PREDICTING USERS' PERSONALITY FROM SOCIAL MEDIA USING LINGUISTIC AND SOCIAL NETWORK FEATURES

This Dissertation is presented to
Islamic University of Technology(IUT)

By

Ahmed Al Marouf
Student ID - 161041001

In Partial Fulfillment of the Requirements for the Degree of
Master of Science in Computer Science and Engineering

Supervisor
Dr. Md. Kamrul Hasan
Professor, Department of CSE, IUT

Department of Computer Science and Engineering (CSE)
Islamic University of Technology (IUT)
Organization of Islamic Cooperation (OIC)
Gazipur-1704, Bangladesh
November, 2019

Recommendations of the board of examiners

The thesis titled "Predicting Users' Personality from Social Media using Linguistic and Social Network Features" Submitted by Ahmed Al Marouf, Student No. 161041001 of Academic Year 2016-2017, has been found satisfactory and accepted as partial fulfilment of the requirement for the Degree of Master of Science in Computer Science and Engineering (M.Sc. Engg. CSE) on November 14, 2019.

1. ………………………………………

Prof. Dr. Md. Kamrul Hasan — Chairman (Supervisor)
Department of Computer Science and Engineering
IUT, Board Bazar, Gazipur-1704, Dhaka, Bangladesh

2. ………………………………………

Prof. Dr. Muhammad Mahbub Alam — Member (Ex-Officio)
Head
Department of Computer Science and Engineering
IUT, Board Bazar, Gazipur-1704, Dhaka, Bangladesh

3. ………………………………………

Prof. Dr. Abu Raihan Mostofa Kamal — Member
Department of Computer Science and Engineering
IUT, Board Bazar, Gazipur-1704, Dhaka, Bangladesh

4. ………………………………………

Prof. Dr. M. Sohel Rahman — Member (External)
Department of Computer Science and Engineering
Bangladesh University of Engineering and Technology, Dhaka, Bangladesh

Declaration

I hereby declare that this dissertation entitled "**Predicting Users' Personality from Social Media using Linguistic and Social Network Features**" was carried out by me for the degree of Master of Science in Computer Science and Engineering, M. Sc. (Engg.) in CSE, under the guidance and supervision of Prof. Dr. Md. Kamrul Hasan, Islamic University of Technology, Gazipur.

The findings put forth in this work are based on my research and understanding of the original works and they are not published anywhere in the form of books, monographs or articles. The other books, articles and websites, which I have made use of are acknowledged at the respective place in this thesis.

For the present thesis, which I am submitting to the University, no degree or diploma or distinction has been conferred on me before, either in this or in any other University.

(Signature of the Candidate)
Ahmed Al Marouf
Student No: 161041001
Session: 2016-2017
November, 2019

Dedication:

"This dissertation is dedicated to my parents, teachers and my wife for all their continuous support, love and inspiration"

On the authority of Abu Hurayrah (may Allah be pleased with him), that the Messenger of Allah, Prophet Muhammad (peace be upon him) said:

"Whoever believes in Allah and the Last Day, let him speak good or remain silent"

Source:

Sahih Bukhari 5673, Sahih Muslim 48

This is a foundational principle for social media. It is hard because the entire purpose of social media is sharing and discussion. We often come to regret things we post, realizing too late that silence would have better served us. (Source: Fiqh of Social Media)

Acknowledgements

First I express my heartiest gratefulness to Almighty God for His divine blessing keeping me in sound mind and health during the work which makes it possible to complete this dissertation successfully.

I am really grateful and wish my profound indebtedness to Dr. Md. Kamrul Hasan, Professor, Department of Computer Science and Engineering (CSE), Islamic University of Technology (IUT). The door to Dr. Kamrul's office was always open whenever I ran into a trouble spot or had a question about my research or writing. He consistently allowed this work to be my own, but steered me in the right direction whenever he thought I needed it. I would also like to extend my cordial gratitude towards Mr. Hasan Mahmud, Assistant Professor of CSE department, for his continuous support in any questions. Thank you for answering my foolish questions. Their endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior draft and correcting them at all stage have made it possible to complete this thesis.

I would like to express my heartiest gratitude to Prof. Dr. Muhammad Mahbub Alam, Head**,** Department of CSE, IUT, for creating scope to work willingly in the department and also want to thank other faculty members and the staffs of CSE department of IUT.

Finally, I must acknowledge with due respect the constant support and patience of my parents and in laws. The trust and believe my father had on me, encouraged me more than anything in this world. Without his tremendous mental support, this thesis would never have finished.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

SNS	Social Networking Sites
BFFM	Big Five Factor Model
RQ	Research Question(s)
OCEAN	Openness-to-Experience, Conscientiousness, Extraversion, Agreeableness and Neuroticism
MBTI	Myers-Briggs Type Indicator
DISC	Dominance, Influence, Steadiness, Conscientiousness
IPIP	International Personality Item Pool
NEO-FFI	Neuroticism-Extraversion-Openness Five-Factor Inventory
BFI	Big Five Inventory
TIPI	Ten Item Personality Measure
MTurk	Amazon’s Mechanical Turk
LDA	Linear Dirichlet Allocation
TF-IDF	Term Frequency- Inverse Document Frequency
MLP	Multi-Layer Perceptron
LSTM	Long Short Term Memory
GRU	Gated Recurrent Unit
CNN-1D	1-Dimensional Convolutional Neural Network
POS	Parts-of-Speech
LIWC	Linguistic Inquiry and Word Count
MRC	Medical Research Council
SPLICE	Structured Programming for Linguistic Cue Extraction
SVR	Support vector Regression
SVM	Support Vector Machine
PCA	Principle Component Analysis
LDA	Linear Discriminant Analysis
CFS	Correlation-based Feature Selection
IG	Information Gain
SU	Symmetrical Uncertainty
CHI	Chi-squared test ($\chi 2$)
PCC	Pearson Correlation Coefficient
NLTK	Natural Language Tool Kit
NB	Naïve Bayes
DT	Decision Tree
RF	Random Forest
SLR	Simple Linear Regression

Abstract

Social media such as Facebook, Twitter, Google+ etc. has become a huge repository of textual data and images as each of the users' are creating posts, sharing views or news, capturing the moments via photos etc. User generated textual data such as statuses can be considered as the essential language to communicate in social media with others. Predicting personality traits from these social media data is a sophisticated task performed in computational social science. Among several personality prediction models, the Big Five Factor Model is one of the widely used personality traits hypothesis used by computational psychologists. The five traits that are centered for identifying ones personality are Openness-to-experience (O), Conscientiousness (C), Extraversion (E), Agreeableness (A), and Neuroticism (N). The first four traits are considered as positive traits and the only negative personality trait is neuroticism. In this thesis, we have focused on predicting these personality traits utilizing linguistic & social network features and identifying the prominent features using feature selection algorithms for each of the traits separately. We have evaluated the efficiency of machine learning techniques using the extracted features. To determine the most prominent features for individual personality traits and features that are commonly found in every personality traits, manual and automated feature selection has been applied. It is anticipated that the analysis reported in this study can be applied to develop personalized recommendation systems in social media, predicting personality disorder and identifying the trust issues in social media.

Keywords: Social Media, Computational Personality Prediction, Personality Traits, Psycholinguistic Features, Social Network Features, Automated Feature Selection Algorithms.

CHAPTER 1

INTRODUCTION

In this modern era of the internet, people are connected to the world through different social networking sites (SNS) like Facebook, Twitter, LinkedIn, WhatsApp, Google+ etc. Each user of the social networking sites is considered as an entity of the network. As social network provides a platform to share the personal views and news of its user, it has become a huge text repository of different activities of individuals. Each entity has a profile in the SNS and each profile also contains some demographic data such as name, age, gender, hometown, educational information, professional information, marital status, contact information etc.

As a part of society, people are engaged in offline and online socializing with the help of social media. Social media nowadays works as a proficient medium of interaction between its users and users are modifying the way of daily life activities because of social media. Apart from the demographic data each user creates an activity log starting from the day one of joining the SNS. 1.40 billion people on average log onto Facebook daily and are considered daily active users for December 2017 [1]. The quantitative rate of generating textual, image and video data in online social networking sites is rapidly increasing. The facilities provided by the SNS's are encouraging its users' to get accessed and connect with the peoples from different corners of the world. The concept of getting connected to the same kind of persons has evolved nowadays. People these days, don't want to meet and greet with fake accounts of SNS anymore. The idea of similar personality comes in the social media as having the same personality traits means that these persons can become friends, as they can mingle easily. Using the online behavior or personality traits, different recommendation systems such as community recommendation [3], friend recommendation [4], and community detection [5] could be approached. In the context of human computer interaction, social media are playing significant roles as people are interacting through social media every day.

"What is my personality? What are my personality traits? What does my personality says?" these are some common questions people ask to the psychologists or even themselves. People also try to estimate strangers' personality before starting a new personal or professional relation. Psychologists are emphasizing on understanding the personality of a human being from their demographic information such as age, gender, home location, occupation, his/her hobbies, problems he/she is facing in regular life, job life satisfactions, marriage life satisfaction etc. Though people try to understand their own personality, but they hesitate to answer such personal questions to the psychologists. Nowadays predicting personality is one of the challenging task that a psychiatrist have to face every day. It is evident that after third session of a patient with a psychologist only around fifty percentage of his/her personality can be predicted. And, psychologists have to guess the other fifty percentage of that patient before psychological helpings.

Usually psychologists take a personality test of the patients and try to understand his/her personality from the test. The test taker have to answer a pre-set questionnaire, which are set by the psychologists (generally known as International Personality Item Pool or IPIP) [25]. There is an underlying scoring system from each questions/items. To understand one's overall personality specific set of personality traits are need to know. Therefore, after years of research psychologists have come up with personality models, which are essential to predict overall personality of individuals.

One of the limitation of traditional personality prediction system is 'test-takers need to answer a lot of questions'. But the main limitation is that 'test-taker need to answer the questions honestly and have to answer all the questions'. Because of this time-consuming process, often test takers skip attending personality tests in different platforms. As the traditional process involves filling up self-assessment reports or online surveys, test-takers hesitate to answer the questions honestly. Therefore, predicting personality without asking direct questionnaires could be considered as a challenging task.

As social media has become an online repository of user-generated data, therefore, we try to utilize those profile data, especially natural language data (texts). Under the umbrella of computational social science, computational personality prediction has become a significant research domain. Though the idea of predicting personality is an ancient concept, utilizing and exploring computational approach for predicting personality is relatively newer in the area of computational psychology. From the state-of-the-art works, we found that researchers have focused on improving the accuracy of predicting personality traits, rather than finding a fixed set of features which can be utilized for personality traits. This research gap has been addressed in our proposed mechanism and feature selection (manual and automated) has been introduced.

In this thesis, we try to use users profile data especially the status updates and social network features to predict personality traits of an individual user. Majority of the data are text, therefore applying text analytics algorithms to find and use the linguistic features (traditional and psycholinguistic features) will be the primary features to reach the target. Then, utilizing the social network features to understand the effect of them is another latent target in this research. Investigating the most prominent features among these linguistic and social network features is one of the main focus and contribution of this thesis. To help the psychologist community to predict personality more accurately from the social media platforms, this research could be used as reference. We have tried to investigate and disclose some of the interesting observations related to the individual personality traits.

1.1 Problem Statement

Predicting users' personality from digital footprints of social media is a challenging task as the context of identifying personality traits in social media is not trivial. Users behave differently in social media and real life. Therefore, the user generated content such as status updates in social media may provide enough evidential reflection of personality as SNS user posts statuses based on his/her current situation, a recent political or popular event, hyped topics etc. For example, during an election of his/her country, he/she may posts positive or negative reviews/opinions about a political party. These type of status may have contextual trend, as other friends of the users may also be involved in posting similar statuses. Considering trend user may post his/her political views. Users are creating trend as well as following different trends to become popular or socially accepted by their friends in social media. Moreover, each user have different perceptions and different interest category to be triggered to update statuses.

Personality traits are those properties of a user that could be considered as biometrics. Each and every human in the world have different personality. Therefore, the effect of the personality also implies on the social networking sites. Using the status updates users usually shows their view and news of many things, which have high expectancy of personality involvements. Each user’s personality can be expressed using the personality models such as Big Five Factor Model (BFFM) [2], RIASEC Model [18] Myers-Briggs Type Indicators [19],etc. Among these models BFFM have been widely used by the psychologies as well as computational psychology researchers. BFFM have five factors to be measured and predict user’s personality. The five factors are: OCEAN: Openness to Experience (O), Conscientiousness (C), Extraversion (E), Agreeableness (A), and Neuroticism (N). [2]

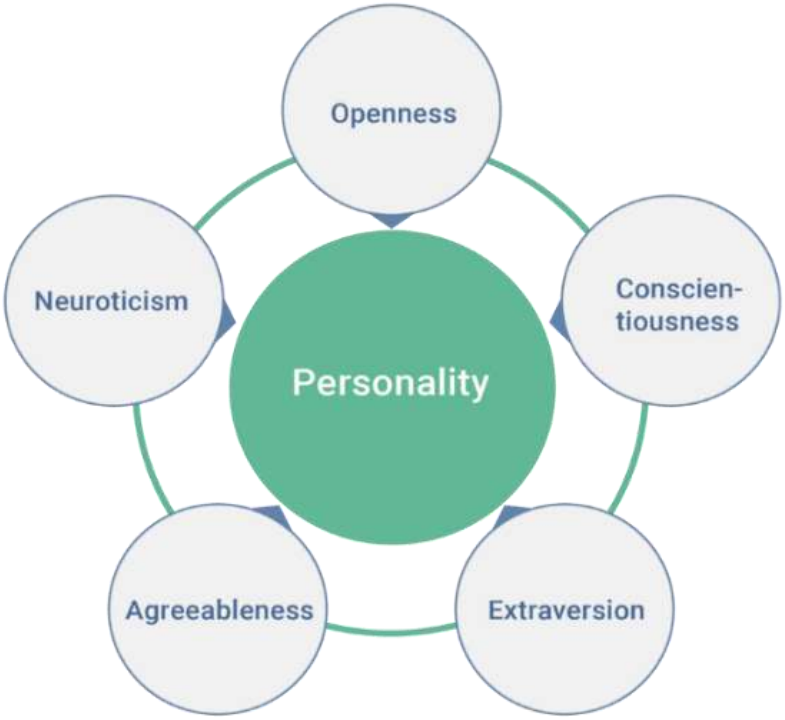


Fig 1.1: Big Five Factor Model [2].

The personality traits mentioned in BFFM, covers tentatively all the areas of human personality including the positive and negative psychology. This BFFM model is also known as OCEAN model. OCEAN is the abbreviation with the first letters of each trait names. Except for neuroticism, the other four personality traits are considered as positive traits and only neuroticism is considered as negative traits. This OCEAN model is widely used in computational social science and have been used by the psychologists to predict personality using computational tools. The following Table 1.1 illustrates the description of those five personality traits. The description tends to describe the relevant psychological action verbs which related to each traits.

Table 1.1: Brief Description of BFFM Personality Traits

Name of the Personality Trait	Brief Description
Openness-to-Experience (O)	Openness is a general appreciation for art, emotion, adventure, unusual ideas, imagination, curiosity, and variety of experience. People who are open to experience are intellectually curious, open to emotion, sensitive to beauty and willing to try new things. They tend to be, when compared to closed people, more creative and more aware of their feelings.
Conscientiousness (C)	Conscientiousness is a tendency to display self-discipline, act dutifully, and strive for achievement against measures or outside expectations. It is related to the way in which people control, regulate, and direct their impulses. High conscientiousness is often perceived as being stubborn and focused. Low conscientiousness is associated with flexibility and spontaneity, but can also appear as sloppiness and lack of reliability.
Extraversion (E)	Extraversion is characterized by breadth of activities (as opposed to depth), surgency from external activity/situations, and energy creation from external means.[39] The trait is marked by pronounced engagement with the external world. Extraverts enjoy interacting with people, and are often perceived as full of energy. They tend to be enthusiastic, action-oriented individuals.
Agreeableness (A)	The agreeableness trait reflects individual differences in general concern for social harmony. Agreeable individuals' value getting along with others. They are generally considerate, kind, generous, trusting and trustworthy, optimistic, helpful, and willing to compromise their interests with others.
Neuroticism (N)	Neuroticism is the tendency to experience negative emotions, such as anger, anxiety, or depression. It is sometimes called emotional instability, or is reversed and referred to as emotional stability.

We try to develop a predictive system by feature-fusion of linguistic and social network features for predicting SNS user's personality from status updates. The traditional way to predict personality, someone need to answer a set of questionnaire, usually 10 to 100 item questions. To solve the scenario, we have extracted linguistic and social network features from the SNS users and proposed a predictive approach to predict his/her personality.

To investigate the impact of using the psycholinguistic features for each of the personality traits, we have designed and implemented an experiment to using the linguistic features (traditional and psycholinguistic features, separately). Different scenarios have been designed by segmenting the feature sets such as only emotional features, only cognitive features etc. The efficiency of the prediction is taken into consideration through the performance metrics such as precision, recall, f-score and accuracy. Depending on the accuracy measurement given by the model, we can say if the model is providing better solution or not.

One of the contribution of this thesis is to find the effects of linguistic (e.g. no. of words, no. of sentences etc.), psycholinguistics (e.g. emotional affect, cognitive words, social relationships etc.) and social network features (e.g. network size, betweenness centrality, density, brokerage, transitivity etc.) to predict personality of a SNS user more precisely. As social media is built upon the mega structure of node to node connected graph, the structural impacts of this social network features are not addressed by the computational psychology researchers'. Therefore, the usability of social network features in a computation model was not present in the works of psychologists. The linguistic features are considered to be profoundly rich data source as status updates of social media users reflects hi/her personality significantly. Again, the usage of psychological words reflects users' personality in a broader sense.

To investigate the impacts, we have designed and implemented an experiment to predict users' personality using all the features we have extracted for the model. Then, we have applied several feature selection algorithms to determine which features are closely related to the class. The impact and effectiveness of social network features could be depicted by this experiment. In this experiment, we have analyzed the features to determine the most prominent features for individual personality traits and features that are commonly found in every personality traits.

Problem Statement:

"To develop a personality predictive system through positive & negative traits of SNS user, exploring the prominent linguistic and social network features"

The problem statement could be divided into following steps:

- To predict personality of social media user using the linguistic (traditional and psycholinguistic features) and his/her social network features.
- To identify the most prominent features for individual personality traits through exploiting manual and automated feature selection algorithms.

The primary objective is to utilize the linguistic and social network features to predict the personality traits of a SNS user. The online behavior, thus the user-generated contents are useful for reaching this specific goal.

The secondary objective involved a development of predictive model based on the machine (supervised) learning, so that we can predict personality computationally. In this computational approach, we have integrated the feature selection process, which is a significantly popular research domain in computer science.

1.2 Thesis Organization

The rest of the thesis is organized as follow. Chapter 2 presents the background study of social media mining for personality prediction, computational personality prediction, feature selection algorithms overview etc. Chapter 3 presents our proposed methodology elaborately with all the steps such as data pre-processing, feature extraction, feature selection and classification model. Chapter 4 provides the experimental findings and analysis. Finally, chapter 5 concludes with the future scopes and opportunities.

CHAPTER 2

BACKGROUND STUDY

In this chapter, we have presented the related works and reviews regarding our thesis. This chapter the computational approaches used in state-of-the-art research are described, including the psychological personality models, personality questionnaires, ground-truth datasets, feature selection algorithms and machine learning models applied for the prediction task. Finally, the background limitations and challenges is presented at the end of this chapter.

2.1 Social Media Mining for Personality Prediction

At first, we try to investigate if social media is the right place to take data from and infer the personality traits from them. Analysis of social media data is referred to social media mining and our research goal is to mine social media for personality prediction. Therefore, we get similar and convincing works are performed over social media.

Social media analysis is one of the largest area of human computer interaction (HCI) and huge number of researchers' are contributing to this area recently. For recommending products, promotional features etc. to a specific user in social networking sites it is important to understand his/her preferences. Therefore, predicting social media user's personality could be a good approach to reach his/her interest. Without any doubt, it is possible to track users' digital footprints from social network data generated by user himself [6]. As social networking sites are the place where people use to share their status, news, views with the help of structured or unstructured languages, textual data could become an effective resource to find personality traits [7].

Nowadays social interaction has reached to a new dimension because of emergence of online social networking sites and easy access of these sites. People positively accepted the feature of sharing his/her own thoughts through statuses or tweets. Therefore, millions of textual data are produced by all the users' on daily basis. Facebook provides many features or activity support to its users such as, writing status, sharing others posts, giving reactions (like, love, angry, sad, happy, wow) to others' posts etc. which could be separated into two type namely, user generated or user supported. User generated contents contain inner or hidden information about the user such as personal choices, opinions, his/her behavior towards any issues, especially personality representations. Therefore, the personality traits could be predicted from the user generated data.

There are several methods proposed by researchers' in [8-10] utilizing different parameter based method to predict personality.

It is evident that user generated contents could be an effective data source to build a predictive model [13]. The status updates posted by the SNS users have influence of culture and personal issues. The structures of various languages actually influence on identity, culture and diversity of persons [14]. Therefore, Facebook status has become a research tool to the researchers for identifying the personality [15]. Data collection is the first step to create any predictive model. After collecting the data, pre-processing should be applied to get a clean data. Feature extraction and feature selection is applied afterwards to identify the most relevant features. Those features are trained to a classification model and testing is performed afterwards. Hence, from this literature we have been provided a ground to work on this problem further and apply different mechanisms to achieve a better performance from the computational model. Hence, we have identified that there are some limitations and challenges present in the state-of-the-art works. The limitations and challenges of literature are addressed below.

- *Offline vs. Online Personality*: The personality that we wanted to predict is based on the online behavior of the SNS user, not the real life behavior. As it is evident, the online and offline behavior of a particular user is different, which could be tracked by digital footprints [4], we have worked on the online behaviors only.
- *Lack of ground-truth dataset*: Because of security and privacy concerns, social media users usually don't agree to share their profile data such as photos, status or chat histories. Therefore, it is not possible to extract those data without their concerns. The Facebook Graph API can extract only the publicly available data of a user. Similar problem is present for Twitter, as web crawlers cannot get all the profile data unless given permission from the user. With limited number of data, it is tough to build a machine learning model.

- *Finding an appropriate psycholinguistic dataset*: Psycholinguistic dataset are closed vocabulary word dataset which are categorized to find the exact category of the language. Hence, finding was also a research challenge. There is no publicly available psycholinguistic database, but some paid databases are present in the state-of-the-art. This challenge is also addressed in the literature review section elaborately.
- Many researchers have tried to find the appropriate classification algorithm to solve the problem. But, which features are experimentally good enough to predict with better accuracy is not focuses. In our work, we have kept focus on finding those features which have high impact on each of the personality traits.

In the next section, we have elaborately described the computational personality prediction problem and literature reviews of each steps of this prediction task.

2.2 Computational Personality Prediction

Computational personality prediction problem in the context of social media can be defined as 'predicting the personality traits from user profile information using computational features rather than asking a set of questionnaire'. Usually for understanding own personality people try to take online or offline personality test. The traditional personality prediction systems depends on a set of questionnaire to be answered honestly by the test taker. For predicting personality traits computationally researchers' have utilized the machine learning techniques such as supervised/ unsupervised learning models, classification algorithms to classify the traits.

