Modeling and Predicting Psychological Attributes and Preferences of Users from Social Media Usage

by

Md. Saddam Hossain Mukta

DOCTOR OF PHILOSOPHY

Department of Computer Science and Engineering

BANGLADESH UNIVERSITY OF ENGINEERING AND TECHNOLOGY

DHAKA-1000 BANGLADESH

August 2018

PhD Thesis

Modeling and Predicting Psychological Attributes and Preferences of Users from Social Media Usage

A thesis submitted to the Department of Computer Science and Engineering in partial fulfillment of the requirement for the degree of

DOCTOR OF PHILOSOPHY
IN
COMPUTER SCIENCE AND ENGINEERING

by
Md. Saddam Hossain Mukta
Student ID 1013054001F

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
BANGLADESH UNIVERSITY OF ENGINEERING AND TECHNOLOGY
DHAKA 1000, BANGLADESH

August 2018

The thesis titled "**Modeling and Predicting Psychological Attributes and Preferences of Users from Social Media Usage**" submitted by Md. Saddam Hossain Mukta, Roll No. 1013054001F, Session October 2013, to the Department of Computer Science and Engineering, Bangladesh University of Engineering and Technology, has been accepted as satisfactory in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Computer Science and Engineering and approved as to its style and contents on August 4, 2018.

i

Board of Examiners

1. _____
Dr. Mohammed Eunus Ali — Chairman (Supervisor)
Professor
Department of CSE, BUET, Dhaka.

2. _____
Dr. ~~Md.~~ Mostofa Akbar — (Ex-officio)
Professor and Head
Department of CSE, BUET, Dhaka.

3. _____
Dr. M. Kaykobad — Member
Professor
Department of CSE, BUET, Dhaka.

4. _____
Dr. Abu Sayed Md. Latiful Hoque — Member
Professor
Department of CSE, BUET, Dhaka.

5. _____
Dr. Jalal Mahmud — Member
Research Staff Member
Research Manager Personality Analytics
IBM Almaden Research Center,
San Jose, CA, USA.

6. _____
Dr. Rifat Shahriyar — Member
Assistant Professor
Department of CSE, BUET, Dhaka.

7. _____
Dr. **Zhe Liu** — Member (External)
Research Staff Member
Personality Analytics
IBM Almaden Research Center, San Jose, CA, USA.

Candidate's Declaration

This is to certify that the work entitled "Modeling and Predicting Psychological Attributes and Preferences of Users from Social Media Usage" is the outcome of the research carried out by me under the supervision of Prof. Dr. Mohammed Eunus Ali in the Department of Computer Science and Engineering, Bangladesh University of Engineering and Technology, Dhaka-1000. It is also declared that this thesis or any part of it has not been submitted elsewhere for the award of any degree or diploma.

Saddam

Md. Saddam Hossain Mukta
Candidate

Dedication

To my Parents

Contents

List of Figures

List of Tables

Acknowledgments

All praises due to Allah, the most benevolent and merciful.

Four years ago, I never anticipated that my research leads me in this position. I believe that I acquired knowledge how to develop, nurture, and accomplish a fundamental research. I would like to express my deepest gratitude to my supervisor Professor Dr. Mohammed Eunus Ali for his guidance on this long journey of Ph.D. and showing me the path of conducting a successful research. He shared his wisdom with me in analyzing subject matters and at the same time valued my thoughts to synthesize those topics. His suggestions encouraged me to identify novel problems, analyze these problems deeply, and solve those insightfully. He taught me how to write academic papers. His support gave me strength at the time of my disappointment. Without his continuous motivation and encouragement, I could not have finished this writing.

A special thanks goes to my co-supervisor Dr. Jalal Mahmud, Research Staff Member at IBM Almaden Research Center and manager of Personality Analytics research group, who helped me whole heartedly during development of my research problems and guided me by suggesting potential machine learning techniques to deal with the research problems. This research is funded by ICT Division, Ministry of Posts, Telecommunications and Information Technology, Government of the People's Republic of Bangladesh. I also convey my heartfelt gratitude to the members, policymakers, and government to introduce such a Doctoral Fellowship scheme with the aim of forwarding the nation in advanced studies and research.

Abstract

In recent times, Social Networking Sites (SNS) such as Facebook, Twitter, Foursquare and IMDb have become major platforms of communications for users in the web. These SNS allow a user to share ideas, thoughts, and opinions with her friends, family and acquaintances. Every day millions of newsfeeds and tweets are posted in these SNS. The contents of these newsfeeds and tweets provide a rich platform for the researchers to identify cognitive and psychological attributes such as personality, values, and preferences of involved users. Several studies have been conducted to identify these attributes from social media usage. These studies predict psychological attributes (i.e., personality and values) independently by analyzing the contents of the social media usage. None of the earlier research investigates how these psychological attributes combinedly influence users' behavior and decision making process. Our research will take a step forward to investigate whether these psychological attributes derived from social media interactions correlate, change, and influence one another in real life. This dissertation particularly addresses the following problems by analyzing the social media interactions: i) identifying *values* from multiple interaction features: we develop a technique for computing unified value score of Facebook users from their different interaction features such as *statuses, page-likes,* and *shared-links*, ii) predicting the change of values: we build a hybrid time series based machine learning model to capture the change of values from users social media usage, iii) identifying psychological groups: we identify psychological group of users from their interactions in an egocentric social network, and iv) predicting users' preferences from psychological attributes: we conduct two case studies for predicting users' *eat-out* and *movie* preferences from their psychological attributes inferred from social media usage.

To address the above research problems, we use users' social media interactions and psychological attributes to build different models. These models take psycholinguistic attributes as independent variables and real life preferences as dependent variables. We conduct experiments with 726 Facebook users, 731 Twitter and Foursquare users, and 330 IMDb users to build and validate our machine learning models. Our models achieve moderate to strong prediction potentials in predicting users' psychological attributes from their social media usage. These models also outperform the current baseline approaches and our validation results are consistent with the real world.

Keywords: Social Networks; Data Analytics; Personality, and Basic Human Values

Chapter 1

Introduction

Social Networking Sites (SNS) have touched the lives of millions of people around the world. People share their thoughts, and activities through the newsfeeds and tweets of these social networks. The contents of these newsfeeds and tweets provide an opportunity to the researchers to identify cognitive and psychological attributes such as personality [128], values [43], and preferences [79] of involved users. Identifying these psychological attributes can benefit a number of real life applications such as prediction of social dynamics, selection of career paths, identification of the transition of fashion or trends, and prediction of customers' buying behavior. This dissertation will take a step forward to find out the correlation between different psychological attributes, i.e., *personality* and *values* and investigate the impact of these attributes on users' preferences from social media usage. In short, the main objective of this dissertation is fourfold: i) finding human psychological attributes, i.e., *values*, from different social media interaction features, ii) capturing the dynamics of *values* from social media usage, iii) finding users' *social role identities* and group of users based on their psycholinguistic signals, and iv) finding correlations between different psychological attributes, i.e., *personality* and *values*, and users' preferences in real life.

1.1 Background and Motivation

In recent years, social media sites such as Facebook and Twitter have become popular platforms for users to express feelings, share happiness, and interact with others. Thus, users' *newsfeeds* and *tweets* have become a major source of information for predicting their be-

havior. We can extract users' key psychological attributes such as sentiment [116], preferences [113], values [43], and personality [80] by analyzing these *newsfeeds* and *tweets*. In recent studies [155, 194], researchers find that what people *say* and *write* reflect their individual characterization. Motivated by the above studies, in this dissertation we capture users' psycholinguistic attributes by analyzing their *newsfeeds* from different social media word usage and investigate whether different psychological attributes correlate, change, and influence one another in real life.

Though there have been a plethora of works [43, 80, 113] in social network data analysis, we find several missing aspects that need to be addressed. We discuss these missing aspects and related works into four different sections. First, we discuss studies on different psychological attributes, i.e., values, and SNS usage and investigate whether there is any gap in the previous studies. Second, we discuss whether existing studies present any research on the detection of temporal change of values from SNS usage. Third, we describe related works and gaps on social role identification and grouping of similar users in an egocentric social network from their psycholinguistic patterns. Fourth, we discuss the existing research works and gaps on users' preferences from their psychological attributes derived from social media interactions.

1.1.1 The Psychological Attributes and SNS Usage

Basic Human Values (or, *values* in short) signify the importance of different things such as power, security, tradition, success, happiness, and social status in our life. Our *values* define who we are and what we do. Society, culture, religion and life experiences build up the priorities of numerous *values* in individuals. The priorities differ from one individual to another which results in diversities of an individual's actions in different situations. Schwartz [175] presents ten root-level *values*: *self-direction, stimulation, hedonism, achievement, power, security, conformity, tradition, benevolence,* and *universalism*. The author also maps the ten root-level *values* into five higher-level *values*: *self-transcendence, openness-to-change, self-enhancement, hedonism, and conservation*. We summarize the core ideas of the five higher-level values from Schwartz [175] as follows. *Self-transcendence* denotes one's concern regarding the safety and well-being of others. *Openness-to-change* describes a person's

independence in thought and action. *Self-enhancement* signifies interest of an individual for recognition, supremacy, and control over others. *Hedonism* refers to the self-contentment and satisfaction for one-self. *Conservation* describes one's desire to reliability, peacefulness, and self-restriction. These values provide predictive and descriptive power in the analysis of opinions and actions of people's lives. Thus, identifying *values* has many potential real-life applications that include selection of career paths, prediction of customers' buying behaviors, and detection of life style changes such as fashion and trend.

Personality is another important human psychological attribute which describes pattern of thought, feeling and action of an individual [214]. *Personality* of an individual is commonly defined using five psychological traits: *openness, conscientiousness, extraversion, agreeableness, and neuroticism*, also known as *Big5* [108]. The *Big5* model is one of the well-studied topics in personality research [144]. Sociology studies show that *Big5* traits are consistent across gender, age, and languages [80, 109, 214]. The *Big5* model is distinguished as follows [109], *openness-to-experience* trait has tendency to reflect ideas, innovation and appreciation for values. People with high *conscientiousness* trait prone to be cautious, meticulous and a tendency to seek achievement. *Extraversion* trait has tendency to seek excitement, show positive emotions. People of high *agreeableness* trait tend to be sympathetic, trusted and merciful to others. *Neurotic* individuals show negative emotions such as anxiety, inhibition, anger, and depression. Identifying personality traits has many real-life applications that include finding similar acquaintances in workplaces, observing the social dynamics, parental approval in friends searching, accepting friend requests, etc [23].

Several studies have been conducted to identify *values* from social media usage [34, 43]. In a recent study, Chen et al. [43] identify five higher-level *values* from user's pattern of word use by analyzing her *user generated* contents (i.e., statuses and comments) in Reddit, an online news sharing community. Boyd et al. [34] also identify *values* from a crowd-sourced and a Facebook dataset by using a natural language processing (NLP) technique. They consider only the user generated content (i.e., *statuses*) while identifying values. Hsieh et al. [102] find out relation between user's reading interest and his values in Twitter. They compute models for three different *values*, such as *universalism*, *achievement*, and *hedonism*.

The authors in previous studies [34, 43] do not consider both of the *user generated (UG)*

content and *user supported (US)* content (i.e., *shares* and *up-votes*) in their study, while identifying values of a user. The above studies do not also consider the *silent users* who are significant part of the total Facebook users. Thus, existing approaches suffer in lack of accuracy due to ignoring *silent users* [1] in their studies [82]. The US content can play an important role to identify *values* of the *silent* users. Hence, we consider both *UG* and *US* contents to predict values accurately which is missing in the existing studies of *value* computation from the social media usage.

1.1.2 The Dynamics of Basic Human Values

Values are considered relatively unstable during adolescent period. Several psychological studies [17, 21, 84, 138] also show that values change over time. Bardi et al. [17] propose a value change structure of an individual. Bible et al. [21] find that values and ethical behavior change through experience and education. Steinberg et al. [186] show that during adolescent period, people adapt different psychological changes due to relation with parents, puberty, self-development, and peer relations. Value priority of a person may change over time due to various factors that include technologies [55] [120], events [19], and environments [73], etc.

Existing studies [34, 43] show that values of a person can only be identified from her social network usage. These studies ignore one important aspect is that though values change over time, these studies have not captured the change of values. A large number of applications can be benefited from the *value change* identification that include shifting product selection preference, detection of life style changes (e.g., fashion, trend), and transition of economics and business. Thus, a recommender system can produce customized suggestions on products to the SNS users based on their changing value preferences.

1.1.3 The Social Role Identities and the Psycholinguistic Group

Group (i.e., community) identification is a well studied topic in social media research [71]. Previous studies identify groups based on interaction [57], dense clique [2], label propagation [161], etc. In these studies, researchers largely avoid *psycholinguistic signals* and *cues* to identify group of users. Therefore, it is interesting to explore whether we can extract

[1] Users who do not generate contents themselves frequently.

different groups such as *social role group*, and *personality trait based group* by analyzing users' psycholinguistic features from an egocentric social network (i.e., Facebook) [72].

Roles are important features that characterize the behaving patterns among the members in a society. Depending on social identities and situations, roles guide us to behave differently with different people. Though SNS have put their efforts to make their sites as close as possible to the real life interactions, users in these SNS are still connected with each other with so called *friends/followers* relationships which are not representatives of the real life role identities (e.g., father and son are *friends* to each other in Facebook). However, we observe that interactions among the members inside SNS vary depending on their role identities in real life. For example, we may *post*, *comment*, and *tag* a class fellow instantaneously with a sarcastic meme, whereas we are likely to give advice, share affection and talk about health concerns with a family member in Facebook [37].

In previous studies [6, 121, 193, 206], the authors largely ignore psychological features to identify the role identities. Majority of the existing works identify the social roles by graph analysis and text searching techniques which fail to distinguish the psycholinguistic relationships among the users. Motivated by the above findings, in this research, we identify *social role identities* of two users from their interactions by using psycholinguistic signals (i.e., word use pattern) in an egocentric network.

Understanding these role identities from social network word usage can benefit a number of real life applications. For example, one may want to find out the influence by a specific role member (e.g., family members can persuade to take a decision). One may also be interested in identifying how a person is intimate with particular role members (e.g., a person feels free to discuss his personal issues with his academic members). Marketers, and third party apps can be benefited by automatically identifying these role identities in real life.

In recent years, several works on identifying *Big5* personality traits of users from social media usage have also been appeared [80,173]. People are increasingly using SNS to interact with others. The way people interact with other users usually varies depending on their personality traits. According to a sociology principle [131], people with similar personality often interact with each other more frequently. A group of strongly connected people with similar personality traits is termed as a *homophily*.

McPherson et al. [131] described in a sociology study that *homophilies* can be identified based on different cognitive attributes of humans such as attitude, belief, and behavior. Though we find several studies of personality identification from social media interactions [80, 173], identifying *homophilies* from an egocentric social network interactions has not been studied until recently. In this dissertation, we investigate whether we can discover users' *personality trait based homophilies* from their psychological signals, i.e., word use. Finding the personality trait based homophily in an egocentric network has a number of real-life applications that include finding similar acquaintances in workplaces, observing the social dynamics (e.g., group behavior), and parental approval in friends searching [23].

1.1.4 The Psychological Attributes and User Preferences

Values and *Personality* have been extensively studied across wide range of disciplines [105, 113, 208]. *Personality traits* are largely describe our innate characteristics, while *values* are learned adaptation which is strongly influenced by the environment [147]. In brief, personality guides our instinct thoughts while values control our judgmental process. However, independently both *personality* and *values* can influence one's decision or choice differently in real life [105, 113, 202, 208].

Several studies [9, 113, 124, 163] have been conducted to predict one's behavior *independently* from *personality traits* and *values*. For example, personality traits influences on eating style and food choices [113], artistic and scientific creativity [70]. Several studies also show the influence of personality on consumer buying behavior [200], career choice [167], and political attitudes [25]. Weaver et al. [208] find that personality influences movie preference of the user since different movies are capable of giving rise to different ranges of stimuli, which are directly connected to human psychological state. On the other hand, *values* are the factors that determine who will value which thing the most in her life. Values have also profound impact on behavior and decision making [163], reading habits [9], buying products [202], etc. In the light of above discussion, it is also essential to investigate how combinedly both *personality* and *values* influence user behavior and actions in different ways.

In this dissertation, we manily investigate how psychological attributes, i.e., *personality* and *values*, influence users' decision making process in real life by using two different case

studies: i) *eat-out behavior*, and ii) *movie genre and rating preferences*, from their social media usage.

In the first case study, we observe that people may visit different categories of restaurants, i.e., *cheap, moderate, expensive,* and *very expensive*, based on their interest, socio-economic factors, etc. For example, one may visit an expensive or very expensive restaurant, because of its attractive ambiance, decoration, good flavor and presentation of the foods. On the other hand, one may visit cheap restaurants because they cater foods quickly. In this study, it is interesting to explore whether we can predict users' *eat-out* preferences from their psychological attributes solely derived from their social media usage.

In another case study, we investigate how both *personality* and *values* influence users' decision making process in movie selection. Since user tastes vary depending on her *personality* and *values*, movie genre choices and rating preferences might also vary depending on these psychological attributes. The main benefit of our approach is that, by using our proposed model, one can determine how a user will rate a movie of a specific genre by knowing her *personality* and *values* from tweets and decide whether to recommend the movie.

Several works have been produced to identify user behaviors from social media usage. Silva et al. [184] use Foursquare *check-ins* data from Twitter to find out cultural differences (or similarities) among countries, cities, and regions from their food and drink habits (i.e., check-ins). Wagner et al. [205] show culinary preference (food consumption behavior) based on gender by analyzing *Flickr* images. Said et al. [170] show correlation between users' health statuses and their food preference in social media. On the other hand, we also find that authors of several important studies [42, 95, 97, 212] on predicting movie preferences do not consider psychological attributes, i.e., personality/values. To the best of our knowledge, we find no research that bridge the gap of finding correlations between users' psychological attributes and their preferences from social media usage.

1.2 Research Problems

In view of the limitations of the previous studies, in this dissertation we will investigate the following research problems:

i) Identifying *value* from multiple interaction features: Building different models to predict values for each content types, i.e., *user generated*, and *user supported*. Then, constructing a *weighted linear ensemble model* to build a unified value score from those independent models. Developing a separate model also for computing values of *silent* and *active* users.

ii) Identifying the change of *values* from SNS usage: Investigating the *change of values* from social media usage. Then, developing a hybrid technique to predict the *change of values* from the competitive baseline techniques. Finally, evaluating the accuracy of the hybrid model with a real world scenario through questionnaire.

iii) Predicting social role identities: Identifying social role identities from users' psycholinguistic attributes during their interactions. Then, building an integrated model by using different psycholinguistic techniques, i.e., *open* and *closed* vocabulary.

iv) Identifying Big5 trait based homophilies: Building a computational model to find *homophiles* based on psychological attributes, i.e., Big5 personality traits. Validating the personality trait based *homophilies* in real life.

v) Predicting users' preferences from psychological attributes: Building a data fusion framework between multiple social media sites, i.e., Twitter, IMDb, and Foursquare, to predict users' preferences from their psychological attributes. Evaluating the framework by using different datasets and investigating the efficacy of the technique.

1.3 Solution Overview

In this thesis, we identify different psychological attributes and find the correlations of these attributes with users' preferences from social media usages. Then, we build model to predict users' psychological attributes, and preferences from social media interactions. We take the following strategies to solve our four research problems:

i) Identifying *values* from mulitple interaction features: In this research problem, we identify *values* from different interaction features of SNS usages. We first extract users' contents of different interaction features, i.e., *statuses, page-likes,* and *link-sharing* through a Facebook application. We conduct the 21-items portrait value questionnaire (PVQ) test among these Facebook users for collecting ground truth data on value scores. Then, we compute psycholinguistic analysis of users' each content type by using LIWClite7 [152], a student version of the content analysis method. We compute correlation between *LIWC categories* of users' word use in each content type and *PVQ* test result of these Facebook users. We predict five *higher-level values* by using linear regression models from different types of contents. We construct a weighted linear ensemble model to build a unified value score by integrating the models of different types of contents. Finally, we also build a linear regression model to identify values of *silent* users by analyzing the contents of *page-likes,* and *link-sharing*.

ii) Identifying the change of *values*: In this research problem, we describe the steps to capture the change of *values* from social media usage. We first build linear models for value computation from *LIWC categories* of Facebook statuses and PVQ scores of the users. Then, we divide all the statuses of Facebook users into 6-months time interval for building the *value change* model. We predict value score for each of the time interval statuses by using the models that we built initially. We predict the value score of each user on a future date from her past predicted value scores by using different independent time-series prediction techniques. We apply a weighted hybrid value prediction model by combining the independent techniques. We compute the statistical accuracy of the predicted value scores on a future date and the PVQ test result of the users in real life.

iii) Predicting social role identities and Big5 trait based homophilies: In this research problem, we identify users' *social role identities* from their psycholinguistic pattern of word use in social media interactions. First, we collect all the *comments* of alters (i.e., Facebook friends) and ego that they made on different objects (i.e., statuses, photos, and check-ins, etc.) We separate all the comments for each ego-alter/alter-ego friendship connections. Then, egos manually annotate each set of comments for ego-alter friendship connections by a *role identity*, i.e., *academic member, family member, professional member, friend*, and *acquaintance*.

We compute correlation between LIWC and MEH categories of words and the *role identities*. Next, we build classification models to predict the role identity between an alter and ego by using these linguistic features. Finally, we evaluate our model with different datasets.

We also find group of similar users based on their *personality traits* in an egocentric network. First, we analyze the statuses of the Facebook users by using both *closed* and *open* vocabulary based approaches. We conduct a standardized 44 items IPIP [107] personality test to compute the ground truth data on Big5 [108] personality score. Then, we build independent linear regression model by using the psycholinguistic scores of *open* and *closed* vocabulary based approaches and the ground truth data of those users personality. We combine these two psycholinguistic models and use a linear weighted ensemble model to compute the final personality score [183] [106]. To validate, we develop a questionnaire based novel validation technique that enables us to validate the identified *homophilies* in real-life.

iv) Predicting users' preferences from psychological attributes: In this research problem, we mainly conduct two different case studies: i) *eat-out preferences*, and ii) *movie genre and rating preferences*.

In our first case study, we collect tweets of users who use Foursquare links of restaurant check-ins in their tweets. From these *check-ins*, we categorize the restaurants into four categories: *cheap, moderate, expensive*, and *very expensive*, based on the food price. We compute *Big5 personality traits* and *values* of the users by using *IBM Watson personality insights API* [2]. Arnoux et al. [14] shows that IBM API for personality outperforms the state of the art techniques. Later, we compute pearson correlation between personality traits and value dimensions with the users' frequency of their visits in different categories of restaurants. Then, we build linear regression models from the correlated *Big5 personality traits* and *values* to predict one's *eat-out* preference.

In our second case study, we collect tweets of users who also write reviews, and rate a movie in IMDb. We compute *Big5 personality* traits and *Schwartz value* dimensions of users from tweets by using *IBM Watson personality insights API*. Next, we collect the movie genre, stories, reviews, ratings of these users in IMDb. Later, we compute pearson correlation between *personality traits* and *value dimensions* of the users' with the movie genre and

[2]https://personality-insights-livedemo.mybluemix.net/

rating. Then, we build an ensemble model that combines both personality and values of a user to predict her movie genre and rating preferences.

