INTERIM PROJECT REPORT ON

Personality Prediction through Social Media Posts

SUBMITTED BY

Jay Naik (403046)
Ajinkya Pingale (403061)
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

Jay Naik (403046)

Ajinkya Pingale (403061)

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
(Head of Computer Engineering 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.

Jay Naik
Ajinkya Pingale
Manoj Nandha
Akash Misal

Contents

List of Figures

Abstract

Our focus for this project is using machine learning to build a classifier capable 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

Problem Statement

One of the Major Platform today to express your emotions , expression , feelings , views about something or about any product is Social Media .People can put forward their opinion on social media , since active social media user base around world is more than that of billions , it can be used to judge what people think about an issue or product . People put forth what they feel , therefor their indivisual post can be judged as a reflection of their personality , but how to judge a personality of a person based on social media post and that too of say large data say thousands , and produce analysis on it . Therefor the problem statement is to predict personality of a person using his socail media posts .

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

Problem Definition

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.

3.1 Data Extraction

This phase involves identifying and collecting data sets deemed suitable for the application being developed. Twitter has been the major source for social media data. Data sets for training and testing are collected from Twitter using suitable API's. Other sources of data include the publicly available datasets viz, myPersonality and essays.

This is a very important phase in the pipeline as it decides the efficiency of the other steps down in line. Preprocessing involves the standard steps of Case Conversion, Stop-words Removal, Punctuation Removal, Stemming, Lemmatization, POS Tagging. After the relevant tweets are fetched and pre-processed, features relevant for prediction are extracted and selected. User's profile data is the most common feature set used by most researchers. Among other features used are - number of followers, number of following, twitter posts, linguistic features etc. The features extracted are dependent on the prediction to be made.

3.2 Classifications

This phase includes modules for model training, testing and prediction. After the selected model has been trained and tested using different data sets , an unseen (new) dataset is presented to the trained model for prediction, in order to classify and predict the personality trait of new user(s). Several standard algorithms have been used by authors like Support Vector Machine, K Nearest Neighbour, Multinomial Naive Bayes, Naive Bayes, multi task regression, incremental regression algorithm etc.

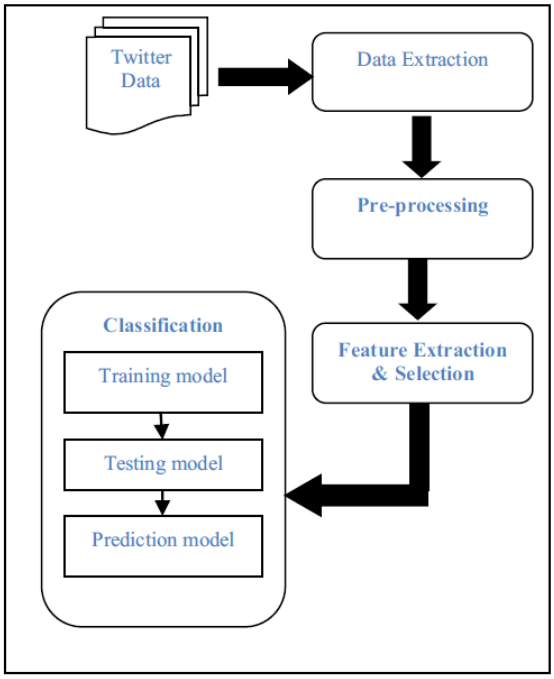

Figure 3.1: Basic Model Workflow

Chapter 4

Concepts Required for the Project

PREPROCESSING : The dataset obtained from myPersonality was pre-processed before it proceeded to the feature selection and training stage. To pre-process the dataset, we employed OpenNLP . First, we used tokenization in order to separate the last word of each sentence with punctuation and an aggregation of the same words. Next, we removed URLs, symbols, names, spaces and lower cases. Since many of the words in LIWC and SPLICE linguistic features share common stems, the relationship between personality and stemmed words could be negatively affected. For instance, in the case of tenses, such as present or past tense, verbs stemming would make it impossible to distinguish between particular tenses . Hence, the correlation analysis in the pre-processing part of our experiment does not apply stemming, and all the words are left unstemmed.