Personality prediction is more likely to be performed manually asking a pre-set questions. Modifying the manual task into a computational methodology is quite a challenging task. In this sub-section of the chapter, we try to cover the popular personality prediction models, international personality questions set, ground-truth datasets and computational methods for predicting personality from social media. The following fig. 2.1 shows the state-of-the-art system for computational personality prediction. In the later sections the detail description of each step is presented.

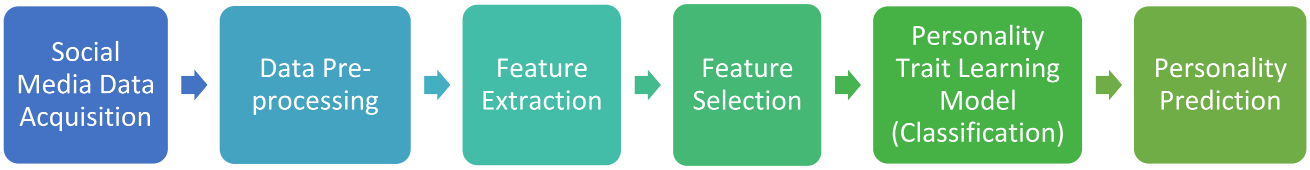

Social Media Data Acquisition
Data Pre-processing
Feature Extraction
Feature Selection
Personality Trait Learning Model (Classification)
Personality Prediction

Fig. 2.1. Computational Personality Prediction System.

2.2.1 Social Media Data Acquisition

Social media data are becoming richer and richer day by day, as the amount of user-generated data is increasing in a huge margin. For analyzing these user-generated data such as images, texts etc. computationally becoming challenging task because of the huge volume. Therefore, collecting unstructured data from social media itself is a complex task. Some of the constraints in this work are set by the social media authorities, as they share only public posts with the researchers. Using only public posts for predicting personality may lack sufficient data to build a better model.

Researchers' have collected data manually or developed Facebook API to collect the Facebook status updates (both public and private posts). To collect this data, the participant users need to give consent to the researchers and allow the procedures through the Facebook application. Researchers have performed personality test on the same users to generate personality scores for individual personality traits. Then, the labeling is performed from the scores in binary class. Therefore, the datasets have the personality traits of model in binary (yes/no) class.

For formalizing such personality prediction dataset, we need to follow a specific personality prediction model. Personality prediction model provides the individual traits defined with specific pre-set questions (also known as IPIP). IPIP stands for International Personality Item Pool. These pools are questions/items which are used to identify if a person have a specific personality trait or not. In the following sub-sections, detail about some of the existing personality prediction models, IPIP and ground-truth datasets are described.

2.2.1.1 Personality Prediction Models

The leading researchers in psychology had defined personality in their own way. “Personality is defined as the characteristic set of behaviors, cognitions, and emotional patterns that evolve from biological and environmental factors” [11]. Understanding the personality of individuals is quite necessary for developing several personalized application such as recommendation systems. “Personality is the dynamic organization within the individual of those psychophysical systems that determine his characteristics behavior and thought” and Weinberg and Gould defined personality as “the characteristics or blend of characteristics that make a person unique” [12]. From both the definition one thing is common, which is uniqueness of the individual and consequently adopt an idiographic view. For predicting or recognizing human personality different theories have been adopted by the psychologists such as type theory, trait theory, psychodynamic theory, behavioral theory and humanist [12].

In the literature, there are various types of personality prediction models which are based on pre-set questionnaires. The process of predicting personality from question set could be described as the subject or person has to answer around 50 to 100 questions about him\her. From the answer sets a mathematical model is developed to find specific score metrics. From that score a certain type of personality is being predicted for that person. As personality of each person could be different from others, but for generalization psychologists have come up with personality prediction models such as Big Five Factor Model (FFM) [22], Myers-Briggs Type Indicator (MBTI) [19], Type A and Type B personality theory by Mayer Friedman [16], RIASEC vocational model by John L. Holland [18], DISC (Dominance Influence Steadiness Conscientiousness) [26] etc.

Type A and Type B personality hypothesis, presented by Mayer Friedman, describes two contrasting personality types. In this hypothesis, personalities that are more competitive, highly organized, ambitious, impatient, highly aware of time management and/or aggressive are labeled Type A, while more relaxed, less 'neurotic', 'frantic', 'explainable', personalities are labeled Type B. The two cardiologists who developed this theory came to believe that Type A personalities had a greater chance of developing coronary heart disease [16]. Following the results of further studies and considerable controversy about the role of the tobacco industry funding of early research in this area, some reject, either partially or completely, the link between Type A personality and coronary disease. Nevertheless, this research had a significant effect on the development of the health psychology field, in which psychologists look at how an individual's mental state affects physical health [17].

John L. Holland's RIASEC vocational model, commonly referred to as the Holland Codes, stipulates that six personality types lead people to choose their career paths. Holland originally labeled his six types as "motoric, intellectual, esthetic, supportive, persuasive, and conforming." He later developed and changed them to: Realistic (Doers), Investigative (Thinkers), Artistic (Creators), Social (Helpers), Enterprising (Persuaders), and Conventional (Organizers)." In this circumplex model, the six types are represented as a hexagon, with adjacent types more closely related than those more distant. The model is widely used in vocational counseling [18].

Apart from these two models, the most popular personality prediction models are MBTI and FFM. MBTI gives overall 16 types of personality combinations. It is an introspective self-report questionnaire with the purpose of indicating differing psychological preferences in how people perceive the world around them and make decisions [19-21]. Though the test superficially resembles some psychological theories it is commonly classified as pseudoscience, especially as pertains to its supposed predictive abilities. The 16 types are typically referred to by an abbreviation of four letters—the initial letters of each of their four type preferences (except in the case of intuition, which uses the abbreviation "N" to distinguish it from introversion). For instance: ESTJ: extraversion (E), sensing (S), thinking (T), judgment (J) and INFP: introversion (I), intuition (N), feeling (F), perception (P). These abbreviations are applied to all 16 types.

The Big Five personality traits, also known as the OCEAN model, is a taxonomy for personality traits [22]. It is based on common language descriptors. When factor analysis (a statistical technique) is applied to personality survey data, some words used to describe aspects of personality are often applied to the same person. For example, someone described as conscientious is more likely to be described as “always prepared” rather than “messy”.

Table 2.1 shows the FFM personality traits and the related adjectives for the people having high and low scores in these personality traits [23-24]. Among these test Big Five personality test has been widely accepted among the test-takers. This model has proven to provide very close predictions to its users. Because of the similarity found with themselves with the result of the test, test-takers tend to use the online tools which internally uses this model.

Table 2.1: OCEAN Model and Related Adjectives

Personality Trait	People with high score	People with low score
Openness-to-Experience (O)	Imaginative, Creative Curious, Sensitive	Down-to-earth, Conventional, Uncurious
Conscientiousness (C)	Careful, Dependable, Self-Disciplined	Negligent, Lazy, Disorganized, Late
Extraversion (E)	Outgoing, Talkative, Sociable, Assertive	Loner, Quite, Passive, Reserved
Agreeableness (A)	Courteous, Good-natures, Empathic, Caring	Suspicious, Critical, Ruthless
Neuroticism (N)	Anxious, Hostile, Depressed	Calm, Even-tempered, Comfortable, Unemotional

2.2.1.2 International Personality Item Pool (IPIP)

International Personality Item Pool (IPIP) [25] are the items or questions to be answer to devise a scoring mechanism for traits identification. Depending on the behavior of test taker on different issues of practical life, these items are presented.

Using the IPIP questionnaire the quantitative method has been adopted for the problem and many variations of the question sets were used for developing a better ground-truth dataset. This manual procedure of taking answers of a set of questionnaire could be easily adopted. But, the main limitation of this process is the test takers need to answer the questions honestly.

Many online personality testing sites like 16Personality[1], 123test[2], Personality Perfect[3], PsychCentral Personality Test[4], Open Source Psychometrics Project[5], See My Personality[6], Discover My Profile[7] by University of Cambridge etc. are very popular for identifying precise personality reviewed by the test-takers. The reviews are analyzed from each of the websites and

1. https://www.16personalities.com/
2. https://www.123test.com/
3. https://www.personalityperfect.com/
4. https://psychcentral.com/personality-test/
5. https://openpsychometrics.org/
6. http://www.seemypersonality.com/
7. https://discovermyprofile.com/

found positive comments delivered by the reviewers. Literature provides evidential proof that computational personality prediction provides better results than manual paper-based methods. Therefore, the acceptability of these online personality tools is much higher than manual questionnaire based personality testing. Hence, this encourages to apply automated personality prediction from social media. It is evident that computational personality judgments are more accurate than those made by humans [27].

The history of personality prediction goes a long way as researchers have tried to optimize the number of questions being asked to the test taker. Usually high volume of questions are asked and the answers are analyzed to predict personality precisely. But answering these questions could be time consuming as well as tiring for the test takers. Therefore, asking a minimum number of questions to get a better prediction could be a challenging task. Researchers' have come up with various number of questions or items. NEO Five Factor-Inventory (NEO-FFI) [28] is a 60-item personality measure model. A similar models were proposed by researchers in psychology area for the personality prediction task.

Depending on scores determined by the International Personality Item Pool (IPIP), the computation of personality traits are performed. Depending on the number of IPIP items considered for prediction, there are several models proposed by many researchers. The 50-item IPIP Five Factor Model (FFM) [29] proposed by Goldberg, 44-item Big Five Inventory (BFI) [30] proposed by John, 40-item Big Five Mini-Markers [31] proposed by Saucier, 20-item Mini-IPIP [32] proposed by Donnellan, 10-item Personality Inventory (TIPI) [33] proposed by Gosling are existing models in the literature. Short form of item sets are also proven effective in some cases [34]. Although there are many scoring systems adopted for this particular problem, each of them have own advantages to be used. The myPersonality dataset [35-36] collected from Facebook users and used 100-item IPIP questionnaire set. For our experimentations, we have used the widely myPersonality dataset and particularly the status updates of 250 users. The dataset contains around 10000 status updates which could be utilized for our problem.

The traditional way to determine the personality scores for predicting personality traits is discussed on Appendix A.

2.2.1.3 Ground-truth Datasets for Personality Prediction

In this sub-section of the chapter, we have discussed on the existing ground-truth datasets formalized from social media data for predicting personality. Social media websites provide a unique opportunity for personalized services to capture various aspects of user behavior. Besides users' structured information contained in their profiles, e.g., demographics, users produce large amounts of data about themselves in a variety of ways including textual (e.g., status updates, blog posts, comments) or audiovisual content (e.g., uploaded photos and videos). Many latent variables such as personalities, emotions and moods — which, typically, are not explicitly given by users, but can be extracted from user generated content [37-39].

myPersonality [40] was a popular Facebook application introduced in 2007 allowing its users to take a number of psychometric tests, including a standard Five Factor Model questionnaire [41]. Users received feedback on their scores and could option to donate their scores and Facebook profile data to research. Data for over 6 million myPersonality users is available to researchers at: http://mypersonality.org/. It contains scores on more than 20 psychological tests, demographic profiles, and Facebook profile data including status updates, Likes, social networks, views, work and education history and much more. In the later chapter of proposed mechanism, the statistical properties of *myPersonality* dataset is provided.

The *Twitter dataset* consists of a small set of 102 Twitter users, labeled with gold-standard self-assessed personality types in the range of [−0.5, 0.5]. Users have been recruited by means of a Twitter advertising campaign in different languages and their personality types have been assessed with the 10-item personality test (BFI10) [42], which is available in the selected languages[8]. In addition to personality types, they collected age and gender of the Twitter users, and a set of other metadata about them. This Twitter dataset has become available as part of the PAN2015 competition[9]. The reason that it is small is because manually labeling text (tweets) with personality scores to obtain ground truth data is expensive and to the best of our knowledge, no other publicly available datasets of tweets exist that have been labeled with personality scores.

[8] https://www.ocf.berkeley.edu/ johnlab/bfi.htm

[9] http://www.uni-weimar.de/medien/webis/events/pan-15

The *YouTube Vlog dataset* was collected by Biel et al. in 2011 [43-44], and consists of 404 vlogs. For each vlog, 25 audio-video features are available, as well as a raw text speech transcript corresponding to the full video duration, the gender of the vlogger, and personality impression scores. The personality impressions consist of Big Five personality scores that were collected using Amazon's Mechanical Turk (MTurk) crowd sourcing platform and the Ten-Item Personality Inventory (TIPI). MTurk annotators watched one-minute slices of each vlog, and rated impressions using a personality questionnaire.

The Big Five personality impression scores are available for each user over all the five traits in the range of [1, 7]. The audio-video features were automatically extracted from the conversational excerpts of the vlogs and aggregated at the video level. The video features were extracted from the vloggers body activities and include 4 features: the entropy, median, and center of gravity in horizontal and vertical dimensions.

2.2.2 Data Pre-processing

Data pre-processing is an intermediate step performed generally before the main data processing or feature extraction is performed. For computational personality prediction using textual data, simple natural language processing (NLP) related tasks are performed, such as sentence detection, removing unnecessary spaces, symbols, URLs, names in the texts, stemming of words, applying tokenization and Parts-of-Speech (POS) tagging etc. Based on the feature extraction criteria these pre-processing steps are performed [8-11].

Sentence detection process usually takes input of huge paragraphs and detects number for sentences in the given text. Whereas removal of unnecessary spaces, symbols, URLs need to provide a set of symbols or names which must be removed. Stemming of words outcomes the original word ignoring the past or future tense version of the word. POS tagging outputs the different parts of speech tags along with the words such as adjectives, nouns, verbs etc. Which steps should be performed depends on the research considerations.

2.2.3 Feature Extraction

In computational personality prediction, feature extraction is considered to be one of the most highlighted area to contribute. In the context of computational personality prediction, generally in this step from relevant features are extracted. Feature extraction starts from an initial set of measured data and builds derived values (features) intended to be informative and non-redundant, facilitating the subsequent learning and generalization steps, and in some cases leading to better human interpretation. Features could be numeric or nominal values, which are useful for further steps of prediction.

State-of-the-art works have tried to establish different set of features extracted from texts, such as topic words of the status [8], time-base features and social network features [9], sentiment features [10] etc. For extracting traditional textual or linguistic features general high-level programming language such as python, JAVA can be useful. In literature, psycholinguistic features from Linguistic Inquiry and Word Count (LIWC) and Structured Programming for Linguistic Cue Extraction (SPLICE) features [11] etc. are also used. For extracting psycholinguistic features, researchers have developed psycholinguistic tools using closed vocabulary approach. The widely used psycholinguistic tools are discussed below.

Psycholinguistic Tools used:

For extracting relevant psychological features from texts psycholinguistic tools are utilized. These software tools are developed for easier experimentations. LIWC [55], MRC [56], SPILCE [57], are widely used psycholinguistic tools. Developed by Pennebaker and Francis, a word list based text analysis tool, LIWC extracts 93 features consisting standard counts (word counts, words longer than six letters etc.), personal concerns (occupation, financial issues, health etc.), psychological processes (cognitive, emotional, perceptional and social processes) and other features (punctuation counts, swear words etc.) [55].

On the other hand, MRC [56] features are computed using Medical Research Council's psycholinguistic database which consists over 150,000 words with linguistic and psycholinguistic features of each word. MRC includes very interesting latent features of text such as Kucera-Francis written frequency [58], Brown verbal frequency [59] etc.

SPLICE (Structured Programming for Linguistic Cue Extraction) extracts 74 features related to linguistic. Upon the input of textual data SPLICE [57] outcome various features including the quantities (number of characters, sentences, words etc.), Parts of Speech features (number of nouns, noun ratio, verb ratio, adjective ratio etc.), immediacy (number of passive verbs, passive verb ratio), pronouns, positive self-evaluation, negative self-evaluation, influence, deference, whissel (imagery, pleasantness, activation), text complexity, spoken word features, tense, SentiWordNet features and readability scores. Among these three widely used closed vocabulary psycholinguistic tools, for our work we have used LIWC. LIWC consists a psycholinguistic dictionary in backend which contains huge number of words, synonyms and antonyms in different psychological categories. LIWC is proven to be useful in the context of personality traits prediction.

- ***Textual features***: Extracting textual features is a common approach in natural language processing and hence, it could be applied in this contest as the status updates are also textual data. Some of the textual features can be count of characters, words, structures and function words etc. [11].
- ***Psycholinguistic features***: Psychologists have formalized several psycholinguistic tools (e.g. LIWC) for counting the number of psychological words used by a specific user. These psycholinguistic databases are used to determine the word counts in several categories. Some of the interesting features of psycholinguistic can be count of emotional, cognitive, perception, self-focus related words [11]. The detail about the psycholinguistic tools are discussed in section 2.2.6.
- ***Social network features***: Social network features are the feature extracted from the underlying graph structure of social media. Each user is considered as a node of the graph and the relevant features can be extracted from the graph. Some of the social network features are friend network size, betweenness centrality score, transitivity score etc. [9]. These features have insightful meanings which can be mapped into computational personality prediction. The importance of social network features are extensively discussed in chapter 4.
- ***Time-based features:*** Time based features are numeric features extracted from the data associated with the status updates. The time of updating the status is provided as a timestamp with the dataset and the time related features are extracted from them. Some of the time-based features can be frequency of status updates per day, number of statuses posted between 6-11 AM etc. [9].

- ***Topic Modeling:*** Topic modeling algorithms such as Linear Dirichlet Allocation (LDA) has been applied to infer the topic words of a given text [8]. Alternative of LDA could be applied for inferring the insight topic of a given text. This types of topic modeling algorithms are widely used in natural language processing and also can be adopted to computational personality predictions.
- ***Sentiment Features***: Usually sentiment features are extracted from the status, comments, reviews or opinions to understand the positive, negative, neutral sentiments [10, 53, 54]. Several works in the context of computational personality prediction has been performed, but the impact of these sentiment information does not carry much reflection of users personality.
- ***Others***: Finding word polarities [8], Parts-of-Speech (POS) tagging, lexicon based features are explored in [53]. A review on the existing feature used for personality prediction is presented in [54] which covers the above-mentioned features descriptions.

Based on these extracted features, existing works have manually selected features to feed into classification or learning models. The detail of the above-mentioned feature extraction and computational methods are described in section 2.2.6 elaborately.

2.2.4 Feature Selection

For finding the closely related features or most prominent features, we have to apply feature selection algorithms. As we are dealing with a special type of textual data which are considered as unstructured shot texts, sometimes noisy. Working with this type of texts, the traditional natural language processing feature selectors or text mining features are not very effective, as reported in [38]. At a certain point, more features or dimensions can decrease a model's accuracy since there is more data that needs to be generalized, which is known as the curse of dimensionality [103].

Therefore, the use of features selection in this domain is relatively new, but efficient. There are two types of feature selection approach: Manual and Automated. Manual feature selection have been adopted by few researchers [9, 11] and the effect of manual feature selection on improving the overall accuracy has been reported by them. While manual feature selection, with the extracted features from the feature extraction step, researchers' have used plug-n-play method to test the effectiveness of the features. For each category of selected features, accuracy and supporting evaluations are performed. Though, alternative of manual feature selection could be automated feature selection based on feature-feature correlation values, class-feature correlation values, information gain etc. To the best of our knowledge, state-of-the-art works does not explored the effectiveness of automated feature selection methods in this context. Therefore, applying automated feature selection may result prominent features which can be utilized to improve the overall accuracy of the prediction system. Hence, the automated feature selection methods are explained with the thresholding criteria in appendix C.

Feature selection is a wide research domain in the context of data mining to mine the most relevant features in a huge feature vector. While working with huge feature vectors it is quite complex task to test all the possible feature subsets manually. Therefore, the automated feature selection algorithms can be useful in this regard. The automated feature selection could be divided into several types, such as filter-based, wrapper-based feature selection etc.

All these feature selection methods provide a ranking generated based on the relevance between feature and class. From the relationship scores such as correlation, information gain values or best-fit subset, a ranking of the features are created. And, finally the high rank features are selected for the task. There is no generic feature selection method which can be said to be best. Depending on the domain of research and feature extracted, different algorithm may perform better. The following are some of the widely used automated feature selection methods.

- Principle Component Analysis (PCA) [96, 97]
- Linear Discriminant Analysis (LDA) [98]
- Correlation-based Feature Selection (CFS) subset evaluator [60, 61]
- Information Gain (IG) [64, 65]
- Symmetrical Uncertainty (SU) [67]
- Chi-squared test (CHI) [70, 71] and
- Pearson Correlation Coefficient (PCC) [72]

Depending on the above mentioned algorithms the number of features to be selected for the problem could be determined. For comparing the feature selection criteria’s we have experimented with various scenarios for computational personality trait prediction. Detail about the feature selection algorithms can be found in Appendix C.

2.2.5 Personality Trait Learning Model

In computational personality prediction, each personality trait is considered separately and the performance of classification model has been evaluated. The personality trait learning models are the regular classification models used in machine learning systems. As each personality traits prediction system is considered as binary classification, state-of-the-art works have adopted the widely used classifiers, which works better in several research domains for binary classification.

Therefore, binary classifiers such as Naïve Bayes (NB), Decision Tree (DT), Random Forest (RF), k-Nearest Neighbor (k-NN), Logistic Regression (LR), Gradient Boost (GB), Support Vector Machine (SVM) etc. are used as traditional machine learning model.

In [11], deep learning based algorithms such as multilayer perceptron (MLP), 1-D Convolutional Neural Network (CNN), GRU etc. are utilized for learning model.

2.2.6 Computational Methods for Personality Prediction

Table 2.2: Computational Methods and Approaches

Features Used	Feature Selection Used	Classification Method Used	Evaluation
Linear Dirichlet Allocation (LDA) and Term Frequency-Inverse Document Frequency (TF-IDF) [8]	No	Support Vector Regression (SVR) SVR-Linear, Poly, RBF Decision Tree	MSE= 0.0017 (SVR) for Conscientiousness
Activity & demographic information, SentiStrength [10]	No	Linear Regression, SVM	RMSE-0.651 Using all features for Openness
Time-based and social network feature [9]	Yes (Manual fusion)	Support Vector Machine (SVM), kNN and NB	Accuracy: 63% for SVM and kNN Extraversion Trait
LIWC features [11]	Yes (Manual fusion)	Deep learning algorithms MLP, CNN-1D, LSTM, GRU	Accuracy: 70.78% (MLP) for myPersonality Accuracy: 74.17% (LSTM + CNN-1D) for Bahasa

In this sub-section, we have described the computational methods adopted in the state-of-the-art researches. The usage of psycholinguistic databases for extracting psycholinguistic features are noticeable in existing works. Therefore, we are studied some of the psycholinguistic tools, which are used for extraction of relevant psychological features.

The background study has been conducted on the basis of the following keypoints.

- Dataset used in the work
- Features extracted from the user-generated data
- Feature selection algorithms used
- Classification method used
- Evaluation metrics

As mentioned in Section 2.2.3, the number of publicly available datasets with ground-truth personality score or class data is very less. Therefore, we have utilized one of the widely used dataset myPersonality in our work and compared our work with the above-mentioned works as they have used the same datasets.