1.4 Contributions

This dissertation has following key contributions:

1. **A novel value identification approach from different interaction features.**
    i. We propose an integrated model to compute values from different types of Facebook interaction features, i.e., *user generated* and *user supported* contents.
    ii. We develop an effective technique for value computation of *silent* and *active* users.
2. **A new time-series based technique to identify temporal change of values.**
    i. We propose a *weighted hybrid time-series* based technique to capture the *value priority changes* over different time intervals, i.e., 6-months.
    ii. We provide a novel validation technique of the hybrid model with a real world scenario through questionnaire and compare our technique with other baseline techniques.
3. **Frameworks for identifying psychological groups.**
    i. We present a technique to find out correlations between users' psycholinguistic attributes and social role identities in Facebook from their interactions.
    ii. We develop an integrated model that shows a high prediction accuracy while identifying the role identities from different psycholinguistic techniques, i.e., *open* and *closed* vocabulary based models.
    iii. We propose a technique to identify multiple personality trait based *homophilies* in an egocentric network for group recommendation.
    iv. We develop a novel interview based *homophily* validation technique to measure the accuracy of our framework.
4. **A hybrid technique to compute users' preferences.**

i. We propose a novel technique to exploit the data fusion of multiple social media to predict users' *eat-out* preference from their *psycholinguistic attributes*, i.e., personality and values.

ii. We also present a new approach to predict users' *movie genre* and *rating* preferences from their psychological attributes derived from multiple social media sites.

1.5 Organization

In this section, we give preview of different chapters in our thesis:

- **Chapter 2** describes preliminaries and related works for this dissertation. The chapter presents different psychological attributes such as values, personality. The chapter also describes computational models that we use for this dissertation. Later, the chapter discusses about related works for these psychological attributes, their dynamics, and correlations with users' preferences from social media interactions.

- **Chapter 3** presents users' value identification technique from different interaction features. The chapter describes the approach to identify unified value scores from different interaction features. The chapter also shows a technique to identify values of active and silent users.

- **Chapter 4** focuses on identifying the change of value of users from their social network usages. The chapter predicts the change of values by using different baseline time-series techniques. The chapter also describes an integrated technique to identify the change of values from different baseline techniques. The chapter presents an evaluation section at the end.

- **Chapter 5** describes the techniques to automatically derive a user's role identities with other connected users in an egocentric network. The chapter describes the technique for five different social role identities from users' psycholinguistic attributes obtained from their interactions.

- **Chapter 6** describes the approach to identify group of similar users in an egocentric network based on their personality traits. The chapter also presents the validation

technique of the group in real life. The chapter describes a group recommendation technique based on users' personality.

- **Chapter 7** describes an approach to identify users' eat-out preferences from their personality traits and value scores obtained from social media interactions.
- **Chapter 8** first describes a technique to predict users' movie genre preferences from their personality traits and value scores derived from the social media interactions. Later, the chapter also describes another technique to predict users' rating behavior from their personality traits, values scores, and story lines.
- **Chapter 9** concludes the thesis and describes different avenues of the extensions of our research, respectively.

Chapter 2

Preliminaries and Related Works

In this chapter, we describe background and related works pertaining to our dissertation. Section 2.1 briefly discusses different psychological attributes such as *personality*, and *values* from the socio-psychological perspective. Section 2.2 provides a short description on major machine learning models which are largely used in social network analysis. Then, Section 2.3 focuses on the most seminal and relevant works which explore important psychological attributes by analyzing the contents of social networks using different machine learning models.

2.1 Psychological Attributes

In socio-psychological literature, different psychological attributes such as *individual difference* [132], *intelligence* [188], *personality* [109], and *values* [176] have been appeared. Among these psychological attributes *Schwartz value* (Subsection 2.1.1), and *Big5 personality traits* (Subsection 2.1.2) are well studied topics. In this section, according to different psychotherapists and socio-psychologists, we briefly describe the background of these two psychological attributes which will be focused mainly in this dissertation.

2.1.1 Basic Human Values

Our goals, actions, beliefs and behaviors depend on our values. Our values define who we are and what we do. Values signifies the importance of different aspects such as power,

security, tradition, success, happiness, social status, etc., in our life. Society, culture, religion and life experiences shape up the priorities of numerous values in individuals. The priorities differ from one individual to another, which results in diversities of an individual's actions in different situations. Schwartz et. al. [176] present ten root-level values by analyzing data collected from more than seventy countries with divergent cultures, languages, and customs. They also portray the discrepancies and likenesses between these values. We can summarize the core ideas of these ten root-level values from the writings of several theorists and researchers as follows:

1. **Self-Direction** signifies an individual as independent, innovative, confident, deciding and controlling his own aims.

2. **Stimulation** characterizes one's preference to challenges, excitements, thrills and varieties in life.

3. **Hedonism** refers to the contentment, self-indulgence and satisfaction for oneself.

4. **Achievement** represents a person's interest to be socially recognized and the motivation, talents and accomplishments for acquiring the recognition.

5. **Power** indicates one's attraction for authority and social status, supremacy, control over other humans and resources.

6. **Security** specifies reliability, safety and peacefulness of individuals, their associations and the society.

7. **Conformity** denotes responsibility, loyalty, obedience towards others and avoid actions that can harm others and the society.

8. **Tradition** symbolizes respect and reverence for norms and customs of society, cultures or religions.

9. **Benevolence** means one's concerns for others wellbeing and happiness with whom he interacts most (e.g., family, friends, primary groups etc.).

10. **Universalism** stands for the indulgence, responsibility and accountability regarding the safety and wellbeing of all the people in world.

Schwartz et. al. [176] map ten root-level *values* into four higher-level value dimensions. Schwartz [174] shows that the *hedonism* value is related to both openness-to-change and self-enhancement value dimensions. In the recent study of Chen et al. [43], the authors consider the *hedonism* as a separate higher-level value, so that they can measure the scores of *openness-to-change, self-enhancement*, and *hedonism* values distinctly. Therefore, the five higher-level value dimensions can be obtained from ten root-level values: self-transcendence includes benevolence and universalism values, openness-to-change encompasses stimulation and self-direction values, self-enhancement encompasses achievement and power values. Conversation includes security, conformity and tradition, and hedonism remains as it appears in the root-level.

2.1.2 Big5 Personality Model

Big5 model is one of the well-studied topics in personality research [144]. Big5 model has five personality traits: *openness, conscientiousness, extraversion, agreeableness,* and *neuroticism*. Sociology studies show that Big5 traits are consistent across gender, age and languages [80] [214] [109] to analyze human personality. Big5 model is distinguished as follows [109], *openness-to-experience* trait has tendency to reflect ideas, innovation and appreciation for values. People with high *conscientiousness* trait prone to be cautious, meticulous and a tendency to seek achievement. *Extraversion* trait has tendency to seek excitement, show positive emotions. People of *agreeableness* trait tend to be sympathetic, trusted and merciful to others. *Neurotic* individuals show negative emotions such as anxiety, inhibition, anger, and depression.

2.2 Computational Models

In this section, we describe different machine learning models that we will apply to predict users' behavior, and preferences from their psycholinguistic attributes. We use two major learning paradigms: i) *Supervised learning*, and ii) *Unsupervised learning*. We only discuss the machine learning techniques of different paradigms that are relevant to our dissertation.

2.2.1 Supervised Learning

Supervised learning is a machine learning process that takes an input and develop a function which produces an output based on example input-output pairs [169]. In supervised learning, a guide provides a label for each training pattern, and we seek an efficient algorithm to reduce error [62]. In our dissertation, we use *classification*, *linear regression*, and *time-series* prediction by using supervised learning paradigm.

Classification: *Classification* is a predictive modeling that constructs a mapping function *(f)* from input variables *(X)* to labeled output variables *(y)*. *Classification* is a two step process: i) *the learning step* is a process where a classification model is being built, and ii) *the classification step* predicts the class labels for the given data points [89].

The *learning step* (or training phase) of a classification technique constructs the classifier by analyzing a training set built from data tuples and their corresponding class labels. A tuple, X, is represented by an n-dimensional attribute vector, $X=(x_1,x_2, x_3,...,x_n)$, where different values of n describe n data attributes, respectively, $A_1, A_2,..., A_n$. Each tuple, X, is assumed to belong to a predefined class as determined by another data point attribute called the *class label* attribute. The class label attribute is discrete and categorical.

In the context of classification, data tuples can be named as samples, instances, or data points. Each training tuple is provided a corresponding class label, this step is also known as supervised learning, since each attribute of each training instances describes which class they belong to.

The predictive accuracy of the classifier is estimated by different performance metrics [90]. After building classification model, we compute confusion matrix which describes the performance of the classification model. We also compute precision, recall to measure the performance of the classifier. If we measure the performance of the classifier with the training dataset, then the classifier likely to show optimistic performance; hence the classifier tends to be *overfitted*. Therefore, we build a test set which is made up of test tuples and their associated class labels. We use a *cross-validation* technique to evaluate the predictive model by partitioning the original dataset into training set and test set.

Linear Regression: *Linear regression* model builds a continuous-valued function that predicts missing or unavailable numerical data points rather than (discrete) class labels [89]. The model is a common statistical technique for modeling the relationship between independent variables and real valued dependent variable. We are interested to learn a linear function $h : R^d \rightarrow R$ that best approximates the relationship between our variables (say, for example, predicting the rent of a house as a function of it's size and rent) [181]. Figure 2.1 shows an example of a linear regression predictor for d = 1.

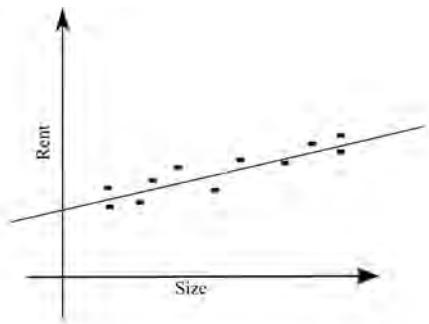

Figure 2.1: Linear regression for d = 1. The x- axis denotes size of the house, and the y-axis shows the rent of the house.

We define a loss function for regression, therefore we determine the difference between the predicted data points *(h(x))* and actual outcome (*y*). We usually apply the squared-loss function, namely

$$l(h, (x, y)) = (h(x)y)^2$$

For this loss function, the empirical risk function is called the *Mean Squared Error*, namely,

$$L_{\mathrm{s}}(h) = \frac{1}{m}\sum_{i=1}^{m}(h(x_{\mathrm{i}}) - y_{\mathrm{i}})^2$$

Feed Forward Neural Network: We also apply *artificial neural network (ANN)* based predictive modeling in this thesis. Majority of the statistical models are parametric and need a clear understanding on statistics. In contrast, ANN is non-parametric model and can be implemented easily the classification and regression problems. A *feedforward* neural network [62] is a network where connections between the neurons are acyclic. In this network,

the neurons are generally arranged in layers, and each layer consists of one or more neurons that receive their inputs from a previous layer or layers directly and send their output to the next layer or layers directly. It means signals in this network constantly move "*feedforward*" from one layer to the next layer(s). Therefore, the network is called *feedforward* network.

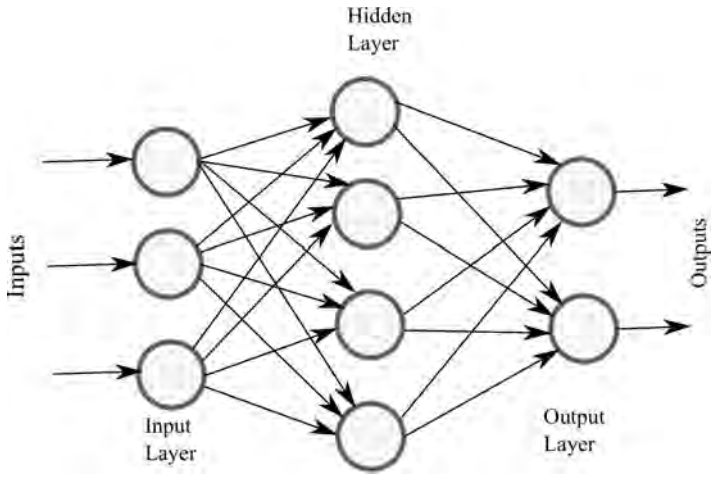

Figure 2.2: A simple feedfoward neural network.

The properties of the *feedforward* networks are summarized as follows [62].

- There are no feedback connections that connect from output layer to input layer. The output of a feedforward neural network is solely dependent on the current input.
- The neurons of *feedforward* networks are usually organized into layers. The first layer receives the input and the last layer produces the final output.
- There may be one or more layers between input and output layers which are not accessible from the external world, therefore the layers are called *hidden layers*.
- Based on the connection to the neurons of the next layer, we name the feedforward networks as fully connected or partially connected.
- In a same layer, the neurons are not connected with each other.

The two-layer neural network usually works with three parameters: *input, weight* and *activation functions*. We can expand the two layers function into *n* layers of processing units. In such a network, there are *(n-1)* hidden layers. The output of the *j-th* neuron in the

first hidden layer is [62] as follows:

$$h_{\mathrm{p}}{}^{(n-1)} = f^{(\text{n-1})}(\sum_{n=0}^{M_{\text{n-2}}} w_{\mathrm{kj}} h_{\mathrm{j}}{}^{(n-2)})$$

Time Series Models:

In this part of our dissertation, we will discuss the time series models. We investigate the dynamics of users' psychological attributes. As computation models, we apply the following machine learning models to capture the dynamics of values. We use three different time-series models: i) Hidden Markov Model (HMM), ii) Autoregressive Integrated Moving Average (ARIMA), and iii) Long Short-term Memory (LSTM).

a) Hidden Markov model (HMM) [62] is a formal foundation for building a stochastic model from a series of observations. Stochastic process of value evolution can be represented as a sequence of random variables $X_1, X_2, \ldots, X_n$.

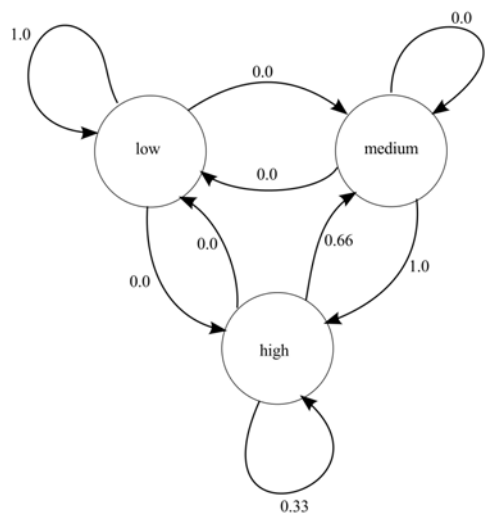

Figure 2.3: Markov chain distribution for stock market prediction.

We consider a sequence of states at successive times. The state at any time t is denoted P(t). A certain sequence of length N is denoted by P^{N} = P(1), P(2), ..., P(n) as [62].

Our model for the prediction of any sequence is described by transition probabilities $Pr(P_{\mathrm{j}}(\mathrm{n}+1)—P_{\mathrm{i}}(\mathrm{n})) = \mathrm{a}_{\mathrm{ij}}$. Figure 2.3 describes the stock market prediction for *high, low,* and *medium* market index. In brief, the Markov property [76] is formally given in equation

where $x_i \in S$ the value of the corresponding random variable and S is the state space.

$$P\{X_{i+1}—X_i=x_i,\ X_{i-1}=x_{i-1},..\}= P\{\ X_{i+1}—X_i=x_i\} \tag{2.1}$$

b) *Autoregressive Integrated Moving Average (ARIMA)* [32, 156, 217] is a widely used linear time-series forecasting model among the researchers in statistics and econometrics. To predict the time-series forecasting, we need to apply differentiation with the non linear transformation to make the data stationary. A random time-series variable is stationary if its statistical properties are all constant over time. A stationary series has no trend, it varies around its mean, has a constant amplitude, and it wiggles in a consistent fashion. Since the time-series data is stationary, it's autocorrelation remains constant over time. Therefore, the ARIMA forecasting has a linear equation that has a stationary pattern. The predictors consist of lags of the dependent variable or lags of the forecast errors.

If the predictors built from the lagged values of *Y*, it is a pure autoregressive. The equation is just a special case of a regression model. The model could be fitted with standard regression software. The Lags which is part of the stationarized series are called "autoregressive" terms. The lags which is the part of the forecast errors are called "moving average" terms, and the differenced time series part which is made to be stationary called "integrated" version of a stationary series.

The general forecasting equation can be stated as follows:

$$\hat{y}_{\mathrm{t}} = \mu + \phi_1 y_{\mathrm{t\text{-}1}} + \ + \phi_{\mathrm{p}} y_{\mathrm{t\text{-}p}} - \theta_1 e_{\mathrm{t\text{-}1}} - ... - \theta_{\mathrm{q}} e_{\mathrm{t\text{-}q}}$$

Here the moving average parameters (θ's) are defined so that their signs are negative in the equation, following the convention introduced by Box and Jenkins [32]. Often the parameters are denoted there by AR(1), AR(2), , and MA(1), MA(2), etc..

c) *Long Short-term Memory (LSTM)* [98], a novel recurrent network architecture that also uses an appropriate *gradient-based learning* algorithm. An LSTM is well-suited to predict the time-series data given time lags of unknown size and duration between important events. The LSTM model persists for a long period of time. A common LSTM unit is consist of a *cell*, an *input gate*, an *output gate* and a *forget gate*. The cell can remember values over

arbitrary time intervals, thus LSTM includes the term "*memory*". Each of the three gates are similar to the conventional artificial neurons. They are used in feedforward neural network.

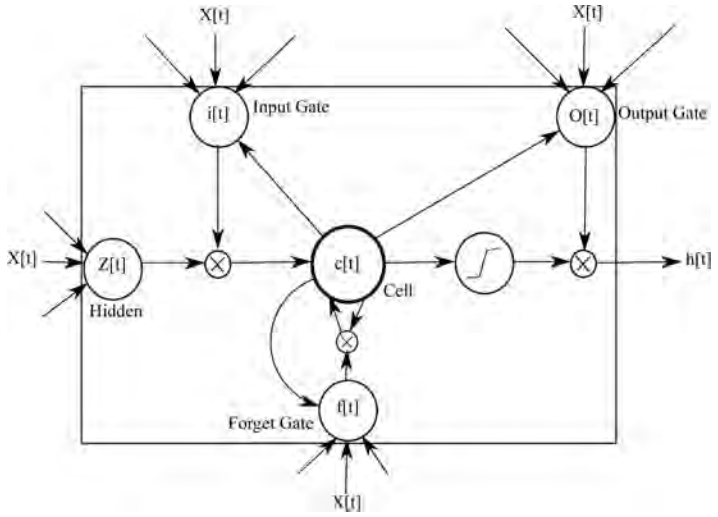

Figure 2.4: The Long Short Term Memory Architecture.

LSTMs can resolve with the exploding and vanishing gradient problem that we suffer during the training of traditional Recurrent neural networks (RNNs). The weights of the neural network receive an update which is proportional to the gradient of the error function in each iteration of training. A potential problem may arise when the gradient becomes vanishingly small, effectively prevents the weight from changing its value. In the worst case, the problem may fully stop the neural network from further training.

2.2.2 Unsupervised Learning

Human has inherent ability to distinguish objects by analyzing their appearances such as apples from oranges, trains from cars, and tigers from lions. We even possess the ability to discriminate these objects without knowing their names. This innate property of human characteristics is called unsupervised learning. Actually, unsupervised learning is a technique which guides us to learn without any teacher, instead we learn by self-organization.

Clustering: Clustering is a widely used technique in data analytics domain. To group similar data points based on different attributes, people use the clustering technique. In brief, *clustering* is a technique of organizing different pattern vectors of X into groups whose

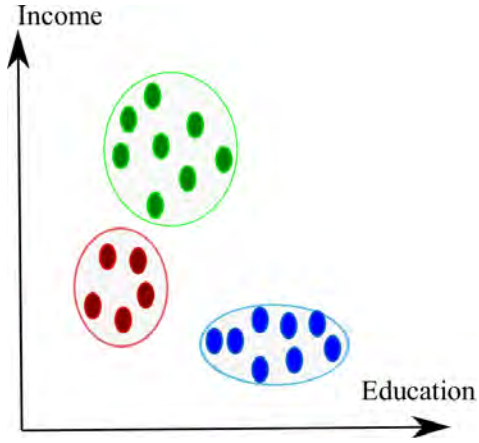

Figure 2.5: Clustering on data points.

elements have similarity based on any attribute [62, 89]. To measure the similarity (or dissimilarity) between two *n-dimensional* pattern vectors, *x(p)* and *x(q)*, the *Euclidean distance*, *D(x(p), x(q))*, is calculated as follows:

$$D(x(p), x(q)) = \sqrt{\sum_{k=1}^{n} (x_{\mathrm{k}}(p) - x_{\mathrm{k}}(q))^2} \tag{2.2}$$

Since distance is a good measure of dissimilarity, we take in account the Euclidean distance between pattern vectors in the same cluster to be reasonably less than the distance between pattern vectors in different clusters. Figure 2.5 represents three different clusters.

2.3 Related Works

In this section, we discuss previous seminal works related to the four primary research problems of this dissertation. In particular, we consider the related works to the problem of identifying values from different interaction features, predicting the change of values, psychological group identification from an egocentric network, and finding users' preferences from their psychological attributes derived from social media usage.

2.3.1 Value Identification

In this subsection, we discuss about the related works of identifying values from social media interactions which is the first research problem of our dissertation. A few recent studies [34,

43] have been appeared to identify values from social media usage. The authors identified values from user generated content in Reddit [43], and Facebook [34].

Chen et al. [43] identify five higher-level values from user's patterns of word usage in social media. Authors analyze the user generated contents (i.e., statuses and comments) of Reddit, an online news sharing community, by analyzing words using Linguistic Inquiry and Word Count (LIWC) tool. They also conduct portrait value questionnaire (PVQ) test among the Reddit users. Later, authors compute the *pearson* correlation analysis between LIWC categories and PVQ scores. Authors predict the value scores by computing the linear regression. To investigate the prediction potential, authors also propose classification techniques. They do not consider other interaction features such as *page-likes*, and *shared-links*. Boyd et al. [34] also identify values from a crowd-sourced and a Facebook dataset using natural language processing (NLP) technique. They only consider user generated content (i.e., statuses) while identifying values. They do not consider other interaction features, i.e., *page-likes*, in Facebook for computing values. Hsieh et al [102] find out relation between user's reading interest and his values by Twitter. They compute regression analysis for three different values, such as *universalism, achievement,* and *hedonism*. Authors [82] characterize the behavior of *silent* users (lurkers) in Twitter. They use Naive Bayes (NB) and Support vector machine (SVM) to learn their classifiers. They also discuss the reasons behind lurking such as: i) personal privacy and safety concern; ii) no need to post, and iii) shyness over public posting. They conduct a systematic study on lurkers and their behavior in Twitter over time.

However, the above studies have the following limitations: i) They do not consider multiple interaction features in their works. To identify values from different features, we find different subset of linguistic features. Since linguistic feature subsets are different for each value, the linear regression models are also different. We cannot apply their techniques, as different features attribute to different value dimensions in different ways, ii) Silent users are significant part of the total Facebook users. Authors in previous studies do not take into account *silent* users during value computation, and thus suffer in lack of accuracy due to ignoring *silent* users in their studies.

