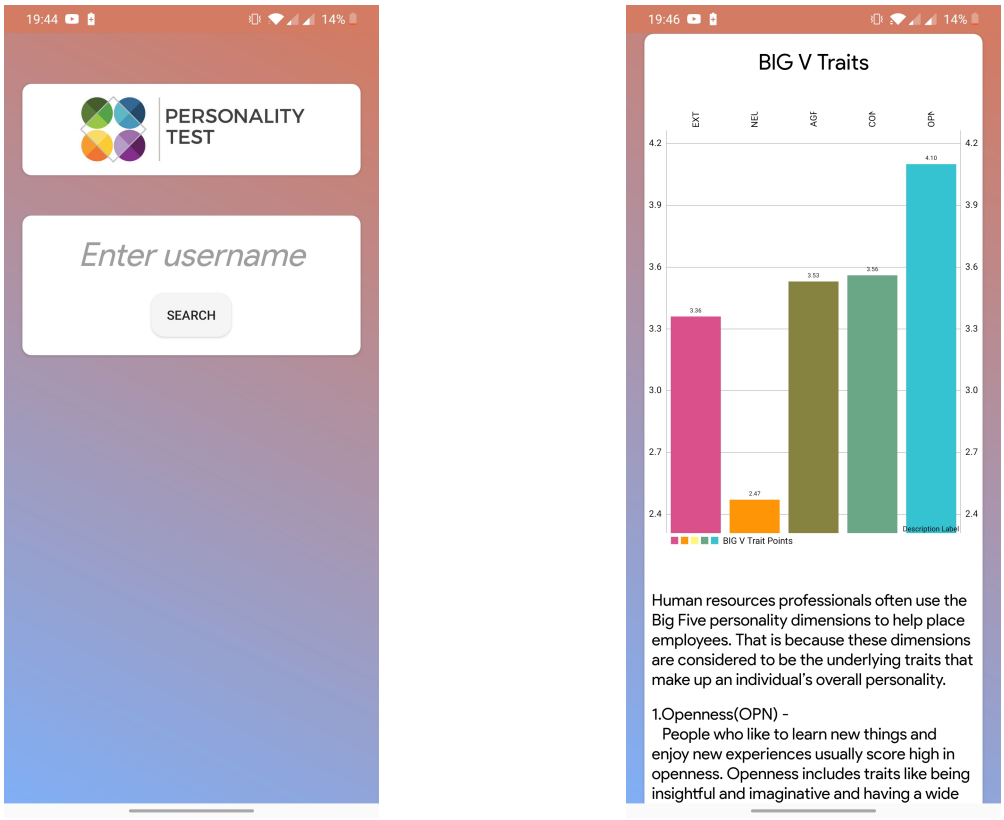


Figure 9.12: Application ScreenShot 1

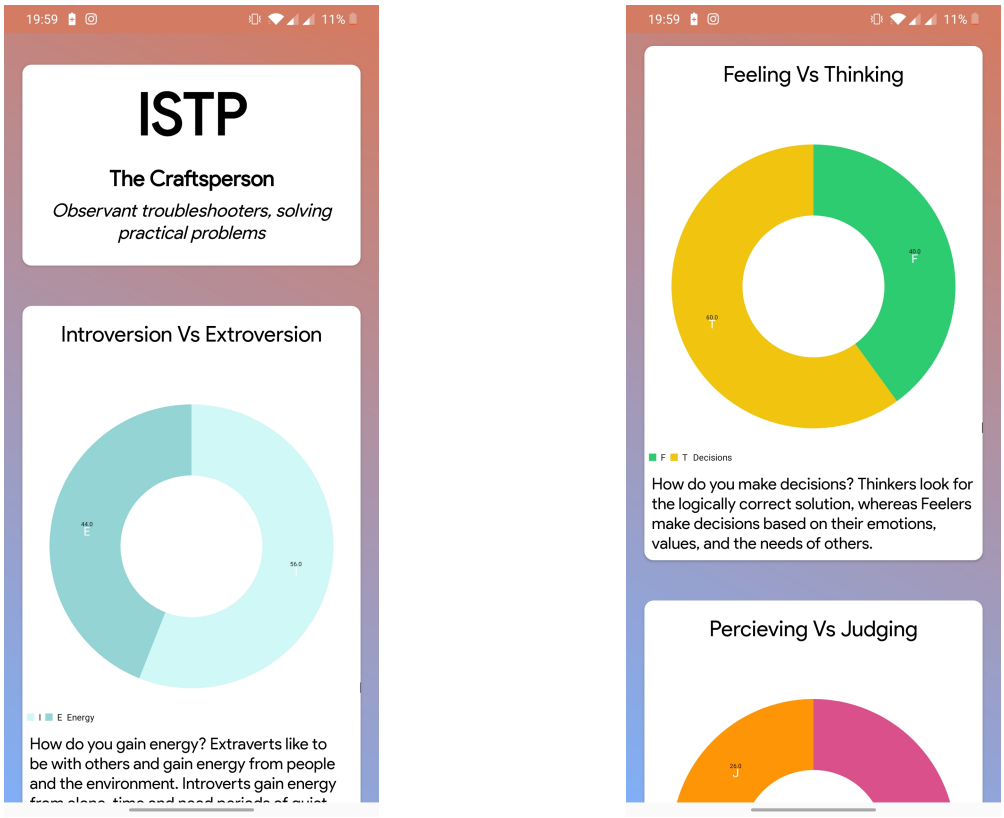


Figure 9.13: Application ScreenShot 2

Chapter 10

Conclusion

10.1 Conclusion

Personality identification using social network analysis is a relatively new domain within machine learning research. Since its introduction, however, it has drawn increasing attention by the research community with applications in wide variety of domains. Traditionally the only way personalities were identified was through questionnaire based personality tests that the subjects used to undergo. The surveyed techniques for automatic identification from online social networking profiles have yielded promising outcomes. Yet, many challenges and opportunities exist. Surveying this topic, we listed some challenges and insights that constitute promising research directions.

10.2 Future Work

10.3 Applications

Scope of this Project is in wide range of different buisnesses and different industries as it can widly provide public consensus such as:-

1. Recruitment –
Employers can use PP techniques to gain a deep understand-ing of the applicants. This helps them to find the qualified personnel they really need.

2.Counseling –
personality can be used as an important assessment in career, relationship and health counseling.

3.Online marketing –
a user's predicted personality can be used by online mar-keter's to personalize their message and presentation to suit individual preferences.

4.Corporate –
for targeted advertising and marketing, employee recruitment, ca-reer and health counseling.