Personality prediction using topic modeling:

P. Howlader et al. [51] have used myPersonality Dataset and tried to utilize the Linear Dirichlet Allocation (LDA) for topic modeling words. The LDA algorithm outputs specific number of words analyzing the input texts given. These output words are tend to be the topic of the inputted text. In this work, from the topic words they have generated Term Frequency-Inverse Document Frequency (TF-IDF) features and performed two experiments separately. For experiment-1, they have used LDA and TF-IDF as feature vector and for experiment-2, they have used LIWC. LIWC features has shown better performance than LDA and TF-IDF. Machine Learning Model used in this paper are Support Vector Regression (SVR), Decision Tree (DT). The performance has been measured using Mean Squared Error (MSE).

Table 2.3: MSE values from [51]

Personality Traits	Actual Personality Score	MSE
Openness	3.912669	0.0093 (DT)
Agreeableness	3.562032	0.0120 (DT)
Extraversion	3.585833	0.0036 (SVR)
Conscientiousness	3.453712	0.0017 (SVR)
Neuroticism	2.772607	0.0068 (SVR)

Personality prediction using time-based and Social Network features:

Farnandi et al. [9] have proposed a personality prediction system using time-based and social network features using myPersonality Dataset. They have extracted time-based features from the status updates and social network features from the given dataset.

The have used manual feature fusion using Time and SN features for performance evaluation. They have experimented using the time-based features first, then only with the SN features and finally merging both the features to generate feature vectors. Different results has been highlighted as they got decent accuracy using these features. Machine learning model they utilized are Support Vector Machine (SVM), k-Nearest Neighbor (k-NN) and Naïve Bayes (NB) and for evaluation they used Accuracy (ACC). The accuracy shown in [9] is given in Table 2.4.

Table 2.4: Performance evaluation (Accuracy) from [9]

Personality Traits	Accuracy (Classifier)	Features Used
Openness	61% (SVM)	Time + SN features
Agreeableness	53% (k-NN)	Time-based features
Extraversion	62% (SVM)	SN features
Conscientiousness	54% (k-NN)	Time + SN features
Neuroticism	56% (k-NN)	SN features

Personality Prediction using LIWC and SPLICE:

Tendra et al. [11] has utilized the psycholinguistic tools such as LIWC and SPLICE for extracting features and applied both machine learning and deep learning algorithms to predict the personality. Machine learning model they have used area Support Vector Machine (SVM), Naïve Bayes (NB), Logistic Regression (LR), Gradient Boosting (GB) and deep learning algorithms used in this paper are Multilayer Perceptron (MLP), Long-Short Time Memory (LSTM), Gated Recurrent Unit (GRU), CNN-1D, LSTM + CNN-1D. This paper could be considered as the most recent work done in this area and they have claimed highest accuracy of all time 70.78% for openness using MLP. The following Table 2.4 shows the accuracy mentioned in the paper, which we have compared at chapter 4 with our approach.

Table 2.5: Accuracy values from [11]

Personality Traits	Accuracy using machine learning (Classifier)	Accuracy using deep learning (Algorithm)
Openness	63.20% (GB)	70.78% (MLP)
Agreeableness	63.20% (GB)	59.13% (CNN-1D)
Extraversion	68.80% (SVM)	65.39% (MLP)
Conscientiousness	59.20% (NB)	63.26% (GRU)
Neuroticism	60.40% (LR)	64.52% (GRU)

The correlation between the usage of Facebook, thus social media and personality has been studied in [45-46]. In [45] the study shows that the correlation is higher for neuroticism and extraversion trait, but average for the other traits. Different literature works establishes the relationships between the personality and social media uses such as personality of popular social media users [47], influence of personality from Facebook usage, wall posting [48], by mining social interactions in Facebook [49], capturing personality from photo or photo related posts in social media [50] etc.

A huge feature set (725 features) has been analyzed in [53] considering basic linguistic features, POS-tagger parameters, AFINN (Lexicon list) parameters, H4Lvd parameters. A review of emerging trends of personality prediction from online social media is performed by V. Kaushal and M. Patwardhan in [54]. They listed different categories of features such as linguistic features (LIWC features, POS tags, Speech acts, sentiment features), non-linguistic features (structural, behavior, temporal features) and social network features. Based on the features used for identifying personality traits the methodologies have modified.

In literature, determining the prominent features and their relationships that affects the accuracy to predict personality are still unexplored. Researchers' have performed experiments using manual feature fusion, rather than considering the effectiveness of automated feature selection techniques. Also, the manual fusion of features are not elaborately studied. The studies represents the best classifier to be used for individual personality traits, but not focusing on the prominent features that should be used for specific personality traits.

2.3 Scopes and Contribution to Knowledge

The contribution to the existing knowledge and statement of significance are listed and discussed in this section. In order to achieve the primary and secondary research goals or objectives, we have designed experimentations following the data mining step sequences. The main contributions of this thesis are addressed here.

- One of the main contribution of this thesis is to utilize the linguistic and psycholinguistic features extracted from the Facebook status updates. The feature extracted are 93 in terms of number. Working with this high feature dimension is one of the contribution to the literature.
- In the state-of-the-art works, manual feature selection process is not elaborately studied. Researchers' tried to understand the impact of their proposed features using manually inputting them into traditional classification algorithm. In this thesis, we have considered 16 cases for manual feature selection and analyzed which feature set are prominent for which personality traits.
- In the literature, to the best of our knowledge, the automated feature selection algorithms are not applied to determine the most relevant features for each of the personality traits. In our thesis, we have applied several automated feature selection algorithms to extract the specific features which are important and impactful for the individual personality traits. We have analyzed the features to determine the most prominent features for individual personality traits and for determining the commonly found feature for all the personality traits.
- Using computation personality prediction models, the state-of-the-art paper shows around 69% accuracy found for Extraversion using deep learning based multilayer perceptron algorithms. From our experimental models we have found accuracy of above

70% of the same trait. Not only extraversion, but also for the other four personality traits, our proposed mechanism of using automated feature selection has performed better than the state-of-the-art models. A better performing computational model could be used by the psychologists to predict individual personality traits more accurately.

- Our experimental outcomes served as an alternative source of knowledge and attempt to fill the gaps of the traditional personality prediction reports and surveys.

CHAPTER 3

PROPOSED MECHANISM

In this chapter, we have described the proposed methodology in detail for the experiments. The proposed method and experimental method, both the terms could be used in this chapter having the same meaning. The data collection or dataset used, data pre-processing, feature extraction, feature selection and applying the classification model are the steps for overall methods.

3.1 Proposed Personality Prediction System

For the experimental analysis, we have designed a common mechanism for testing the performance of each of the feature selection algorithms. The proposed method consists data acquisition, data pre-processing, feature extraction, feature selection and classification as depicted in Fig. 1. For each of the five personality traits, we are going to apply the proposed method. The rest of the section elaborately discusses on the steps of experimental methods.

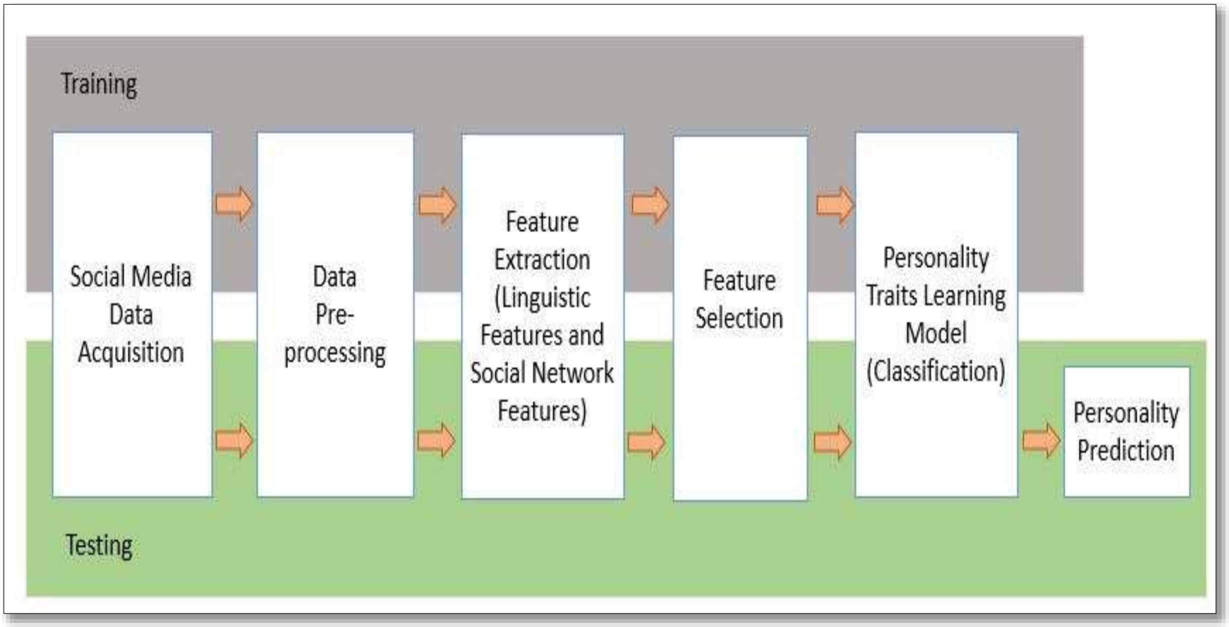

Training
Social Media Data Acquisition
Data Pre-processing
Feature Extraction (Linguistic Features and Social Network Features)
Feature Selection
Personality Traits Learning Model (Classification)
Personality Prediction
Testing

Fig. 3.1. Overview of proposed mechanism

Figure 3.2 shows the elaborated steps performed for each steps in the proposed mechanism.

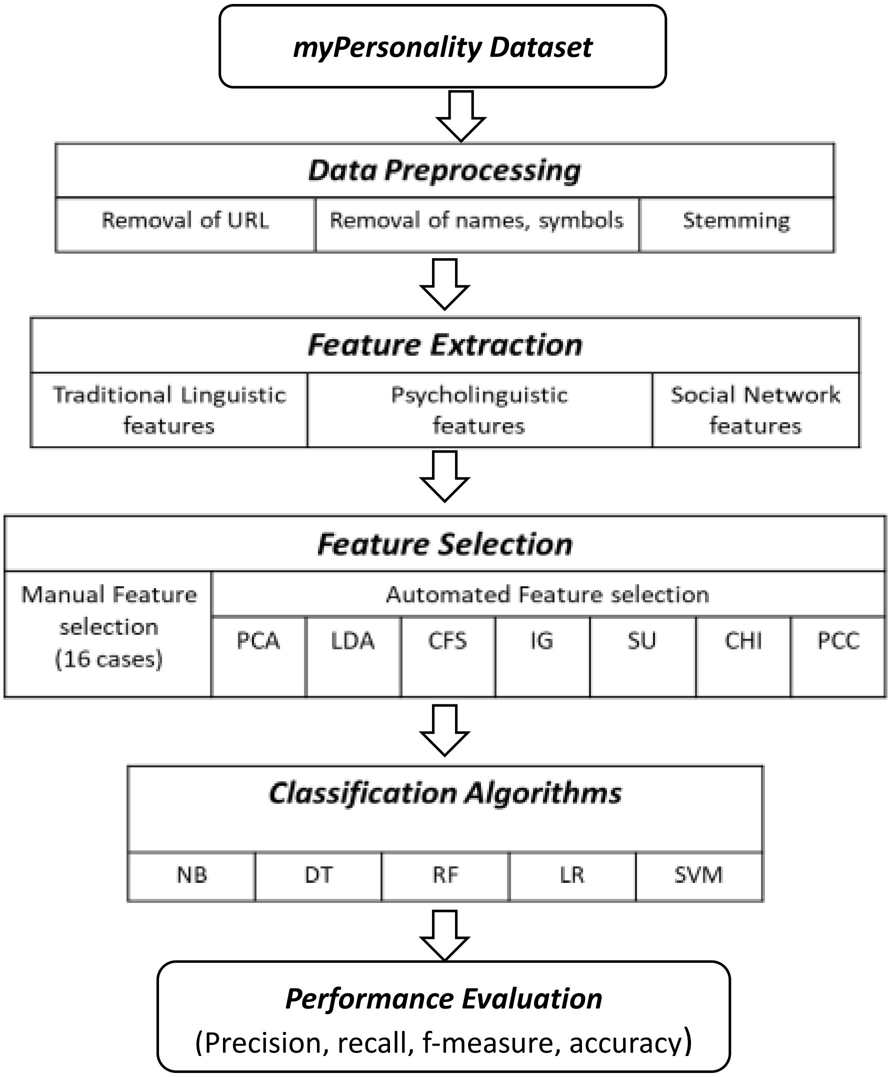

myPersonality Dataset
Data Preprocessing
Removal of URL
Removal of names, symbols
Stemming
Feature Extraction
Traditional Linguistic features
Psycholinguistic features
Social Network features
Feature Selection
Manual Feature selection (16 cases)
Automated Feature selection
PCA
LDA
CFS
IG
SU
CHI
PCC
Classification Algorithms
NB
DT
RF
LR
SVM
Performance Evaluation
(Precision, recall, f-measure, accuracy)

Fig. 3.2. Proposed mechanism elaborated.

3.1.1 Data Collection

One of the main challenging task in our thesis, was to find appropriate and complete dataset for this research problem. We need to have a ground-truth dataset which includes the user-generated contents such as status updates and the labeled class data. Thus, the lack of publicly available datasets, we have worked on two different datasets. The dataset descriptions are given briefly in the following.

myPersonality Dataset

For our experiment, we have used the *myPersonality* dataset [35, 36] which consists the status updates, social network features, ground-truth personality traits scores as well as classes. The traits used in the dataset are formalized in Big Five Factor Model (FFM). For each of the five personality traits: openness to experience (O), conscientiousness (C), extraversion (E), agreeableness (A) and neuroticism (N), the personality score and the class value (yes or no) is given in the dataset. The dataset contains 250 users around 10,000 status updates and it is considered as a ground-truth dataset for personality prediction.

Table 3.1: Properties of myPersonality Dataset

Properties	**Values**
No. of Users'	250
No. of Status Updates	9917
Avg. No. of Status per user	39.668
No. of Personality Traits	5 (Openness to experience, Conscientiousness, Agreeableness, Extraversion & Neuroticism.)
Personality Labels	Scores and Classes
Personality Score	Min-1.25, Max-5, Avg-3.437103
Personality Class	Yes, No

The Table 3.1 shows the properties of the subset dataset. For each of the personality traits. The dataset is labeled with the classes having yes or no values. Therefore, using this dataset, we have a binary classification problem to be solved. The class distribution of the *myPersonality* dataset is demonstrates in Table 3.2.

Table 3.2: Class Distribution of *myPersonality* Dataset

Personality Traits	**Class Value**	
	Yes	**No**
O	176	74
C	130	120
E	96	154
A	134	116
N	99	151

Predicting basic human values [87] and churners [88] from social media is also a relevant work to do for generating ground-truth datasets. Apply magic Sauce application is a web application developed by Cambridge psychometrics Centre to predict psychological traits from digital footprints of human behavior. Their models are based on over 6 million social media profiles and matching scores on psychometrics tests. They have published their methods in the Proceedings of the National Academy of Sciences [89] and proven to predict someone better than their friends or partners [90]. Apply Magic Sauce open for any researches that want to use their API to help them collect information on psychological characteristics based on Big Five Personality without inconveniencing the participants with personality questionnaires. Thus, after using *applymagicsauce*, we have the dataset having the status information of each user along with the personality traits class label.

3.1.2 Data Pre-processing

All the status of the dataset are in English language and follows every steps of pre-processing. The pre-processing step consists the removal of URLs, names, symbols, unnecessary spaces, stemming. The reason behind removing these aspects are from commonsense. As we know URLs (Unified Resource Locator) are used in social media mainly for commercial advertisement or promotional activities. Names in status could reflect the privacy concern regarding someone. Symbols or unnecessary spaces are used for making attractive statuses. Therefore, above mentioned facts are considered as valid reasons to eliminate these text-parts from the actual status. The pre-processed data are fetched to the next step for feature extraction purpose. The operations are performed using NLTK package [80] library. NLTK package gives the necessary elements and objects to do the pre-processing.

3.1.3 Feature Extraction

In this step, the extracted features are in two categories: linguistic features and social network features. We have extracted the traditional linguistic features and psycholinguistic features as well.

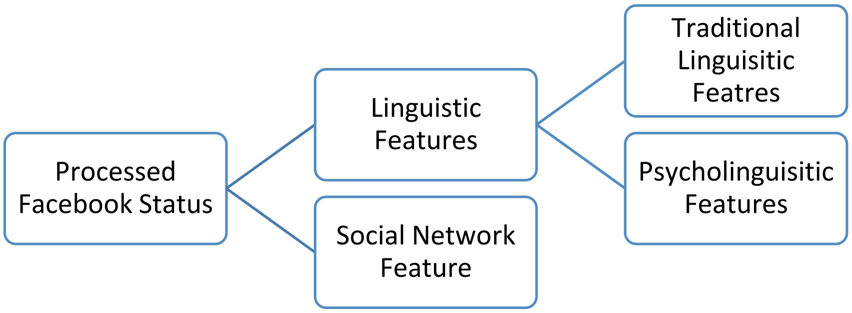

Processed Facebook Status
Linguistic Features
Social Network Feature
Traditional Linguisitic Featres
Psycholinguisitic Features

Fig 3.3. Types of features extracted from the Facebook status updates.

Traditional Linguistic Features:

The traditional linguistic features are textual features which could be divided into four types: character based, word based, structural and function words. The list of traditional features considered for our study are shown in Table 3.3.

Table 3.3: Traditional Linguistic Features

Feature No.	**Feature Description**
Character-level Features	
F1	No. of Characters
F2	No. of Punctuations
F3	No. of Special Characters
F4…F29	No. of individual alphabets (a, b, c,…z)
F30	Total no. of Alphabets
Word-level Features	
F31	No. of Words
F32	No. of words with 1 character
F33	No. of words with 2 character
F34	No. of words with 3 character
F35	No. of words with 4 character
F36	No. of words with 5 character
F37	No. of words with 6 character
F38	No. of words with 7 character
F39	No. of words with 8 character
F40	No. of words with 9 character
F41	No. of words with 10 character
F42	No. of words with 11 character
F43	No. of words with 12 character
F44	No. of words more than 12 character
F45	Avg. Word Length
Structural Features	
F46	No. of Sentence
F47	Avg. Sentence Length in terms of Character
F48	Avg. Sentence Length in terms of words

Function Words	
F49	No. of Function Words
F50	Percentage of Noun
F51	Percentage of Pronoun
F52	Percentage of Verb
F53	Percentage of Adjective
F54	Percentage of Adverb
F55	Percentage of Preposition
F56	Percentage of Conjunction
F57	Percentage of Interjection

For extracting the linguistic features we have applied LIWC [55] on the pre-processed textual data. LIWC gives total 93 features having psycholinguistic and traditional linguistic categorical features. All the features are integer or fractional values meaning the percentages of words in specific categories.

Among 93 features, only 29 could be considered as psycholinguistic features divided into five categories namely, emotional affect, cognitive process, self-focus, social relationships and perceptions, which are demonstrated in Table 3.4.

Table 3.4: Associated Features in Psycholinguistic Cues

Feature No.	**Feature Description**
Emotional Affect	
F58	Affect
F59	Positive emotion
F60	Negative emotion
F61	Anxiety
F62	Anger
F63	Sad
Cognitive Process	
F64	Cognitive process
F65	Insight
F66	Cause
F67	Discrepancy
F68	Tentative
F69	Certain
F70	Different
Social Relationships	
F71	Social
F72	Family
F73	Friend

F74	Female
F75	Male
Self-focus	
F76	Self-focus
F77	Work
F78	Leisure
F79	Home
F80	Money
F81	Religion
F82	Death
Perceptions	
F83	Perception
F84	See
F85	Feel
F86	Hear

Apart from the psycholinguistic features another 65 different linguistic features are extracted using LIWC. The linguistic features are word count, analytical word, tone, word per sentence, no. of six-letter words, no. of articles, different punctuation symbols (period, comma, colon, semi-colon, question mark, exclamatory mark, dash, quote, apostrophe, parenthesis etc.) etc. The percentage of function words or parts-of-speech such as percentage of noun, pronoun (personal pronoun and impersonal pronoun), preposition, adverb, conjunction, verb, adjective, comparative, interrogative words etc.

Table 3.5: Social Network Features

Feature Number	***Feature Description***
F87	network size
F88	betweenness
F89	n-betweenness
F90	density
F91	brokerage
F92	n-brokerage
F93	transitivity

Social Network Features:

The second type of feature category is social network features. In social networking sites, the architecture is build up on a graph. Each of the user is considered as one of the node of this huge graph. The edge between this nodes could be considered as the friend or connection between users. Therefore, the social network works as a huge graph.

Network Size defines the number of friends, connections or followers in social networking sites. Using this feature we may predict if the user has decent number of friends or not. Having smaller number of friends may lead to a characteristics of introvert user and vice versa.

Betweenness centrality is the measure to determine the central nodes within a graph, whereas betweenness centrality demonstrates how many times a node behaved as a connector along the shortest path between two other nodes. This measure is useful in assessing which nodes are central with respect to spreading information and influencing others in their immediate neighborhood. *Normalized Betweenness centrality* is the normalized value of betweenness centrality. *Density* is the measure of network connections. Density demonstrates the potential connections in a network that are actual connections. *Brokerage* refers to the nodes embedded in its neighborhood which is very useful in understanding power, influence and dependency effects on graphs. A broker could be considered as the communicator between two different nodes. *N-brokerage* is the normalized parameter of brokerage which is the measure of brokerage nodes divided by the number of pairs. *Transitivity* is the measurement which could be defined as FOF (Friend-of-Friend) concept of social media such as Facebook. The idea of FOF is “when a friend of my friend is my friend”. In the context of network or graph theory, transitivity is measured based on the relative number of triangles or triads present in the graph comparing to the total number of connected triples of nodes.

Moreover, in the myPersonality dataset, social network features are extracted from this huge graph. In the later chapter, from experiments, we have found that these features are closely related to the behavior and personality of a user.

3.1.4 Feature Selection

In this thesis, both manual and automated feature selection approaches has been adopted. As described in section 3.1.3 feature extraction, we have extracted three types of features namely:

traditional linguistic features, psycholinguistic features and social network features. Using these features we have experimented using both manual and automated feature selection methods.

3.1.4.1 Manual Feature Selection

For manual feature selection process, we have designed different case of input features using these features from combination and individualism. In Table 3.6, we have shown, how the features were manually selected to create 16 different input cases.