In this thesis, we propose a technique to identify values by analyzing the content of two types of interaction features: *user generated* content (i.e., *statuses*), and *user supported*

content (i.e., *page-likes*, and *link-sharing*). We construct a weighted linear ensemble based machine learning model to build a unified value score by integrating the models built from *statuses, page-likes,* and *link-sharing*, which can better predict human values than that of models built from individual interaction feature. We also build machine learning model that can identify values of *active* and *silent* users from their social media interactions.

2.3.2 Change of Values Identification

In this subsection, we discuss about the related works on identifying the change of values which is our second research problem. Several socio-psychological studies [17, 21, 84, 138] show that *values* change due to different factors. However, to date we find that no research has been investigated whether the values change over time from social media usage. Bardi et al. [17] show that *values* change over time. They show that *values* are likely to change among adolescents due to cognitive, biological, and psychological factors. They also use the PVQ to obtain users' value scores. In this paper, the authors conduct the PVQ among students at 9-months interval. Bible et al. [21] also show that *values* and ethical behavior change by conducting an experiment among business major students. A longitudinal study [166] also shows that the pattern of mean-level personality changes over time where mean-level change refers to absolute change in individual's certain psychological attribute, i.e., *personality* and *values*. The authors show that few personality traits change across different life courses among the young adults (age 20 to 40).

Since capturing the change of values is a time series problem, we investigate the major machine learning based temporal modeling techniques in the current literature [76, 98, 217]. To predict the future values, we find a number of studies for sequential time-series forecasting. We first examine the ARIMA [32] model which is widely used for linear time-series forecasting problem. The model is powerful for incorporating both stationary and non-stationary properties of a time-series data. ARIMA builds a pre-assumed linear structure of the model [217]. The model cannot capture and predict nonlinear pattern of a real world dataset. Hidden markov model (HMM) [76] is another well-used technique for modeling time-series data. HMM model predicts future data based on probability distributions over sequence of observations. The model is expensive in terms of memory and computation

time for dynamic programming [199]. The model also needs to train with a large number of seeds.

In another study, Connor et al. [50] use recurrent neural network (RNN) to predict time-series data. RNN has feedback connections from the output neurons, and sends back the error to the input neurons [98]. RNN does not work well for long lags due to error back flow problem. Long short term memory (LSTM) [98] is a potential RNN based architecture that uses gradient based algorithm to control error back flow problem. LSTM can recall long lags data. Zhang [217] develops a hybrid time-series prediction model by using both ARIMA and simple neural network models. Since real life dataset contain both linear and nonlinear properties, author builds such a model that can capture every pattern of real life problems. However, author does not describe clearly how the model deals with time lags. In this research problem, we use a weighted hybrid time-series model by using both ARIMA and LSTM models to capture the linear and nonlinear patterns accurately. Our hybrid model performs better than the hybrid model of Zhang [217], because we solve different limitations of simple neural network based approach by using LSTM based technique.

2.3.3 Social Group Identification

In this subsection, we mainly present the related works that are relevant to our third research problem: *identifying social roles* and *group of similar users* based on their psychological signals from social media interactions. First, we describe the related works on identifying different role identities such as friend, family member, academic member from users' word use pattern in an egocentric network. Then, we also discuss the previous works on finding group of similar users based on their psychological attribute, i.e., personality traits, from social media interactions.

Social Role Identities:

Social role identities are important features that characterize the behaving patterns among the members in a society. Depending on social identities and situations, the roles guide us to behave differently with different people. Several studies [6, 121, 219] have been appeared on automatically identifying social role identities from social media usage. Leuski [121] identifies key roles such as professor, researcher, and graduate students among people in a

research group from their email messages. Author first tags manually email messages based on *speech act*. For example, when he finds content like "We are going to do", he annotates the email by a speech act "Plan". Author uses a total of 8 different speech acts from different email messages. Based on the speech act types and occurrences, author builds classifier to identify different roles. Zhao et al. [219] investigate social roles (e.g., advertiser, company supporter, etc.) by analyzing online activities in *Linkedin*. They manually label social roles by domain experts. Authors build classifier by using both textual (e.g., biography) and categorical information (e.g., educational level) to predict social roles.

Burke et al. [37] present a quantitative study on *parent-child* relationships in Facebook across different stages of life and gender. Authors observe that parents shift their language with their grown-up children. They also show that *mother-children* relationships differ from the father-children relationships. Aiello et al. [6] present a technique to identify different domain of social interactions, namely *status, knowledge,* and *social support* groups. Authors understand the process of social tie construction by the theoretical foundation of *Social Exchange Theory* [26]. They use *non-negative matrix factorization* (NMF) technique to cluster messages to the corresponding interaction group. Tang et al. [193] distinguish social relationships by using a pairwise probabilistic graph model from *publication, email,* and *mobile* dataset. In their probabilistic graph model, they identify colleagues, friends, and family role members. For example, two users who made a number of calls in the working hours might be friends; while two users who call in the evening might be family members to each other.

McAuley and Leskovec [130] describe that SNS are growing enormously, therefore organizing SNS members manually by a customized list (e.g., Google+ circle, Facebook friend list) becomes a difficult task. Authors propose an automated circle exploration approach by discovering common aspects (e.g., family members, university friends) of a group of people. Tang, Lou, and Kleinberg [192] develop a framework to classify relationships from five different social networks such as Espinions, and Slashdot. They use a probabilistic model by using *transfer based factor graph* for predicting social ties. Wang et al. [206] analyze *advisor-advisee* relationships from a computer science bibliographic network. They propose a semi-supervised Time-constrained Probabilistic Factor Graph (TPFG) to identify the advisor-advisee relationships.

According to the notions [78, 190], people use different psychological signals to interact with different social groups (i.e., role identities). Therefore, identifying these signals is important because these signals can distinguish members of different social roles effectively. In previous studies, authors largely ignore these psychological features to identify role identities. Majority of the existing works identify social roles by graph analysis and manual text searching techniques which fail to distinguish the psycholinguistic relationships among the users. For example, family members are likely to use *affection* related words (i.e., love, happy) while professional members are prone to use *achievement* related words (i.e., win, earn) during commenting to ego/alter. Frequency of interactions with both family and professional members might be same, but they have different role identities with ego which is difficult to identify by using previous techniques. Our approach is able to discover psycholinguistic features of Facebook interactions that correlate with which social roles. By using our technique, we can identify multiple roles (e.g., family members can be colleagues to each other) differently among users in Facebook than other techniques.

Homophily Identification

Homophily is a group of people who are strongly connected with each other. We share our intimate personal information in social networks [173]. These networks become ideal place for testing our social phenomena (e.g., homophily) [131] [52]. Bisgin et al. [23] investigated homophily upon interests of individuals such as genre of music, blog, tags, categories, and friends they like. They studied similarity of members by computing Jaccard similarity co-efficient. Adali et al. [3] described person's social circle of similar people ($homophily$) from different action features (i.e. network bandwidth, message content, pair behavior, etc.) of a public network like Twitter. Gilbert et al. [77] defined a predictive model to find out *weak* and *strong* tie among the users of a social network. Trusted family and friends are categorized as strong tie. The authors experimented with 2184 friendships for collecting data with a random subset of friends by *Greasemonkey*, an automated script which was executed at client side. They identified 74 Facebook variables such as predictive variables (intensity, intimacy, duration, structural, emotional support and social distance), structural and dependant variables. Finally, the score scale lie 0-1 which is capable of finding the weakness or strengthens of ties. It is possible to find out tie of individual and clique, but they did not

focus on extracting tie strength upon any psychological attribute, i.e., personality traits.

Valerio et al. [12] explained relation to human behavior and dynamics of interaction over time. They observed a large percent of weak ties and turnover among users in Twitter can be modeled how society is changing. According to anthropology and psychology literature, they modeled different types of egocentric networks, such as support clique (5 members), sympathy (15 members) group , affinity (50 members) group and active network (150 members). The authors found different psychological circles which may evolved over time. They also discovered similarity among the alters and ego from different perspectives (i.e. sphere of sentiment, intensity of activities). However, the authors did not consider similarity among members inside the circles in terms of personality. Crandall et al. [52] described that people are similar to their neighbor in a social network for two different reasons: i) *social influence* guide them to adopt similar behavior of their neighbors and ii) people likely to establish new relationship who are already similar to each other. The authors use *wikipedia* dataset to find out homophily of users based on users' activities (i.e., editing an article) and interactions (i.e., talk to a person). McAuley and Leskovec [130] described that SNS are growing enormously and it is cumbersome to customize member list (e.g., Google+ circle, Facebook friend list) manually. Authors proposed an automated circle exploring approach by discovering common aspect (e.g., family members, university friends) of a group of people. These circles could be nested, hierarchical and overlapping of members, but the authors did not extract any circle based on the psychological or cognitive aspects of members. Hamid et al. [88] identified cohesion circle with limited data, such as number of mutual friends, mutual groups and common apps they are using, sometimes this mechanism works but in reality it is unable to find out psychological tie among the members.

Kafeza, et al. [110] detected influential communities based on a specific topic (i.e., #SocialNetworks) and personality traits of users. They classified personality traits into four basic categories, such as *popular, energetic, conversational,* and *multi-systemic*. Later, they extended their work with Big5 model in [111]. Along with personality, they used basic Twitter metrics (e.g., tweets, retweets, hashtags, etc.) to extract influential communities. However, the results of these studies [111] [110] have following limitations, such as i) Authors extracted topic-based tweets of users for a time interval. This method suffers in collecting

limited features to measure personality accurately; ii) Authors extracted highly active community in a public network which possess members with strong personality profile based on Big5 (even if two members are completely disconnected from each other). However, in this problem, we find community of an ego who are also likely to be same in the real world (e.g., Facebook). iii) It is unclear whether the members inside identified communities are similar in their behavioral aspects. Kafeza et al. [111] mainly focus on identifying influential community where they consider personality traits as an additional parameter. Since the authors identify community in an open network, the members may be disconnected in real-life. It remains unclear to what extent the members inside the identified communities are similar to the real-world situations. In our work, we investigate whether members are likely to possess similar personality traits in the real-world if they are discovered in the same homophily. We introduce a novel statistical reliability analysis technique to show a high similarity among the homophily members. Note that, both [111] and our work consider personality trait while identifying a community.

In the light of the above discussion, we find that our approach differs from the existing studies in different aspects. None of these approaches discovered personality trait based homophily with ensemble of multiple personality identification techniques in an egocentric network, which is one of the main goals of our work. Our framework is able to discover five different homophilies based on five personality traits. We devise a novel interview based technique to evaluate the strength of our homophily structure. To the best of our knowledge, our approach is the first such innovative technique to asses homophily. In this research, we have bridged the gap between personality trait based homophily formation and approval of homophily members in the real world.

2.3.4 User Preferences

In this subsection, we investigate the important works from social media interactions that are related to our fourth research problem. We first describe the previous works which investigate users' preferences from single and multiple social media sites. Then, we present the related research pertaining to our case studies: *prediction of eat-out and movie preferences* from users' psychological attributes derived from social media usage.

Wong et al. [211] investigate the political alignment of a person based on his tweets and retweets. The authors assign a meaningful numerical score to compute the political propensity of a user. They analyze the tweeting and retweeting behavior of the user by *SentiStrength* [196], a lexicon based sentiment analysis package. The studies in [49, 79] and [151] are also similar examples where political orientation of humans are studied using social media analysis.

Bollen et al. [31] show that social and economic events have profound impact on public mood. The authors also represent that group behavior can be predicted with social media. Park et al. [149] investigate the diversity in the usage of *emoticons* in various cultures. *Emoticon* is a strong aspect of informal communication, therefore, information about usage of emoticons can reveal different psychological attributes about a person. Researchers also determine the age of a person based on the language he uses in her tweets. Nguyen et al. [141] predict the age categories, life stages and exact age based on a person's tweets. Different studies also show analysis on personal attributes such as name [126], gender [36, 125], and education [201] from a social media such as Twitter.

Nofsinger [143] shows that financial decisions are strongly driven by human emotions and mood. In the study [30], authors investigate that emotions can profoundly affect individual behavior and decision-making ability. The study also finds several significant correlations between public mood and stock market condition. The psychological studies of [54] and [112] also demonstrate that human decision-makings are intensely influenced by emotions.

Foursquare is another social networking site which allows users to share their locations with friends through check-ins. By analyzing Foursquare data, it is possible to reveal many interesting insights about the pattern of human mobility and diurnal activities. Several studies such as [45], and [145] analyze the spatial properties of data shared in location-based services such as Foursquare.

An emerging research trend is to discover interesting human behavioral and preferences by combining multiple social networking sites such as Twitter and Foursquare. There are many users in Twitter who share Foursquare check-ins in their tweets. Silva et al. [184] use Foursquare check-in data from Twitter. The authors analyze the food and drink habits of

people from different locations. The locations are categorized into drink, fast food and slow food places. Cultural preference of different cities around the world were analyzed. The study extracts individual food and drink habit. Our proposed approach differs from [184], since we analyze the user's pattern of word use in tweets and predict her eat out preferences according to the price category of restaurants. The authors [184] do not compute any correlation between text of tweets and eat (drink) habit. They do not predict any drinking or eating behavior by analyzing users' tweets.

Psychological Attributes and Eat-out Behavior:

We now describe the related works on the first case study: users' *eat-out* preferences from their psychological attributes. Several studies show that personality traits have association with food consumption preference and behavior. Blades [24] describes that what we eat and why is a complex issue. The issue involves numerous factors that include sociological, nutritional, biological, and psychological aspects. The author also presents that the qualities of foods itself such as flavor, appearance, texture, color and psychological factors including behavior, mood, attitude, etc. influence people during food selection. Keller et al. [113] show that consumers' eating style and food choices are directly and indirectly connected with their personality. Heaven et al. [94] show in an experiment that *neuroticism* and *conscientiousness* personality traits have straight link with the eating behavior of an individual. They conduct the experiment among a total of 167 psychology students. However, none of the previous studies find any association between personality traits and eat-out behavior from social media use.

Several hospitality studies [99, 105] show that consumers' restaurant experience is directly linked to their values. Jensen et al. [105] show consumer values in restaurants' meal experience from customers' point of view. The authors present that different categories of experience such as excellence, harmony, emotional stimulation, acknowledgment, context, and ambiance enhance customer's eating preferences in restaurants. Holbrook [99] assumes that consumers' product purchase and service preference linked to efficiency, excellence, aesthetics, politics, esteem, etc. that are derived from their values. Ha [86] suggests different marketing strategies for three restaurant segments (i.e., fast-food, casual, and fine dining) that have connection with consumer's dining values.

We also find few psychological studies [61,100,159] where authors show that values have link with food consumption behavior. Povey et al. [159] observe that values play important role in the decision to adopt vegetarian and vegan diet. Both vegetarians and vegans select their dietary choice based on health and humanitarian reasons. Dreezens et al. [61] explore whether peoples' attitude towards food product are related to the *values* they adhere to life. They find the link between value and genetically modified food and organically grown food. Homer et al. [100] show that values like *hedonism* and *stimulation* have direct effect on nutrition consumption behavior. Steptoe et al. [187] show that political values reflect the acceptability of the country of origin of the food. In another study [172], Schlegelmilch et al. show that environmental issues such as ecological aspect are important factors for buying organic food. However, we do not find any prior study that finds association between values and eat-out behavior from social media interactions. In this thesis, we find associations between users' eat-out preferences with their personality traits, and values independently and combinedly from social media interactions.

From the above discussion, we find that recent studies of social media usage still miss few major aspects such as value identification from multiple interaction features, temporal modeling of value change, psychological group formation, etc. We observe that several socio-psychological studies cover some of the aspects such as change of values, and group formation. However, till date no study has been made on these missing aspects from social media usage. Throughout the dissertation, we will investigate these aspects by analyzing the social media content.

Psychological Attributes and Movie Preferences:

We now describe the previous works on our second case study: users' movie preference from their psychological attributes. A number of studies on different psychological attributes such as personality [80], values [43], and sentiments [116] have been conducted from social networking sites such as Facebook and Twitter. Several studies are also available that have been conducted on the impact of personality and values independently on human behavior and actions [9, 69, 163, 180, 182].

Selfhout et al. [180] find that personality influences friends selection in social media. For example, people who have higher *extraversion* score tend to make more friends than oth-

ers whereas people who have higher *agreeableness* score are likely to be chosen as friends. Shaver et al. [182] show that personality traits are significantly associated with attachment styles. The authors relate personality with relationship status, length, commitment and satisfaction. A study in [41] analyzes personality and individual differences. Rentfrow et al. [164] also examine that people with higher *extraversion* score prefer cheerful music with vocals whereas people with higher openness to experience tend to prefer artistic and intricate music. *Openness-to-experience* people do higher purchases of corporately named products while extrovert people do that for national brands as stated in [209]. Personality can also be related to job performance [18], team performance [140], counterproductive behaviors [171], entrepreneurial status [218], etc.

Values also affect choices and drive people in performing actions that are important to them. Feather [69] conducts a study that finds the correlation between the attractiveness of choice options and values. Payne et al. [150] describe that a decision maker arrives at a choice through a variety of choice rules that are strongly influenced by values. Values have also impact on behavior and decision making [163], reading habits [9], buying products [202], etc.

Recently, several studies have also been conducted on identifying movie preferences from IMDb [42, 95, 97, 212]. Ho [97] finds movie genres from movie synopsis and Wortman et al. [212] identify movie genres from IMDb plot key words. Dooms et al. [59] conduct a study that finds structured tweets (e.g., 'I rate matrix 9/10 http://www.imdb.com/title/tt0133093/ #IMDb') containing movie rating of IMDb. Later, they predict movie rating in another paper [60] from the same dataset, namely *movietweetings*[1].

Existing recommendation systems use many machine learning techniques to recommend movies or items to users. Herlocker et al. [95] use preference matrix of users and finds other users with similar preferences or interests to recommend items to users. Burke [38] emphasizes on the analysis of attributes of items to recommend items. Some predict user rating, others predict mean rating of a movie. Armstrong et al. [11] conjecture that there exists a relationship between a movie's average user rating and its various attributes i.e., Director, Actors, Box Office Gross, etc. Chapphannarungsri et al. [42] use *movie feature vector* con-

[1] https://github.com/sidooms/movietweetings

sisting of movie genre, release period, awards and User Preference Vector consisting of user rating history to predict user rating behavior. Several studies are being conducted to improve the efficiency of movie recommender systems. Different attributes have impact on user rating behavior, such as actors, actresses, director, movie genre, personal interest, etc. Existing recommender systems use these attributes, user's previous history, similar users' histories to recommend movies to users [133, 216].

However, the above studies have the following limitations: (1) They do not consider psychological attributes, i.e., personality/values, in predicting users' movie genre preferences and rating behavior, (2) None of the previous approaches combine psychological attributes with IMDb features to improve the models of movie rating significantly. Therefore, limitations of the previous studies motivate us to combine the psychological attributes to investigate users' effects on movie genre preference and rating behavior in real world.

Chapter 3

Predicting Values from Different Social Media Interactions

In this chapter, we address the research problem of identifying psychological attribute, i.e., *values*, from different interaction features of social media. We mainly analyze the interactions of three different features, i.e., *page-like, shared-link,* and *statuses*. We discuss about our data collection process, methodology, and value identification process from different features in Sections 3.2, 3.3, and 3.4, respectively. Section 3.5 describes about the construction process of our hybrid model to compute unified *value* score from different interaction features. Section 3.6 shows the strength of the models for *active* and *silent* users. Finally, Section 3.7 discusses the findings of our experiment.

3.1 Introduction

In recent times, Social Networking Sites (SNS) have become a major platform for the users to share opinions, views, and thoughts with their connected friends by using different types of interactions such as statuses, page-likes, link-sharing and comments. Among all the existing SNS, in the recent years, Facebook has gained tremendous popularity due to its wide variations of interaction features. The contents of these interactions in SNS provide a rich platform for the researchers to develop different applications that include identifying cognitive and psychological attributes such as values [43], personality [80], preferences [162], and emotion [29] of involved users. In this chapter, we are the first to identify human values of

Facebook users from different types of interaction features.

Basic human values (or, *values* in short) represent a set of criteria such as conservation, hedonism, etc. that are used by individuals to take different actions in their daily lives. Values can be categorized into five higher-level dimensions: self-transcendence, self-enhancement, conservation, openness-to-change and hedonism [178]. These values provide predictive and descriptive power in the analysis of opinions and actions of people's lives. Thus identifying values has many potential real-life applications that include selection of career paths, prediction of customers' buying behaviors, and detection of life style changes such as fashion and trend.

Since interactions of users in SNS resemblance their real-world charectirstics, these interactions facilitate us with the contents that can be used for identifying values. We can categorize the contents of SNS based on social networking activities into two types: user generated (UG) and user supported (US) contents[1]. When a user creates or writes a content by himself, we define the content as a *user generated content*, e.g., statuses and comments of a user. In contrast, when a user expresses her positive association with a content by a particular social network activity such as page-likes and link-sharing, we term the content as a *user supported content*.

A recent study [43] identifies five higher-level values from a user's pattern of word use collected from her statuses and comments (UG content) in Reddit. In [43], authors do not consider US contents such as *shares* and *up-votes*, while identifying values of a user. Hsieh et al. [102] show the correlation between a user's reading interest and her value scores from tweets. None of these techniques consider both UG and US contents while identifying values. This limits the applicability of existing techniques in many scenarios, particularly for the silent users who do not generate contents themselves. For example, in Facebook, many people are not interested in generating contents such as statuses and comments due to i) privacy and safety issues, and ii) shyness of generating contents in public [160] [82]. However, we observe that most of these people interact in Facebook using other interaction features such as page-likes and shared-links. In a psychology study [185], author observes that peo-

[1]For the sake of simplicity, we would refer UG and US for "user generated" and "user supported" respectively throughout the chapter.

ple generally *like* in Facebook because it is quick, easy and wordless interaction feature. In another study [210], author finds that 37% people share to let other know what they believe and 29% people share why they support a cause, organization and belief. Thus, US content can play an important role to identify the values of such users. We consider both UG and US contents to identify values for all types of users in Facebook.

The motivation of our work comes from a key observation that different types of social media contents contribute differently in identifying different human values. Authors in [15] [104] showed that some psychological attributes such as neurotic and conscientiousness personality traits, of users are difficult to predict from UG content in social network. In another study [115] authors successfully predicted highly sensitive personal attributes such as ethnic origin, political views, religion, and substances use (e.g., alcohol) from Facebook *likes* (i.e., US content). Facebook *likes* also express users positive association with online contents such as product, restaurant, sports or music. In a study [82], authors defined users as *lurkers* or *silent users* who generate a little content, and prefer to consume other users' content in SNS. Understanding lurkers is important for recommender systems and targeted advertisement. In Facebook, it is observed that such users (*lurkers*) mostly interact with other by US content rather than UG content. Hence, we need to consider both UG and US contents to predict values accurately.

In this chapter, we propose a technique to identify values by analyzing three types of interaction features: statuses, page-likes and link-sharing, of 567 Facebook users. First, we compute correlation between users' word use in Facebook by Linguistic Inquiry and Word Count (LIWC) [152] tool and portrait value questionnaire (PVQ) [175] test result of these Facebook users. Then, we predict five higher-level values using linear regression models from different types of contents. We also apply different linear classification models to predict human values and find out the suitable classification model for each type of content. Next, we construct a weighted linear ensemble model to build a unified value score by integrating the models of both UG and US contents, which can better predict human values than that of models built from individual contents.