FEATURE EXTRACTION : A user's behaviour on social networks is mutually affected by the presence and behaviour of other users. These interactions can have an impact on the transition of new information or behaviours through the groups. There are many potential applications for understanding how such behaviours arise and spread . In our study, all the information from the dataset can be categorized into two groups. The first group is the text features extraction which reflects a user's language habits on Facebook and contains an expressions count and a topics count. To analyse the content of Facebook status texts, we use two dictionaries, namely, LIWC and SPLICE. The second group is the social interaction behaviour analysis, which contains networksize, density, brokerage and transitivity. This information reflects a user's basic social network behaviour on Facebook.

LIWC, or the Linguistic Inquiry and Word Count dictionary, is widely used in psychology studies [26]. In our study, we use it to extract 85 linguistic features from the texts including five subcategories such as standard counts (e.g., word count, words longer than six letters, number of prepositions), psychological processes (e.g., emotional, cognitive, sensory, social and emotional processes), relativity (e.g., words about

time or tense verbs), personal concerns (e.g., occupation words such as job, majors, financial issues or health), and other linguistic dimensions (e.g., counts of various types of punctuation, swear words) . For the text analysis, we chose LIWC2015 which is designed to analyse individual or multiple language files quickly and efficiently. In comparison to LIWC 2007 and LIWC 2001, it attempts to be transparent and flexible in its operation, allowing the user to explore word use in multiple ways

FEATURE SELECTION : Generally, there are two main reasons why feature selection is important for building a model. First, it reduces the high dimensionality of the dataset by removing the features not essential for training, improving the generalization of the model and reducing the training time. Second, the model gains a better understanding of the features and their relationships to the response features. Additionally, it improves the accuracy of the learning algorithms and reduces the processing requirements. To measure the strength of the linear relationship between two variables and to examine features important for personality traits prediction, we used the Pearson correlation analysis, Eq. (1), as the standard feature selection method. Pearson correlation is a measure of the linear correlations between two variables, and we used it to predict the relationship between the personality scores and extracted features. For a pair of variables (x; y), the linear correlation coefficient r is given by the formula: where x and y are sample means given by the relations n

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}},$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

in the above equation represents the sample size, and xi, and yi describe the single samples indexed with i, where the value of r lies between -1 and 1 inclusive. If x and y are completely correlated, r takes either the value of 1 as a positive correlation or -1 as a negative correlation. If x and y are completely independent, r is zero

Chapter 5

Scope of the project

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 understanding 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 marketer's to personalize their message and presentation to suit individual preferences.

4.**Corporate** – for targeted advertising and marketing, employee recruitment, career 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 studies undertaken on personality and crime.

Chapter 6

System Overview

The following steps are followed to detect the personality of person through social media posts:

1. Extract the tweets from the dataset

2. Remove Unwanted or Irrelevant Tweets that should be filtered using pre-proessing tehniques.

3. Extract key-word vectors so that it could be fed into psychological classification models.

4. Learn through key-word vetors or Label distribution so that classification model can be made

5. Build a Classification Model based on this so that we can compare output.

6. Various Model could be used in conjucture so that we can get better result and accuracy,for ex - Linear Regression , Support Vector Machine, Random Forest method could be used side by side

7. Calculate Accuracy , Confusion Matrix so that a better model could be selected

Chapter 7

Hardware and Software Requirements

7.1 Hardware

1. GPU VRAM 4GB minimum
2. CPU minimum 6th gen
3. RAM – 8GB

7.2 Software

1. Windows or Ubuntu
2. Python 3.6
3. Anaconda
4. Pip

Chapter 8

Feasibility Study

8.1 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.

8.1.1 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.