Table 3.6: Manual feature selection cases

Case	Traditional Linguistic Features (57)	Psycholinguistic Features (29)					Social Network Features (7)	No. of Features
		Emotional Affect (6)	Cognitive Process (7)	Social Relationship (5)	Self-Focus (7)	Perceptions (4)		
C1	✓	✓	✓	✓	✓	✓	✓	93
C2	✓							57
C3		✓	✓	✓	✓	✓		29
C4							✓	7
C5	✓						✓	64
C6		✓						6
C7			✓					7
C8				✓				5
C9					✓			7
C10						✓		4

Case	Traditional Linguistic Features (57)	Psycholinguistic Features (29)					Social Network Features (7)	No. of Features
		Emotional Affect (6)	Cognitive Process (7)	Social Relationship (5)	Self-Focus (7)	Perceptions (4)		
C11		✓					✓	13
C12			✓				✓	14
C13				✓			✓	12
C14					✓		✓	14
C15						✓	✓	11
C16		✓	✓	✓	✓	✓	✓	36

As depicted above, the manual feature selection cases considers the combination of all features (C1), individual category of features (C2, C3, and C4) and combination of the features (rest of the cases). For understanding the impact and effects of each type of features, we have designed these cases and the performance of these feature combinations are given in next chapter.

3.1.4.2 Automated Feature Selection

Automated feature selection algorithms are used to find the essential or important features from a set of feature vector. In features extraction step, we have collected the prominent features, each feature vector containing 93 features.

$$F = \{F1,F2,F3,\ldots,F93\} \tag{3.1}$$

This feature vectors are used to find the optimal number of essential features using the features selection methods.

In our proposed approach, we have applied seven different types of automated feature selection algorithms namely: Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), Correlation-based Feature Selection (CFS) subset evaluator, Information Gain (IG), Symmetrical Uncertainty (SU), Chi-squared test (CHI) and Pearson Correlation Coefficient (PCC). The common steps that we have gone through for each of the above mentioned automated feature selection algorithms could be depicted as the following Fig. 3.4.

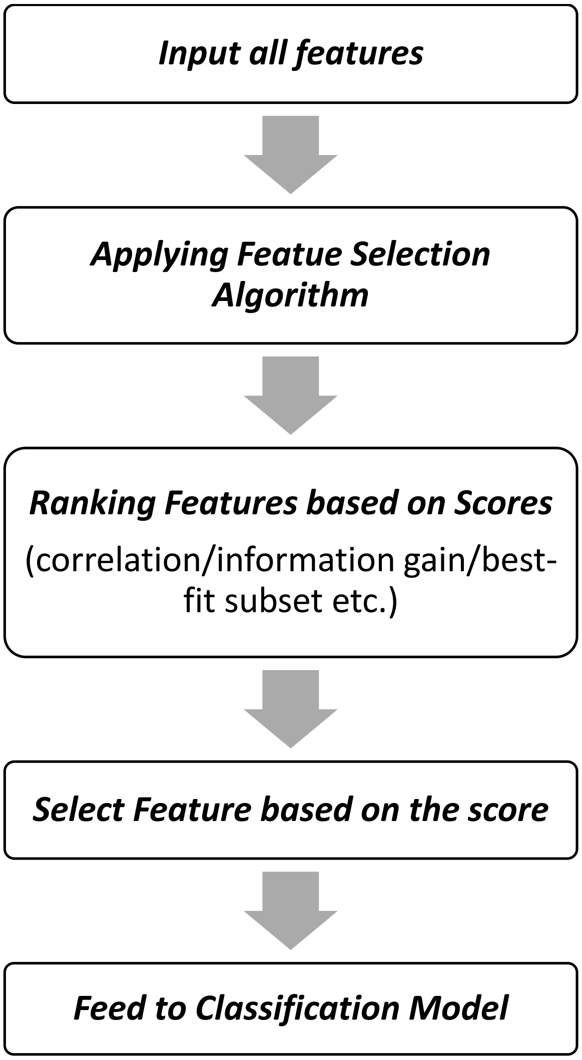

Input all features
Applying Featue Selection Algorithm
Ranking Features based on Scores
(correlation/information gain/best-fit subset etc.)
Select Feature based on the score
Feed to Classification Model

Fig. 3.4. Steps of automated feature selection algorithms.

The mentioned seven different types of feature selection methods are adopted, generated the ranking of the features based on specific scores such as correlation value, information gain ratio, best-fit subset values etc. Then based on the scores the features were selected and the selected features are feed in to the classifiers. The performance matrices are determined for each classifiers to evaluate the experimental method. Finally, the highly accurate feature selection algorithm is identified.

3.1.5 Classification Model

Based on training-testing mechanism, we have applied the most popular classification algorithms. For our experiments, the problem can be defined as a binary classification problem. Therefore, we have exploited the traditional machine learning based classification algorithms which are widely used for binary classification. The classifiers, which we have used are:

- Naïve Bayes (NB) [81],
- Decision Tree (DT) [82],
- Random Forest (RF) [83-84],
- Logistic Regression (LR) [86] and
- Support Vector Machine (SVM) [85]

The usability of these classification model for binary classification in various research domain is the main reason of using them. Details about the above-mentioned classification algorithms can be found in Appendix D.

Chapter 4

EXPERIMENTAL RESULT & DISCUSSION

In this chapter, we have presented the elaborated analysis of the experiments performed for this thesis. The experiments are designed in this chapter and the experimental setup required to perform the experiments are described. Then, the performance metrics are described with the required formulas. The result analysis for the experiments are illustrated, separately.

4.1 Experiments

The research contributions are presented in chapter 1 while describing the problem statement of the thesis. From the contribution point of view, we have design and come up with two different experiments for this thesis.

We have studied the literature and found personality prediction systems in literature using profile data such as status updates, likes, comments, share options, even profile pictures. The user-generated contents are useful to solve this problem without asking a huge set of questionnaire.

Experiment-1 is designed for manual feature selection using all the linguistic (traditional and psycholinguistic) and social network features. 16 different input cases are analyzed in this experiment. The experimental setup used for this experiment is similar to the next experiment, except from the feature selection technique used. Different scenarios have been designed by segmenting the feature sets such as only emotional features, only cognitive features etc. We have considered different scenarios by plug-n-play different feature sets and classification algorithms applied on them. The results are illustrated according to the scenarios.

Experiment-2 is designed to understand the impact of automated feature selection using the linguistic features incorporating the psycholinguistic features and social network features to apply on the supervised learning model. Through this experiment, we try to understand the effects of SN Features like network size, betweenness centrality, density, brokerage, transitivity etc. We have applied seven automated feature selection algorithms to determine which features are closely related to the class.

The positive impacts are evaluated by the performance metrics such as precision, recall, f-score, accuracy and AUC curves. To find the impact of social network features on the prediction system, we have applied the feature selection algorithms namely, PCA, LDA, CFS, CHI, IG, SU and PCC based feature selection. The selected features are then passed into the prediction model and evaluated the system using the same performance metrics. In this experiment, we have analyzed the features to determine the most prominent features for individual personality traits and features that are commonly found in every personality traits.

4.2 Experimental Setup

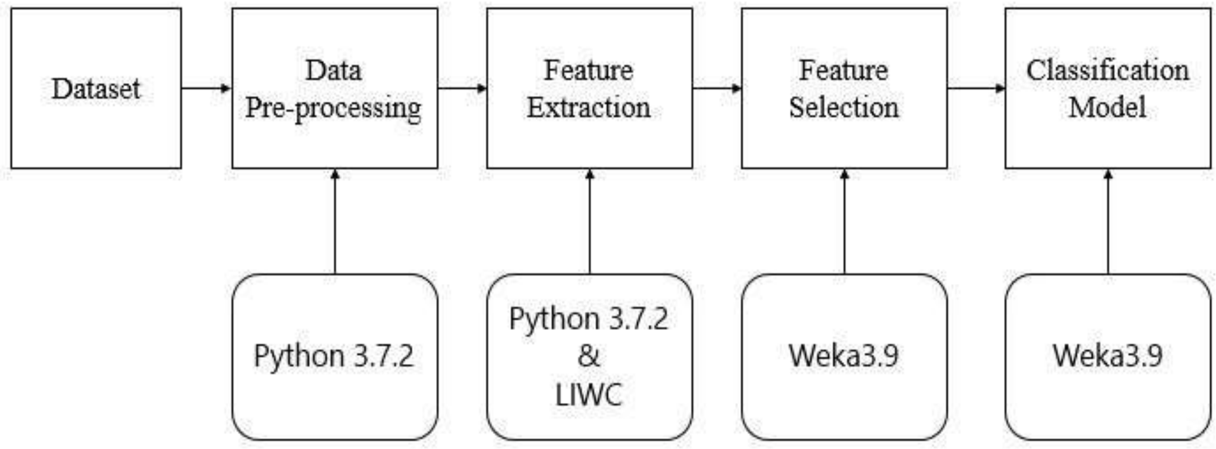


Fig 4.1: Experimental Setup for personality prediction system

The experimental setup has been designed according to the ideal pattern recognition process including the data acquisition, data pre-processing, feature extraction, data processing and classification models. The setup requires different types of software tools and mechanisms.

To evaluate performance and effectiveness of our experiment, we applied several software and open source tools as illustrated in fig 4.1. The experiments were carried out in a computer with the following configurations.

Operating System: Windows 10

Processor: Intel Core i5

RAM: 6 GB

Internet: 3G Connection of Local Operator

The following Table 4.1 shows the name of the package, descriptions and URL links used for the experiments. This open source packages are used in our experiment as they are being used widely by the machine learning researchers.

Table 4.1: Open source software/packages used for experimental analysis

Package Name	Descriptions	URL Link
Python 3.7.2	Python is a widely used high-level, general-purpose, interpreted, dynamic programming language. Its design philosophy emphasizes code readability, and its syntax allows programmers to express concepts in fewer lines of code than would be possible in languages such as C++ or Java. The language provides constructs intended to enable clear programs on both a small and large scale.	https://www.python.org/downloads/release/python-372/
LIWC	LIWC (Linguistic Inquiry and Word Count) is a text analysis program. It calculates the degree to which various categories of words are used in a text, and can process texts ranging from e-mails to speeches, poems and transcribed natural language in either plain text or Word formats.	http://liwc.wpengine.com/
WEKA 3.9	Waikato Environment for Knowledge Analysis (Weka) is a suite of machine learning software written in Java, developed at the University of Waikato, New Zealand. It is free software licensed under the GNU General Public License. Weka contains a collection of visualization tools and algorithms for data analysis and predictive modeling, together with graphical user interfaces for easy access to these functions.	https://www.cs.waikato.ac.nz/ml/weka/index.html

	Weka is a collection of machine learning algorithms for data mining tasks. It contains tools for data preparation, classification, regression, clustering, association rules mining, and visualization.	

4.3 Performance Evaluation Metrics

For evaluating prediction systems, the widely used performance metrics are precision, recall, f-score and accuracy. As we have applied different binary classification algorithms, the evaluation metrics are kept same for all the algorithms. In this sub-section, we are going to describe these performance metrics with the help of formulas to calculate them.

In the context of our experiments, True Positive (TP) = Personality traits are actually positive and predicted positive; True Negative (TN) = Personality traits are actually negative and predicted negative; False Positive (FP) = Personality traits are actually negative but predicted positive and False Negative (FN) = Personality Traits are actually positive but predicted as negative. For each of the personality traits that we are working on are considered to be examined using these same metrics. Using these four metrics the other metrics such as precision, recall, f-score and accuracy are measured.

Precision is the ratio of true positives to the cases that are predicted as positive. It is the percentage of selected cases that are correct.

$$Precision\ (P) = \frac{TP}{TP+FP} \qquad (3.2)$$

Recall, also known as sensitivity, is the ratio of true positives to the cases that actually positive. It is the percentage of corrected cases that are selected.

$$Recall\ (R) = \frac{TP}{TP+FN} \qquad (3.3)$$

F1-score is the mean of Precision and Recall. It takes both false positives and false negatives into an account. F-measure is calculated as:

$$F1 - score = \frac{2*Precision*Recall}{Precision+Recall} \qquad (3.4)$$

Accuracy is the considered to be the base metric for any kind of prediction system. The percentage calculated over the equation are within the range of zero to hundred percent and the more it is the better.

$$Accuracy\ (ACC) = \frac{TP+TN}{TP+TN+FP+FN} \tag{3.5}$$

4.4 Result Analysis of Experiment 1

In this experiment, we tried to emphasis on the hypothesis of generating a machine learning model for personality prediction system without asking prior questionnaire to the personality tester. For this experiment, we have used only the Facebook status updates of the users and extracted the linguistic features from them. The two types of linguistic features: 57 tradition linguistic features and 29 psycholinguistic features are extracted from the status updates. Total 93 features of different category are extracted from the given text using LIWC. Among them only 29 features are related to psycholinguistic. Therefore, we have only counted them for our experiments. The input features passed into the experimental model gives different outcomes and different scenarios.

For case 1, input features considered for this experiment are total 93 features (57 traditional linguistic + 29 psycholinguistic features + 7 social network features). The mentioned classification algorithms area applied over the model and the performance metrics are calculated for each of the personality traits. The performance metrics are shown in Table 4.2 for the following scenario only for the extraversion trait.

Credentials of Case 1:

Input Features: 93 features (57 traditional linguistic + 29 psycholinguistic features + 7 social network features)

Feature Extractor: Python for traditional linguistic and LIWC for psycholinguistic features

Classification Models: NB, DT, RF, SLR and SVM

Table 4.2 illustrates the precision, recall, F1-score and accuracy generated from the five different classification models namely Naïve Bayes (NB), Decision Tree (DT), Random Forest (RF), Simple Logistic Regression (SLR) and Support Vector Machine (SVM). The classifiers which gives highest accuracy among the five, are highlighted in bold. In case 1, for openness RF classifier gives the better accuracy (69.26%) over the other classifiers. For conscientiousness trait, SVM gives 54.51%, for extraversion RF gives 68.03% and for both agreeableness & neuroticism traits NB gives highest accuracy of 57.79% & 60.66%, respectively. We can see, for different personality traits different classifiers are showing better performance in terms of accuracy.

Table 4.2: Performance metrics of case 1 for OCEAN Model

Personality Trait	Classifier	Precision	Recall	F1-score	Accuracy (%)
Openness-to-experience (O)	NB	0.618	0.467	0.474	46.72%
	DT	0.579	0.598	0.587	59.84%
	RF	0.485	0.693	0.57	**69.26%**
	SLR	0.484	0.689	0.568	68.85%
	SVM	0.567	0.676	0.582	67.62%
Conscientiousness (C)	NB	0.517	0.520	0.481	52.05%
	DT	0.509	0.512	0.504	51.23%
	RF	0.473	0.475	0.472	47.54%
	SLR	0.535	0.537	0.530	53.69%
	SVM	0.544	0.545	0.539	**54.51%**
Extraversion (E)	NB	0.620	0.537	0.527	53.69%
	DT	0.574	0.570	0.572	56.97%
	RF	0.679	0.68	0.651	**68.03%**
	SLR	0.619	0.635	0.615	63.52%
	SVM	0.373	0.611	0.463	61.07%
Agreeableness (A)	NB	0.582	0.578	0.548	**57.79%**
	DT	0.471	0.471	0.471	47.13%
	RF	0.508	0.512	0.508	51.23%
	SLR	0.513	0.516	0.514	51.64%
	SVM	0.544	0.549	0.537	54.92%
Neuroticism (N)	NB	0.584	0.607	0.565	**60.66%**
	DT	0.544	0.590	0.503	59.02%
	RF	0.552	0.582	0.544	58.20%
	SLR	0.562	0.598	0.493	59.84%
	SVM	0.559	0.598	0.455	59.84%

The variation of precision and recall values are also mentionable here. Fig 4.2 represents the result of precision-recall curve (PRC), illustrated as line graph, for five different personality traits and classifiers.

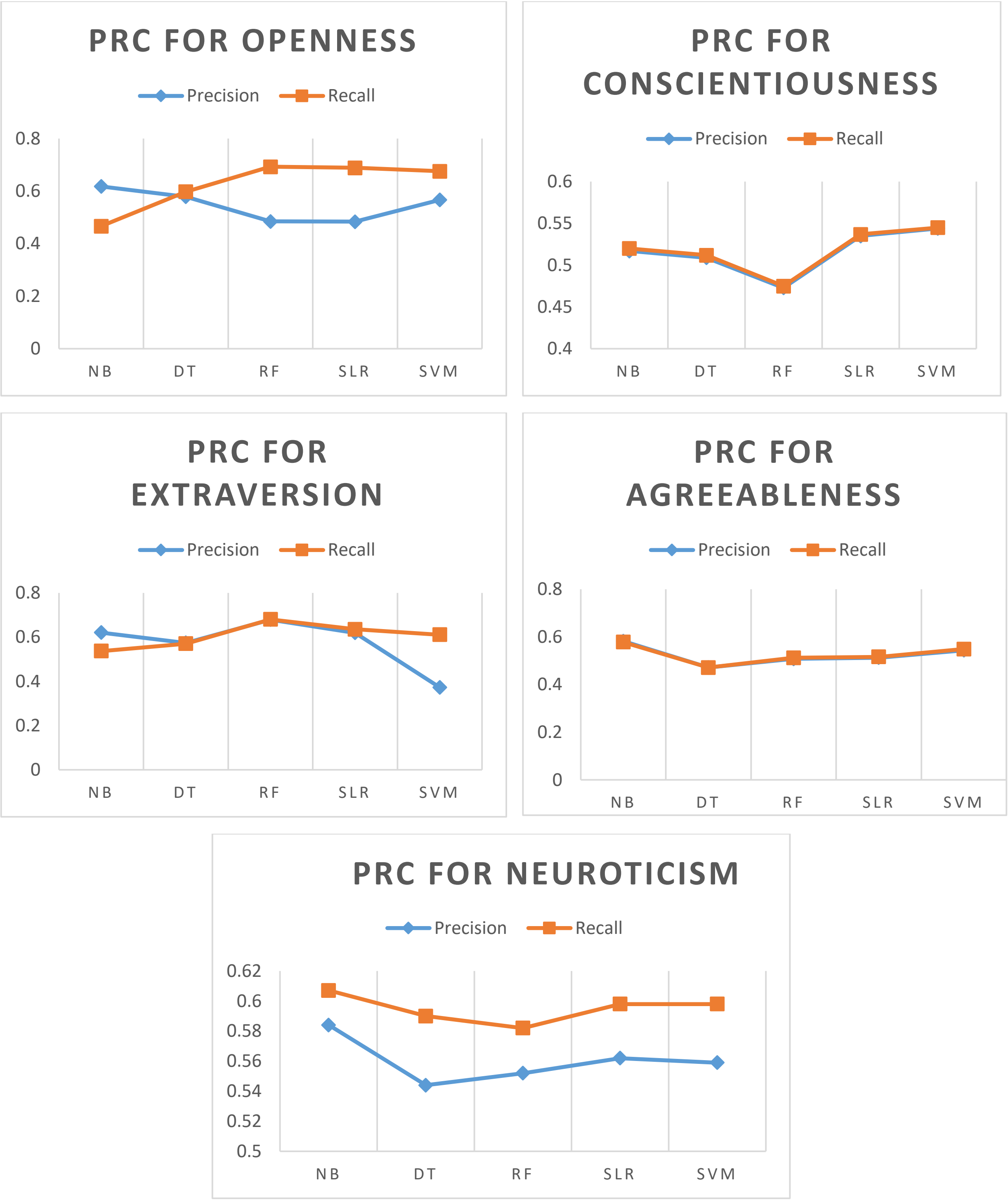


Fig 4.2: Precision-Recall Curve (PRC) for five different personality traits

Fig 4.3 shows the performance of classification algorithms in terms of accuracy for each of the traits. It is to be mentioned that, for different personality traits different classifiers are proven to perform better.

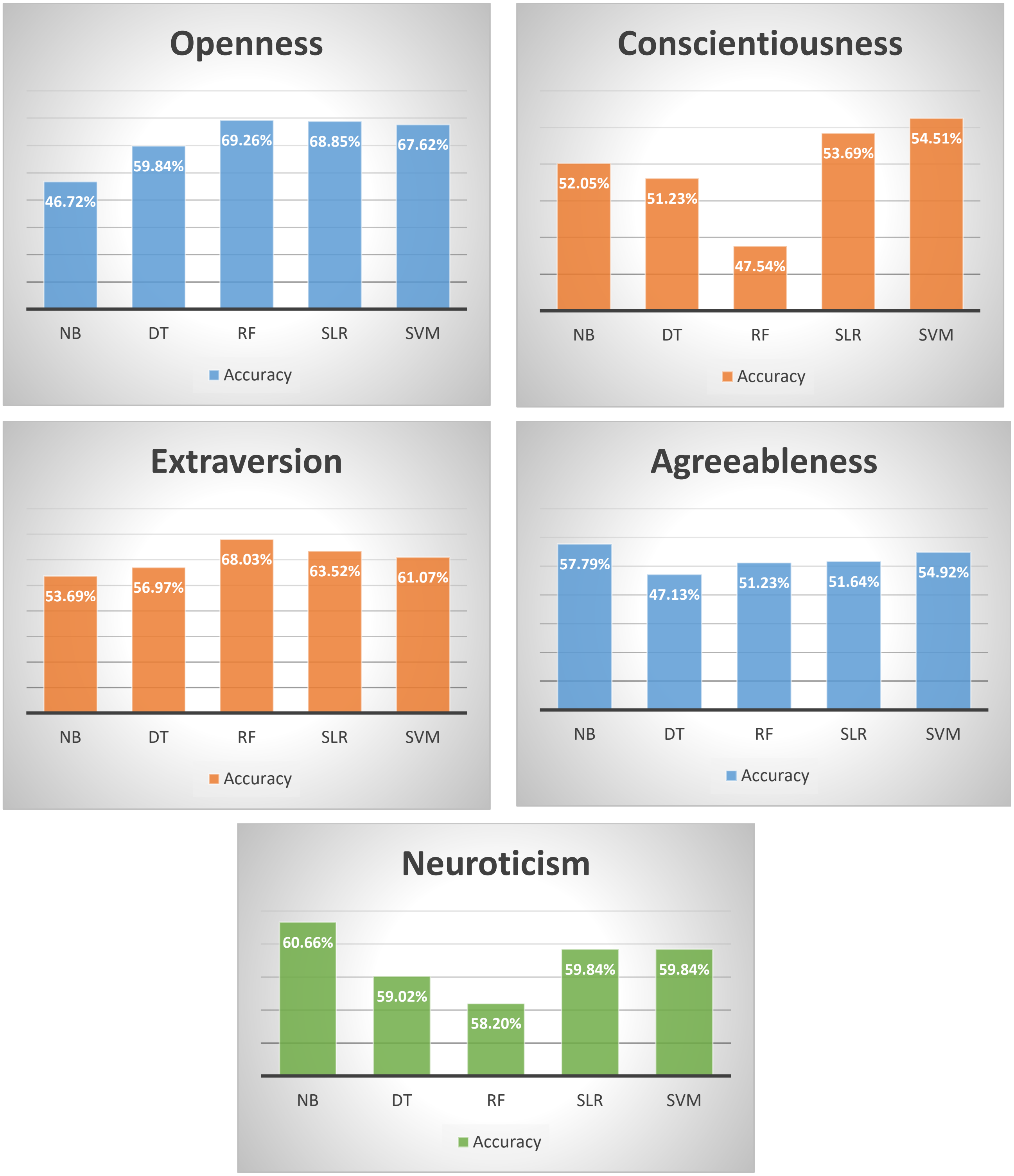


Fig 4.3: Performance of Classifiers in terms of accuracy for five traits

We tried to understand the effectiveness and sufficiency of psycholinguistic features extracted from status updates of social media users to predict the personality traits. The following Table 4.3 shows the accuracy obtained in first 4 cases for each of the personality traits.