In this chapter, we model silent users who give less status updates (i.e., 2-3 times in every two months time interval), but interact with others by liking pages and sharing links

frequently (i.e., 3-5 times in a week). In particular, we build separate models of values by using different types of contents for silent and active users. Our experimental results show that we can build better value models of Facebook silent users by using US contents than that of active users.

In summary, we have the following contributions:

- We are the first to consider different types of Facebook interaction features, i.e., user generated and user supported contents to compute five higher-level values.
- We compare strength of the models that are built from different types of contents and show which features can best predict which values.
- We build an integrated model that combines different types of interaction features through a weighted linear ensemble technique, which improves the prediction accuracy significantly.
- We compare the strength of value models between silent and active users, and show which features can be effectively used for which models.

3.2 Data Collection

We have invited 645 users to collect Facebook data through posts on Facebook, emails to relevant mailing lists, personal messages, and word of mouth communication. We send invitation through different channels for collecting users' Facebook data only. We create a Facebook application to access users' time-lines. Among 645 Facebook users, 582 users (male=316, female=266) authenticated the application to read their time-lines. The rest 63 users have not shown interest to share their time-lines through the application.

Among the 582 users, we identify 15 users as *churners* [82] [146], who stoped using Facebook for some reasons. We observe their pattern of Facebook usage. We find that these users do not use Facebook for at least last 7 months (according to the paper [82]). Thus, we discard the churners from our dataset. Finally, we consider our dataset with a total of 567 (582-15) Facebook users. We collect a total of 62902, 51066 and 17408 English statuses (tag: ”message”), page-likes (tag: “about”) and shared-links (tag: “description”), respectively. We

collect users' data over a time span of 8 years (January 2007- December 2015). Table 3.1 presents maximum, minimum and average word counts for each type of content.

Table 3.1: Maximum, minimum and average word counts of three different types of contents.

Content type	**Maximum**	**Minimum**	**Average**
Statuses	12168	7	1046.912
Page-likes	1887	3	1234.12
Shared-links	8949	2	498.62

We have conducted 21-item PVQ test among these 567 users as the ground truth data on the value scores. These users are the members of student and different professions (e.g., engineer, doctor, and banker), and aged between 20 and 55 years. Thus our dataset comprises different age groups, professions and gender within same ethnicity. All users are asked to fill out the survey questionnaire via an experimental web page. In these questionnaire, a user is presented a set of statements, where each statement corresponds to a single value dimension. These statements identify the participant's value dimensions without directly asking him about a particular value [175].

Average survey scores for five higher-level values are ranged from 0.542 to 0.769 on a normalized scale of 0-1. We also calculate Cronbach's alphas of PVQ answers for five higher-level values that ranged from 0.428 to 0.553. Cronbach's alpha [53] estimates the internal consistency of reliability of test scores. Although we find that some Cronbach's alpha scores are low, these low scores are acceptable according to the study of Schwartz [178] [43].

3.3 Methodology

In this chapter, we identify values from users' pattern of word usage through different types of interactions in Facebook. First, we extract users' UG and US contents through the Facebook application. Later we preprocess the emoticons and remove noisy special characters. For each type of content, we find out LIWC [152] category of words independently. Then, we find out best subset of LIWC categories (independent variables) and the scores of the

21-item PVQ test (dependent variable) for each type of content separately.

Next, we build linear regression models that predict value dimensions using LIWC category of words from different types of contents. Figure 3.1 shows our methodology to build value models from different types of contents. We compare the strength of different models that are built from different contents. To improve the accuracy, we also combine all the models to compute a unified value score. Different modules of our methodology are described as follows:

- *Data Preprocessing.* We extract users' data of three interaction features through Facebook application and discard irrelevant fields such as *creation time*, *user id*, curly brackets *("{")*, etc. For processing *emoticons*, we use the technique of [213]. We find emoticons in the sentence level. We replace *emoticons* with a corresponding text of emotional sense (e.g., happy, sad, angry, haha, etc.).
- *Linear regression models.* In this module, we present the construction process of different linear regression models for each type of content. We also compute the strength of these models with the R^2/adjusted-R^2 measures.
- *Ensemble of the models.* In this module, we describe how to build a weighted linear ensemble model to determine a unified value score from different types of contents.

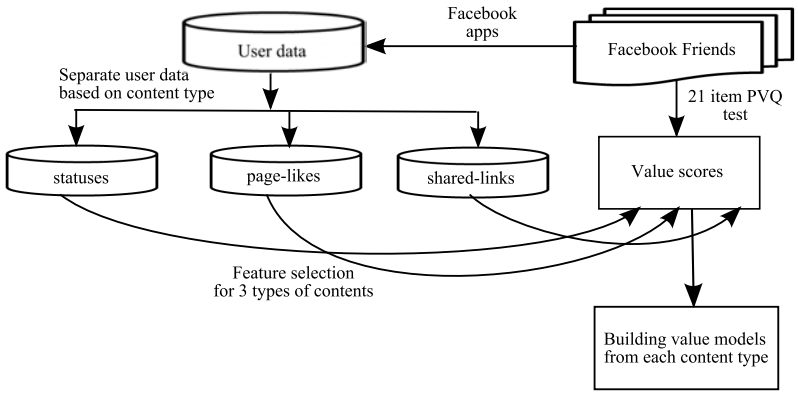

Figure 3.1: Predicting value models from different types of contents.

3.4 Building Models of Values

In this section, we build different types of value models from different interaction features. First, we identify best subset of LIWC categories for each type of content and for each

human value. Later, we build regression models to identify values for each type of content. Next, we investigate prediction potential of our models using machine learning classification techniques.

3.4.1 Feature Selection

Users' way of expressing thoughts and ideas in social network might be different. Some may like to express their thoughts by writing statuses while others may convey their thoughts through page-likes or link-sharing. It is difficult to predict all value dimensions of a user from one type of interaction feature, as not every type of interaction feature reveals every value dimension. Thus, we need to first identify features from different types interactions that are suitable for predicting different value dimensions.

In this section, we use PVQ scores of the 397 (70% of the total dataset) Facebook users as ground truth (dependent variable) data on value scores and best subset of LIWC categories are used as independent variables. We first select the best subset of LIWC categories (predictors) using forward selection approach [56] for each type of content using *leaps* [127] R package implementation. *Leaps* package performs an exhaustive search to find out best subset of LIWC categories using an efficient *branch-and-bound* algorithm. For example, we find family, affect, human, negemo, anx, feel, etc. LIWC categories of words for computing self-transcendence values from users' statuses. Again, we find social, cogmech, insight, tentat, work, etc. LIWC categories of words to compute self-transcendence values from users' page-likes. We also find affect, posemo, bio, assent, etc. LIWC categories of words to compute self-transcendence values from users' shared-links. For each type of content and for each human value, we find a total of 15 (3×5) different subsets of LIWC categories using *leaps* R package implementation. We empirically find that our models (according to Subsection 3.4.2) perform the best by using the subsets of 15 LIWC categories of words. We observe that our models deteriorate the performance when we increase the features than that of these 15 LIWC categories of words.

3.4.2 Regression and Classification Models

We build linear regression models to predict the score of values from each type of content. To build linear regression models, we consider 15 different subsets of LIWC features as our independent variables that are identified in Subsection 3.4.1. We consider PVQ scores as our dependent variable. Figure 3.1 presents the technique of building linear regression models from different types of contents. For each type of content and for each type of value, we have a total of 15 different regression models. A potential problem arises when collinearity is found between values and LIWC features. When there is a perfect linear relationship exists among independent variables, the outcome for a regression model cannot be unique. To remove collinearity among independent LIWC features, we have computed lasso penalized linear regression using *glmnet* R package [43] [92]. This technique reduces the coefficients to a low value or zero, thus the model does not get over fitted. Finally, for each content type, we perform a linear regression analysis with our best subsets of selected LIWC categories with a 10-fold cross-validation with 10 iterations.

Table 6.2 shows the results of our regression analysis. The R^2 and adjusted-R^2 of our models are reasonably moderate across all the values for each type of content. These models show that self-transcendence and openness-to-change values are predictable more accurately through UG content (i.e., status) than US contents (i.e., page-likes). In contrast, hedonism, self-enhancement and conservation values can be predictable through US content than UG content more accurately.

The above linear regression model suffers from the following limitation. Sumner et al. [189] suggested that computing mean squared error (MAE) and root mean squared error (RMSE) for error measure in regression analysis are not adequate. In particular, when the majority of the individuals are around the mean of unimodal distribution, these error measures can often mask large errors.

Table 3.2: Strength of the linear regression models of five higher level values.

	Status (UG)		**Page-likes (US)**		**Shared-links (US)**	
Values	R^2	Adjusted-R^2	R^2	Adjusted-R^2	R^2	Adjusted-R^2
Self-Tran.	18.8%	14.1%	14.2%	10.3%	14.4%	11.2%
Openn.	22.7%	18.9%	16.8%	12.7%	17.8%	14.1%
Hedonism	13.3%	10.09%	13.6%	11.1%	12.3%	8.7%
Self-Enhan.	13.3%	11.3%	12.8%	8.3%	16.8%	12.7%
Conserv.	20.1%	16.3%	20.6%	17.71%	11.9%	8.1%

Since value dimensions have unimodal distributions, RMSE and MAE suffer in lack of investigating strength of a prediction model. To overcome this limitation, we apply different supervised binary machine learning algorithms on our dataset. We classify above-median level as *high* class label and below-median as *low* class label value dimension. We have experimented with several classifiers that include Logistic Regression, Naive Bayesian, Adaboost, Random Forest and RepTree classifiers. For each type of content (i.e., status, page-likes or shared-links), we have applied these classifiers to understand the prediction performance of different value dimensions.

Table 3.3: Best classifiers of different types of contents.

Values	**Best classifying content**	**Best Classifier**	**AUC**	**TPR**	**FPR**
Self-Tran.	status	Adaboost	0.66	0.65	0.34
Open.	status	Naive Bayes	0.69	0.71	0.28
Hedonism	page-like	Naive Bayes	0.59	0.61	0.40
Self-Enhan.	shared-link	Naive Bayes	0.60	0.62	0.39
Conserv.	page-like	Logistic Reg.	0.67	0.66	0.32

Table 7.7 presents the best classifier, content type, its true positive rate (TPR), true negative rate (TNR) and area under the ROC curve (AUC) for computing each of the value dimensions [68]. TPR defines how many samples are correctly classified as *positive* among all positive samples and FPR defines how many samples incorrectly classified as *negative*

among all negative samples available during the test. We conduct the performance of the classifiers using AUC values under the 10-fold cross validation. Performance of the classifiers were conducted using AUC values under the 10-fold cross validation. The curve is plotted the TPR against the FPR at different threshold.

Table 6.2 and 7.7 show the strength of the value models that are built from different types of contents. We observe that our regression models achieve moderate performance. We notice that our classifiers also achieve moderate improvement over random chances. We find the best regression models and classifiers for self-transcendence and openness-to-change values with the content type *statuses*. Again, we observe that regression models and classifiers for hedonism and conservative values show the best performance with the content type *page-likes*. We also notice that we get the best regression model and classifier for self-enhancement value from the *shared-link* content type.

3.5 Ensemble of Models

In this section, we combine three different linear regression models (the models that we have described in Section 3.4) to increase the accuracy of our prediction model. Among our built models, one feature may predict a better value score than the other. For example, predicting self-enhancement value dimension (according to Table 6.2), shared-links show the strongest (the R^2-16.8%) and page-likes show the weakest strength (the R^2-12.8%). Since every feature contributes to the value score based on their strength (weaker or stronger), to find out final value score, we combine all the relations obtained from the previous steps.

It is necessary to prioritize the features based on their importance, as we compute value scores from multiple interaction features in Facebook. For example, some may think that *statuses* can reveal a value of a person more accurately, while other may emphasize on *page-likes* to determine the value score correctly. Ordering among interaction features associates different weights to compute values. Weight signifies the relative importance of a particular feature/content type. To build our ensemble model, we perform the following two steps: i) computing weights of each content type using neural networks, and ii) combining the models with a weighted linear ensemble technique. Figure 3.2 presents the architecture

of our weighted average ensemble value model. From 30% of our total dataset, we learn the weights for each type of content and from the rest 70% of our total dataset, we build individual value models. Finally, we build an ensemble of value models that are derived from different types of contents using their corresponding weights.

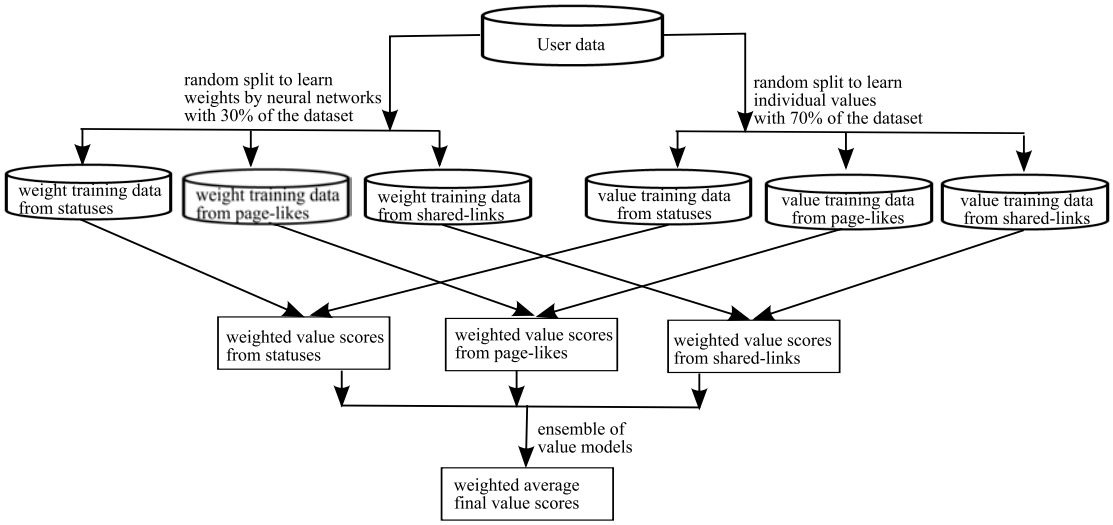

Figure 3.2: Ensemble of value scores derived from models of different types of contents.

3.5.1 Learning Weights from Neural Networks

In this subsection, we determine the *weight* of each content type to determine a value dimension. To this end, we model a neural network using the data of 170 (30% of the total dataset) Facebook users. The neural network build regression model to predict value from each content type. We model our network with three types of contents and five types of values; we build in a total of 15 (3×5) neural networks using R *caret* package implementation [117]. For a single neural network, we use nine input neurons in the input layer, five neurons in the first hidden layer, three neurons in the second hidden layer and one output neuron in the output layer. For each value, we take LIWC scores as input and gives a *value* prediction score as output. We consider the *strength* (the R^2) of the regression model as the *weight* of the content type to determine a value dimension.

We first select the best subset of LIWC features for each content type and value using R *leaps* package implementation [127] by forward selection approach. Then, we normalize the LIWC scores in the interval [0,1] with max-min normalization technique to get better precision. We keep 90% of the data in the training set and the rest are in the test set using

10-fold cross validation with 10 iterations. For each content type and value, we compute the strength of different models. Table 6.7 presents the strength (the R^2) of our neural network based linear regression models that will be used as *weights* of our ensemble models in the next Subsection, 6.5.1.2.

Table 3.4: Weights (the R^2) derived from Neural Networks.

	Status (UG)	**Page-likes (US)**	**Shared-links (US)**
Values	**R^2**	**R^2**	**R^2**
Self-Tran.	19.1%	12.2%	14.3%
Openn.	20.7%	15.0%	13.8%
Hedonism	14.7%	15.1%	9.07%
Self-Enhan.	10.3%	16.1%	18.9%
Conserv.	17.8%	18.2%	13.9%

3.5.2 Weighted Linear Ensemble

In this subsection, we build a weighted linear ensemble model from different types of contents of 397 (70% of the total dataset) Facebook users [183]. We have already built different models from US and UG contents that are described in Section 3.4. Since, we train different neural networks that produce weights, we compute weighted linear ensemble score using the weights in Table 6.7. Finally, we build our weighted linear ensemble model using the weights generated from another dataset, thus our models do not get over-fitted.

Table 3.5 shows the R^2 strength of weighted linear ensemble models from two different US (page-likes and shared-links) contents. We also observe the ensemble models by combining three different types (both US and UG) of contents. We gain better result by ensembling US contents than that of a single US content. We achieve self-transcendence and self-enhancement value scores as the highest and the lowest R^2 strength, respectively. We notice that people with high self-transcendence values generally like pages for humanity such as save the children or movie pages like “The Revenant”. On the other hand, people with high self-enhancement values like pages and share links less frequently through social network. People with high self-enhancement values sometimes like page and share links

regarding technical tutorial, success stories, etc. We also find that we find highest strength of 5 different value dimensions by using ensemble models.

Table 3.5: R^2 strength of ensemble model by integrating different types contents.

	Status (UG)	**Page-likes (US)**	**Shared-links (US)**	**Ensemble of US Content**	**Ensemble of US and UG Contents**
Values	R^2	R^2	R^2	R^2	R^2
Self-Tran.	18.1%	17.9%	16.7%	23.4%	24.6%
Openn.	18.3%	16.1%	17.3%	21.9%	26.7%
Hedonism	12.9%	13.2%	11.5%	17.4%	18.9%
Self-Enhan.	14.8%	15.4%	15.1%	17.2%	19.2%
Conserv.	20.2%	20.9%	10.2%	22.8%	25.2%

Table 6.8 presents the strength of our ensemble models and performance of the respective classifiers that are built from both US and UG contents. We observe that our models obtain a substantial improvement in prediction potential over single feature based value identification models (according to the Table 6.2). Note that, we randomly split our dataset of 567 Facebook users into two parts, 30% (170 Facebook users) of the dataset for learning weights and 70% (397 Facebook users) of the dataset for building the ensembles. Splitting the dataset is somewhat similar to cross validation where we learn from one dataset and apply on another dataset. If we learn weights (i.e., contribution of different content type) from 70% dataset and then again apply the ensemble on the same dataset, this would be like doing training/testing on the same dataset. Thus, we keep the training and testing dataset separate while building ensemble.

Table 3.6: Strength of ensemble model by integrating different interaction features.

Values	**R^2 of linear regression**	**Best classifier**	**AUC**	**TPR**	**FPR**
Self-trans.	24.3%	RepTree	0.71	0.70	0.29
Openness.	35.2%	Naive Bayes	0.78	0.77	0.25
Hedonism	15.1%	Naive Bayes	0.61	0.58	0.41
Self-enhance.	17.7%	RepTree	0.62	0.623	0.39
Conserv.	31.8%	Naive Bayes	0.74	0.735	0.26

Table 3.7: R^2 strength of active and silent users of different values with different types of contents.

	statuses (UG)		**page-likes (US)**		**shared-links (US)**	
Values	active users	silent users	active users	silent users	active users	silent users
Self-Tran.	16.5%	9.5%	15.3%	17.2%	12.1%	13.3%
Openn.	19.3%	13.1%	16.1%	18.5%	14.5%	15.1%
Hedonism	11.1%	7.1%	12.2%	16.9%	13.8%	13.7%
Self-Enhan.	12.5%	10.5%	14.5%	15.1%	17.1%	17.7%
Conserv.	17.3%	13.7%	17.1%	19.7%	11.1%	12.5%

We use ANNs and regression techniques while learning weights and learning values, respectively. We also train the weights for our model with other techniques such as log-odds ratio [168]. However, we obtain the best result of our ensemble model when we train weights with the ANNs and train values with the regression models.

3.6 Silent vs. Active Users

In this section, we compare active users (who give regular status updates) with silent users (who give less updates through statuses) in Facebook. According to Gong et al. [82], users who contribute little to the online content and opt to be quiet most of the time are *silent users*. We consider the users as *silent users* in our dataset who update statuses on average 2/3 times during every two months time interval. They frequently like different Facebook pages such as business, brand, organization and celebrity. They like pages on average 3/5 times in a week. They also share links on average 2/3 times in every two weeks. Though these silent users generate less content, they like pages and share links regularly. Thus, page-likes and shared-links are vital interaction features to identify the values of silent users effectively in Facebook. We show that value scores of silent users are more predictable by US contents. From our dataset of 567 Facebook users, we have a total of 155 silent users who update statues irregularly. These users are regular in liking pages and sharing links. We randomly pick 155 active Facebook users from the rest of 412 (567-155) Facebook users. We find that these active users likely give statuses on an average 5 times in a week. We observe that our active users' generate 14 times more content in online than that of the silent users.

Table 3.7 presents the strength of the models that are built from the data of active and silent Facebook users. In the table, we present the best result among several random trails of active 155 Facebook users. We observe that active users achieve significant better strength of value models than that of silent users by using UG contents. In contrast, silent users gain better strengths in value prediction models than active users by using US contents.

3.7 Discussion

Our work is the first study to identify values from different types of contents (i.e., UG and US) in Facebook. We observe in Table 7.7 that US contents such as page-like and shared-link achieve better prediction potential than UG content (i.e., status) for hedonism, self-enhancement and conservation values. We find that best classifiers for hedonism and conservation value dimensions can be built from page-likes. Again, we notice that best classifier for self-enhancement value can be built from shared-links. On the other hand, we can build better prediction potential for self-transcendence and openness-to-change values from user statuses. In this chapter, we have demonstrated that strength of the value prediction models could be improved (see Table 6.2 and 6.8) by building ensemble of the interaction features. In particular, the strength of self-transcendence, openness-to-change and conservation values are increased by 29.25%, 55.06% and 58.20%, respectively than that of the highest scoring model built from single interaction features. We achieve slight improvement of prediction strength for hedonism and self-enhancement values through ensemble technique.

From Table 6.2 and 7.7, we again observe that people with particular high value scores use particular type of interaction features (i.e., statuses, page-likes and shared-links). Some people convey their thoughts regularly by updating statuses while some may express their opinion by liking pages or sharing links in Facebook. Using a particular interaction feature depends on one's selection and satisfaction. Thus, we find better value models from those interaction features that are used frequently. For example, it is likely that the users who write long statuses, they tend to have high score in self-transcendence and openness-to-change value scores. They usually write about public awareness (i.e., well-being) and interesting insights from different observations. Again, people with high hedonism scores usually like

Facebook pages such as fashion, gadget, restaurants, sports, and music that largely represent fun, enjoyment and pursuit of happiness in one's life. Similarly, people with high conservation value score like Facebook pages with heritage, religion and awareness (e.g., health tips). On the other hand, people with high self-enhancement value score have less propensity to share information in the social media through status updates frequently. Because social media foster procrastination and distract from other activities [104]. But they tend to share information (e.g., career counseling, IT tutorial) through different links in Facebook. In a previous study, it is shown that a moderate personality prediction strength can be achieved from myPersonality [40] dataset (N=250) with minimum and average word count of 1 and 585.004, respectively. In another well cited study [80], authors successfully predict personality from Facebook with a sample size of 279 Facebook users. Therefore, the size (N=567) of our dataset is sufficient to predict value models from social media usage.