8.1.2 Software Feasibility

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

8.2 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.

8.3 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.

8.4 Operational Feasibility

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

Chapter 9

Design

9.1 UML Diagrams

9.1.1 Use-Case Diagram

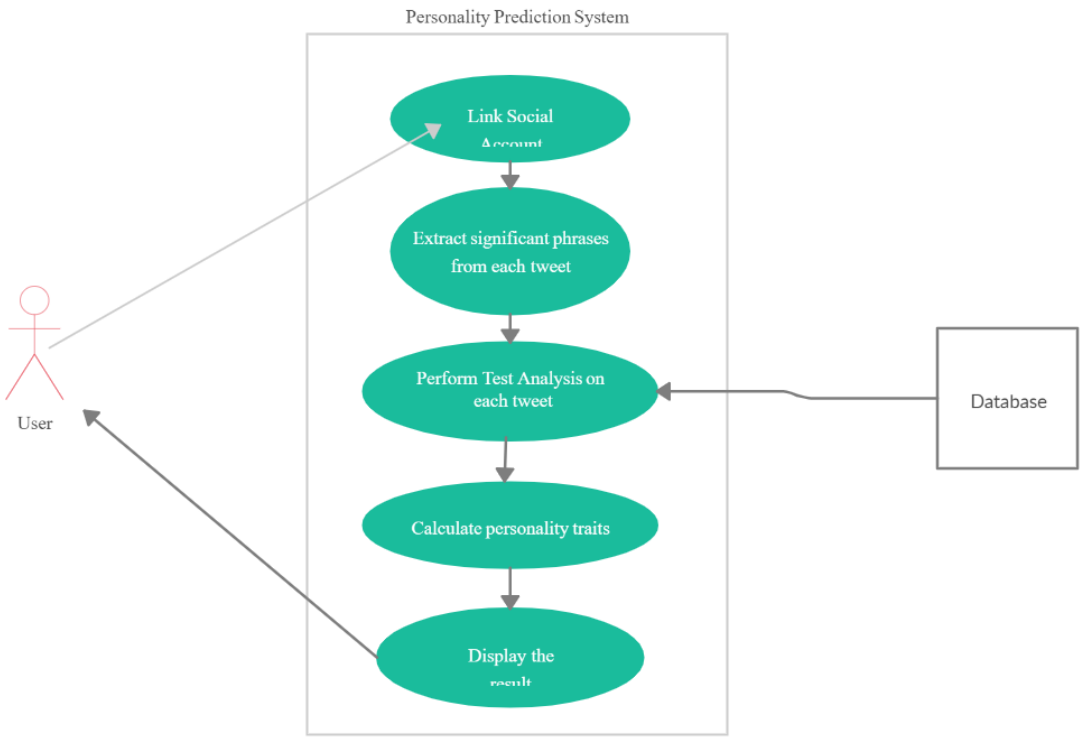

Figure 9.1: Use-Case diagram

9.1.2 Class Diagram

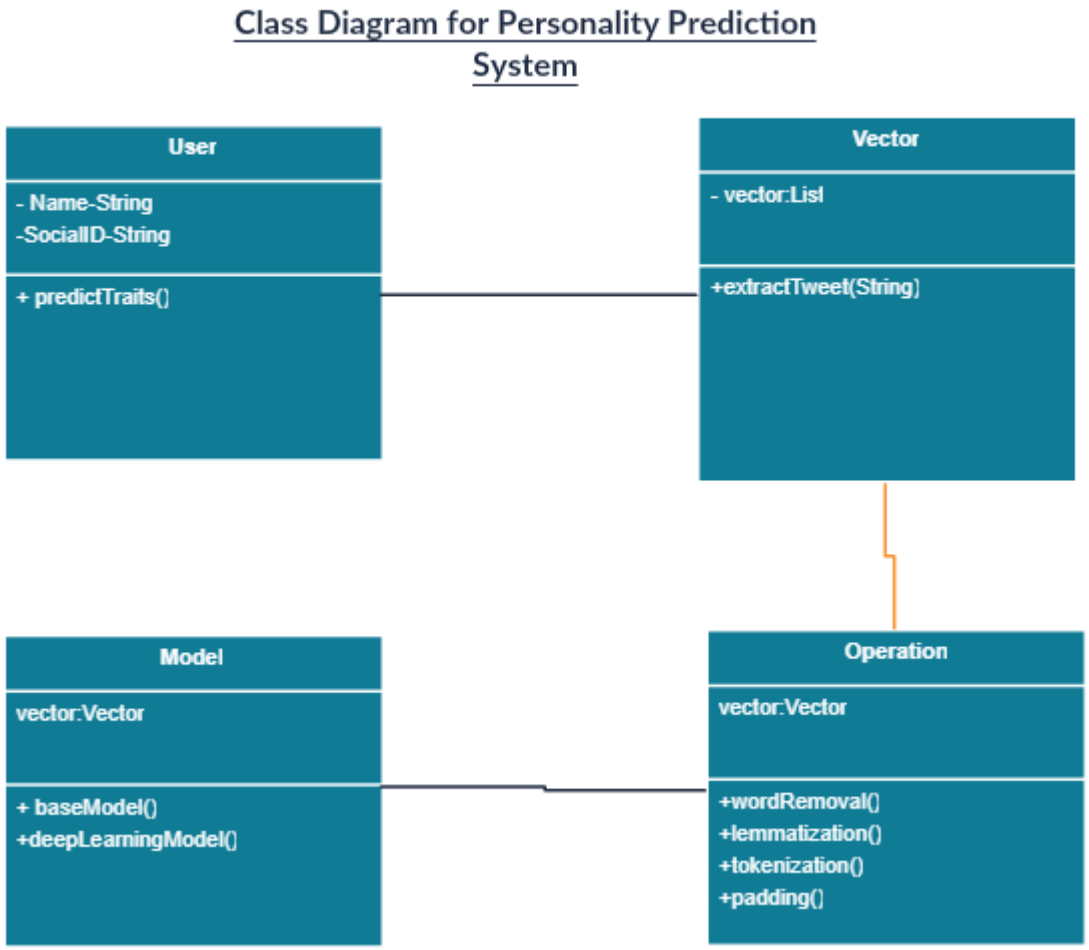