Table 4.3: Accuracy of C1, C2, C3 and C4 input cases of Experiment 1

Input Cases	Classifier	O	C	E	A	N
C1 (All features)	NB	46.72%	52.05%	53.69%	**57.79%**	**60.66%**
	DT	59.84%	51.23%	56.97%	47.13%	59.02%
	RF	**69.26%**	47.54%	**68.03%**	51.23%	58.20%
	SLR	68.85%	53.69%	63.52%	51.64%	59.84%
	SVM	67.62%	**54.51%**	61.07%	54.92%	59.84%
C2 (Only Traditional Linguistic)	NB	53.48%	59.05%	54.66%	52.10%	60.01%
	DT	57.85%	52.20%	57.18%	54.14%	59.01%
	RF	**59.97%**	**53.74%**	60.02%	57.23%	58.02%
	SLR	57.14%	51.13%	58.89%	58.19%	59.83%
	SVM	54.43%	52.24%	**61.19%**	**59.79%**	**60.12%**
C3 (Only Psycholinguistic)	NB	38.52%	**60.24%**	47.54%	54.50%	**60.65%**
	DT	69.67%	51.22%	61.06%	52.86%	59.01%
	RF	67.21%	54.50%	57.79%	**59.83%**	58.20%
	SLR	68.85%	54.50%	60.06%	56.55%	59.84%
	SVM	**69.77**%	56.96%	**62.21%**	54.09%	60.25%
C4 (Only SN Features)	NB	69.67%	53.69%	58.61%	53.28%	55.74%
	DT	69.67%	48.36%	65.98%	52.46%	58.61%
	RF	69.57%	48.77%	**68.85%**	50.41%	58.20%
	SLR	69.26%	**58.20%**	67.62%	56.97%	**62.70%**
	SVM	**69.67%**	48.77%	67.62%	**57.38%**	59.43%

From this four cases several findings can be put together. The findings are stated below and shown in Fig. 4.4.

- Combining all the features (case 1) declines the performance than using individual feature categories.

- Social network (SN) feature performs better than traditional or psycholinguistic features for Openness, Extraversion and Neuroticism. For Conscientiousness and Agreeableness traits psycholinguistic features are performing better.

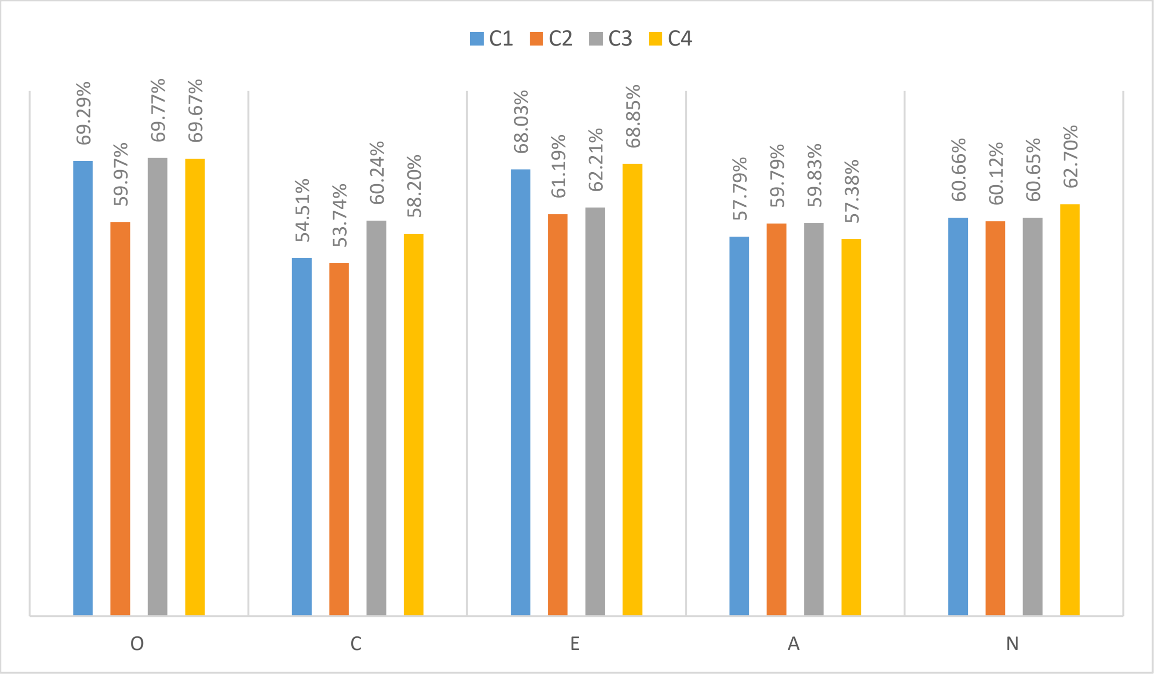


Fig. 4.4. Comparative analysis between manual feature selection case C1, C2, C3 and C4

As we use the LIWC tool for extracting the features, we get 93 features incorporating different textual features and psycholinguistic features. For this experiment, we have used only 29 psycholinguistic features.

Category of psycholinguistic features are:

- Emotional Affect
- Cognitive Process
- Self-focus
- Social relationships
- Perceptions

Therefore, for the rest of the cases were designed based in the involvement of psycholinguistic features. Case 6 to case 9 are designed using the psycholinguistic feature categories separately.

From Case 10 to case 16 are designed using the psycholinguistic feature categories along with the SN features to find out if SN features have positive impact on the personality traits.

In Table 4.4, the rest of the cases of manual feature selection utilizing these psycholinguistic features are reported. The comparison on the accuracy are depicted in Fig 4.4.

Table 4.4: Accuracy of different cases of input for Experiment 1

Input Cases	Classifier	O	C	E	A	N
C5 (Traditional Linguistic + SN)	NB	46.31%	52.05%	58.61%	55.33%	62.70%
	DT	68.85%	50.41%	61.48%	47.13%	57.79%
	RF	64.34%	48.36%	67.62%	52.87%	62.52%
	SLR	68.44%	**55.33%**	**68.85%**	53.69%	**63.11%**
	SVM	**69.67%**	50.41%	61.26%	**56.56%**	59.84%
C6 (Emotional Affect)	NB	65.16%	50.40%	44.26%	56.55%	57.78%
	DT	69.67%	51.22%	61.06%	**54.09%**	59.42%
	RF	61.06%	**55.73%**	58.19%	52.04%	**60.65%**
	SLR	68.85%	55.32%	**61.07%**	52.86%	57.78%
	SVM	**69.67%**	54.50%	61.06%	52.86%	59.73%
C7 (Cognitive Process)	NB	55.32%	53.27%	59.59%	**54.91%**	**62.70%**
	DT	69.67%	51.63%	60.65%	53.27%	59.42%
	RF	62.29%	**54.91%**	53.68%	52.45%	49.18%
	SLR	69.67%	52.45%	60.65%	51.63%	59.83%
	SVM	**69.67%**	53.27%	**61.07%**	52.45%	59.83%
C8 (Social Relationships)	NB	65.16%	56.96%	56.55%	50.00%	58.19%
	DT	69.67%	53.27%	61.06%	52.45%	59.81%
	RF	65.98%	57.78%	52.45%	52.86%	54.50%
	SLR	69.60%	56.14%	59.42%	**56.96%**	**59.85%**
	SVM	**69.67%**	**58.19%**	**61.10%**	51.63%	59.83%
C9 (Self-focus)	NB	44.26%	51.22%	44.26%	51.62%	60.25%
	DT	69.26%	51.63%	60.06%	53.27%	61.38%
	RF	65.16%	42.62%	54.91%	50.00%	50.40%
	SLR	69.67%	46.72%	61.00%	**53.68%**	**61.47%**
	SVM	**69.67%**	**51.69%**	**61.43%**	52.45%	60.25%
C10 (Perceptions)	NB	32.37%	56.55%	40.98%	50.00%	57.78%
	DT	69.05%	49.59%	61.06%	**53.27%**	59.79%
	RF	67.21%	47.13%	52.86%	50.81%	54.50%
	SLR	68.85%	51.22%	59.01%	52.45%	**59.85%**
	SVM	**69.67%**	**53.68%**	**62.21%**	53.27%	59.43%
C11	NB	**69.69%**	56.97%	57.38%	56.97%	55.83%
	DT	69.67%	52.05%	65.15%	51.64%	59.84%

(Emotional Affect + SN)	RF	61.89%	56.15%	66.39%	52.87%	**63.39%**
	SLR	68.44%	**61.07%**	65.98%	57.38%	59.43%
	SVM	69.67%	54.92%	**67.62%**	**59.84%**	59.02%
C12 (Cognitive Process + SN)	NB	48.36%	54.10%	57.38%	54.92%	59.02%
	DT	69.67%	48.77%	63.52%	50.41%	55.74%
	RF	65.16%	53.28%	65.16%	52.87%	**62.30%**
	SLR	69.67%	58.61%	65.98%	56.15%	59.84%
	SVM	**69.67%**	**56.97%**	**68.03%**	**58.20%**	60.25%
C13 (Social Relationships + SN)	NB	43.85%	**57.79%**	69.67%	56.15%	56.97%
	DT	69.67%	50.82%	63.52%	53.69%	59.02%
	RF	65.98%	52.87%	65.16%	52.87%	**63.11%**
	SLR	68.44%	57.38%	65.57%	57.38%	62.30%
	SVM	**69.67%**	56.56%	**68.03%**	**59.02%**	59.43%
C14 (Self-focus + SN)	NB	40.98%	52.05%	56.56%	53.28%	**63.11%**
	DT	68.44%	48.36%	64.34%	51.23%	57.38%
	RF	65.98%	48.77%	66.80%	52.87%	59.84%
	SLR	68.44%	**52.46%**	66.80%	55.74%	58.20%
	SVM	**69.67%**	50.82%	**67.21%**	**57.33%**	60.25%
C15 (Perceptions + SN)	NB	40.97%	58.20%	**68.26%**	50.82%	54.92%
	DT	69.67%	46.31%	65.98%	53.28%	57.79%
	RF	68.98%	48.36%	63.93%	54.10%	62.70%
	SLR	69.26%	**59.02%**	67.21%	57.38%	**64.75%**
	SVM	**69.67%**	56.15%	67.62%	**57.79%**	59.43%
C16 (Psycholinguistic + SN)	NB	40.98%	61.07%	55.33%	53.28%	**60.25%**
	DT	66.39%	49.18%	49.59%	50.66%	55.33%
	RF	67.21%	59.02%	53.28%	54.75%	58.19%
	SLR	69.26%	59.83%	55.74%	55.57%	56.56%
	SVM	**69.27%**	**59.84%**	**58.20%**	**56.80%**	59.43%

In Table 4.4, the supervised model is run for the five classifiers inputting these psycholinguistic features separately in C6 to C10. Then, these features are combined with the social network features (SN features) and considered as case C11 to C15. Finally, all the features are inputted together (C16), to understand the effect in accuracy.

Considering the five classification algorithms, different algorithms have performed better in different cases. Experimental results shows that, the SVM classifier performs better for openness-to-experience trait. Again, the RF outperforms other classifiers for conscientiousness using the emotional affect and cognitive features. On the other hand, the SVM outperforms other classifiers for conscientiousness using social relationships, self-focus and perception features. The most unstable performance is shown for agreeableness traits as different classifiers are proven to work better for different input features.

For each of the traits except for neuroticism, all features combination is performing better than the individual features. As the number of features lacks, the relevant information to predict the personality trait also lacks behind. Therefore, adding all the features together demonstrates higher accuracy. This situation is common for extraversion, openness-to-experience, agreeableness and conscientiousness traits.

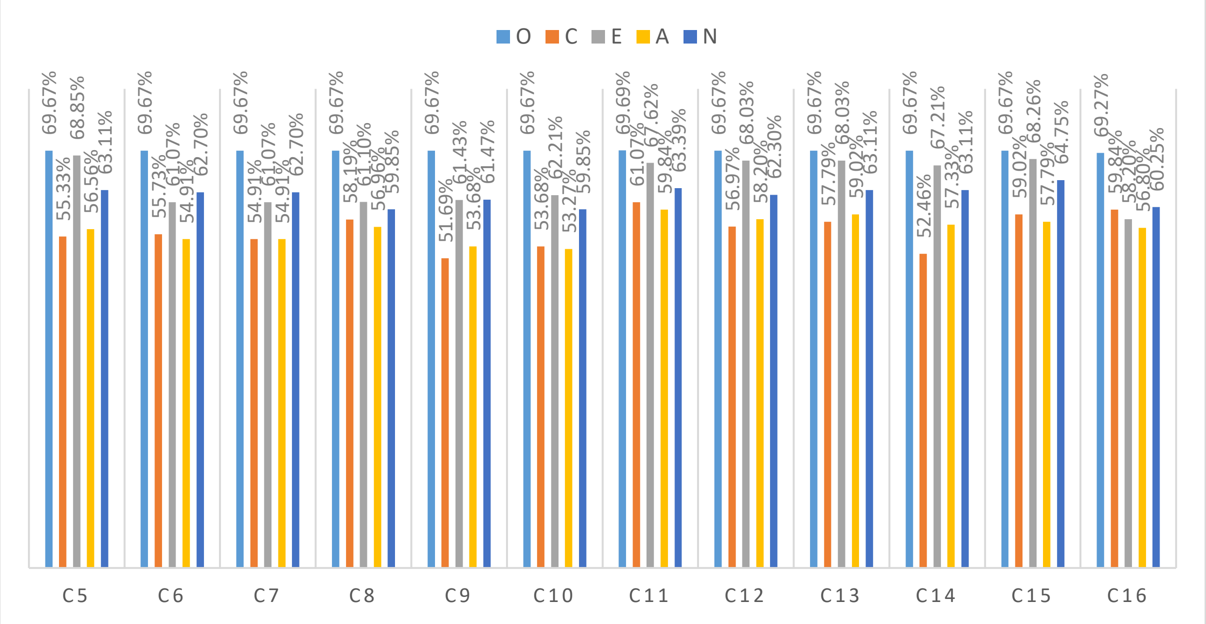
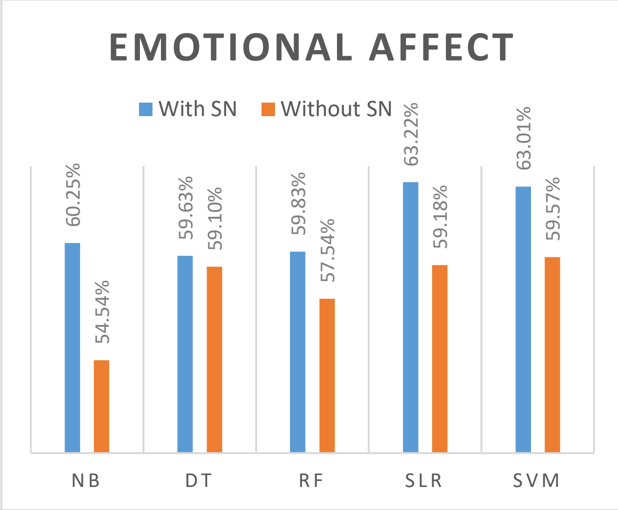


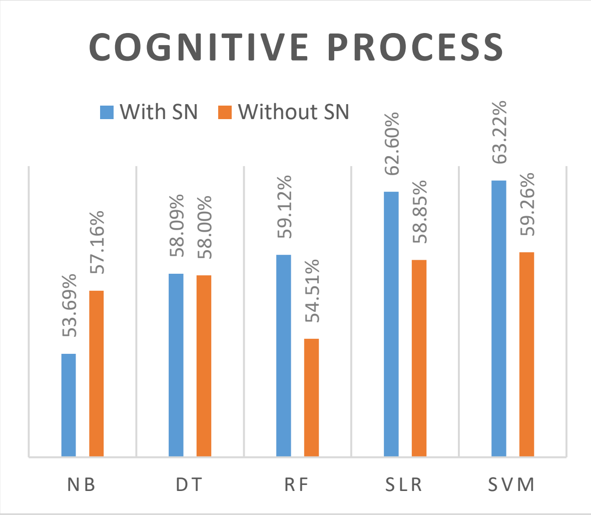
Fig. 4.5. Accuracy of classifiers for Experiment-1 (C5 to C16)

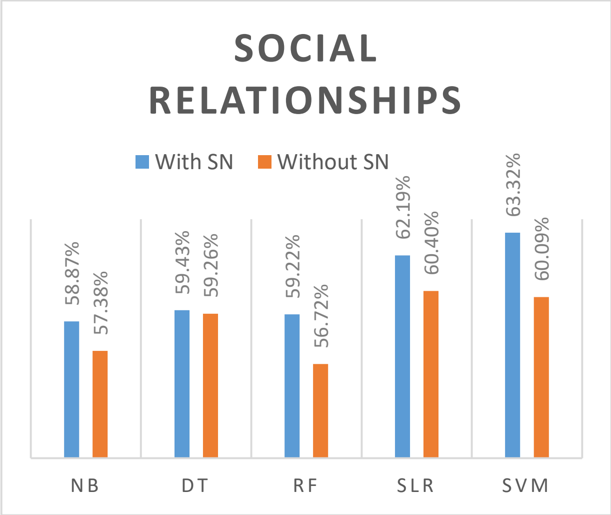
An interesting observation is made for Neuroticism trait. Individual cues are showing better performance than the collective approach. The highest accuracy for each of the psycholinguistic cues are highlighted in bold and maximum (62.70%) between the highest accuracies is found from the cognitive process cues and social network features.

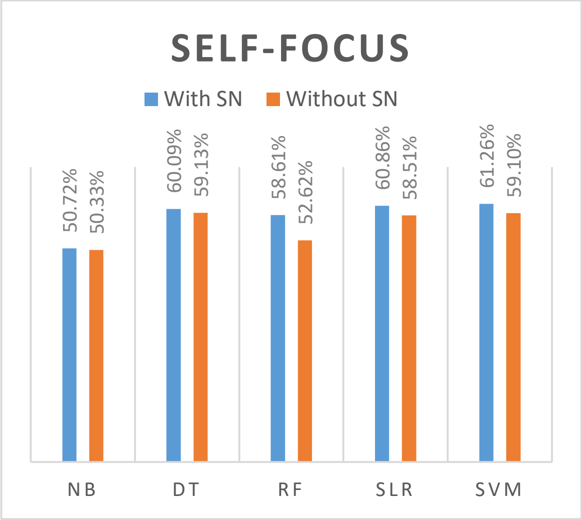
For emotional affect RF gives the maximum accuracy and for cognitive process NB gives the maximum accuracy. SLR performs consistently better for the other psycholinguistic cues.

Self-focus features shows promising performance (61.48%) having smaller difference than the cognitive process. Therefore, we can derive that cognitive process would be more useful psycholinguistic cue to predict neuroticism from Facebook statuses.


EMOTIONAL AFFECT


COGNITIVE PROCESS


SOCIAL RELATIONSHIPS


SELF-FOCUS

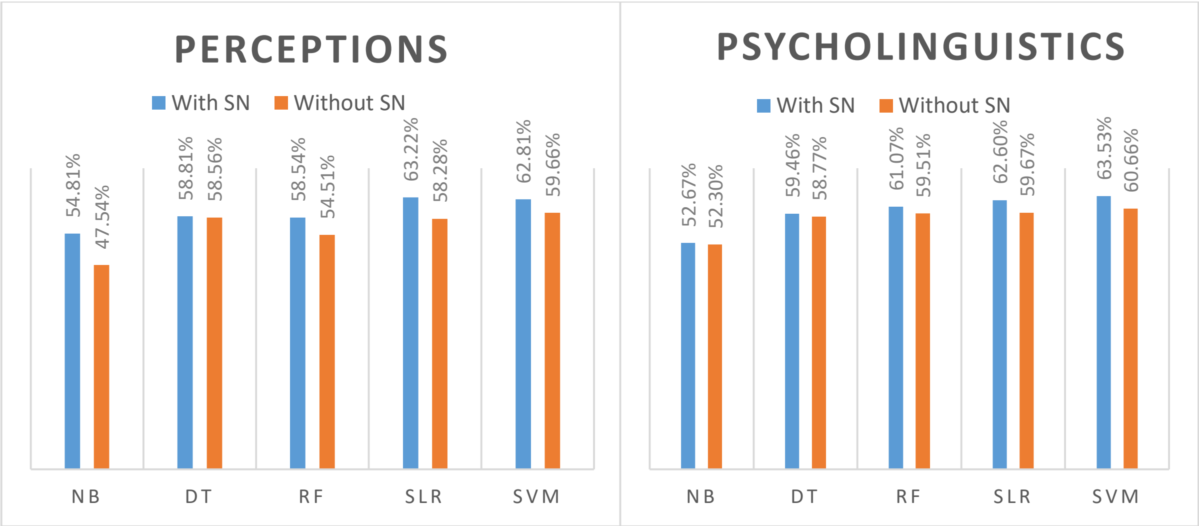


Fig. 4.6. With and Without SN features accuracy measures of classifiers for Extraversion trait.

For understanding the impact of social network features, we have designed the above-mentioned eight cases. Each case integrates the SN features with the individual psycholinguistic features and collective features.

From the above cases, we observed that the accuracy value improved when we use the social network features with the psycholinguistic features. The social network features alone are giving quite good margin of accuracy, but if you add the social network features with the individual psycholinguistic features such as emotional affect, cognition, social relationships etc. the accuracy outperforms the previous scenario. To visualize the impact of social network features, the above-mentioned Fig. 4.5 is used.

From Fig 4.5, we can see, for every case the blue bar (representing with SN features) is higher than the orange bar (representing without SN features). This proves the impact of social network features with every individual psycholinguistic feature and also with all the psycholinguistic cues together.

Experimental Findings from Experiment-1:

In Experiment-1, we have manually selected feature combinations as input case and try to find out the effectiveness of individual feature categories and found some interesting outcomes. The findings from experiment-1 can be stated as below:

- Combination of all features showing less accuracy than the individual feature categories.
- Social network (SN) feature performs better than traditional or psycholinguistic features for Openness, Extraversion and Neuroticism. For Conscientiousness and Agreeableness traits psycholinguistic features are performing better.
- Impact of SN features is mentionable, as in every cases of using SN features are giving better accuracy than the cases of not using them (shown in fig. 4.5).
- Neuroticism trait is highly correlated with the Cognitive words (62.70%) and decent correlation with self-focus words (61.47%)
- Social relationships features are showing better accuracy for openness trait in every classifier used.

4.5 Result Analysis of Experiment 2

In this experiment, we have applied automated feature selection algorithms for understanding the impact or effectiveness of features. To investigate the most prominent features for each personality traits, these automated feature selection algorithms play vital role. Basically, the automated feature selection algorithms are used for selecting bunch of features from a huge set of features, using the strength of relationships between the class and the features. Features which are higher in strength of relationships appears first in these filter-based approaches.

For this experimentation, we have used seven filter-based feature selection algorithms using ranker method as intermediate algorithm. The algorithms have different parameters as explained in Appendix C.