Our approach has several limitations. For simplicity, we have conducted our experiment with three interaction features ignoring some other features such as comments, photo captions, etc. We cannot collect users' comments. Users may write comments on the objects such as photos and statuses of other users who have not authorized their timeline through the Facebook application. We do not also read photo captions of users. Users' may have photo captions with two different types of photos: i) self-tagged photos and ii) tagged photos by others. Photo captions that are tagged by other users may not be supported (negative association) by the tagged user (the user who is being analyzed), since a Facebook friend can tag by default to any of his friends without permission. Thus, photo caption is such an interaction feature that may contain neither US nor UG content. However, we can compute more fine-grained values using other interaction features (e.g., user comments and replies if these were accessible completely). In our study, we have demonstrated that it is possible to achieve better accuracy to compute values using both UG and US contents. We find limited prediction potential strength according to Table 6.2 for few values and content type (e.g., hedonism and conservation values by shared-link content). We overlook hashtags used in users' statuses, because people generally use random customized hashtags (e.g., #we_dont_care, #ramadan, etc.) in Facebook. Some of the hashtags also contain local linguistic terms. Thus, it is difficult to harmonize and normalize a diverse set of hashtags. Since we use LIWC to analyze

our data, this approach usually correlates text corpus with a fixed set of words whereas a lexicon based (open vocabulary) [34] approach analyzes all the texts of user data.

Since we work on different interactions features on Facebook, and Facebook is a closed network, we collect the interactions data of a community that are known to each other. In our experiment, we use *judgmental sampling* technique [129], because we first identify most productive Facebook friends who might response in our survey actively. We build our dataset from different age groups and professions.

3.8 Summary

In this chapter, we are the first to identify five higher-level values from different types of contents among Facebook users. We have demonstrated which types of interaction contents can better predict which human values by using linear regression models and a wide varieties of classification methods. We have also built a unified prediction model by combining the values obtained from different interaction features. To produce a unified prediction model, we have integrated different linear regression models with a weighted linear ensemble technique and showed that prediction potential can be improved significantly using our ensemble technique. We have also shown that the value models of silent and active users works differently and can be derived from different content types.

In this chapter, we have conducted the study on value computation from multiple social media interaction features. In the next chapter, we will investigate whether the value scores of an individual change over time from the social media usage.

Chapter 4

Identifying the Change of Values from Social Media Usage

In this chapter, we explore the second research problem of our thesis. Primarily, we investigate whether *values* change over time from the socio-psychological point of view. Section 4.2 describes the data collection process from social media users. Later, Sections 4.4, and 4.5 describe the value building and transition models, respectively. Section 4.6 demonstrates an experiment to show how the change of *values* affects in our decision making process in real life. Finally, Section 4.7 discusses the results and implications of our experiment.

4.1 Introduction

In the contemporary modern world, Social Networking Sites (SNS) have become a vital platform where millions of people share ideas, thoughts and opinions with their friends, family, and acquaintances. These SNS data bring a wide range of opportunities for the researchers to identify cognitive and psychological attributes such as Basic human values [43], personality [47, 80], behavior [10], and sentiment [148]. In this chapter, we focus on identifying the change of Basic human value priorities of users from their social network usages.

Basic human values (or, *values* in short) represent a set of criteria such as security, self-enhancement, etc. that are used by individuals to take different actions. According to the study of Schwartz [175], *values* can be categorized into four higher-level dimensions: self-transcendence, self-enhancement, conservation, and openness-to-change. Schwartz de-

scribes that hedonism, one of the ten root-level values, shares elements of both *openness-to-change* and *self-enhancement* higher-level value dimensions. The author also describes that *hedonism* can be categorized 75% of the time under the *openness-to-change* value dimension. However, in a recent study [43], the authors consider five higher-level value dimensions, where *hedonism* and *openness-to-change* are considered as two separate values. Their motivation was to separate the effect of *hedonism* and *openness-to-change* from each other. *Values* influence user behavior and actions such as reading habits, buying products, and technology use [16, 63].

A recent study has been conducted for predicting higher-level *values* from user's word use in Reddit, an online social news sharing community [43]. Hsieh et al. [102] show the correlation between user's reading interest and his values from tweets. In another study, Boyd et al. [34] identify values from Facebook by using an open vocabulary based technique. None of these approaches can capture the value priority change of users over time, which is our main focus in this chapter.

Values may differ from one person to another. The ordered set of values form a system of value priorities. High priority to a value induces intense and rigid pursuit of that value [175]. For example, a person who emphasizes on achievement is likely to possess high score on self-enhancement value dimension. Similarly, a person who provides a high priority in helping the society is likely to possess a high score on self-transcendence value dimension.

Recent psychological studies [17, 21, 84, 138] show that value changes over time. Bardi et al. [17] propose a value change structure of an individual. They find that values change over time among adolescents. People assign different weights on value priorities over time to make a choice against the situational changes. Bible et al. [21] find that value and ethical behavior change through experience and education. The *value* priority of a person may change over time due to various factors that include technologies [55, 120], events [19], and environments [73], etc. For example, a person might have more preference to get a job quickly (i.e., hedonism) after the completion of his undergraduate studies (i.e., an event), but his value priority can be shifted to pursue higher studies (i.e., self-enhancement) by observing other fellow members of his class (i.e., environment).

In this chapter, we propose a technique to identify five higher-level value changes from

social network usages. Motivated by the work of Chen et al. [43], first we build a linear regression model from Linguistic Inquiry and Word Count (LIWC) [152] scores of Facebook statuses and portrait value questionnaire (PVQ) [175] test results of the Facebook users. Then, we capture our *value* prediction results from different time intervals (e.g., 6-months interval) by using a weighted hybrid time-series model. In our hybrid time-series model, we combine both autoregressive integrated moving average (ARIMA) [32] and long short-term memory (LSTM) [98] techniques, that capture both linear and non-linear patterns effectively from the data. We have observed that the ARIMA model produces large residual errors while predicting *values*, and thus we apply the LSTM model as our next step to predict these residual errors. We compare our hybrid model with the hidden markov model (HMM) [76] based value change technique, which works as the baseline model in our experiment. We have observed that our weighted hybrid model performs better than other independent time-series based value change prediction models. We validate the value priority changes of Facebook users through questionnaire in real life. To show the correlation of value change and real world decision making process, we conduct a detailed experimental study of users' movie watching preference.

In our experimental study, we have used Facebook statuses of a total of 726 users. We build the linear regression model for predicting *values* from the statuses and PVQ scores of 388 active Facebook users. Then, we build and validate our hybrid time-series model with a total of 338 active Facebook users for the duration of an average of 7.5 years, where first 70% time interval data of each user is used for training and the remaining 30% time interval data is used for testing. We observe that on average our hybrid model reduces root mean square error (RMSE) by 41% and 34.49% than the LSTM and the ARIMA models, respectively. Our model also reduces the RMSE by 58.35% than the HMM based baseline technique. We also observe that our hybrid model captures the value changes accurately for *self-transcendence, openness-to-change,* and *conservation* value dimensions in real life.

A large number of applications can be benefited from our value priority change identification that include major or career path selection, prediction of customers' buying behavior or shifting of product selection preference, and detection of life style changes (e.g., fashion, trend). For example, the identification of value priority change of customers may help

business owners to do the targeted marketing based on the changing life styles (e.g., fashion, trend), and product preferences. Thus, a recommender system can produce customized suggestions on products to users based on their changing value preferences.

In summary, our contributions in this chapter are as follows:

- We are the first to propose a technique to identify temporal value changes from social media usage.
- We develop a weighted hybrid time-series based technique to capture the value priority changes over different time intervals.
- We validate our hybrid technique with a real world scenario through questionnaire and compare our technique with other baseline techniques.
- We demonstrate how the change of values relates to the users' movie watching preferences.

A preliminary version of this work has appeared in [135] where we predict the value transition with the HMM based technique only. We find that the HMM based technique sometimes cannot capture unforeseeable events accurately from our dataset. Therefore, the model does not perform well for building the *value* transition model. In this chapter, we investigate further two potential time series techniques: ARIMA and LSTM. First, we build value transition model by using the ARIMA and the LSTM techniques independently. Later, we find that the ARIMA and the LSTM models perform better for our dataset with linear and non-linear properties, respectively. We also find that both of these models outperform the HMM based technique to predict the value transition. Then, we combine both of the ARIMA and the LSTM based techniques by using a weighted hybrid time-series based technique.

4.2 Data Collection

We have collected Facebook statuses of a total of 726 users for our experiments, where statuses of 388 Facebook users are used for building value models, and then statuses of 338 Facebook users for over 6 years are used for building the temporal value change models.

Dataset for value building model: First, we develop a Facebook application that accesses the users' status updates. Later, we have invited 750 users to allow us to collect Facebook statuses through the posts on Facebook, relevant mailing lists and word of mouth technique. Among 750 Facebook users, 388 users (male=210, female=178) authenticated the application to read their time-lines. The users are members of student and professional community. We recruited users who are aged between 15 and 28 years, as they go through different stages of life such as high school, college, and starting a job. This age group (15-28) is likely to change *values* compared to the age group of 35-50 years. We have collected 105,252 Facebook English statuses as of March 20, 2016. The maximum, minimum and average word counts of the collected statuses are 6786, 150 and 894.26, respectively. The statistics describes aggregated word counts for all the users of the value building dataset. We have conducted 21 items PVQ [175] test on a 6 point Likert scale among these users to collect the ground truth data on value scores. The PVQ describes short statements about different people that identify participant's value dimensions without directly asking him about a particular value [175]. The users are asked to fill out the survey questionnaire through an experimental web page.

Dataset for value change model and validation: We have also invited a new set of 695 Facebook users. We have collected a dataset of 338 (male=196, female=142) active Facebook users by using our Facebook application to build value change model and validated the model of value priority change in real life. These users are aged between 17 and 26 years and they are also members of student and professional community. Since, our model works on time-series data, we select the users who update statuses regularly (i.e., weekly 5-7 statuses) in Facebook during the last 6-10 years. Each time interval (6-months) data contains on an average of 158.21 statuses. The maximum, minimum and average number of time intervals of each user are 19, 14 and 15.5, respectively, in their time-lines. Again, each individual has 4193, 63 and 1494 maximum, minimum and average word counts, respectively, for each time interval. For the value change model, first 70% time interval data of each user is used for training and the remaining 30% time interval data is used for testing. We conduct the PVQ test to validate the predicted value scores of these users in real life.

4.3 Solution Overview

In this chapter, we identify temporal changes of value scores from Facebook statuses of 338 users. To identify the temporal changes of *values*, we perform the following steps:

1. We first build linear models for value computation from LIWC categories of Facebook statuses and PVQ scores of 388 users.

2. We divide all the statuses of 338 Facebook users into 6-months time interval for building value change model.

3. We predict value score for each of the time interval statuses by using the models that we built in Step 1.

4. We predict value score of each user on a future date from her past predicted value scores by using the ARIMA and the LSTM techniques.

5. We apply a weighted hybrid value prediction model by combining both the ARIMA model and the LSTM model.

6. We compute the statistical accuracy of the predicted value scores on a future date and the PVQ test result of the users in real life.

7. Finally, we compare our weighted hybrid model (ARIMA+LSTM) with independent ARIMA model and LSTM model, and the baseline HMM based technique.

4.4 Building Models of Values

In this section, we build a linear regression model to predict users' value dimensions from their Facebook statuses. We build our regression model by using the cross validation, and regularization techniques. These techniques can control over-fitting and then enable us to build a machine learning model accurately [43, 215].

We build the linear regression model to predict *values* from psycholinguistic scores obtained through LIWC scores of 388 users' Facebook statuses and PVQ scores as a ground

truth data on *values*. LIWC 2007 identifies 74 different features from text statuses into different categories, where each category contains hundreds of words [152]. These categories include standard counts (word count, words longer than six letters, dictionary words, and function words), psychological processes (social, affective, cognitive, sensory, and perceptual processes), relativity (words on time, the past, and the future), personal concerns (work, achieve, leisure, and health), and other dimensions (counts of various types of punctuation, and swear words). We exclude the categories that are non-semantic, e.g., proportion of long words, and filler.

Table 4.1: Strength of linear regression models and correlation coefficients of five higher-level value dimensions.

Values	**R^2 of linear regression**	**Correlation coefficient**
Self-trans	19.09%	0.4370
Open.	27.63%	0.5257
Hedonism	15.41%	0.3926
Self-enhance.	10.23%	0.3199
Conserv.	21.26%	0.4611

We use PVQ scores of users as dependent variable (ground truth) on value scores and a subset of LIWC categories as independent variables. We select the best subset of LIWC categories (predictors) by using the forward selection approach [56] for each of the value dimension using *leaps* [1] R package implementation. Leaps package performs an exhaustive search to find out the best subset of LIWC categories by using an efficient branch-and-bound algorithm. For example, we find *anx, cogmech, percept, feel, tentat, money,* etc. LIWC categories of words for computing the *openness-to-change* value from users' statuses. We also find *human, affect, posemo, negemo, percept, feel*, etc. LIWC categories of words for computing the *self-transcendence* value from users' statuses. In this way, we find five different subsets of features for computing five different *values* from LIWC categories of users' statuses. A major noise we find during the value prediction is the collinearity between LIWC

[1] https://cran.r-project.org/web/packages/leaps

categories. Therefore, by following the methods of Yarkoni et al. [215] and Chen et al. [43], we use *Lasso penalized* linear regression by using *R glmnet* implementation package [74]. This technique reduces the coefficients to a low value or zero, and thus the model does not get over fitted. We compute the linear regression model for each of the selected LIWC feature subsets and the value scores with a 10-fold cross-validation with 10 iterations. We show that the R^2 values are moderate across all the value dimensions.

Though we obtain the moderate strength in our regression models, we can rely on the prediction result. From the value study of Chen et al. [43], we find that the R^2 strength of different value dimensions is ranged from 13.8% to 18.2%. In contrast, we observe in our experiment that the R^2 strength of different value dimensions is ranged from 10.23% to 27.63%. We find that our experiment results show even better strength for *self-enhancement, openness-to-change,* and *conservation* value dimensions than that of previous study of *values* [43].

Note that, Boyd et al. [34] also propose another value computation technique from social media usage. In our experiment, we use the technique of Chen et al. [43] that needs less intervention from users. On the other hand, in the technique of Boyd et al. [34], we need to collect the following information from users: i) users' self-descriptive values, ii) last seven day behavior, iii) Schwartz Value Survey (SVS) [123], and iv) users' Facebook statuses. Since the authors of the both techniques of Boyd et al. [34] and Chen et al. [43] compute *values* from language inference, we select the approach of Chen et al. [43] that needs to attend in the PVQ test only by the users.

4.5 Modeling the Transition of Values

In this section, we build a model for predicting the transition of value priorities over a continuous time-series dataset. First, we segment the statuses of 338 Facebook users by 6-months time interval starting from their account creation date. In the well cited socio-psychological study of Bardi et al. [17], the authors investigate the change of *values* among the students within a time span of 9 months. Since we train machine learning models to predict value transition, we use a moderate time span (6 months) which is also substantial duration to

change the *values* among adolescents. Our dataset has a total of Facebook statuses of 111 months (January 2007- March 2016). In the dataset, we find that the average sizes of our training instances are 9.25, 12.33, and 18.5 for 1 year, 9 months, and 6 months time intervals, respectively. Therefore, we can build a better time-series based machine learning model by using the 6-months time interval than that of 9 months and 1 year time intervals. Later, we compute the linguistic features from user's each time interval statuses by using the LIWC. We predict the value score from the statuses of each time interval by using the regression model that we have built in the previous section. After that, we apply different competitive time-series techniques to predict the value transition based on the past observations of time interval statuses. In the next subsection, we identify an appropriate machine learning model to capture the value transition of users from the dataset.

4.5.1 Selecting an Appropriate Model

Since real world dataset may contain both linear and non-linear patterns, a single model solely may not able to capture different patterns effectively from the dataset. ARIMA [32] is a linear model and it performs better than other competitive models over the linear dataset; however, the ARIMA model cannot capture nonlinear patterns from the data. On the other hand, among the non-linear models, a simple neural network model cannot use any lag *values* (i.e., past observations) explicitly in their architecture [217]. Another popular non-linear model, the RNN suffers in vanishing error problem [98]. In contrast, the non-linear model, LSTM [98] overcomes the above limitations, and thus we use the LSTM as our model to capture the non-linear patterns from the dataset. For simplicity and brevity of the presentation, in this chapter we only demonstrate the *openness-to-change* value transition model. Without loss of generality, the transition models for other value dimensions can be built by using our proposed technique.

4.5.2 Predicting Value Transition with ARIMA Model

The ARIMA model predicts future results from the past observations and errors [217]. In this subsection, we apply the ARIMA in our *openness-to-change* value change prediction dataset.

Stationarity: A time-series data is called stationary if certain statistical properties such as mean, variance, etc. of the data remain constant over time. Since these properties are constant, it follows a regular pattern over time. However, a real time-series data has both stationary and non-stationary (random) properties. To predict a time-series data in the long-term, we need to transform the non-stationary parts of the dataset into stationary to construct a regular pattern.

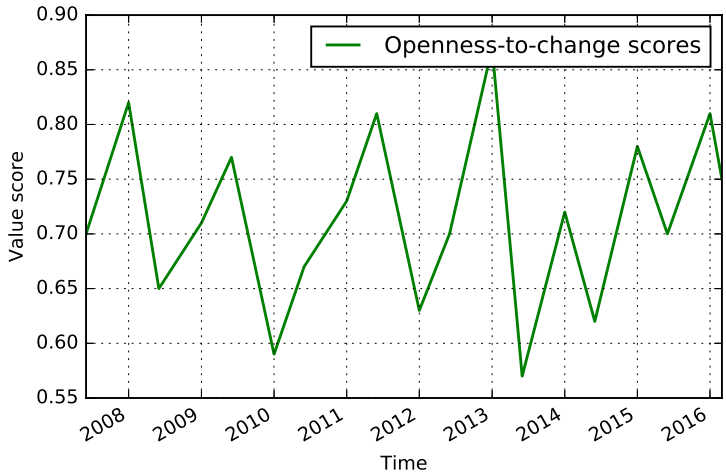

Figure 4.1: The *openness-to-change* value score distribution for a user.

We first convert the non-stationary parts into stationary by using the log transformation technique [32]. Then, we extract different components such as seasonal, trend, and residuals from our time-series data [165]. In particular, we separate out residuals from our time-series data. The residual is measured as the difference between the observed value and the predicted result based on previous observations. Since we identify different unforeseeable events in our value transition dataset, these residuals may represent the unforeseeable events. We use the Python *statsmodels* package implementation [179] of ARIMA to find out the linear patterns from our dataset.

Analysis: After applying the ARIMA model, we find that our initial *openness-to-change* value transition model is non-stationary. In the model, we do not also find seasonality in our dataset as seasonal factors (i.e., summer or winter) are not likely to influence one's value priorities. We do not observe any trend for value dimensions in our value transition dataset. It cannot be guaranteed that certain *values* increase or decrease over time. We also

identify random noise due to some unforeseeable events (e.g., death of a family member) and influence of friends and colleagues.

Figure 4.1 presents the *openness-to-change* value score distribution of a user by 6-months interval from 2007 to 2016 (i.e., 19 time intervals). The figure shows the *openness-to-change* value scores of two different halves in a year. We notice that the *openness-to-change* value scores range between 0.57 to 0.89 in different quarters in the duration of 10 years (2007-2016). For each user, we separate 70% and 30% of the time interval value scores into the training and the test datasets, respectively. To predict time-series data from the past observations, our time-series data need to be stationary and autocorrelated with itself. We check the stationarity property of the user time-series data of the *openness-to-change* value transition by using the well known *augmented Dickey-Fuller (ADF) test* [32]. We find that the p-value of the test is 0.117, and the value is not significant. Therefore, we cannot reject the null hypothesis, and our dataset is not completely stationary. Thus, we conclude that the user's value transition data has non-stationary parts also.

We transform non-stationarity part of the dataset into stationary by changing the parameters of lag-time, degree of differencing, and order of moving average. We find the best result from the ARIMA (4,1,2) model where the model has 4 lags, mean 1, and variance 2. We select the best model that has minimum AIC (-28.72) and BIC (-27.51) scores (according to Box et al. [32]). Figure 4.2 presents the actual value scores (green line) of the next 3 years (next 6 intervals), and the predicted *openness-to-change* value score (the blue line) for the next 3 years (6 intervals). We observe that the ARIMA model cannot predict value changes closely for few points (e.g., June 2013, and June 2014) due to the existence of non-linear patterns in the dataset.

From the above outcome of the ARIMA model, we may describe the following intuitive reasons for the existence of non-linearity in the dataset. Users in our collected dataset are students of universities in an open-credit hour system, and young corporate service holders such as private bankers, software developers, and telecommunication engineers who might change their job frequently. Universities in an open-credit hour system have 2/3 semesters in a year. When the students start a new semester, they get acquainted with many new students. They exchange ideas and thoughts among themselves about their observations and findings.

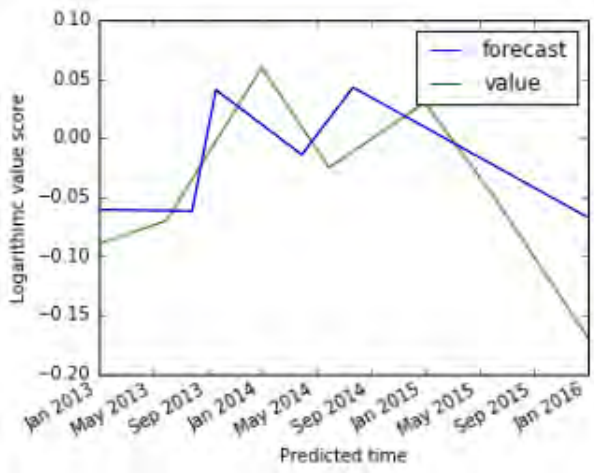

Figure 4.2: Change prediction for the *openness-to-change* value of the user (ARIMA).

They also frequently update statuses through the social media about their new friends and environments. Similarly, the young service holders likely to change their job frequently due to some reasons such as low job security, salary freezing, and willingness to seek new challenges in life. In summary, majority of the users in our dataset pass through different turbulent stages of life. Therefore, we may find several non-linear patterns of the value change among these users which are difficult to model with the ARIMA only.

Results: We observe that our model shows 0.0468 *p-value* with 95% confidence interval. We have achieved 0.65 root mean squared error (RMSE) by using the ARIMA model, which is better than the random baseline. We find that the RMSE for random baseline ARIMA model is 0.671. However, we find high random errors while using the ARIMA model in our dataset.

We apply the LSTM based approach to minimize the random errors observed in the ARIMA model due to non-linear patterns of the dataset.

4.5.3 Predicting Value Transition with the LSTM Model

In this subsection, we apply the LSTM model, which has a deep network architecture, to capture the non-linear patterns in our time-series dataset.

The LSTM model: We can compute the LSTM based time-series prediction by using the window-size architecture, as this architecture performs better than other LSTM based architectures for our problem setting. It is also evident [207] that the window-size LSTM

method works better for a small training dataset. In our case, time-series data of each person are also limited in numbers. For example, we find several users who are active for the whole duration from January 2007 to March 2016 (19 time intervals). We also find that few users have only 14-15 time intervals of time-series data.