Figure 9.2: Class diagram

9.1.3 Activity Diagram

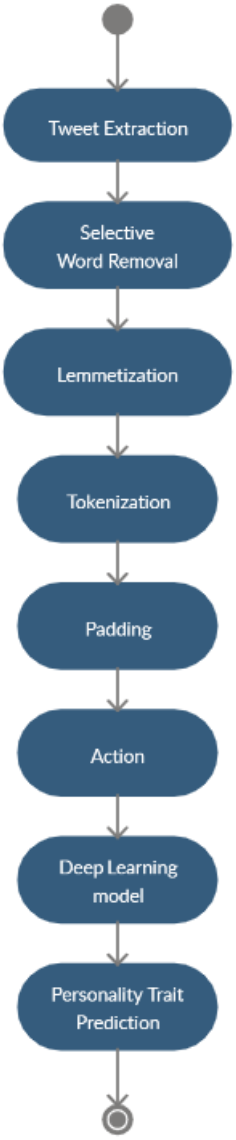

Figure 9.3: Activity diagram

9.1.4 Sequence Diagram

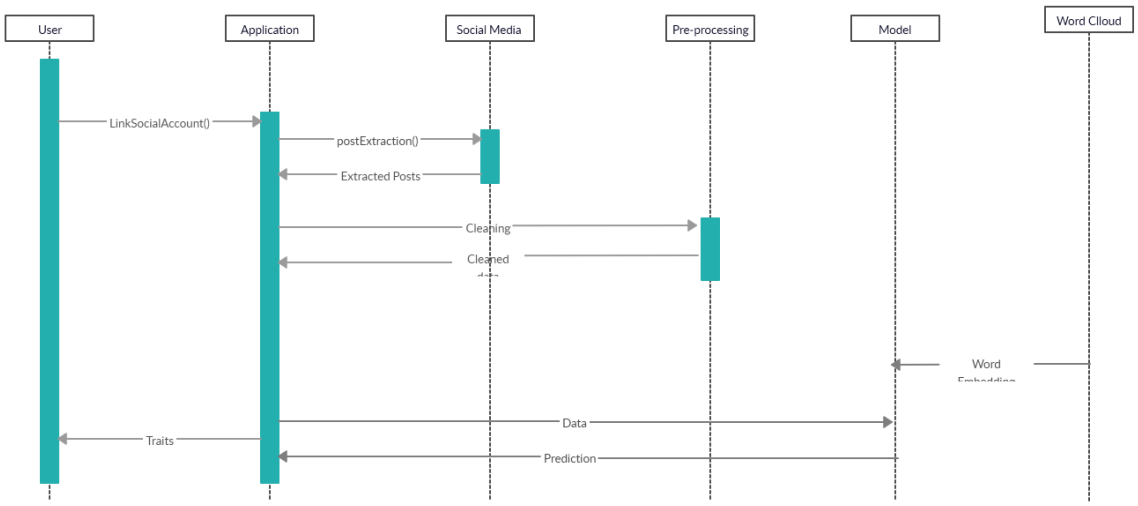

Figure 9.4: Sequence diagram

9.1.5 Architechtural Diagram

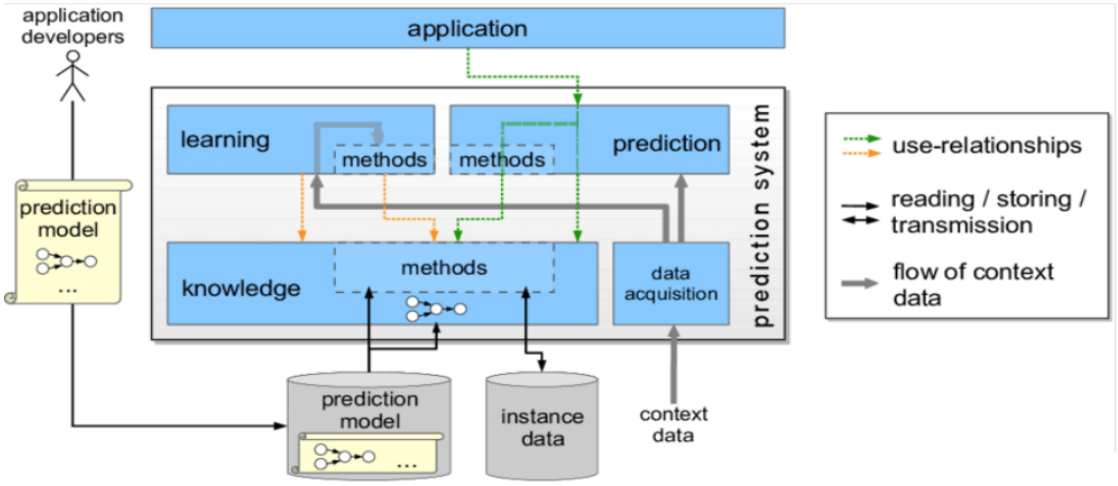

Figure 9.5: Architechtural diagram

9.2 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

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

M = vocabulary , J(θ) = Cost function

Chapter 10

Time line Analysis of the Project

Phase	Time	Deadline
Phase formation and initial domain discussion	1 month	August
Problem statement finalisation	1 month	September
Requirement analysis	1 month	October
Analysis/Design	2 months	December
Implementation	2 months	February
Deployment	1 month	March
Testing	1 month	April
Report writing	1 month	May

Chapter 11

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.

Bibliography

[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