Table 4.5 shows the accuracy measurement of the five classification algorithms applied on the selected features. The feature selection criteria are different for each of the FS algorithms. Table 4.5 contains the best percentage and the classifier that gives the best accuracy.

For analyzing the prominent features among all 93 features, we have applied feature selection algorithms. From Table 4.5, we found that the Pearson's Correlation Coefficient gives highest accuracy (72.13%) for extraversion using Naïve Bayes classifier. Also, the accuracy found for other four traits is higher with the PCC-based selected features. This reflection, as PCC based selected features have outperformed the other methods can be visualized from the following fig. 4.7.

Table 4.5: Accuracy measurement of classifiers on selected features applying FS algorithms for five personality traits

Feature Selection Algorithm	Selected Features				
	O	C	E	A	N
PCA	66.84% (SVM)	58.20% (NB)	64.74% (NB)	54.95% (SVM)	59.35% (SLR)
LDA	68.75% (RF)	59.21% (NB)	62.85% (RF)	56.14% (SVM)	60.24% (NB)
CFS	69.67% (RF)	59.02% (RF)	66.80% (RF)	56.56% (NB)	63.16% (NB)
IG	68.27% (SVM)	59.02% (RF)	68.85% (RF)	56.56% (NB)	63.52% (SLR)
SU	67.67% (SVM)	59.02% (RF)	68.03% (RF)	60.25% (SVM)	61.48% (SLR)
CHI	69.67% (SVM)	60.65% (RF)	67.63% (RF)	60.66% (SVM)	63.52% (SLR)
PCC	**70.08%** **(RF)**	**66.39%** **(NB)**	**72.13%** **(NB)**	**61.89%** **(NB)**	**64.75%** **(NB)**

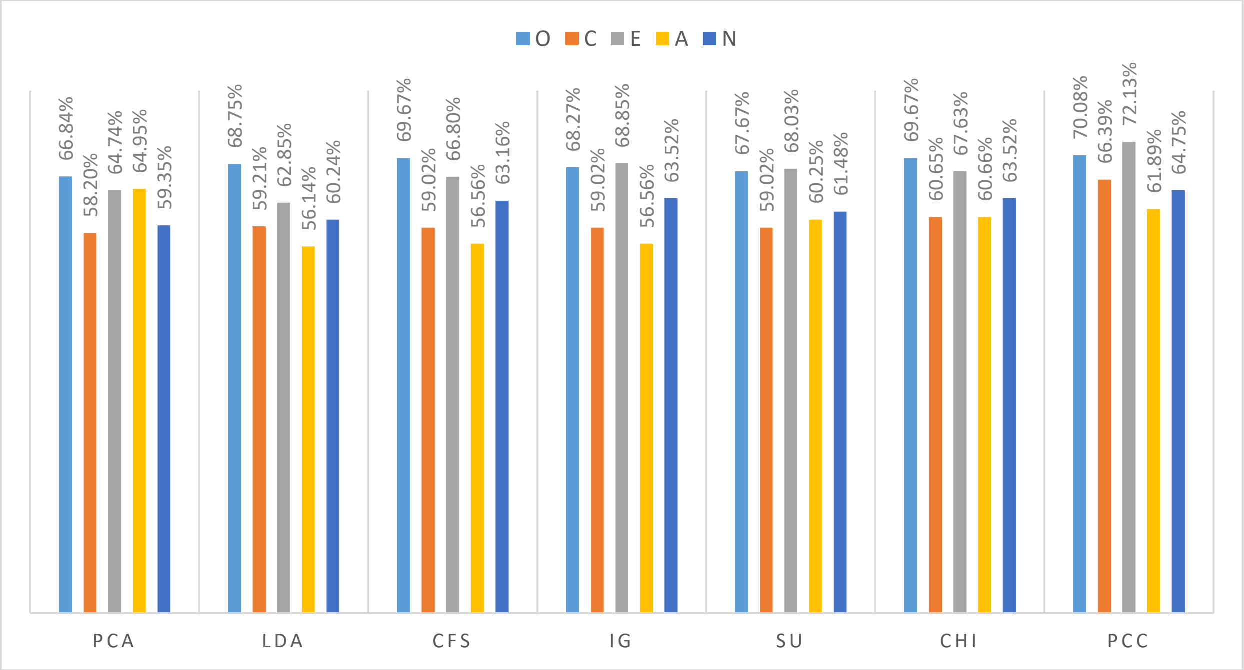


Fig. 4.7. Comparative analysis between automated feature selection algorithms in terms of accuracy.

Table 4.6: Win-Draw-Loss Table of automated feature selection algorithms

FS Methods	PCA	LDA	CFS	IG	SU	CHI	PCC
PCA	--	1-0-4	0-0-5	0-0-5	0-0-5	0-0-5	0-0-5
LDA	4-0-1	--	1-0-4	1-0-4	2-0-3	0-0-5	0-0-5
CFS	5-0-0	4-0-1	--	1-2-2	2-1-2	0-1-4	0-0-5
IG	5-0-5	1-0-4	2-2-1	--	3-1-1	1-1-3	0-0-5
SU	5-0-5	3-0-2	2-1-2	1-1-3	--	1-0-4	0-0-5
CHI	5-0-0	5-0-0	4-1-0	3-1-1	4-0-1	--	0-0-5
PCC	5-0-0	5-0-0	5-0-0	5-0-0	5-0-0	5-0-0	--

For a better understanding of the comparison, we have computed the win-draw-loss as constructed in [100-102] into Table 4.6 describing how many times each method has won against the other methods. Form the data, we can see, PCC has won against all other methods each time and not even drawn with any method. Therefore, we had analyzed the features selected using the PCC method.

The selected features which are determined by applying the Pearson correlation based feature selection method gives very promising insights about the personality traits. Here, we have considered the features in set representation and found interesting combinations for different traits. For each of the traits the sets are named using their initial such as E for extraversion, N for Neuroticism and so on. We have tried to determine the common features from the intersection of these sets and collective features from the union of these sets.

O = {informal, feel, affect, conj, filler, conj, focuspast, swear, allpunc, period}

C = {network-size, betweenness, n-betweenness, density, brokerage, sad, friend, social, feel, you, colon, power, percept, male, family, anx, affiliation, differ, conj}

E = {network-size, betweenness, n-betweenness, density, brokerage, n-brokerage, transitivity, conj, they, I, filler, drives, Authentic, Dash, interrog, reward, body}

A = {parenth, transitivity, conj, we, social, nonflu, they, adverb, n-betweenness, swear, quote, informal, she/he, word-per-sentence(WPS), differ, male}

N = {network-size, betweenness, density, conj, brokerage, transitivity, relig, number, comma, differ, work}

$U = E \cup N \cup A \cup C \cup O$

= *{network-size, betweenness, n-betweenness, density, brokerage, n-brokerage, transitivity, conj, they, I, filler, drives, Authentic, Dash, interrog, reward, body, relig, number, comma, differ, work, sad, friend, social, feel, you, colon, power, percept, male, family, anx, affiliation, differ, informal, feel, affect, conj, filler, focuspast, swear, allpunc, period}*

Common features among the personality traits could be found from the following sets.

$E \cap N$ = *{network-size, betweenness, density, brokerage, transitivity, conj}*

$E \cap A$ = *{n-brokerage, transitivity, conj}*

$E \cap C$ = *{network-size, betweenness, n-betweenness, density, brokerage, conj}*

$E \cap O$ = *{conj}*

From the above mentioned sets, we can depict that the social network features are playing influential role in high accuracy predictions. The seven social network features could be found in each of the trait sets showing the influence, except for Set O.

Table 4.7: Ranking of Features and associated correlation coefficient value for Extraversion

Serial Number	**Coefficient Value**	**Feature Number**	**Serial Number**	**Coefficient Value**	**Feature Number**
1.	0.1042	F56	13.	0.0763	F52
2.	0.0947	F36	14.	0.0757	F32
3.	0.0939	F42	15.	0.0756	F55
4.	0.0909	F41	16.	0.0745	F49
5.	0.0891	F44	17.	0.0737	F34
6.	0.0858	F39	18.	0.0703	F35
7.	0.0824	F37	19.	0.0605	F57
8.	0.0823	F31	20.	0.0598	F45
9.	0.0817	F40	21.	0.0595	F53
10.	0.0792	F38	22.	0.0501	F50
11.	0.0791	F43	23.	0.0309	F54
12.	0.0787	F33	24.	0.0125	F51

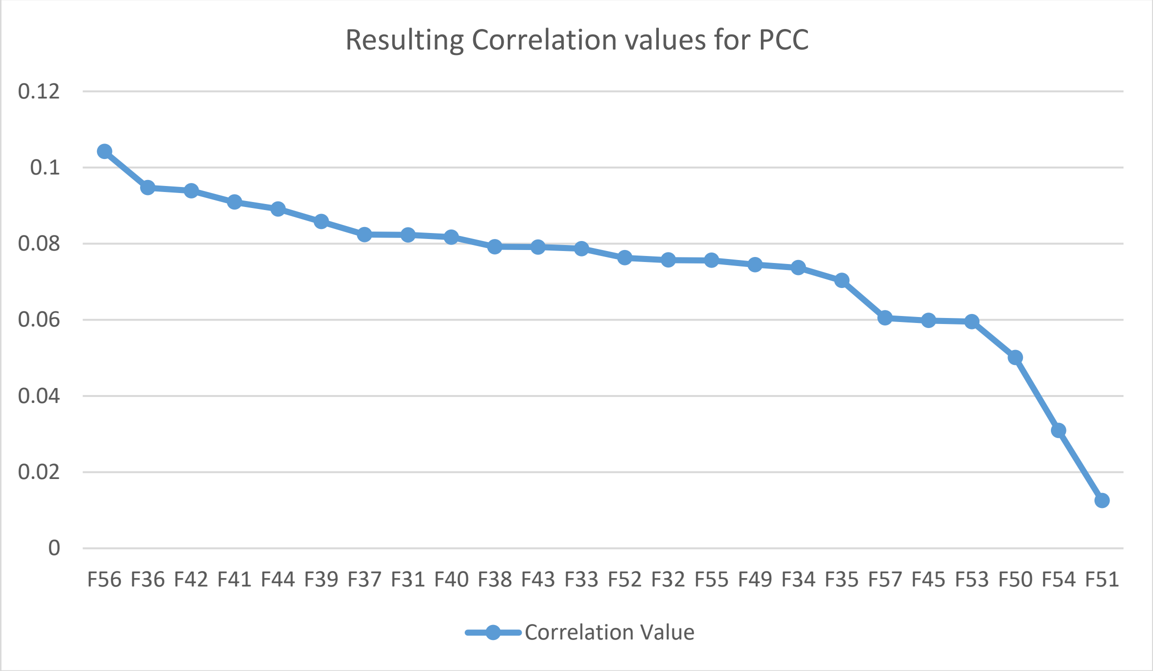


Fig. 4.8. Feature-class correlation values using Pearson's Correlation Coefficient (PCC)

Therefore, openness-to-experience (O) trait have lesser correlation with the social network features. The Universal (U) set represents the set having union of all the sets, which contains 51 distinct features. Apart from the social network features, prominent word-level features and function words are also determined in this thesis. The word-level features and function words are described in chapter 3.1.3. Applying ranker as search method and PCC based feature selection algorithm for the extraversion trait is listed in Table 4.7. The ranking is in order (top to bottom) and the associated coefficient value is plotted in Table 4.7.

As we can see from Fig. 4.8, we have determined the feature-class correlation index for each of the features and only selected the features having ρ > 0.01, where ρ is the correlation value between class and features. Therefore the higher correlated features are selected.

It is notable that the feature F56 is one of the high ranked feature among the 24 features for both the IG and PCC based feature selection. This F56 feature is none other than the percentage of conjunction words. The similar experimental findings are noted for the other four personality traits. Therefore, we can say, the impact of correlation between the word-level features and personality traits are in middle range, but interestingly the same feature of F56 is found to become high ranked feature.

Considering the PCC based feature selection algorithm the ranking of features are determined for each of the traits. Then, we have taken the first five highly correlated features for each of the traits and denote them as set representation. *Corr_E*, *Corr_C*, *Corr_A*, *Corr_O* and *Corr_N* are the first five highly correlated features set of EXT, CON, ARG, OE and NR traits, respectively.

Corr_E = {per_conjunction, no. of words with 5 char, no. of words with 11, no. of words with 10, no. of words more than 12 char}

Corr_C = {per_verb, per_conjunction, no. of words more than 12 char, per_adj, per_adverb}

Corr_A = {per_conjunction, no. of words with 5, no. of words with 10, no. of words with 11, no. of words with 12}

Corr_O = {no. of words with 12, per_preposition, per_conjunction, per_interjection, no. of words more than 12 char}

Corr_N = {per_adjective, per_preposition, per_conjunction, avg. word length, per_noun}

$$Corr\_E \cap Corr\_C \cap Corr\_A \cap Corr\_O \cap Corr\_N = \{per\_conjunction\}$$

From the correlated feature sets it is determined that the common feature which is highly ranked and correlation with all five personality traits is “F56” which is percentage of conjunction words. Therefore, there is a special relationship between the conjunction words and the personality traits.

A list of widely used conjunction words are given in the Table 4.8. These words are also used by the Facebook users while expressing their critical thoughts via status updates.

As we found this high correlation between Percentage of conjunction words and personality traits, this relationship is not trivial to find out. The reason behind the correlation could be tested in different application area. A similar work [99] has been performed on the flexibility in writing style and physical health. The writing sample of 124 students and 59 maximum security prisoners were taken into account and Latent Semantic Analysis (LSA) has been applied. They found that personal pronouns are mostly used while writing traumatic memories which were related to positive health outcomes. The secret relation of conjunctions could be declared with the Big Five Personality traits.

Table 4.8: Widely used Conjunction words list

Coordinating Conjunctions	Subordinating Conjunctions			
	Concession	**Condition**	**Comparison**	**Reason**
For	Though	If	than	Because
And	Although	Only if	Rather than	Since
Nor	Even though	Unless	Whether	So that
But	While	Until	As much as	In Order
Or		Provided that	Whereas	Why
Yet		Even if		
So		In case		

Experimental Findings from Experiment-2:

In experiment-2, we have applied seven different automated feature selection algorithm and try to investigate the insight of different features. From the experiment we have found some interesting findings which can be stated as below:

- Social Network Features are the most prominent features as they are highly correlated to the personality traits.
- Among the psycholinguistic features all the "social relationship" features are found in the universal set except no. of female related words.
- Percentage of Conjunction words have high correlation with the all the personality traits.
- The openness-to-experience trait has shown divert results and selected features set does not contain any SN features.
- PCC outperforms the other existing feature selection algorithms for predicting personality.

4.6 Comparative Analysis

In this section, we have summarized the comparative analysis between the experiments performed for our thesis and the comparison with the state-of-the-art methods.

The comparatively higher accuracy demonstrated in experiment-2 than experiment-1 are tabulated in Table 4.9. As experiment-1 consists 16 cases, in the following table only the highest accuracy obtained comparing all the case scenarios are reported. For experiment-2, as we have applied seven automated features selection algorithms and reported that PCC outperforms other feature selection algorithms, we have tabulated the accuracy percentages of PCC in the following table.

Table 4.9: Accuracy measurements of classifiers along with the traits for each experiments

Personality Traits	Experiment-1	Experiment-2
Openness-to-experience (O)	69.67% (SVM)	70.08% (RF)
Conscientiousness (C)	61.07% (LR)	66.39% (NB)
Extraversion (E)	68.85% (RF)	**72.13%** **(NB)**
Agreeableness (A)	59.02% (RF)	61.89% (NB)
Neuroticism (N)	63.39% (RF)	64.75% (NB)

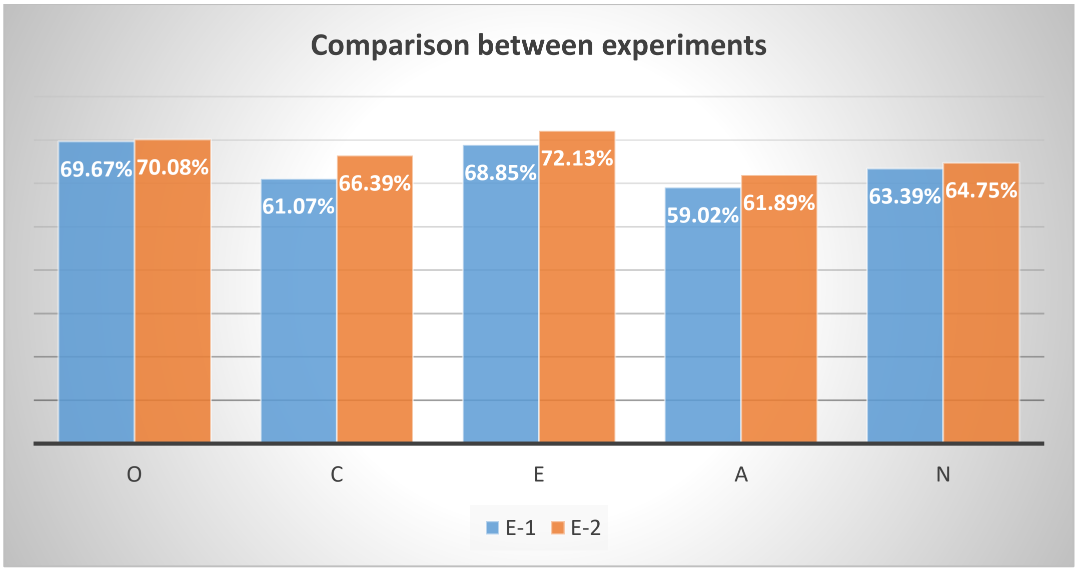


Fig. 4.9. Comparison between the experiments and accuracy measurements.

From fig 4.9, we can see, each time experiment-2 has outperformed experiment-1, which reflects that automated feature selection is better than manual feature selection for predicting personality traits.

Comparative Analysis with the State-of-the-art Methods:

Table 4.10: Comparison with state-of-the-art methods

Personality Traits	**Farnandi et al. [9]**	**Tendra et al. [11] LIWC & SPLICE**	**Tendra et al.[11] Deep learning**	***Our Approach***
Openness-to-experience (O)	61.00% (SVM)	63.20% (GB)	70.78% (MLP)	70.08% (RF)
Conscientiousness (C)	54.00% (k-NN)	59.20% (NB)	59.13% (CNN-1D)	66.39% (NB)
Extraversion (E)	62.00% (SVM)	68.80% (SVM)	65.39% (MLP)	**72.13%** **(NB)**
Agreeableness (A)	53.00% (k-NN)	57.60% (GB)	59.13% (GRU)	61.89% (NB)
Neuroticism (N)	56.00% (k-NN)	60.40% (LR)	64.52% (GRU)	64.75% (NB)

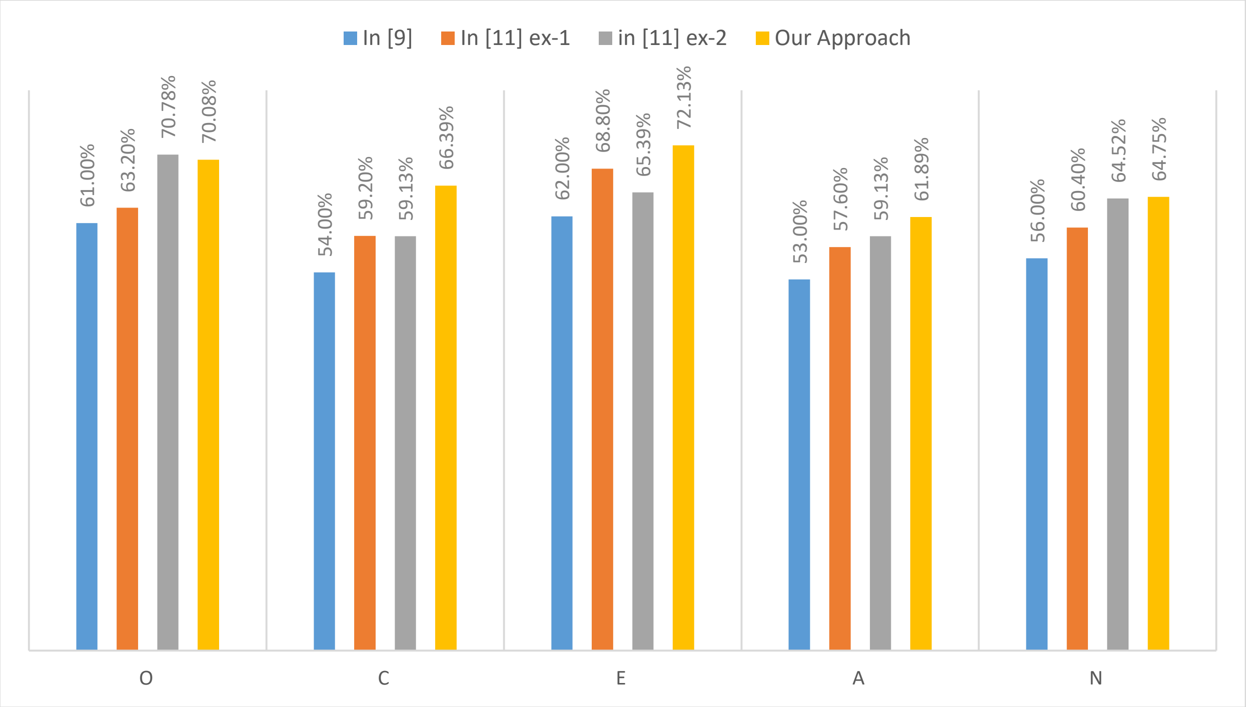


Fig. 4. 10. Comparison with the state-of-the-art systems.

Table 4.10 and fig 4.10 shows the comparison between or proposed approach and the approaches discussed in background study (chapter-2). It is noticeable that our approach is outperforming the other existing approaches by a good margin. In comparison with Tendra [11], they have applied deep learning based approach which shows better accuracy in only one case (for openness trait). Therefore, applying feature selection algorithms have determined the most prominent features to be used and using those features the overall performance has increase, in terms of accuracy.

4.7 Discussion

Theoretical Implications of Predicting Personality Traits

The idea of using social media contents for analysis of different research problem is relatively new idea. With this thesis, we have promoted the idea of using the Facebook status updates for identifying the most relevant features.

The theoretical advancements done in the area of psychology for predicting human personality accurately is quite satisfactory. But for understanding ones personality the amount of data needed was quite large. Though the psychology researchers have worked on it for long time to optimize the data needed to understand ones personality, they could only think about taking questionnaire from the individuals. Ethical questions may arise, as the personality tester must be honest while answering each questions. Because the IPIP items used for questionnaire are quite easy to understand and tester may give bias answers for some good questions. As each of the questions have impact on the final result of the whole system, the prediction will not be as accurate as wanted.

For mitigating this problem, we asked the literature, if it is possible to understand and predict individuals' personality without asking a set of questionnaire. Afterward, we have designed experiments and depicted the outcome in experimental analysis.