We use the Python *theano* based *keras* implementation package to run the window size (e.g., 2, 3) based LSTM technique [46]. Figure 4.3 presents the *openness-to-change* value transition result of a user for two window size neural network based LSTM technique. In the figure, x axis shows the time (i.e., 2007-2016, a total of 19 intervals) of capturing *values* and y axis shows the value scores. For each user, we separate 70% and 30% of the time interval value scores into the training and the test datasets, respectively. The blue line in the figure represents the original value (i.e., *openness-to-change*) scores of the user that are found in our dataset over time. The green and red lines represent the prediction over the training and the test data, respectively.

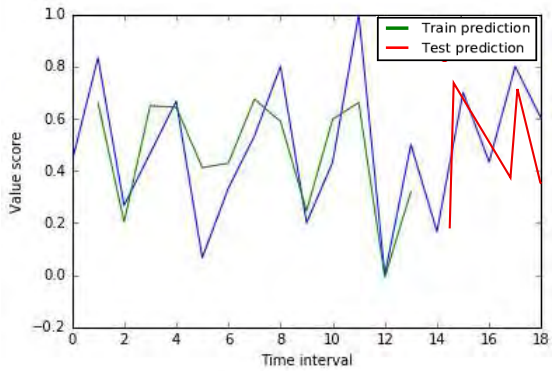

Figure 4.3: Change prediction for the *openness-to-change* value of the user (LSTM).

We set *lookback* and *epoch* parameters as 3 and 2000, respectively. The *lookback* parameter determines the total number of lag values and the *epoch* parameter describes how many times we train the dataset. Though we use 4 lag values in the ARIMA model, we find that the lag value of 3 gives the best result in our LSTM model. We test the LSTM technique with our dataset by varying epochs from 10 to 2000. We find that with a minimum 2000 epochs in our LSTM model, we can adjust weights accurately.

Result: We have achieved 0.61 RMSE for the adjusted weights with 2000 epochs by using our LSTM model. These RMSE scores are better than random baselines. We find that random baseline for RMSE is 0.74 by using our LSTM model.

4.5.4 The Hybrid Value Transition Model

In the previous section, we apply the ARIMA and the LSTM techniques to capture the value transitions of the same set of users. We observe that different techniques produce different RMSEs. The technique that produces less RMSE, is likely to predict more accurately than the other techniques. Therefore, we build a weighted hybrid time-series prediction model. Here, the weight signifies the relative performance of each model. To build a weighted hybrid time-series prediction model, we perform the following two steps: i) compute the weight for each model, and ii) combine the weighted time-series models.

Learning weights: For the value transition prediction, we determine the weights for both models, i.e., ARIMA and LSTM. First, we segment the statuses of the 101 (30% of the users in validation dataset) Facebook users by 6-months time interval and predict the value score for each of the interval. Then, we apply both of the ARIMA and the LSTM models over the time interval value scores to predict the future value score. Later, we calculate the average RMSE scores of all these users' time interval data for the ARIMA and the LSTM models independently. The model that produces less error is considered to have more potential. In our experiment, we obtain that the average RMSE scores for the ARIMA and the LSTM models are 0.65 and 0.61, respectively. Then, we compute the weight for each model by subtracting the RMSE from 1.0. Thus, we find that weights for the ARIMA and the LSTM models are 0.35 and 0.39, respectively.

Figure 4.4: The residual value scores of the user after applying the ARIMA.

Weighted hybrid model: We build a weighted hybrid model for each of 237 (338-101) Facebook users, which is 70% of the validation dataset. We build the weighted hybrid time-series value prediction model in the following form:

$$Y_{\mathrm{t}} = \frac{w_{\mathrm{L}} * L_{\mathrm{t}} + w_{\mathrm{N}} * N_{\mathrm{t}}}{w_{\mathrm{L}} + w_{\mathrm{N}}} \tag{4.1}$$

where, Y_t refers to the future *value* score at time t, L_t and N_t denote the linear and non-linear components of our dataset, respectively and w_L and w_N denote the weights of the linear and non-linear components, respectively.

After computing the weights, we build our hybrid model as follows. We first apply the ARIMA model in our dataset. We predict the change of value scores over the time. Then, we capture the residuals by computing the difference between actual and predicted value scores. Figure 4.4 presents the residual value scores of the user, where x axis shows the time of capturing values and y axis represents the change of value scores. We find that residuals are non-linear in structure. We also observe that a large number of users' value changes have similar non-linear property (according to Figure 4.4) in the dataset. Since ARIMA is a linear model, the residuals of its outcome will more likely have non-linear property. We store these residual value scores. Then, we apply the LSTM technique over our residual dataset. Later, we compute the hybrid *value* score by using the weighted ARIMA and LSTM models (according to the weights that we obtain in the weight learning part). Figure 4.5 presents that our LSTM model can predict the non-linear residual value scores of our dataset with a good accuracy. Blue line in the figure represents the residual values that we find after applying the ARIMA in our experiment data. For each user, we again separate 70% and 30% of the time interval value scores into the training and the testing datasets, respectively. The blue line in the figure represents the original value (i.e., openness-to-change) scores that are computed from our dataset over time. The green and red lines represent the prediction over the training and the testing data, respectively. We observe that our hybrid model can closely predict the non-linear dataset than that of the independent ARIMA and LSTM models. Note that, we randomly split the dataset into 30% and 70% for the learning weights and the training hybrid model, respectively. Splitting the dataset is somewhat similar to the cross validation where we learn from one dataset and apply on another dataset. Thus, we keep the training and the testing datasets separate for weight learning while building the hybrid model.

We show the value score of a single user in Figure 4.5. We predict the future value score of an individual user by using our hybrid model based on the training dataset that are

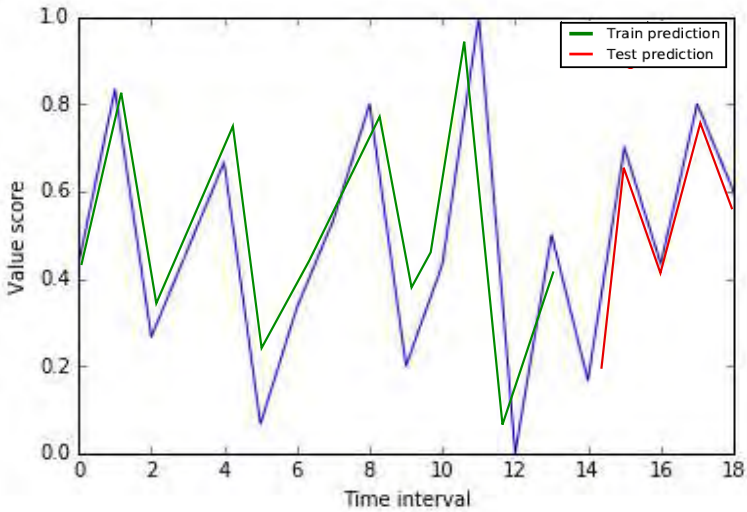

Figure 4.5: Predicting residual value for the *openness-to-change* of the user (LSTM).

constructed by his 70% of the time intervals. Therefore, we cannot show an aggregated value prediction model for the all users in a single figure. We segment our validation dataset into 6-months time interval. Therefore, we have limited number of time intervals for each user. However, since we use the LSTM based model in our hybrid approach, our technique can perform better with the small training dataset [98].

Result: After applying the hybrid value prediction model, we achieve a lower average RMSE score, 0.461, which is a substantial improvement than the previous RMSE. In particular, accuracy of the weighted hybrid model is increased by 41%, and 32.13% than that of the ARIMA and the LSTM models, respectively.

4.6 Evaluation

In this section, first we validate our hybrid model through a questionnaire among the users in real life. Then, we investigate how accurate our hybrid model can capture the value change. Later, we compare the accuracy of our hybrid model with other baseline techniques by using the chi-square ($\tilde{\chi}^2$) test. Finally, we check how the value change prediction results compatible with our real life decision making processes.

4.6.1 Validation through Questionnaire

In this subsection, we first extract users' statuses from 237 (discarding the 101 users' data of weight learning) Facebook users to validate them in real life. We divide the statuses of each user into time interval dataset. By following the experiment of Bardi et al. [17], we divide these statuses based on 6-months time interval. To get sufficient status updates and detection of every value changes, we find that the 6-months interval is suitable for our experiment. We select such a time interval which is close to the interval (i.e., 9 months) considered in the work of Bardi et al. [17] and has a substantial amount of intervals to train the time series model effectively. The minimum and average word counts for each of our time interval dataset are 63 and 1494, respectively. In a previous study of Celli et al. [39], it is also shown that a moderate personality prediction strength can be achieved from the *myPersonality* dataset (N = 250) with the minimum and average word count of 1 and 585.004, respectively.

Then, we predict the users' value scores from each of their time interval data by using our value prediction model (according to Section building models of values). We capture all the changes of value scores by using our hybrid time-series model. Later, we predict user's value score on a future date (i.e., one year later) with our hybrid time-series model. Based on the predicted value scores, we define three different labels: *low* (less than 0.4), *medium* (between 0.4 to 0.7), and *high* (greater than 0.7). On the future date, we conduct the PVQ test among these 237 users.

By following the study of Bardi et al. [17], we conduct the PVQ survey only a single trial among the validation users after 1 year and 15 days (October 25, 2017) when we predict the future (October 10, 2016) value scores. We predict the next two intervals (i.e., 6-months each) and conduct the validation in real life. In another study of Bible et al. [21], the authors only show the value change among the business major students between 2004 and 2010 which is also a single trail. Hence, we show the change of value among the validation users only for a single trail.

Next, we investigate how accurate our model can capture the change of *values* based on the past observations. Towards this direction, we compute correlation by using the *chi-square* ($\tilde{\chi}^2$) test in a single trial between the value levels (i.e., high, medium or low) obtained

Table 4.2: The Chi-square test result between the predicted transition and the PVQ test results for three levels of values. Comparison with other baseline techniques.

Values	**HMM**		**ARIMA**		**LSTM**		**Hybrid (ARIMA+LSTM)**	
	Chi-Sq. Value	Signif.	Chi-Sq. Value	Signif.	Chi-Sq. Value	Signif.	Chi-Sq. Value	Signif.
Self-trans.	3.69	0.489	8.01	0.091	7.53	0.112	12.11	0.007
Open.	7.85	0.128	8.47	0.089	9.15	0.059	15.21	0.004
Hedonism	4.98	0.379	5.80	0.251	6.61	0.143	7.31	0.139
Self-enhan.	2.84	0.682	3.17	0.521	4.41	0.382	5.11	0.259
Conserv.	2.42	0.679	7.25	0.124	8.73	0.081	14.31	0.004

from our hybrid prediction model and the PVQ test levels in real life. We get two categorical variables (the predicted value and the PVQ test levels) each with three different levels, which form a 3×3 contingency table. Table 4.2 shows significant correlation between the predicted value and the PVQ test levels for a single trial by using our hybrid model. According to Lancaster et al. [118], we obtain the significance level *p<0.05 for $\tilde{\chi}^2$>= 9.488 with the degrees of freedom (df) equal to 4 (based on the contingency table), where *df= (#rows-1)×(#col-1)*. By using our hybrid model, we observe that *self-transcendence*, *openness-to-change*, and *conservation* values are significant. Therefore, we reject the null hypothesis and infer that *values* have an effect over time. We also conclude that we can capture the change of *values* in real world [118].

4.6.2 Comparison with the Baselines

For the first baseline, we assume that the transition of *values* can be predictable with the HMM based technique. We select the HMM as a baseline because the model has been successful in analyzing and predicting time-series data in previous studies [85, 103]. For example, the model has been widely used in diverse applications including speech recognition, stock market prediction, and genomic sequence modeling. Then, we also predict our value transition by using the ARIMA model and the LSTM model independently.

Table 4.3: RMSE evaluation for all value dimensions by using different models.

Values	**Avg. RMSE using HMM**	**Avg. RMSE using ARIMA**	**Avg. RMSE using LSTM**	**Avg. RMSE using Hybrid**
Self-trans.	0.79	0.823	0.723	0.49
Open.	0.73	0.65	0.61	0.461
Hedonism	0.813	0.75	0.73	0.52
Self-enhan.	0.83	0.71	0.78	0.571
Conserv.	0.81	0.805	0.645	0.469

The HMM based value prediction (Baseline 1): We present the HMM based value transition model as the first baseline in our experiment. As we describe in Subsection *Validation through questionnaire*, we define the predicted value scores into three different labels: *low* (less than 0.4), *medium* (between 0.4 to 0.7), and *high* (greater than 0.7). We predict whether our value levels are changed (i.e., from *medium* to *high*) over time. The changes of value levels over a continuous time can be represented as a stochastic process with a probability distribution. Such probability distributions are typically determined by our built prediction model of *values* over each of the time-series data.

The HMM model is a formal foundation for constructing a stochastic model over a sequence of observations. The HMM invokes three different states, one for each of the three labels: low, medium and high. The Markov chain model predicts the next state depending on the current state. The HMM maintains hidden states also to capture other information to predict the next state. We assume that our hidden states can capture different unforeseeable circumstances, e.g., particular event, influence of friends, etc. that might influence to the change of our *values*. Thus, we select the HMM as a potential baseline model that can accurately deal with our problem scenario. We have computed the transition matrices of five value dimensions for these 237 validation users by using the R *depmixS4* [204] implementation.

ARIMA based value prediction (Baseline-2): We compute our value transition problem with the ARIMA model independently (according to Subsection *predicting value transition with ARIMA model*). We assume that the value transition can be captured by using only linear time-series technique.

LSTM based value prediction (Baseline-3): We again assume that value transition can be captured accurately by using a non-linear technique. To this end, we use the recurrent neural network based LSTM technique to predict the value change over time. We use similar technique that we used in Subsection *predicting value transition with LSTM model*.

4.6.3 Correlation of Value Change and User Preference

In this subsection, we investigate whether users' change of value priorities relates to their decision-making process in real life. In this experiment, we explore how users' movie watching preferences might be changed due to the change of value priorities. Motivated by the studies of Hsieh et al. [102], and Verplanken et al. [202], we hypothesize a link between the *openness-to-change* value with the movie watching preferences of a particular genre. Later, we compute correlation between the predicted value change and the movie watching preference on a future date.

Hypothesis H1: *High openness-to-change* value dimension has a strong connection with the sci-fi/adventurous movies. An individual who possesses high score in the *openness-to-change* value dimension is likely to give high importance on the futuristic view and active imagination. He likes to experience challenges and excitements in his life. Since sci-fi/adventurous movies contain new ideas and stimuli in the movie contents, it is likely that an individual with high score in the *openness-to-change* value prefers to watch those movies.

Movie change preference experiment: First, we collect a new dataset of 123 (male=70, female=53) active Facebook users through our application as of July 2016. Users are from the same ethnic group, but they are from different educational and professional background. The users are aged between 20 and 28 years. We prepare a questionnaire to find out the movie preferences of these users through interviews. All the recruited participants must have watched some selected movies to attend in the interview. Among the 123 Facebook users, we conduct a semi-structured interview during July 2016. We conduct the interview in different locations (e.g., restaurants, university library, etc.). Majority of the interviews are taken in the face-to-face settings. At the end of the interview, the users are compensated by a small gift. We also take few interviews through Skype for the users who stay in distant locations.

To prove the hypothesis (**H1**) in real life, we select two different movies, namely *Interstellar* (2014), and *Inception* (2010) from sci-fi/adventurous genre movies. *Interstellar* and *Inception* movies have 8.6, and 8.8 rating according to the imdb (Internet movie database) [2]. We confirm that users have watched these two movies. They are asked to rate these two movies in a likert scale of 1-5. In the likert scale, 1 is strongly disinterested, 2 is disinterested, 3 is neither disinterested nor interested, 4 is interested and 5 is strongly interested to watch a particular movie. Based on the answer of the likert scale, we normalize the score between 0 and 1. We use the *min-max* [89] normalization technique. We find that the lowest, average, and the highest rates are 0.59, 0.73, and 0.92, respectively. Though we offer these users to rate in a scale of 1-5, some of them rate in a floating point score such as 2.9, 4.6, etc. Thus, we find a few unusual values (e.g., 0.59) in the normalized score of 0-1.

Then, we segment users' Facebook statuses based on the 6-months time interval. In our dataset, we find 65, 38, and 20 users created their Facebook account in the years of 2007, 2008, and 2009, respectively. Then, we predict users' *openness-to-change* value score for each 6-months interval up to the year of 2016. From the predicted value scores of past several years, i.e., 8-10 years, we predict the *openness-to-change* value score for the future date (October 2017, after more than 2 time intervals) by using our hybrid model. Then, we identify users whose *openness-to-change* value dimension change significantly, i.e., from high to low, or from low to high. We find that value dimensions of 33 users change drastically within one year. Later, we conduct another semi-structured interview among these 33 users about their movie preferences on the future date (October 2017). We also select another two sci-fi/adventurous movies: *Mad Max: Fury Road* (2015), and *The Avengers* (2012). Both of these movies have the *imdb* rating 8.1. We also confirm that these selected 33 users have watched these two movies. Then, we again ask them to rate these two movies in a likert scale of 1-5. We observe that majority of these users rate these two movies differently than they rated sci-fi/adventurous movies on the past date (July 2016). We compute the Pearson's correlation test between the predicted value change and the average rate of these two movies among these 33 users. We find a moderate correlation ($r=0.364^*$, $p<0.0373$, where $N=33$) between these two variables which is significant. Thus, by using our hybrid

[2] http://www.imdb.com/

model, we can identify the change of our real-life preferences that are related to the change of *values*. Following the study of Hsieh et al. [102], we can also hypothesize between other *value* dimensions and genre of movie preferences. For simplicity, we only present a single hypothesis (**H1**) in this experiment.

4.7 Discussion

Our work is the first study to identify the change of *values* from the word usage of Facebook users. Figures 4.2 and 4.3 present the prediction of value change of a user in our dataset by using both ARIMA and LSTM models, respectively. Figure 4.5 presents the prediction of value change by using the proposed hybrid model by combining both of the ARIMA model and the LSTM model. Table 4.2 presents that our hybrid model can accurately capture the changes for *self-transcendence, openness-to-change,* and *conservation* values. We observe that the *openness-to-change* value score (7.85) is nearly close to the significant (9.88) for the HMM based technique. For other value dimensions, the HMM does not give any significant scores while computing the change of value score. On the other hand, we find that *self-enhancement* (8.01) and *openness-to-change* (8.47) values are also close to the significant when we evaluate the change of value scores by using the ARIMA model. Again, we observe that *openness-to-change* (9.15), and *conservation* (8.75) values are also close to the significant (9.88) score while we use the LSTM based technique. Finally, we see that *self-transcendence* (12.11), *openness-to-change* (15.21), and *conservation* (14.31) values show statistically significant during evaluation of the change of value scores by using our hybrid model. Table 4.3 presents that our hybrid model performs better than other value change prediction techniques. We find that the hybrid model achieves less average RMSE for all the value dimensions than that of other value change prediction models. For example, we observe that our hybrid model reduces the average RMSE for the *openness-to-change* value dimension by 58.35%, 41%, and 32.32% than that of HMM, ARIMA, and LSTM models, respectively.

Similar to the study of Chen et al. [43], we find weak prediction potential (according to Table 6.2) for the *self-enhancement* value (10.23%). Thus, our hybrid model cannot capture

the *self-enhancement* value change accurately. People with high score in *self-enhancement* value dimension have less propensity to share information in the social media frequently, as social media foster procrastination and distract from other activities [104]. Though these three baselines show better value change for few values, none of these techniques obtain significant $\tilde{\chi}^2$ score. Thus, these baselines cannot capture value changes accurately in real world.

We observe an interesting insight from our experiment, i.e., when the priority of a certain value goes high, the other value may go down. For example, an individual with high openness-to-change score is likely to possess a low conservation score. Similarly, a person with high self-transcendence score is likely to possess low hedonism and self-enhancement scores. In a well cited study of Golbeck et al. [80], the authors successfully predict personality from Facebook with a sample size of 279 Facebook users. Therefore, the size (N = 388) of our dataset is sufficient to predict the value dimensions from social media usage.

We observe that several value based socio-psychological studies [17, 138, 175] use PVQ to compute *values*. Researchers show that PVQ has good internal reliabilities with the original Schwartz's Value Survey (SVS) [177]. In contrast, Boyd et al. [34] show that value survey based approach has limitations, e.g., people may not truly represent themselves in a self report. Therefore, the authors use language based inference. Following the approach of Chen et al. [43], we also infer *values* from the languages by using a trained model which is built based on the ground truth surveys. However, Boyd et al. [34] use an open vocabulary based approach using the self-descriptive *values*, behavior, and SVS while inferring *values* from languages, whereas Chen et al. [43] use a closed vocabulary based approach by using PVQ only. We assume that both of the approaches are parallel in inferring value dimensions from languages and we choose the closed vocabulary based approach in this chapter. However, computing *values* by using the approach of Boyd et al. [34] is an interesting future direction to investigate the change of *values* over time. We build both of our value building and value transition models by using the dataset of Facebook. However, it would be also interesting to investigate how our approach performs with the datasets of other networks such as Twitter, and Reddit.

Few psychological studies [17, 28, 186] show that *values* and ethical behavior change

between the age of 15-25. Bardi et al. [17] describe that appropriate time to investigate the value change is adolescent, because we observe important changes such as biological, cognitive, and social in this stage of life. Steinberg et al. [186] show that during adolescent period, people adapt different psychological changes due to relation with parents, puberty, self-development, and peer relations. In another study, Blonigen et al. [28], the authors show that personality might be changed from the age 17 to 24. This age span represents a turbulent period of adjustment due to significant life changes. In a recent study, Tufft [197] also states that adulthood begins at the age of 25. Therefore, we also collect the dataset of Facebook users between the ages of 15 to 25 to investigate the change of *values*.

Our approach has also several limitations. We find limited prediction potential for *hedonism* and *self-enhancement* value dimensions according to Table 6.2. To predict users' value priority transition, users should be regular in Facebook, so that their continuous time-series data are available for capturing the value priority change. During building the value prediction model from LIWC categories, we ignore few important words such as *selfie, inbox, grrr, bullshit*, etc., since LIWC cannot capture these words. These words might contribute in predicting *values* accurately from languages. In our value change and movie preference experiment (Section 4.6.3), we could improve our correlation results by eliminating covariance of the average movie scores.

4.8 Summary

In this chapter, we have investigated whether the change of value priorities can be identified from social network word usages. We have proposed a weighted hybrid time-series based machine learning model to capture the change of *values* of a user. We have built our value model with 388 Facebook users and validated the efficacy of the proposed technique through questionnaire with 237 Facebook users in real life. In our experiments, we have found that our hybrid model accurately captures the value priority changes from the social network usage and achieves significantly higher accuracy than that of HMM, ARIMA, and LSTM based models. We have also presented an experiment on how the change of *values* relates to our decision-making process in real life. To the best of our knowledge, this is the first work

that identifies users' value change from their social media interactions. The outcome of this work can be applied to many practical applications such as predicting transition of business strategies, change of colleges or degrees, detection of life style changes, etc.

In this chapter, we have investigated how values change over time from users' social media interactions. In the next chapter, we will investigate by which role(s) two users are connected with each other from their interactions in an egocentric social network.