5.Psychological Profiling –
of user's is a useful tool for job satisfaction, career progression, selling preferences in different interfaces etc.

6.E-commerce/ E-learning –
can benefit by a user interface that adapts the interaction according to the user's personality.

7.Recommendation Systems –
performance of such systems can be enhanced to attract more user's.

8.Determining Antisocial Behavior –
personality traits have been found to have a close correlation with antisocial behavior. This has been revealed by the stud-ies undertaken on personality and crime.

Chapter 11

Bibilography

[1] Vishal Kaushal and Manasi Patwardhan "Emerging Trends in Personality Identification Using Online Social Networks," ACM Transactions on Knowledge Discovery from Data, Vol. 12, No. 2, Article 15. doi: Available: http://dx.doi.org/10.1145/3070645

[2] S. Jiang, W. Min, L. Liu and Z. Luo, "Multi-Scale Multi-View Deep Feature Aggregation for Food Recognition," in IEEE Transactions on Image Processing, vol. 29, pp. 265-276, 2020. http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8779586&isnumber=8835130

[3] Miheal M. Tadesse,Bo Xu and Liang Xang , "Personality Predictions Based on User Behaviour on the Facebook Social Media Platform" IEEE Conference 2018. http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8710854&isnumber=8710814

[4] Di Xue,Zheng Hong and Liang Gao "Personality Recognition on Social Media With Label Distribution Learning," 2017 IEEE Conference Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8601243&isnumber=8601084