One of the main theoretical implication of predicting personality traits is to find the impact of linguistic features for the system. For our methodology, we have kept separate feature extraction step, where we have extracted the linguistic features of two different category; traditional linguistic features and psycholinguistic features. The linguistic features could be considered as stylometric features also. Stylometry is the study of analyzing the writing style of individuals. For analyzing this particular style researchers have worked on the writing pattern, writing capabilities in terms of time to write specific number of words, handwriting recognition etc. The main way to study human writing style is to identify the boundary of vocabulary used by that person. Thus the number of words, number of character and such features may become useful to analyze the problem.

The traditional linguistic features considered for this thesis could be divided into several parts: character-based features, word-based features, structural features and function words/parts-of-speech based features. These features could be considered as stylometric features as well, as in literature these features are used for authorship attribution, style identification etc. works.

The psycholinguistic features from LIWC have been utilized for many work such as email spam detection, similar author identification etc. Thus using this tool for predicting personality from social media contents is relatively new in the literature. We have adopted this trend and used LIWC features, especially psycholinguistic cues such as emotional effect, cognitive process, social relationships, self-focus and perception features. The word counts for each of the psychological cues will help to understand the theoretical vocabulary boundary of the person. From these numerical features, the supervised learning model has been developed.

With the help of social network features such as network size, betweenness centrality etc. the interaction performed by a single node of social entity in social media could be understand. A social media user is nothing but considered as one of the social entity/ node of a huge graph. The edges are the connections or friends in social media and the weight of the edges could be considered as the interaction strengths [3]. Based on these evidence, we have experimented using the social network features.

Practical Application of Proposed Approach

Practical implications of computational personality traits prediction is quantitatively higher. The proposed system could be utilized in many real-life problem scenarios and for different types of practical research problems, as well. In this sub-section, the practical implications of predicting personality traits are discussed.

- *Personalized recommendation systems*: Personalized recommendation systems in SNS such as, friend recommendation [4], community recommendation [3], community detection [5] etc. could be developed utilizing the proposed prediction system. In this era of business intelligence, data mining and machine learning, people don’t want to see unwanted products in recommendation while shopping online, or unknown faces in the “people you may know” section of the Facebook. For building personalized solutions for above mentioned different types of recommendation systems, the proposed personality prediction systems could be useful.
- *Predicting personality disorder*: If we can understand individuals’ personality traits of different time throughout the life, we can understand the trend of changes or modifications in places. If we can develop a system which can predict the order of changes throughout the whole life of similar personality peoples, we may come up with

the idea of predicting the personality disorders [91, 92]. This could be very useful for early identifying mental disorders before time.

- *Finding specific personality type SNS users*: for specific recommendations such as real-life events similar personality type SNS users may group together. Even for identifying the future life partners' personality, the proposed personality prediction systems may become useful.
- *Identifying misbehavior*: Identifying misbehavior from a users' regular behavior according to his/her personality [93]. Very much similar as identifying personality disorder, the sequence or behavior of an individual could be detected using the digital footprints quite easily. Identifying the noisy data may detect the misbehavior in social media. This has become a very buzz topic in the area of social media mining recently.
- *Identifying trust issues in SNS*: Trust has been redefined in perspective of social media and digital platforms [94-95]. Personality trait scores could be utilized to identify the trust issues and concepts from social networking sites. Individuals using such networks to connect to their friends and families, governments and enterprises have started exploiting these platforms for delivering their services to citizens and customers. However, the success of such attempts relies on the level of trust that members have with each other as well as with the service provider. Therefore, trust becomes an essential and important element of a successful social network. Before judging the trust issues from social media, it is importance to understand and predict individuals' personality. Personality traits may provide necessary information nuggets to develop a model over it.
- *Detecting character assassination*: Detecting character assassination [96] through detection of troll comments and users personality could be done. Trolls are usually the SNS users who utilizes the facilities of SNSs and posts bad comments, accuses someone online with false information, starts rumors or controversies, spreads fake news or influencing people to viral posts etc. Each of the above mentioned actions could be considered as character assassination. It is common scenario that actresses of renowned media background are accused falsely. The celebrities sometimes try to block or ban those commenters manually or ignore them. Commenting or sharing false news about someone could be considered as direct character assassination which is punishable

crime. Identifying these troll comments may lead us to identify character assassination from social media. By detecting the character assassination, we may recommend a list of commenters who must be banned to specific user accounts.

Therefore, the practical application areas mentioned above shows the practical implications of this thesis.

CHAPTER 5

CONCLUSION

In the context of social media, understanding the personality of user is a significant task, as the online behavior of a person is quite different than his/her real-life personality. In social media platform, people are more talkative and expressive, while in real-life they don't want to share any views on the political agenda or movements. Hence, the idea of using "personality as a feature" in e-commerce websites is getting popular these days. The e-commerce websites are giving facilities of personalized feelings to its users by recommending items closely related to the customer choice.

5.1 Summary of the Study

In this thesis, we have tried to develop a predictive system which will automatically predict personality through positive & negative traits of SNS user, exploring the prominent linguistic and social network features. Moreover, we have proposed a methodology to predict personality traits more accurately (in terms of accuracy). The overall accuracy of personality traits in the literature were below seventy percent, where our proposed methodology outperforms the literature works and gives 72.13% accuracy for extraversion trait. The overall accuracy improved because of applying feature selection algorithm after feature extraction process. Therefore, we can declare the following statements from our research findings.

Social Network Features are the most prominent features as they are highly correlated to the personality traits. Among the psycholinguistic features all the "social relationship" features are found in the universal set except no. of female related words. The influence of punctuations are at a decent level, as Dash, comma, parenth, quotes, colon, period, allpunc features are present in the universal set. The important function words (personal pronoun, interrogative, adverb and conjunction) are present in the U set and have good correlation with the relative classes.

5.2 Implication of Future Works

For future scopes with this area of research, there are many opportunities and challenges. In this section, we have highlighted some of them.

- *Personality Prediction from Bangla Status:* Though there are many works found in literature related to psycholinguistic features such as LIWC for English written status updates. But, there is no work present in Bangla. As Bangla language has a complex structure, the identification of proper words is a challenging task. But, similar methods as proposed in this thesis could be adopted for Bangla Status Updates also.
- *Applying Modified Feature Selection Algorithms:* For more accuracy improvement, any modified feature selection algorithms could be applied on the dataset. Nature-inspired evolutionary algorithms could be applied to devise feature selection criteria.
- *Applying Different Types of Features:* For our work, we have focused on linguistic and social network features. We can use different set of features such as stylometric features or we have apply deep learning algorithms directly to find the hidden correlation between the features and outcome better accuracy.

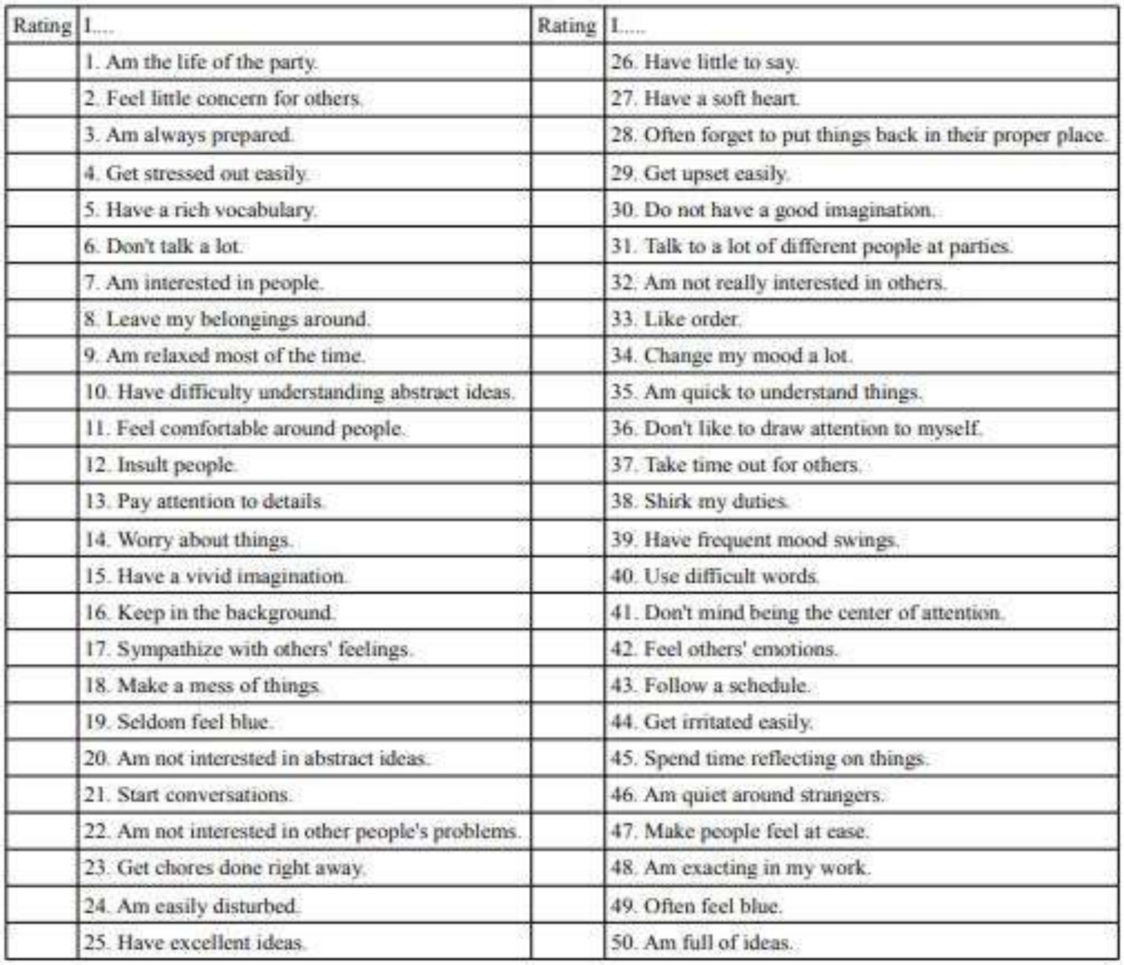
APPENDICES

Appendix A

International Personality Item Pool

The International Personality Item Pool (IPIP) is a public domain collection of items for use in personality tests. It is managed by the Oregon Research Institute. The pool contains 3,329 items. These items make up more than 250 inventories that measure a variety of personality factors, many of which correlate well to better-known systems such as the 16PF Questionnaire and the Big Five personality traits. IPIP provides journal citations to trace those inventories back to the publication as well as correlation tables between questions of the same factor and between results from different inventories for comparison. Scoring keys that mention the items used for a test are given in a list form; they can be formatted into questionnaires.

Example of some IPIP:



Rating	I....	Rating	I.....
	1. Am the life of the party.		26. Have little to say.
	2. Feel little concern for others.		27. Have a soft heart.
	3. Am always prepared.		28. Often forget to put things back in their proper place.
	4. Get stressed out easily.		29. Get upset easily.
	5. Have a rich vocabulary.		30. Do not have a good imagination.
	6. Don't talk a lot.		31. Talk to a lot of different people at parties.
	7. Am interested in people.		32. Am not really interested in others.
	8. Leave my belongings around.		33. Like order.
	9. Am relaxed most of the time.		34. Change my mood a lot.
	10. Have difficulty understanding abstract ideas.		35. Am quick to understand things.
	11. Feel comfortable around people.		36. Don't like to draw attention to myself.
	12. Insult people.		37. Take time out for others.
	13. Pay attention to details.		38. Shirk my duties.
	14. Worry about things.		39. Have frequent mood swings.
	15. Have a vivid imagination.		40. Use difficult words.
	16. Keep in the background.		41. Don't mind being the center of attention.
	17. Sympathize with others' feelings.		42. Feel others' emotions.
	18. Make a mess of things.		43. Follow a schedule.
	19. Seldom feel blue.		44. Get irritated easily.
	20. Am not interested in abstract ideas.		45. Spend time reflecting on things.
	21. Start conversations.		46. Am quiet around strangers.
	22. Am not interested in other people's problems.		47. Make people feel at ease.
	23. Get chores done right away.		48. Am exacting in my work.
	24. Am easily disturbed.		49. Often feel blue.
	25. Have excellent ideas.		50. Am full of ideas.

Appendix B

Social Network Features

Network Size defines the number of friends, connections or followers in social networking sites. Using this feature we may predict if the user has decent number of friends or not. Having smaller number of friends may lead to a characteristics of introvert user and vice versa.

NS(v)=Total no.of edges of V

Betweenness centrality (g(v)) of a node v in a given graph could be determined using equation 9. Centrality is the measure to determine the central nodes within a graph, whereas betweenness centrality demonstrates how many times a node behaved as a connector along the shortest path between two other nodes. This measure is useful in assessing which nodes are central with respect to spreading information and influencing others in their immediate neighborhood.

$$g(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

Normalized Betweenness centrality is the normalized value of g(v) with respect to the minimum and maximum values of g. The formula to determine the n-betweenness is equation (10).

$$normal\big(g(v)\big) = \frac{g(v) - \min(g)}{\max(g) - \min(g)}$$

Density is the measure of network connections. Network density could be measured using the formula (11). Density demonstrates the potential connections in a network that are actual connections.

$$Density = \frac{Actual\ Connections}{Maximum\ Possible\ Connections}$$

Brokerage refers to the nodes embedded in its neighborhood which is very useful in understanding power, influence and dependency effects on graphs. A broker could be considered as the communicator between two different nodes.

Five types of brokering is available in literature namely, coordinator, consultant, gatekeeper, representative and liaison. It is possible that different types of broker is present in a simple social network graph. The general concept of brokerage could be depicted as Fig. A.B.1.

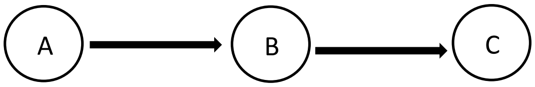


Fig. A.B.1. Broker B between A and C.

In a graph, if A is connected to B and B is connected to C, but A and C are not connected to each other, then A needs B to communicate with C. Thus, B is the broker node here.

Table A. B. 1: Different categories of Broker Nodes

Type of Broker	**Equation to calculate brokerage**	**Conditions/ Criteria**
Coordinator	Counts the no. of times B is a broker and G(A) = G(B) = G(C)	All three nodes belong to the same group
Consultant	Counts the no. of times B is a broker and G(A) = G(C), but G(B) ≠ G(A)	The broker belongs to one group and the other belong to a different group
Gatekeeper	Counts the no. of times B is a broker and G(A) ≠ G(B) and G(B) = G(C)	The source node belongs to a different group
Representative	Counts the no. of times B is a broker and G(A) = G(B) and G(C) ≠ G(B)	The destination node belongs to a different group
Liaison	Counts the no. of times B is a broker and G(A) ≠ G(B) ≠ G(C)	Each node belongs to a different group

The description of five different types of broker nodes are illustrated in Table 3.5. The equations used in the Table IV are considering node B as broker and G(X) denotes the group that node x belongs to. It is presumed that A→ B→ C, thus A be the source node gives information to B, the broker node, who gives the information to C (the destination node).

N-brokerage is the normalized parameter of brokerage which is the measure of brokerage nodes divided by the number of pairs. The equation could be derived as below.

$$n - brokerage = \frac{no.\ of\ broker\ nodes}{no.\ of\ pairs}$$

Transitivity is the measurement which could be defined as FOF (Friend-of-Friend) concept of social media such as Facebook. The idea of FOF is “when a friend of my friend is my friend”. In the context of network or graph theory, transitivity is measured based on the relative number of triangles or triads present in the graph comparing to the total number of connected triples of nodes.

The idea of transitivity is depicted in Fig. A.B.2 and the equation to calculate the transitivity (T(G)) is (13), as follow.

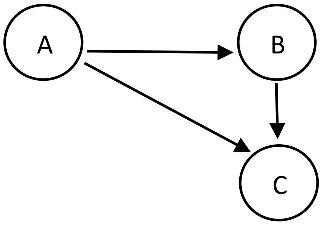


Fig. A. B. 2. Idea of transitivity

As shown in Fig. 3.4, A is friend with B, B is friend with C and A is also friend with C. The relationships between them builds a triad.

$$T(G) = \frac{3 * no.\ of\ closed\ triples\ in\ G}{no.\ of\ connected\ triples\ of\ vertices\ in\ G}$$

Appendix C

Feature Selection Algorithms Overview

Principal component analysis (PCA) is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables (entities each of which takes on various numerical values) into a set of values of linearly uncorrelated variables called principal components [96, 97]. PCA contributes to reduce the dimension by stepping to the following steps: Consider the whole dataset consisting of *d+1* dimensions and ignore the labels such that the new dataset becomes *d* dimensions; compute the mean and covariance matrix of the whole dataset; compute eigenvectors and the corresponding eigenvalues; sort the eigenvectors by decreasing eigenvalues and choose k eigenvectors with the largest eigenvalues to form *d x k* dimensional matrix *W* and use this *W* matrix to transform the samples onto the new subspace.

Linear discriminant analysis (LDA) [98], or discriminant function analysis is a generalization of Fisher's linear discriminant, a method used in statistics, pattern recognition and machine learning to find a linear combination of features that characterizes or separates two or more classes of objects or events. The resulting combination may be used as a linear classifier, or, more commonly, for dimensionality reduction before later classification. LDA is also closely related to principal component analysis (PCA) and factor analysis in that they both look for linear combinations of variables which best explain the data. LDA explicitly attempts to model the difference between the classes of data. PCA on the other hand does not take into account any difference in class, and factor analysis builds the feature combinations based on differences rather than similarities. Discriminant analysis is also different from factor analysis in that it is not an interdependence technique: a distinction between independent variables and dependent variables (also called criterion variables) must be made.

Correlation-based Feature Selection (CFS) subset evaluator [60, 61] is a feature selection algorithm which finds the subset of features via individual predictive ability of each feature along with the degree of redundancy between them. CFS ranks the features subsets according to a correlation based heuristic evaluation method. The subset evaluation function is given in equation (1), where M_S is the heuristic merit of the feature subset S containing k features. $\overline{r_{cf}}$ is the mean of feature-class correlation ($f \in S$) and $\overline{r_{ff}}$ is the average feature-feature inter-correlation.

$$M_s = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}} \qquad (1)$$

CFS subset evaluator is used in different contexts of research such as to predict students' performance [62], selecting features for sentiment classification [63].

Information gain is one of the widely used feature selection method in different research problems including text categorization [64, 65]. Various research fields have utilized the inner mechanism of information gain such as computer vision, text classification [65, 66] etc. Information gain outcomes a ratio value calculated by equation (2), where, *values(a)* denotes the set of all possible values of features $a \in Attr$. $Attr$ is the set of all features, H be the entropy and $x \in T$ denotes the value of specific example *x* for $a \in Attr$. The largest information gain be the smallest entropy.

$$IG(T,a) = H(T) - \sum_{v \in vals(a)} \frac{|\{x \in T | x_a = v\}|}{|T|} . H(x \in T | x_a = v) \qquad (2)$$

In the context of statistics, uncertainty coefficient or entropy coefficient is the measure of nominal association. The ***symmetrical uncertainty (SU)*** [67] Attribute Evaluator is one kind of correlation finder which evaluates the importance of a feature by measuring the *SU* with respect to the class. This feature selection process is not only used for imagery data such as hyperspectral images [68], but also used with the nature-inspired optimization algorithms such as ant colony optimization [69]. The *SU* is determined using equation (3) where, *H(C|F)* be the conditional entropy considering C the Class and F the Feature & *H(C)* be the single distribution of class C. The algorithm outcomes ranking of the most relevant features.

$$SU(C,F) = \frac{2 * (H(C) - H(C|F))}{H(C) + H(F)} \qquad (3)$$

$$H(C) = -\sum_{x} P_C(x) log P_C(x) \qquad (4)$$

$$H(C|F) = -\sum_{x,y} P_{C,F}(x,y) log\, P_{C,F}(x,y) \tag{5}$$

Chi-squared test ($\chi 2$) [70, 71] is used in statistics for determining the association between variables or features. Depending on the difference between the expected frequencies (*e*) and the observed frequencies (*n*) in one or more features in the feature set, the chi-squared value is determined. Depending on the value of the parameter, we can decide the number of features to be selected for a system. The equation for calculating chi-squared value is in equation (6) where, r and c be the number of row and column of the feature table.

$$\chi 2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(n_{ij} - e_{ij})^2}{e_{ij}} \tag{6}$$

Pearson Correlation Coefficient (PCC) [72] is considered as one of the most efficient and widely accepted feature selection algorithm. In PCC, the value of covariance between the class and feature is been determined. The standard deviations of the class & feature are calculated to find the coefficient value (*ρ)*. The coefficient could be used as an efficient parameter to determine the feature sets. The calculation of (*ρ*) is performed using equation (7) as given below. *cov(X, Y)* is the covariance between X, Y where X or Y be the class value and *σ* is the standard deviation in equation (7).

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} \tag{7}$$

The correlation value is distributed between -1 and +1, where 1 is total positive correlation, O is no linear correlation and -1 means negative correlation. The coefficient is invariant under separate modifications in scale and location in the two variables, which could be considered as a key mathematical property of PCC. Therefore, the PCC has been used in diversified research problem for the same purpose of feature selection. In [73] K. J. Kim et al. presented a correlation analysis for DNA microarray datasets such as Leukemia, Colon and Lymphoma. They utilized the ensemble classifiers to get the highest accuracy on each of the datasets. PCC has been utilized in image processing such as tissue classification from CT images [74]. The implication of PCC for noise removal in the context of signal processing is presented in [75]. They provide experimental justification of using PCC on signal data. The statistical perspective of using PCC has been presented in [75] which focuses on the medical research domain. A practical application of PCC has been demonstrated in [75] using the sample data of 780 women attending their first antenatal clinic visits. In the context of natural language processing (NLP), PCC has proven to work better for many applications such as neurolinguistic approach to NLP using medical text analysis [76], automated classification of radiology reports for acute lung injury using machine learning and NLP [77], finding strong correlation between text quality and complex network features [78], and automated plagiarism detection using NLP [79].

Appendix D

Classification Algorithms Applied

Naïve Bayes (NB) classifier relies on the Bayes' theorem that is a straightforward to develop. NB has no sophisticated unvarying parameter estimation that makes it significantly essential for large-scale datasets. NB classifier assume that the impact of the value of a predictor (x) on a given class (c) is not dependent on the values of alternative predictors [81].

As the name suggests, for decision making ***decision tree (DT)*** is used in the context of data mining classifications. For organizing numerical and categorical data and performing classification on large datasets, decision tree has been proven beneficial as it is possible to validate a training model using statistical tests [82]. In this paper, we have applied the variation of decision tree namely, J48, which is built-in in the Weka.

Decision tree based ensemble classifier ***random forest (RF)*** [83-84] is applied in our work. RF is intrinsically suited for multiclass problems which also works well with a mixture of numerical and categorical features. RF internally builds separate multitude decision trees while training, therefore, it outcomes the class that is the mode of the classes and/or mean regression of each trees. Therefore, it is proven to be one of the better ensemble classifier.

Support vector machine (SVM) has become one popular yet essential classifier used in various sections of data mining such as, medical data mining, image data processing, bioinformatics etc. Though it is proven to work better in many cases, the training process is slower in some context. Therefore a fast algorithm for training the SVM is introduced in [85] namely sequential minimal optimization (SMO). Applying quadratic programming (QP) optimization problem by breaking the problem into sub-problems SMO minimizes the time required to train the model. For our work, we have used Weka version of SMO having poly kernel.