Chapter 5

Predicting Social Role Identities from Social Media Interactions

In this chapter, we investigate a part of our third research problem for this dissertation. We identify different role identities based on users' psycholinguistic signals used in SNS usage. We describe methodology and data collection process in Sections 5.2, and 5.3, respectively. Then, we describe taxonomy of different role identities in Section 5.4, and discuss the preparing training dataset in Section 5.5. We present independent and hybrid model building processes to identify role identities in Sections 5.7, and 5.8. Finally, we evaluate our hybrid model with different dataset in Section 5.9.

5.1 Introduction

A person plays multiple roles in a society, e.g., a person has a *family member* role identity with his wife, *professional member* role identity with his colleagues and *academic member* role identity with his class fellows. These roles are important features that characterize the behaving patterns among the members in a society. Depending on social identities and situations, the roles guide us to behave differently with different people. For example, we respect teachers, care family members, share our feelings with friends and build positive relationship with colleagues [22]. In this research, we are the first to develop techniques to automatically derive a user's role identities with other connected users from their psycholinguistic attributes obtained from their social media interactions.

Social networking sites (SNS) have been transformed into virtual societies where users express their feelings, share opinions, and socialize with friends, families, and co-workers. Though SNS have put their efforts to make their sites as close as possible to the real life interactions, users in these SNS are still connected with each other with so called *friends/followers* relationships which are not representatives of real life role identities (e.g., father and son are friends to each other in Facebook). However, we observe that interactions among members inside SNS vary depending on their role identities in real life. For example, we may post, comment, and tag a class fellow instantaneously with a sarcastic meme, whereas we are likely to give advice, share affection and talk about health concerns with a family member in Facebook [37]. Thus, it might be possible to identify individual role identities among the SNS members from their word usage patterns.

A number of studies have been conducted to identify role identities among the members in SNS. Authors in [192, 193, 206, 219] largely identify users' role identities by analyzing features such as interaction frequency and time, text, etc. None of these works identify the role identities between ego and alters by analyzing the patterns of word use of Facebook comments. Note that Facebook is an egocentric social network where a person (an ego) is connected with other friends (the alters) [72].

In this chapter, we primarily connect our work with an important theoretical notion of psychology: *social identity theory* [191]. According to the notion, we follow different norms while we interact with the people of different groups. We consider that members of each role construct a particular group. Therefore, members in a group are likely to exchange identical psychological signals and cues among themselves during interaction [191]. Motivated by the above findings, in this chapter, we investigate how people interact with others differently, and identify role identities of two users from their social interactions.

Understanding the role identities from social network word usage can benefit a number of real life applications. For example, one may want to find out the influence by a specific role member (e.g., family members can persuade to take a decision). One may also be interested in identifying how a person is intimate with particular role members (e.g., a person feels free to discuss his personal issues with his friend). Again, it is possible to investigate the changes of values of a person with respect to different role identities, e.g., *openness* with the

students and *conservative* with the family members. Marketers, and third party apps can be benefited by automatically identifying role identities in real life. Recommender systems can also produce customized suggestions to the SNS users for different preferences and behaviors of their Facebook friends based on different role identities.

To predict the role identities of user from her Facebook comments, we take the following strategies. First, we collect all the comments of alters (i.e., Facebook friends) and ego that they made on different objects (i.e., statuses, photos, and check-ins, etc.) of each other. We extract these comment data from a total of 351 different Facebook users. We consider that each Facebook user possesses an individual ego network which consists of her Facebook friends. Later, we separate all the comments for each ego-alter/alter-ego friendship connections. We consider the *comments* for a total of 1,135 ego-alter friendship connections in our experiment.

After collecting the comments, we explain the taxonomy of the role identities to the egos. Then, egos manually annotate each set of comments for ego-alter friendship connections by a *role identity*, i.e., *academic member, family member, professional member, friend*, and *acquaintance*. Then, we perform a psycholinguistic analysis on each set of *comment* dyads of each ego-alter friendship connection by using both closed (i.e., LIWC) and open (i.e., MEH) vocabulary based techniques [152, 173]. Later, we compute correlation analysis between LIWC (Linguistic Inquiry and Word Count) categories of words and role identities. We also compute correlation between MEH (Meaning Extraction Helper) words and role identities. Next, we build classification models to predict the role identity between an alter and ego by using these linguistic features.

We find that closed vocabulary based classifiers perform better for *family member* and *acquaintance* role identities than that of open vocabulary based model. We also observe that open vocabulary based classifiers outperform closed vocabulary based approach for *academic member, professional member*, and *friend* role identities. Hence, we combine both of these vocabulary based classifiers and develop a hybrid weighted machine learning model to predict final social role identities. Our hybrid classifiers achieve strong prediction potential (i.e., average area under the curve (AUC) is 0.687) to identify social role identities from social network word usage. We find that the hybrid model improves the prediction accuracy

significantly than that of the independent models (i.e., LIWC and MEH). Finally, we validate the prediction potential of social role identities by using a dataset of Facebook and a dataset of Twitter. We observe that our hybrid model predicts the role identities more accurately from Facebook dataset than those of Twitter dataset. In summary, we have the following contributions:

- Our work is the first study that identifies social role identities from Facebook comments by analyzing users' psycholinguistic attributes.
- We show correlations between users' word use and social role identities in Facebook from their comments.
- We develop a weighted hybrid model that shows a high prediction accuracy while identifying the role identities from social network usage.
- We validate our technique by using the datasets of both Facebook and Twitter.

5.2 Methodology

To predict the role identities of ego-alter/alter-ego friendship connections, we perform the following steps:

- *Data extraction.* We extract users' (alters) comments on different objects of an ego through a Facebook application.
- *Data Preprocessing.* We process our data by discarding irrelevant fields such as creation time, curly brackets *("{")*, and special characters (e.g., #). We replace *emoticons* with a corresponding text of emotional sense (e.g., happy, sad, angry, haha, etc.) [213].
- *Preparing training dataset.* We prepare the training dataset from the ego-alter friendship connections where the egos annotate each friendship connection by using a role label. We consider bi-directional (ego-alter/alter-ego) comment dyads [1] as both set of comments reveal users' role identities.

[1]For the sake of brevity, we would use the term *comment dyads* to refer *bi-directional comment dyads*, i.e., ego-alter and alter-ego, throughout the chapter.

- *Correlation analysis.* We identify the linguistic features (independent variables) from both the LIWC words and the MEH words that are correlated with different role identities (dependent variable).

- *Building classification model.* We build classification models to predict the role identities by using significant linguistic features of both LIWC and MEH words. Later, we build a weighted hybrid classification model by using the independent models (i.e., LIWC and MEH).

- *Validation in real life.* We validate our classification results in real life by using a dataset of Facebook, and a dataset of Twitter.

5.3 Data Collection

We have collected 351 Facebook ego networks for building the classification models to predict the role identities. We have collected the interactions of a total of 1,135 different ego-alter friendship connections to build our models. Then, we have also collected 30 ego networks from Facebook, and 30 ego networks from Twitter to validate our prediction model in real life.

Dataset for building classification model: We have invited 530 Facebook users to collect their comments on different objects through posts on Facebook, relevant mailing lists and word of mouth technique. Each Facebook user possesses an individual egocentric network, where she is an *ego* and her Facebook friends are *alters*. To collect user comments, we create a Facebook application which accesses all the comments of alters and ego.

Among 530 Facebook users, a total of 391 users authenticated the application to read their time-lines. Among these 391 users, a total of 351 (male=193, female=158) users give us authorization that they want to apply their data in building automatic role identification model. These users are members of different occupations such as students, teachers, doctors, engineers, bankers, and government service holders. They are aged between 15 and 45 years. We collect user data (i.e., comments on statuses, photos, shared-links, and check-ins) of alters/ego from 351 ego networks. We collect the data starting from the users' account creation date to August, 31, 2017. Each of 351 Facebook users has an average of 728 Facebook

friends. We collect the user comments of a total of 255,528 (351×728) Facebook ego-alter connections.

According to the study [12], we find out top 150 *active network members* who actively maintain the interactions (i.e., the person who comments) with the respective ego. To find out 150 active users, we use *NodeXL* [2], an open-source network analysis and visualization tool. The rest of the users (excluding the *active network members*) just create the Facebook connections. Table 5.1 shows the statistics of our collected dataset. Each of 351 users (ego) possesses an individual egocentric network and they have diversity in their occupation, age, and gender. We observe that identical diversity in occupation, age, and gender exists among the alters when we extract their individual egocentric network.

Dataset for validation in real life: For validating the social role identities in real life, we collect two datasets of egocentric networks, one from Facebook, and another from Twitter. We invite a new set of users of 45 Facebook users. We collect a dataset of 30 (male=18, female=12) active Facebook users by using our Facebook application to validate the role identities in real life. These users are aged between 17 and 43 years, and they are the members of student and professional community. We collect the data of alters/ego from 30 ego networks. We collect data starting from the users' account creation date to September, 20, 2017. Each of 30 Facebook users has an average of 815 Facebook friends. We collect the user comments of a total of 24,450 Facebook ego-alter connections.

Next, we randomly select 30 Twitter users who have ego networks of moderate size (i.e., 250-500). We select a total of 30 Twitter users, since the size of our Facebook validation dataset is also 30. Then, we collect their ego networks (*follower/following* list) by using the python *tweepy* [3] implementation package. In these ego networks, the number of average following/followers is 395. We extract *tweets* of the users of these 30 egocentric networks as of September 25, 2017. Then, we parse users' *reply* dyads of each ego-alter/alter-ego connections, since user can comment other users by replying in Twitter. We have a total of 11,850 individual ego-alter reply dyads in our dataset.

[2] https://nodexl.codeplex.com

[3] http://www.tweepy.org/

Table 5.1: Corpus statistics

# of ego networks	351
Average # of alters in each network	728
# of active alters in each network	150
Average # of status of ego	128.37
Average # of photos of ego	172.21
Average # of check-ins of ego	27.43
Average # of shared-links of ego	29.17
Average # comments on status	7.11
Average # comments on photos	9.03
Average # comments on check-ins	5.91
Average # comment on shared-links	3.15

5.4 Defining Role Identity

In our experiment, we predict five different role identities, namely *family member, academic member, professional member, friend,* and *acquaintance*, from dyadic comments of ego-alter, and alter-ego. In this section, we describe the nature of different role identities. Then, we justify which role identity can be assigned to which relation of Facebook. Based on the existing studies of social media sites [35, 37, 58, 66], we explain the taxonomy of the role identities to our egos and they *label* the relationships to their alters accordingly. Later, we use the *label* corresponding to the role identity to prepare our training dataset.

Following the study [37], *egos* label family members (i.e., father, mother, siblings, wife, children, etc.) and extended family members (i.e., maternal or paternal uncle, cousin, etc.) with the *family member* role identity.

According to the study [96, 195], many faculty members interact with their students in Facebook. Since students and teachers get familiar with each other in academia, *egos* label the *student-teacher* relationships also as *academic member* role identity. Egos also label *office secretary*, *research staff*, *senior university fellow*, etc. as an *academic member* role identity.

People connect with their coworkers for sharing personal issues, advancing career, and

campaigning a project [58]. In our experiment, egos label users as *professional role* identity who work in the same workplace. In our dataset, we find a few people are coworkers in an academia. Since these people have the colleague-colleague relationships, *egos* label these relationships as *professional role* identity.

A friend is a partner who maintains a high-level of pro-social behavior, and intimacy and shows low-level of conflict, and rivalry [20]. We might make a friend in a number of ways, i.e., *family member, academic member, professional member* and *acquaintance*. In our case, an ego labels an alter as a *friend* if he/she has strong interpersonal bond with the alter; e.g., two users studying in the same class may label each other as *friends*.

Bryant et al. [35] describe that *acquaintance* relationship exists among people who vaguely know each other. They experience with little intimacy. We also find a few people who are connected on Facebook, but they do not know each other well. For example, a neighbor whom we see near staircase, a friend of my friend, or the superstore manager whom we talk hardly. Sometimes, we get/send friend request from/to these people by the automated Facebook suggestions, egos annotate these people as *acquaintance role* identity.

5.5 Preparing Training Dataset

In this section, we find out the active alters who frequently interact with the ego. Then, we compute the psycholinguistic categories of users' comments dyads by using the LIWC and the MEH. Later, these egos assign a label for each of the dyads with an appropriate role identity.

In our dataset, we have 351 different ego networks. We allow maximum 150 alters who are the most active with ego (active network [12]). According to the evolutionary psychologist, there is a limit on the number that an individual can actively maintain a social group size due to the cognitive constraint of human brain [64]. Dunbar predicts the size of the human social group is around 150 - the number is known as *Dunbar's number* [220]. Several studies [13, 81] also validate the linkage between the *Dunbar's number* and the activity patten of human brain in a social network.

According to the previous studies [13,81], we find a total of 52,650 (351×150) individual

connections between ego-alters. Then, we conduct double filtering to select particular ego-alter connections from these 52,650 connections to build our training dataset. In our first phase of filtering, we store both comments of user u (as an alter) to the objects of user v (as an ego) and comments of user v (as an alter) to the objects of user u (as an ego). In the second phase of filtering, we discard the connections where the number of comment dyads on different objects is less than 30 times. We empirically find that at least 30 comments (which have on an average 100 words) are necessary for a training instance to build a reasonable machine learning model to identify the role identities.

We find 1,265 ego-alter connections that contain available comments to build our classification model. Since we predict users' role identities automatically, we take into account the invasion of privacy as a valid concern. Therefore, we take explicit permission from the 351 egos for user modeling or any personalization. When we collect the comment dyads from different objects of 351 egos, we are not aware beforehand that the data of which alters will be found in the comment collection. Hence, we could not take the explicit permission before collecting data from the rest of the 914 (1,265-351) users. However, we request the explicit permission from these 914 users later by using direct Facebook messages between September 2017 and October 2017. Among these 914 users, 130 users do not give us permission and they ask even to delete their data. Therefore, we consider the comments of 784 (914-130) alters. Finally, we have comment dyads of a total of 1,135 (351+784) ego-alter connections to build our classification model.

We create an individual file on each ego-alter connection containing the comments on all objects. Thus, we have a total of 1,135 data files that contain all the comments of 351 Facebook ego networks. Then, we compute linguistic analysis of these 1,135 data files by using both LIWC and MEH.

We first use LIWClite7 [152] a student version of LIWC tool. LIWC analyzes 80 different features of text in different categories. The categories are linguistic processes (word count, words longer than 6 letters, etc.), psychological processes, personal concerns (work, leisure, etc. related words) and spoken categories (assent, non-influences, etc.). The psychological processes is divided into five categories. It includes social process (words related to family, friends etc.), affective process (words related to positive emotion, negative emotion,

etc.), cognitive process (insight, discrepancy, etc. related words), perceptual process (see, hear, etc. related words) and biological process (body, health, etc. related words).

Then, we also compute the linguistic analysis of these 1,135 data files by using MEH. MEH [33] is another analysis tool unlike content coding software (e.g., LIWC) which is highly dynamic and flexible for extracting words from text data. MEH first removes *stop-words* and then apply *lemmatization* technique for text processing. The tool can generate output by producing topic modeling, computing 'n-gram' words, calculating word frequencies, etc. from large corpus of dataset. MEH analyzes available words from the corpus dynamically, and therefore authors [33] name the approach as *open vocabulary* based approach.

We build our training dataset where the linguistic features derived from each data files are input and labels of the role identities are output. The instances of our training dataset contain 5 different classes: *Family member, Academic member, Professional member, Friends,* and *Acquaintance*. Table 5.2 presents that our training dataset contains 252, 216, 140, 318, and 209 instances from *academic members, professional members, family members, friends,* and *acquaintances* role members, respectively. We observe that the *friends* who are aged between 15 and 28 years, remain very active through Facebook, as they go through different stages of life such as high school, college, and university. On the other hand, we find the minimum instances for family (i.e., parents, spouse) role identities, since the *family* members communicate less through Facebook. We observe that extended family members (i.e., cousin) interact more through Facebook than that of family members who reside in the same house. In our dataset, we observe that few common words such as *congratulations*, *thanks*, *welcome*, etc. frequently appear in the comment dyads for all of the role identities. We see that the *acquaintance* role members use majority of the time these common words, because they have less experience to interact with personal and intimate issues. The *friend* role members sometimes also use these common words with issues such as assignment, examination, class test, etc. The *academic* and the *professional* role members use these common words with other topics such as assignment, examination, project, etc. However, *family members* hardly use these generic words. Therefore, we find that there is a variation of frequency in usage of common words depending on the role identities, so the frequency of these words

Table 5.2: Number of training instances of different role identities.

Role Identities	**# of instances**
Academic member	252
Professional member	216
Family member	140
Friend	318
Acquaintances	209

create distinction while we build our training dataset.

5.6 Correlation Analysis

In this section, we compute correlation between users' psycholinguistic features of comment dyads and social role identities. First, we find out the correlation by using the closed vocabulary based approach, LIWC. Then, we compute the correlation by using the open vocabulary based approach, MEH.

We find out relevant LIWC categories to build classification models of the social role identities. Towards this, we compute a multi-level discriminant analysis by using SPSS [4] where selected LIWC categories are the predictors and role identity is the dependent variable. Table 5.3 shows the Fisher's linear discriminant analysis (LDA) coefficient between LIWC categories of comment dyads and social identity roles. Fisher's LDA is a technique that finds the association between the continuous independent variables and the categorical dependent variable. Fisher's LDA function scores are proportional to the coefficients of multiple linear regression with dependent variables, i.e., social roles. Therefore, predictors with larger (>1.0) scores are better predictors. These scores are helpful in deciding which variable affects more during building the classification models [51].

We also identify correlated predictors in the open vocabulary based approach by using Fisher's LDA. MEH extracts hundreds of words from ego-alter comment dyads. We find that comment dyads have different set of *'1-g'* words (i.e. unigrams). In our dataset, we find

[4] https://www.ibm.com/products/spss-statistics

Table 5.3: Fisher's linear discriminant function coefficients between LIWC variables and role identities. These LIWC variables have significant correlation with five different role identities. [for significance level: *p <0.05, **p <0.01]

Category	**Example words**	**Family member**	**Academic member**	**Professional member**	**Friend**	**Acquaintance**
WC		-0.051	16.56**	17.65**	55.13**	-8.164**
Unique		43.12**	21.86**	-1.22	33.45**	-0.991
Dic		0.064	2.03*	53.13**	0.22	0.164
pronoun	I, them, itself	0.027	-0.119	0.155	0.19	0.054
you	You, your, thou	0.063	35.17**	-7.702**	0.12	0.158
ipron	It, it's, those	0.113	-0.044	-0.133	0.24	0.091
article	A, an, the	0.027	2.053*	-0.175	0.16	17.34**
verb	Walk, went, see	-0.025	-0.062	0.021	36.83**	29.12**
past	was, had, got	-0.023	70.98**	74.32**	62.98**	12.54**
present	Is, does, hear	-0.038	28.02**	26.68**	-0.27	0.124
swear	shit, fuck, hell	0.044	25.23**	0.140	0.17	0.069
social	Mate, talk, they, child	0.117	15.4**	18.21**	0.052	0.154
friend	friend, girlfriend, neighbor	0.039	56.70**	48.90**	67.51**	29.12**
humans	people, guy, man	2.15*	0.13	0.56	0.22	0.128
affect	happy, cried, abandon	59.12**	0.83	0.34	0.11	-0.021
posemo	love, nice, sweet	0.033	10.12**	0.15	12.23**	27.18**
negemo	hurt, ugly, nasty	0.129	23.14**	2.11*	52.31**	0.125
anx	worry, crazy, awkward	63.11**	0.239	-0.020	0.69	-0.074
anger	hate, kill, annoyed	13.31**	8.76**	0.127	21.68**	0.028
tentat	maybe, perhaps, guess	17.64**	0.18	0.157	0.09	0.134
certain	all, always, never	43.14**	0.08	0.076	0.012	0.051
feel	feel, soft, hard	1.90	11.96**	8.32**	15.71**	0.152
leisure	game, movie, music	-0.043	0.11	0.11	17.38**	-0.120
money	audit, cash, owe	-0.034	0.12	27.057**	45.32**	-0.085
achieve	earn, hero, win	0.005	51.048**	47.02**	41.21**	0.175

Table 5.4: The top 10 words and frequencies during interaction among the members of different roles in Facebook.

Academic	**Freq.**	**Professional**	**Freq.**	**Family**	**Freq.**	**Friend**	**Freq**	**Acquaintance**	**Freq.**
friend	3.39	greeting	1.50	god	4.01	treat	8.33	birthday	1.41
time	2.27	appear	1.33	everyday	2.23	shit	6.67	happy	1.01
long	1.55	inform	1.18	when	1.33	omg	5.23	awesome	0.99
greeting	1.10	boss	1.17	prayer	1.33	grrr	1.50	best	0.77
grade	0.93	job	1.08	lie	0.80	selfie	1.37	car	0.46
watch	0.75	friend	0.90	talk	0.77	travel	1.17	day	0.46
picture	0.62	interest	0.85	baby	0.77	facebook	0.93	beautiful	0.40
assignment	0.57	grrr	0.80	adore	0.69	inbox	0.76	safe	0.35
write	0.46	change	0.70	uncle	0.67	awesome	0.65	enjoy	0.29
faculty	0.38	http	0.65	safe	0.65	crap	0.40	fun	0.20

455, 370, 291, 610, 115 unique *'1-g'* words for academic, professional, family, friend, and acquaintance role identities, respectively. We also find that 93, 105, 75, 137, 77 unique *'1-g'* words are correlated for academic, professional, family, friends, and acquaintance member role identities, respectively. Due to shortage of space, we do not put the MEH correlation table in this chapter. However, Table 5.4 presents the top 10 words and frequencies of these words during interaction among the members of different role identities.

5.7 Building Model of the Role Identities

We identify role identities from the comment dyads with two different machine learning models: i) closed vocabulary (i.e., LIWC)-based model [80], and ii) open vocabulary (i.e., *MEH*)-based model [173]. Then, we compare the strength of different classifiers to show which model better predicts role identities from the comment dyads of ego and alters.

5.7.1 Classification with Closed Vocabulary based Model

In this subsection, first we use the significant LIWC features as independent variables and the role identity as dependent variable. Then, we run our training dataset with different classifiers that include logistic regression, Support vector machine (SVM), naive Bayesian, Adaboost,

Table 5.5: Best classifiers of different role identities by using the closed vocabulary based approach.

Role identities	**Highest AUC classifiers**	**AUC**	**TPR**	**FPR**
Academic member	Naive Bayes	0.627	0.614	0.293
Professional member	SVM	0.618	0.611	0.369
Family member	RepTree	0.599	0.582	0.448
Friend	RepTree	0.593	0.60	0.398
Acquaintance	Naive Bayes	0.609	0.615	0.411

Random Forest and RepTree classifiers using WEKA machine learning toolkit [87]. We use these six classifiers to predict the role identities from the significant LIWC categories. We conduct our experiment with a 10-fold cross-validation with 10 iterations. Table 5.5 presents the best classifiers for five class labels. Table also shows true positive rate (TPR), false positive rate (FPR) and area under the ROC curve (AUC) to compute each of the role identities by different classifiers [68].