Logistic regression (LR) [86] is a linear logistic model using LogitBoost algorithm. As LogitBoost uses a symmetric model, a sufficient number of iteration is performed in simple logistic regression to train the model. Built-in attribute selection is performed in SLR as an additional advantage. Therefore, for our experiments, we have applied this classifier. We have used 10-fold cross validation to train and test the proposed model for each of the classifiers.

Using these traditional classifiers, we try to find the best working classifier for identifying neuroticism from the problem space.

Bibliography

[1] Top 15 Valuable Facebook Statistics, https://zephoria.com/top-15-valuable-facebook-statistics [Last accessed on 10th April, 2019 at 12:00 PM]

[2] L. R Goldberg,., "The structure of phenotypic personality traits", Journal of American Psychologist, issue 48, pp. 26–34, 1994.

[3] A. A. Marouf, R. Ajwad, M. T. R. Kyser, "Community Recommendation Approach for Social Networking Sites based on Mining Rules", 2nd International Conference on Electrical Engineering and Information & Communication Technology (iCEEiCT 2015), 21-23 May, 2015. 543.

[4] M. M. Hasan, N. H. Shaon, A. A. Marouf, M. K. Hasan, H. Mahmud, and Md. Mohiuddin Khan, "Friend Recommendation Framework for Social Networking Sites using User‘s Online Behavior", IEEE- Computer and Information Technology (ICCIT), December 2015, pp. 539-543.

[5] N. Du, B. Wu, X. Pei, B. Wang, and L. Xu, ‘‘Community detection in large-scale social networks,’’ in Proc. 9th WebKDD 1st SNA-KDD Workshop Web Mining Social Network Analysis, 2007, pp. 16–25.

[6] S. Adal and J. Golbeck., "Predicting personality with social behavior", Proceedings of IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 2012.

[7] F. Alam, E. A. Stepanov, G. Riccardi, "Personality Traits Recognition on Social Network – Facebook", 7th International AAAI Conference on Weblogs and Social Media Workshop on Computational Personality Recognition (Shared Task), pp. 6-9.

[8] HA Schwartz, JC Eichstaedt, ML Kern, L Dziurzynski, SM Ramones, M Agrawal, "Personality, Gender, and Age in the Language of Social Media" The Open-Vocabulary Approach. PLoS ONE 8(9): e73791, 2013.

[9] S Poria, A Gelbukh, B Agarwal, E Cambria, N Howard "Common Sense Knowledge Based Personality Recognition from Text". In 12th Mexican International Conference on Artificial Intelligence, Vol. 8266, 2013, pp. 484-496, 2013.

[10] D Quercia, M Kosinski, D Stillwell, J Crowcroft, "Our twitter profiles, our selves: Predicting personality with twitter". In Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third International Conference on Social Computing (SocialCom), pp. 180–185. 2011.

[11] J Corr, Philip.; G Matthews, "The Cambridge handbook of personality psychology" (1. publ. ed.). Cambridge: Cambridge University Press. ISBN 978-0-521-86218-9. (2009).

[12] . Cherry, “What is Personality and Why it matters?”, [Online] < https://www.verywellmind.com/what-is-personality-2795416>

[13] M. S. H. Mukta, M. E. Ali and J. Mahmud. "User Generated vs. Supported Contents: Which One Can Better Predict Basic Human Values?." International Conference on Social Informatics. Springer International Publishing, 2016.

[14] C. P. Williams, "Language, Identity, Culture, and Diversity", [Online] <https://www.newamerica.org/education-policy/edcentral/ multilingualismmatters/>, February 23, 2013.

[15] M. Kosinski, S. Matz, S. Gosling, V. Popov and D. Stillwell, “Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines.”, American Psychologist. vol. 70 issue. 6, pp. 543, February 2015.

[16] S. Friedman, Howard, B. K., Stephanie "Personality, Type a behavior, and coronary heart disease: The role of emotional expression". Journal of Personality and Social Psychology. 53 (4): 783–792. doi:10.1037/0022-3514.53.4.783, 1987.

[17] H. J. Eysenck, "Type A Behavior and Coronary Heart Disease: The Third Stage". Journal of Social Behavior and Personality. 5: 25–44. 1990.

[18] J. L. Holland, "Award for distinguished scientific applications of psychology:." American Psychologist, Vol 63(8), Nov 2008, 672-674.

[19] I. Briggs with P. B. Myers (1995) “Gifts Differing: Understanding Personality Type.” Mountain View, CA: Davies-Black Publishing. 1980

[20] MBTI basics, The Myers-Briggs Foundation, 2014, Retrieved 18 June 2014.

[21] Myers-Briggs Type Indicator (MBTI), CPP.com, Menlo Park, CA, 2014, Retrieved 18 June 2014.

[22] S Rothmann, EP Coetzer, "The big five personality dimensions and job performance". SA Journal of Industrial Psychology. 29. doi:10.4102/sajip.v29i1.88 (24 October 2003).

[23] Five Factor Model, https://relivingmbadays.wordpress.com/2012/09/15/five-factor-model-of-personality/

[24] The Big Five traits, https://natashafelderpsych220com.wordpress.com/ 2016/05/26/the-big-5/

[25] International Personality Item Pool, [Online] Available at https://ipip.ori.org/

[26] I.B Myers, M.H. McCaulley, N.L. Quenk, A.L Hammer, “MBTI Manual: A Guide to the Development and Use of the Myers-Briggs Type Indicator”, Third Edition. Consulting Psychologists, Palo Alto, CA. 1998

[27] W. Youyou, M. Kosinski and D. Stillwell, "Computer-based personality judgments are more accurate than those made by humans" Proceedings of the National Academy of Sciences (PNAS), 2015.

[28] P. T. Jr. Costa and R. R. McCrae, “NEO-PI-R professional manual. Odessa, FL: Psychological Assessment Resources, 1992.

[29] L. R. Goldberg, “A broad-bandwidth, public-domain, personality inventory measuring the lower-level facets of several five-factor models” In I. Mervielde, I. Deary, F. De Fruyt, & F. Ostendorf (Eds.), Personality psychology in Europe, Vol. 7 (pp. 7-28). Tilburg, The Netherlands: Tilburg University Press. http://ipip.ori.org/newBroadbandText.htm, 1999.

[30] O. P. John and S. Srivastava, “The Big Five trait taxonomy: History, measurement, and theoretical perspectives” In L. A. Pervin & O. P. John (Eds.), Handbook of personality: Theory and research (2nd ed., pp. 102–138). New York: Guilford Press, 1999.

[31] G. Saucier, “Mini-markers: A brief version of Goldberg’s unipolar Big-Five markers” Journal of Personality Assessment, vol. 63, pp. 506–516, 1994.

[32] M. B. Donnellan, F. L. Oswald, B. M. Baird, R. E. Lucas, “The mini-IPIP scales: tiny-yet effective measures of the Big Five factors of personality” Psychology Assessment, June 18, 2006, vol. 2, pp. 192-203. PubMed PMID: 16768595

[33] S. D.Gosling, P. J.Rentfrow, W. B. Swann, “A very brief measure of the Big-Five personality domains. Journal of Research in Personality, 37,504–528, 2003.

[34] J. A. Johnson, “Developing a short form of the IPIP-NEO: A report to HGW Consulting”. Unpublished manuscript. Department of Psychology University of Pennsylvania, DuBois, PA, 2000.

[35] M. Kosinski, S. Matz, S. Gosling, V. Popov and D. Stillwell, “Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines.”, American Psychologist, vol. 70, issue. 6, pp. 543, February 2015.

[36] M. Kosinski, D. Stillwell and T. Graepel, “Private traits and attributes are predictable from digital records of human behavior.”, In Proceedings of the National Academy of Sciences of the United States of America (PNAS), pp. 5802-5805, 2013.

[37] M.D. Back, J.M. Stopfer, S. Vazire, S. Gaddis, S.C. Schmukle, B. Egloff, Gosling, S.D., ”Facebook profiles reflect actual personality, not self-idealization”, Psychological Science 21, 372–374 (2010).

[38] G Farnadi, S. Zoghbi, Moens, M., De Cock, M.: Recognising personality traits using Facebook status updates. In: Proceedings of the WCPR, pp. 14–18 (2013).

[39] Golbeck, J., Robles, C., Turner, K.: Predicting personality with social media. In: CHI'11 Extended Abstracts on Human Factors in Computing Systems, pp. 253–262. ACM (2011)

[40] D.J Stillwell,., M Kosinski,."myPersonality Project Website. myPersonality Project" (2015). [URL] http://mypersonality.org

[41] Goldberg, L.R., Johnson, J.A., Eber, H.W., Hogan, R., Ashton, M.C., Cloninger, C.R., Gough, H.G.: The international personality item pool and the future of public-domain personality measures. Journal of Research in Personality 40(1), 84–96 (2006)

[42] Rammstedt, B., John, O.P.: Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. Journal of research in Personality 41(1), 203–212 (2007).

[43] Biel, J., Gatica-Perez, D.: The YouTube lens: Crowdsourced personality impressions and audiovisual analysis of vlogs. Multimedia, IEEE Transactions on 15(1), 41–55 (2013).

[44] Biel, J.I., Aran, O., Gatica-Perez, D.: You are known by how you vlog: Personality impressions and nonverbal behavior in youtube. In: Proceedings of the AAAI International Conference on Weblogs and Social Media (ICWSM), pp. 446–449 (2011).

[45] Y. Bachrach, M. Kosinski, T. Graepel, P. Kohli and D. Stillwell, "Personality and Patterns of Facebook Usage" Proceedings of the 4th Annual ACM Web Science Conference (WebSci'12), pp. 24-32, June 22-24, 2012, Illinois, USA.

[46] J. Golbeck, C. Robles and K. Turner, "Predicting personality with social media" Proceedings of CHI 11 Extended Abstracts on Human Factors in Computing Systems, pp. 253-262, May 7-12, 2011, Vancouver, B, Canada.

[47] D. Quercia, R. Lambiotte, D. Stillwell, M. Kosinski and J. Crowsroft, "The personality of popular Faebook users", Proceedings of CSCW 2012, pp. 955-964, February 11-15, 2012, Seattle, Washiton, USA.

[48] K. Moore and J. C. McElory, "The influence of personality on Facebook usage, wall postings, and regret", Journal of Computers in Human Behavior, vol.28, issue. 1, pp. 267-274, January 2012.

[49] A. Ortigosa, R. M. Carro and J. I. Quiroga, "Predicting user personality by mining social interactions in Facebook", Journal of Computer and System Sciences, vol. 80, issue. 1, pp. 57-71, February 2014.

[50] A. Eftekhar, C. Fullwood and N. Morris, "Capturing personality from Facebook photos and photo-related activities: How much exposure do you need?", Journal of Computers in Human Behavior, vol, 37, pp. 162-170, August 2014.

[51] P. Howlader, K. K. Pal, A. Cuzzocrea and S. D. M. Kumar, “Predicting facebook-users' personality based on status and linguistic features via flexible regression analysis techniques”, SAC '18 Proceedings of the 33rd Annual ACM Symposium on Applied Computing, pp. 339-345, Pau, France, April, 2018.

[52] T. Tandera, Hendro, D. Suhartono, R. Wongso and Y. L. Prasetio, “Personality Preddiction System from Facebook Users”, 2nd International Conference on Computer Science and Computational Intellignece (ICCSCI), Bali, Indonesia, October, 2017.

[53] D. Markovikj, S. Gievska, M. Kosinski and D. Stillwell, "Mining Facebook Data for Predictive Personality Modeling", AAAI Technical Report, Computational Personality Recognition (Shared Task), 2013.

[54] V. Kaushal and M. Patwardhan, “Emerging trends in personality indentification using online social networks-A literature survey”, ACM Transactions on Knowledge Discovery from Data, vol. 12, issue. 2, Article.15, January 2018.

[55] J. W. Pennebaker, M. E. Francis and R. J. Booth. 2001. Linguistic Inquiry and Word Count: LIWC2001. Erlbaum, Mahwah, NJ (www.erlbaum.com).

[56] M. Coltheart, “The MRC psycholinguistic database” Quarterly Journal of Experimental Psychology 33A, pp. 497–505, 1981.

[57] K. Moffitt, J. Giboney, E. Ehrhardt, J. Burgoon, J. Nunamaker, “Structured programming for linguistic cue extraction” [Online].; 2010. Available from: http://splice.cmi.arizona.edu/.

[58] Kucera and W. N. Francis, “Computational Analysis of Present-day American English” Brown University Press, Providence, 1967.

[59] G. D. A. Brown.. A frequency count of 190,000 words in the London-Lund Corpus of English Conversation. Behavioural Research Methods Instrumentation and Computaters 16, 6 (1984), pp. 502–532, 1984.

[60] M. A. Hall, "Correlation-based Feature Subset Selection for Machine Learning" University of Waikato, Hamilton, New Zealand.

[61] M. Hall and L. A. Smith, “Feature Subset Selection: A CorrelationBased Filter Approach,” Proc. 4th International Conference on Neural Information Processing and Intelligent Information Systems, pp. 855-858, 1997.

[62] M. Doshi and R. K. Chaturvedi, "Correlation based Feature Selection (CFS) Technique to Predict Student Perfromance", International Journal of Computer Networks & Communications (IJCNC), Vol.6, No.3, May 2014.

[63] A. Abbasi, S. France, Z. Zhang and H Chen, "Selecting Attributes for Sentiment Classification Using Feature Relation Networks", IEEE Transactions on Knowledge and Data Engineering, vol. 23, issue. 3, March 2011.

[64] G. Forman, "An Extensive Empirical Study of Feature Selection Metrics for Text Classification", Journal of Machine Learning Research, vol. 3, pp. 1289-1305, March 2003.

[65] Z. Gao, Y. Xu, F. Meng, F. Qi and Z Lin, "Improved information gain-based feature selection for text categorization", Proceedings of 4th International Conference on Wireless Communications, Vehicular Technology, Information Theory and Aerospace & Electronic Systems (VITAE), 11-14 May, 2014.

[66] L. Yu, H. Liu, "Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution" Proceedings of the Twentieth International Conference on Machine Learning, pp. 856-863, 2003.

[67] E. Sarhrouni, A. Hammouch and D. Aboutajdine, "Application of Symmetric Uncertainty and Mutual Information to Dimensionality Reduction of and Classification Hyperspectral Images" International Journal of Engineering and Technolofy (IJET), vol. 4, issue. 5, pp. 268-276, 2012.

[68] S. I. Ali and W. Shahzad, "A feature subset selection method based on symmetric uncertainty and Ant Colony Optimization", International Conference on Emerging Technologies, 8-9 October, 2012, Islamabad, Pakistan.

[69] Pearson, Karl (1900). "On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling" (PDF). Philosophical Magazine. Series vol. 5, issue. 50, pp. 157–175. doi:10.1080/14786440009463897.

[70] M. S. Nikulin, "Chi-squared test for normality", Proceedings of the International Vilnius Conference on Probability Theory and Mathematical Statistics, vol. 2, pp. 119–122, 1973.

[71] B. Jacob, J. Chen, Y. Huang, and I. Cohen, “Pearson correlation coefficient,” in Noise Reduction in Speech Processing. Berlin, Germany: Springer-Verlag, pp. 1–4, 2009.

[72] E. Loper and S. Bird, “NLTK: the natural language toolkit” In Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics, 2002.

[73] K. J. Kim and S. B. Cho, “Ensemble classifiers based on correlation analysis for DNA microarray classification” Neurocomputing, vol. 70, Issues 1–3, December 2006, pp. 187-199.

[74] B. Auffarth, M. Lopez-Sanchez and J. Cerquides, “Comparison of Redundancy and Relevance Measures for Feature Selection in Tissue Classification of CT images” Advances in Data Mining: Applications in Medicine, Web Mining, Marketing, Image and Signal Mining / [ed] Petra Perner, Heidelberg: Springer Berlin/Heidelberg, 2010, pp. 248-262.

[75] M.M Mukaka, “Statistics corner: A guide to appropriate use of correlation coefficient in medical research” Malawi Medical Journal, 2012 September, vol. 24 issue. 3, pp. 69-71.

[76] W. Duch, P. Matykiewicz and J. Pestianc, “Neurolinguistic approach to natural language processing with applications to medical text analysis” Journal of Neural Networks, vol. 21, issue. 10, December 2008, pp. 1500-1510.

[77] I. Solti, C. R. Cooke, F. Xia and M. M. Wurfel, “Automated classification of radiology reports for acute lung injury: Comparison of keyword and machine learning based natural language processing approaches” IEEE International Conference on Bioinformatics and Biomedicine Workshop, 1-4 Nov. 2009, Washington, DC, USA

[78] L. Antiqueira, M. G. V. Nunes, O. N. Oliveira Jr. and L. da F. Costab, “Strong correlations between text quality and complex networks features” Physica A: Statistical Mechanics and its Applications, vol. 373, issue. 1 January 2007, pp. 811-820.

[79] M. Chong, L. Specia, R. Mitkov, “Using Natural Language Processing for Automatic Detection of Plagiarism” In Proceedings of 4th International Plagiarism Conference, Northumbria University, Newcastle upon Tyne, UK.

[80] E. Loper and S. Bird, “NLTK: the natural language toolkit” In Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics, 2002.

[81] I. Rish, “An empirical study of the naive bayes classifier”, In Proceedings of IJCAI-01 workshop on Empirical Methods in AI”, pp. 41–46, Sicily, Italy, 2001.

[82] S. R. Safavian and D. Landgrebe, “A survey of decision tree classifier methodology”, IEEE Transactions on Systems, Man and Cybernetics, pp. 660–674, 1991.

[83] L. Breiman, “Random forests”, Machine Learning, 45:5–32, 2001.

[84] L. Breiman, J. H. Friedman and R. A. Olshen, “Classification and Regression Trees”, Wadsworth, 1984.

[85] J. C. Platt, “Sequential minimal oprimization: A fast algorithm for training support vector machines”, Technical REport MSR-TR_98_14, Microsoft Research, 1998.

[86] Niels Landwehr, Mark Hall and Eibe Frank, “Logistic Model Trees”, pp. 161-205, vol. 95, 2005.

[87] Mukta, M. S. H.; Ali, M. E.; and Mahmud, J. 2016b. “User generated vs. supported contents: Which one can better predict basic human values?” In Social Informatics, pp. 454–470. Springer.

[88] Gong, W., Lim, E.P., Zhu, F., “Characterizing silent users in social media communities”, In ICWSM, 2015.

[89] Kosinski M, Stillwell D, Graepel T. “Private traits and attributes are predictable from digital records of human behavior”, In Proceedings of the National Academy of Sciences of the United States of America; 2013: PNAS. pp. 5802-5805.

[90] Youyou W, Kosinski M, Stillwell D. “Computer-based personality judgments are more accurate than those made by humans”, In National Academy of Sciences; 2015. pp. 1036-1040.

[91] Bodlund, O., Ekselius, L. & Linstrom, E. (1993) Personality traits and disorders among psychiatric outpatients and normal subjects on the basis of the SCID screen questionnaire. *Nordisk Psykiatrisk Tidsskrift*, 47, 425–433.

[92] Coid, J. W. (2003) Formulating strategies for the primary prevention of adult antisocial behaviour: ‘high risk’ or ‘population’ strategies? In *Early Prevention of Adult Antisocial Behaviour* (eds Farrington, D. P. & Coid, J. W.), pp. 32–78. Cambridge: Cambridge University Press.

[93] E. Papalexakis, K. Pelechrinis, and C. Faloutsos, “Spotting misbehaviors in location-based social networks using tensors,” in Proc. 23rd Int. Conf. World Wide Web, New York, NY, USA, 2014, pp. 551–552.

[94] Wanita Sherchan, Surya Nepal, and Cecile Paris. 2013. A Survey of Trust in Social Networks. ACM Comput. Surv. 45, 4, Article 47 (Aug. 2013), 33 pages. DOI:http://dx.doi.org/10.1145/2501654.2501661

[95] Hanknson, P., Witmer, H., (2015), “Social Media and trust, a systematic literature review”, Journal of business and economics, Vol, No (3), pp: 517-524. DOI: 10.15341/jbe(2155-7950)/0..06.2015/010

[96] Jolliffe I.T. Principal Component Analysis, Series: Springer Series in Statistics, 2nd ed., Springer, NY, 2002, XXIX, 487 p. 28 illus. ISBN 978-0-387-95442-4.

[97] Abdi. H. & Williams, L.J. (2010). "Principal component analysis". Wiley Interdisciplinary Reviews: Computational Statistics. 2 (4): 433–459.

[98] Mika, S.; et al. (1999). Fisher Discriminant Analysis with Kernels. IEEE Conference on Neural Networks for Signal Processing IX. pp. 41–48.

[99] Campbell RS, Pennebaker JW (2003) The secret Life of Pronouns: Flexibility in Writing Style and Physical Health. In Journal of Psychological Science, American Psychological Society, Vol. 14, No. 1, January.

[100] S. Mallik and Z. Zhao, "ConGEMs: Condensed Gene Co-Expression Module Discovery Through Rule-Based Clustering and Its Application to Carcinogenesis", Genes, vol. 9, no. 1, p. 7, 2017. Available: 10.3390/genes9010007.

[101] S. Mallik, T. Bhadra and U. Maulik, "Identifying Epigenetic Biomarkers using Maximal Relevance and Minimal Redundancy Based Feature Selection for Multi-Omics Data", IEEE Transactions on NanoBioscience, vol. 16, no. 1, pp. 3-10, 2017. Available: 10.1109/tnb.2017.2650217.

[102] T. Bhadra, S. Mallik and S. Bandyopadhyay, "Identification of Multiview Gene Modules Using Mutual Information-Based Hypograph Mining", IEEE Transactions on Systems, Man, and Cybernetics: Systems, vol. 49, no. 6, pp. 1119-1130, 2019. Available: 10.1109/tsmc.2017.2726553.

[103] Trunk, G. V. (July 1979). "A Problem of Dimensionality: A Simple Example". IEEE Transactions on Pattern Analysis and Machine Intelligence. PAMI-1 (3): 306–307. doi:10.1109/TPAMI.1979.4766926.

Publication List:

Journal:

1. Ahmed Al Marouf, Md. Kamrul Hasan, Hasan Mahmud, "Comparative Analysis of Feature Selection Algorithms for Computational Personality Prediction from Social Media", IEEE Transaction of Computational Social Systems.

Conferences:

1. Ahmed Al Marouf, Md. Kamrul Hasan, Hasan Mahmud, "Identifying Neuroticism from User Generated Content of Social Media based on Psycholinguistic Cues", 2019 2nd IEEE Conference on Electrical, Computer and Communication Engineering (ECCE 2019), CUET, 7-9 February, 2019. (Scopus Indexed)
2. Ahmed Al Marouf, Md. Kamrul Hasan, Hasan Mahmud, "Secret Life of Conjunctions: Correlation of Conjunction Words on Predicting Personality Traits from Social Media using User-Generated Contents", 2019 Springer International Conference on Advances in Electrical and Computer Technologies (ICAECT 2019), Coimbatore, India, 26-27 April, 2019. (Scopus Indexed)