To predict these five class labels, we have three different classifiers that produce the best AUC scores (according to Table 5.5). We use *ZeroR* classifier as baseline. We find the weakest AUC score (0.232) for the *family member* role identity and the strongest AUC score (0.398) for the *academic member* role identity by using our baseline classifier. We also find 0.318, 0.243, and 0.281 AUC scores by using our baseline classifier for *professional, friend,* and *acquaintance* role identities, respectively. For other classifiers, we find that classifier (Naive Bayes) for the academic member role identity achieves the highest accuracy (AUC-0.627) while classifier (RepTree) for the *friend* role identity achieves the weakest (AUC-0.593) accuracy among all the role identity classes. We observe that our closed vocabulary based classifiers achieve better accuracy than random chances to classify role identities from ego-alter comment dyads.

5.7.2 Classification with Open Vocabulary based Model

In this subsection, we again train our dataset with the open vocabulary based approach by using *MEH*. The open vocabulary based approach does not categorize similar words into

Table 5.6: Best classifiers of different role identities by using the open vocabulary based approach.

Role identities	**Highest AUC classifiers**	**AUC**	**TPR**	**FPR**
Academic member	RepTree	0.67	0.673	0.321
Professional member	Naive Bayes	0.65	0.648	0.372
Family member	RandomTree	0.567	0.562	0.441
Friend	RepTree	0.69	0.673	0.324
Acquaintance	RandomTree	0.58	0.5911	0.397

groups. For example, words such as *adult, baby, boy*, etc. do not consider as an individual category of words whereas LIWC considers these words as a group (*human* category). In the open vocabulary based approach, these words act as individual predictors. Thus, we find that a large number of independent variables are collinear with each other [139]. To identify the collinear variables, we identify variance inflation factor (VIF) by using SPSS. Then, by removing the independent variables with large VIFs($>$10), we apply classification techniques by using our dataset. This technique reduces the coefficient to a low value or zero, thus the model does not get over-fitted. We conduct our experiment with a 10-fold cross-validation with 10 iterations.

We use correlated unique '1-g' words as predictors in open vocabulary based models. Table 5.6 presents the best classifiers to predict our role identities by using open vocabulary based models. We find that classifier (RepTree) for *friend* role identity achieves the highest accuracy (AUC-0.69) while classifier (RandomTree) for *family member* role identity achieves the weakest (AUC-0.567) accuracy among all of the role identity classes.

5.8 Model Selection

In this section, we select an appropriate machine learning model to predict the role identities from ego-alter comment dyads. In the study [173], we find that LIWC suffers less expressiveness due to fixed priori of words. In SNS, people use diverse set of words such as local dialects, buzz words (i.e., *selfie*). LIWC cannot capture these words, therefore we fail to

extract few important cues to identify role identities from our comment dyads.

In contrast, open vocabulary based approach considers these words (i.e., *selfie*, *grrr, inbox, facebook*) during linguistic analysis. From Tables 5.5 and 5.6, we observe that none of the independent models have sufficient prediction potential to predict all the five role identities completely. We find that the closed vocabulary based classifiers achieve better prediction potential for *family* and *acquaintance* role identity classification while the open vocabulary based classifiers show better result for classification of *friend, academic* and *professional* role identities. We find that people interact more frequently with *friend, academic* and *professional* members in Facebook than that of *family members* and *acquaintances*. We also observe that *friends, academic* and *professional* members frequently use the non-dictionary words such as *grrr, omg, lol, inbox, http, facebook,* etc. during interaction with each other.

Since each of the closed and open vocabulary based models contributes to compute few role identities effectively, we combine these two models to build better classification model. MEH can also capture phrase level words such as '2-g' words, and '3-g' words. On the other hand, LIWC only analyzes word category, i.e., '1-g' words. We consider that both LIWC and MEH are parallel in inferring the role identities from word use, we ignore the phrase level words for the open vocabulary-based approach in our experiment.

Motivated by the previous studies [106,158], we can obtain better prediction potential by combining different experts. However, we prioritize these machine learning models based on their performance, as we predict the social role identities by using these two models with different accuracies. Performance of an individual classifier represents the weight of the best machine learning model. To combine open and closed vocabulary models, we perform the following two steps: i) computing weights for each of the vocabulary based models, and ii) combining these two vocabulary based models with a weighted linear hybrid technique.

5.8.1 Learning Weights

In this subsection, we determine the weights of different machine learning models to predict a unified social role identities. To this end, we build the classification models by using the data of 340 (30% of the total dataset) instances of social role identities. To compute the weights, we take two vocabulary based machine learning models as input and the best AUC

Table 5.7: Weights (AUC of the best classifiers) derived from the weighted training dataset.

Vocab. based models	**AUC of the best classifiers**				
	Academ.	**Profess.**	**Fam.**	**Frnd.**	**Acquant.**
Closed Vocab.	0.623	0.602	0.611	0.603	0.637
Open Vocab.	0.667	0.64	0.581	0.672	0.602

Table 5.8: Classification result to identify role identities by using both vocabulary based machine learning models.

Role identities	**Best classifier**	**AUC**	**TPR**	**TNR**
Academic member	Naive Bayes	0.71	0.72	0.27
Professional member	RandomForest	0.699	0.71	0.33
Family member	RepTree	0.635	0.60	0.37
Friend	RepTree	0.723	0.697	0.359
Acquaintance	RandomForest	0.67	0.651	0.423

score of the social role identity classifier as output. We conduct our experiment with a 10-fold cross-validation with 10 iterations. Table 8.3 presents the best AUC scores that are computed with different classifiers to predict the social role identities by using 30% of the dataset. We find that the open vocabulary model achieves higher AUC scores for *academic* (0.67), *professional* (0.65), and *friend* (0.69) role identities than that of the closed vocabulary model. In contrast, we observe that closed vocabulary model obtains higher AUC scores for *family* (0.599) and *acquaintance* (0.609) role identities than that of open vocabulary model.

5.8.2 A weighted Linear Hybrid Model

In this subsection, we build a weighted linear hybrid model from previous machine learning models of 795 (70% of the total dataset) instances of ego-alter comment dyads. We have built different classifiers from closed and open vocabulary based models to predict users' role identities that are described in previous sections. Since we train different classifiers that produce weights, we compute the weighted linear hybrid score using the weights (AUC scores of the best classifiers) in Table 8.3. Finally, we build our weighted linear hybrid model

Table 5.9: The *Chi-square* test result between the predicted and the real role identities. Comparison with the baseline models.

Facebook dataset						
Role Identity	**Open.**		**Closed.**		**Hybrid (Open V.+ Closed V.)**	
	Chi-Sq .	Sig.	Chi-Sq.	Sig.	Chi-Sq.	Sig.
Academic member	25.45	0.062	25.41	0.062	**27.92**	0.032
Professional member	**26.59**	0.046	**26.45**	0.048	**28.73**	0.025
Family member	24.61	0.077	23.49	0.101	**26.65**	0.045
Friend	26.09	0.052	24.90	0.071	**27.40**	0.037
Acquaintance	24.73	0.074	24.52	0.078	**26.73**	0.044
Twitter dataset						
Academic member	24.87	0.072	22.65	0.123	**26.74**	0.044
Professional member	23.17	0.109	23.57	0.099	**26.69**	0.045
Family member	19.13	0.261	17.27	0.368	21.03	0.177
Friend	24.98	0.070	23.88	0.092	25.88	0.055
Acquaintance	23.97	0.090	21.59	0.156	26.11	0.052

using the weights generated from another dataset, so that our models do not get over-fitted. Table 8.4 presents the classification results of different role identities by using the hybrid model of two machine learning models. We observe that the average AUC of our classifier is 68.7%, and the baseline accuracy is 53.1% by using *ZeroR* classifier. We also observe that our classifiers largely outperform the random baseline. The accuracy of our random baseline is 20%. We find that our hybrid classifiers achieve higher accuracy than that of independent closed and open vocabulary based classifiers.

5.9 Evaluation

In this step of evaluation, we assume that each of the independent models as a baseline model. We compare our hybrid model with different baseline models by using a dataset of 30 Facebook users and a dataset of 30 Twitter users. Then we compute the statistical significance test ($\tilde{\chi}^2$) to validate the prediction accuracy in real life.

5.9.1 Ground Truth Extraction

To measure the accuracy of our technique, we produce the ground truths of role identities for the validation datasets of Facebook, and Twitter. We communicate with the 30 egos of Facebook ego networks through phone and Facebook direct message to label the ego-alter connections according to the role taxonomy. After double filtering (following Section 5.5), we finally retain a total of 230 ego-alter connections among 24,450 (815×30) Facebook connections to validate the role identities in real life. We also select 30 Twitter users who also help us to label the role identities in real life. Then, we collect their Twitter follower/following networks. Later, we also crawl users' tweets by using the python *tweepy* implementation package. We assume that an interaction between two users occurred if we find a reply (response to a previous tweet) between them. We consider only the bidirectional links because we assume reply dyads of ego-alter and alter-ego in Twitter are similar to the Facebook comment dyads. After double filtering, we retain a total of 272 individual connections among 11,850 connections to validate the role identities.

5.9.2 Validation

In this subsection, we investigate the similarity between the predicted role identities by using our hybrid model and the real role identities. We also predict the role identities of the validation datasets by using our independent models. Then, we compute the *chi-square* ($\tilde{\chi}^2$) test to find the correlation between the predicted role identities obtained from our models and the annotated role identities. We get two categorical variables (predicted and annotated role identities) each with five different types, which form a 5×5 contingency table. Table 7.12 shows the significant correlations between the predicted and annotated role identities. According to Lancaster et al. [119], we obtain the significance level *p<0.05 for $\tilde{\chi}^2$>= 26.30 with the degrees of freedom (df) equal to 16 (based on contingency table), where *df= (#rows-1)×(#col-1)*. Table 7.12 presents that our hybrid model (closed + open vocabulary based models) can capture the role identities more accurately in real life than that of independent models.

5.10 Discussion

Our work is the first study that identifies five key social role identities from the social media interaction by using users' psycholinguistic features, i.e., word usage pattern.

LIWC models: We observe that Table 5.3 presents the correlated feature subsets with different social role identities by using LIWC categories of words. We find that our discovered feature subsets are quite intuitive. We see that *unique, human, affect, anx,* and *certain* categories of LIWC words have correlations with *family member* role identity. Again, we notice that *article, past, present, swear, social, feel,* and *achieve* LIWC categories of words are relevant features for identifying the *academic role* identity. We also find that *social, friend, negemo, feel, money,* and *achieve* LIWC categories of words are important features to predict the *professional role* identity. We observe that *friend, posemo, negemo, money,* and *achieve* LIWC categories of words have strong correlations with *friend* role identity. We find that *article, past, friend,* and *posemo* LIWC category of words are correlated features to predict the *acquaintance* role identity. Table 5.5 presents the classification result to predict social role identities by using LIWC categories of words. We find that *academic member* role identity shows the highest (62.7%) and *friend* role identity shows the weakest (59.3%) AUC scores. In Facebook, *friends* are likely to use *non-dictionary, shortened,* and *abbreviated* English words such as *omg, brb, tia, asap, wc,* and *facebook* during interactions with each other. Hence, we find the weakest prediction potential for the *friend* role identity by using the classifiers of LIWC word categories.

MEH models: Table 5.4 presents the top 10 *'1-g'* (i.e. unigrams) words based on different role identities by using *MEH*. We also observe that these *'1-g'* words are consistent with the findings of the correlation Table 5.3 for different categories of LIWC words. For example, we find that the *friend* LIWC category of words (in Table 5.3) are correlated with the *academic member* role identity. We also observe that *friend* unigram (in Table 5.4) is a high frequency (3.39) word of the *academic member* role identity. However, we observe that few high frequency words such as *lol* (0.62), *grrr* (0.80), *inbox* (0.46), etc. appear by using the MEH in Table 5.4, but we cannot capture these words by using the LIWC. We also find few acronyms such as *laliga, fifa, icc*, and *uefa* by using MEH that appear frequently in our dataset. Therefore, classification results obtained from MEH show different performance

than that of classification results obtained from LIWC to predict role identities. Table 5.6 presents the classification result to predict the role identities by using MEH. We find that *friend* role identity shows the highest (69%) and the *family* member role identity shows the weakest (56.7%) AUC scores.

Independent vs. Hybrid models: We find that the classifiers of *academic member* and *professional member* role identities perform better and the classifiers of *family member* role identity perform weaker than the rest of the classifiers for both of vocabulary based models. Since family members communicate less with each other through Facebook, we find less instances to build our classifier strong. Table 8.4 presents the performance of our hybrid classifier by using both of the machine learning models. We observe that our hybrid model performs better for each of the role identities than that of independent models. We find that *friend* role identity achieves the best classification strength (AUC-0.723). *Friends* are likely to interact with each other frequently for personal, emotional, study, career, foods, social network, and movie related topics by using the *non-dictionary* words. On the other hand, *family members* build the weakest (AUC-0.635) classifier since they interact less with others through social network.

Twitter vs. Facebook: Table 7.12 presents that our hybrid model largely predicts the role identities accurately for Facebook validation dataset. In contrast, our hybrid model correctly predicts only *academic* and *professional* member role identities for Twitter validation dataset. However, our hybrid model is close to the significant (26.11) for *acquaintance* role members for Twitter dataset. We also find that our hybrid model fails (Chi-sq. value 21.03) to identify the *family* member role identity. In a study, Schwartz et al. [173] describe that people share their personal intimate information in Facebook than other social networks. Since Facebook is a closed egocentric network, we can predict all the role identities correctly. In contrast, Twitter is by default an open network, therefore we cannot predict the *family, and friend role* identities effectively. In our experiment, we find that majority of our independent models fail to predict role identities in real life from the social media interactions.

Other role identities and dataset size: We can predict more subtle relationships from the Facebook comments. For example, the *family* member role identity could be more specific such as *spouse, child, father, mother* and so on. Since egos annotate the training data by

multiple labels, by utilizing the multi-class classification, it is also possible to identify these subtle role identities by using our approach. We can also identify multiple roles (e.g., family members can be colleagues to each other) among users by annotating the training instances using corresponding role identities. In previous few well cited studies [40,80] related to psycholinguistic research from social media, we observe that the size of the datasets is ranged from 250 to 300. Therefore, the size of our dataset (N=1,135) is sufficiently large to predict the social role identities by using psycholinguistic features.

Social identity theory and role identification: The *social identity theory* [198] describes one-self on which social group she belongs to. By using the theoretical tool, we perceive users' role membership by analyzing their word usage pattern during social media interactions. We find similar insights from both of the vocabulary based approaches that people use different psycholinguistic word categories with the people of different role groups.

Limitations: Our approach has also several limitations. We find limited prediction potential for few classifiers (according to Tables 5.5 and 5.6). We do not define *hashtags* (i.e., #earthquake), we only consider these tags as words. However, we could define these *hashtags* by using a hashtag dictionary [5] and could use the content of these definitions during model building.

5.11 Summary

In this chapter, we have proposed a technique to infer the interpersonal role identities of users from their comments in Facebook. We have first built a training dataset with 1,135 different ego-alter Facebook friendship connections. Later, we have explained the role taxonomy to the egos and they have annotated these role identities. Then, we have built two different vocabulary based machine learning models to identify the role identities. Later, we have built a weighted hybrid machine learning model to predict role identities from Facebook comments. Finally, we have validated our model with a different dataset of Facebook and a dataset of Twitter. Our hybrid model shows high accuracy in inferring the role identities of users' words usage patterns in Facebook comments. The model also presents moderate

[5] https://www.hashtags.org/definition/

accuracy in predicting social role identities from Twitter dataset.

Psychological attributes play important role for modeling individual user's personalization. These attributes can also play key role to model preferences of a group of users. In the next chapter, we will discuss about personality trait based group (i.e., *homophily*) formation technique and we also investigate to what extent the group members are similar to the real world.

Chapter 6

Personality Trait based Homophily Identification

In this chapter, we investigate the last part of our third research problem. We show a framework to build social group based on users' psychological attributes, i.e., personality. The chapter is organized as follows. Section 6.2 and 6.3 discuss the problem formulation and data collection process, respectively. Section 6.4 shows the personality building model and Section 6.5 reports the model selection process. Sections 6.6 and 6.7 describe the homophily (i.e., group) identification and validation techniques, respectively. Sections 6.8 and 6.9 present the baseline and group recommendation techniques, respectively. Finally, Section 6.11 concludes the chapter with the findings of our experiment.

6.1 Introduction

People are increasingly using Social Network Sites (SNS) such as Facebook and Twitter to share their thoughts with family, friends and acquaintances. Recent research shows that it is possible to capture individual behavioral attributes, e.g., personality, from their interactions in SNS. Personality of an individual is commonly defined using five psychological traits: Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism, also known as Big5 [108]. In this chapter, we first identify communities based on the similarities of Big5 personality traits, which we called *homophilies*, from social media (i.e., Facebook) usage. We also devise a questionnaire/interview based validation technique, which enables

us to verify the identified homophilies in real-life. Moreover, we compare our approach with two baseline homophily identification techniques. Note that, we focus on identifying and validating homophilies in an *ego-centric network* [72], where the ego-centric network denotes social relationship between an ego (a user) and other connected members (Facebook friends) with the ego in the network.

Homophily is a group of people who are strongly connected with each other. Authors in a sociology study [131] described that homophilies can be identified based on different cognitive attributes of humans such as attitude, belief, behavior and so on. In another sociology study [15], authors demonstrated that Facebook reveals actual personality of people. During Facebook interaction, people can not reveal different personality traits apart from their own. Authors [115] also successfully made a predictive analysis from Facebook *likes* about different highly sensitive personal attributes including ethnic origin, political views, religion, satisfaction in life and substances use (e.g., alcohol). Facebook *likes* also express users positive association with online contents such as product, restaurant, sports or music. The above studies motivate us that it is possible to extract actual personality trait based homophilies from social media usage in real-life.

Though a large body of works on identifying the personalities of users from social media usage have been appeared in recent years [173] [80], identifying homophilies in social medias has not been studied until recently. Kafeza et al. [111] first proposed a method to identify influential communities in Twitter based on the users' personality traits. In an open network like Twitter, message or content is made available to everyone who wants to receive it. Kafeza et al. [111] identify influential communities in Twitter. Since, they extract influential communities of the entire network, these communities largely represent celebrities, domain experts, and scientists. However, these people may not be friends to each other in real-life. Thus, it is very difficult to validate whether personality of members are similar to each other in an open network. Hence, authors' [111] approach limits applicability in the practical world. On the contrary, we identify personality trait based communities for each member considering him as an ego in an egocentric network. In an egocentric network (e.g., Facebook), a connection is made between participants when they are agreed to do so. Previous trust among participants leads them to be connected in the egocentric network [157].

When an ego is changed, our approach reshapes community according to his personality traits. Since, the network is closed (i.e. Facebook), it is possible to validate similarity of personality among members of the extracted communities in real-life. We are the first to validate personality traits derived from social network which conforms with the traits in real-life.

Finding the personality trait based homophily in an egocentric network like Facebook has a number of real-life applications. Let us consider the following scenario. A person sends a recommendation to another person (friend) in the Facebook. If the second person sees that the personality of the recommender matches with her own and the recommender is already known to her (trusted) in real-life, then there is a high chance that the second person will accept the recommendation. Other applications that can be benefited from personality traits based homophilies identification in an egocentric network include finding similar acquaintances in workplaces, observing the social dynamics (e.g., group behavior), parental approval in friends searching, accepting friend requests, etc [23].

Our proposed approach has two major steps: (i) identifying homophilies from users' Facebook usage, and (ii) validating the identified homophilies in real-life using questionnaires.

In the first step, we build a homophily prediction model by analyzing the Facebook usage (statuses) of users. In this process, we first analyze statuses of 663 Facebook users by using both closed (e.g., Linguistic Inquiry and Word Count (LIWC)) [77] and open (e.g., Meaning Extraction Helper (MEH)) [173] vocabulary based approaches. Then, we conduct a standardized 44 items IPIP [107] personality test to compute Big5 [108] scores of each user, which we consider as the ground truth data of user personality. Later, we build stepwise forward selection based automatic linear regression model, where a user's Facebook statuses are given as input and Big5 personality scores are produced as output. We observe that LIWC based approach produces moderate strength while predicting personality score due to its less expressiveness (i.e., capture only 4500 dictionary words and word stems). Later, we build our personality prediction model with two linguistic features (i.e., '1-gram' words and topics) using open vocabulary based approach. Next, we compare the strength of these two models, i.e., open and closed vocabulary based approaches. We observe that open vocabulary based approach with two linguistic features (i.e., '1-gram' and topics) outperforms closed vocabu-

lary based approach. Then, we combine these two linguistic feature based models and use a linear weighted ensemble based technique to compute final personality score that produces significant improvement over single linguistic feature (i.e., '1-gram' words or topics) based approach [183] [106]. By using these ensemble based personality scores, we identify homophily of an ego and all connected users (friends) from their Facebook usage.

In the second step, we develop a questionnaire based novel validation technique that enables us to validate the identified homophilies in real-life. For this step, we need to collect the data of Facebook friends who are also friends in real-life. We collect their Facebook usages and construct the homophily network by using our prediction model developed in the first step. Our homophilies correctly cluster 155 Facebook users of validation dataset ranged from 73% to 87%. After that we conduct individual trait based IPIP test on these users, where everyone rates himself and other connected friends in terms of personality scores through different questionnaires. To validate homophily result statistically, we compute *intraclass correlation coefficient (ICC)* [114] and *Cohen's Kappa* [203] values among members. We find several *moderate* and *substantial ICC* scores between our computed personality scores and self evaluating result by the homophily members in real-world. We compute Cohen's Kappa to measure inter-rater agreement between two observers about homophily members. Later, we compute two baseline homophily identification techniques from the existing studies and make a comparison whether our approach can more accurately determine personality trait based homophilies than those techniques. We also propose a group recommendation technique. The technique recommends movie preference to a group of users who are similar to an ego. We find strong correlation between trait based personality scores of homophily members and their movie watching preferences in real world.

In summary, our contributions are as follows:

- We identify multiple personality trait based homophilies in an egocentric network, where five different personality traits (i.e., Big5) are considered as five different homophilies.

- We compare the strength of the models (i.e., open and closed vocabulary) and select the one that predict better personality scores.

- We first build an ensemble based personality identification technique with two different linguistic features, i.e., '1-gram' words and topics.
- We develop a novel interview based homophily validation technique to measure the accuracy of our framework.
- We present experimental result to show the accuracy of our homophilies and compare those homophilies with other baseline techniques.
- We propose a group recommendation technique based on Big5 personality trait.

6.2 Problem Formulation

Let $G=(V,E)$ be an undirected social network where V and E denote the set of vertices (users) and set of edges (social connections), respectively. We consider the social network graph as *Big5* personality graph. Each vertex (user) in G has *Big5* personality attribute set $V.a \in P$, where $P=\{o,c,e,a,n\}$ denotes five different traits of personality and $V.a$ is a real value in the range of [0,1] . In this section, we first give necessary definitions and then state our problem statement.

Definition 1: Egocentric Social Graph. *Let u be an ego, and G be the social graph. Then, an egocentric graph G_u of u can be defined as follows: $G_u=(V_u,E_u)$ be an undirected connected egocentric social network where V_u and E_u denote the set of vertices and set of edges, respectively, connected to ego u. Here, $u \in V_u$ be the ego of the network where $V'=V_u \setminus u$ denotes the alter of the graph.*

Definition 2: Homophily Graph. *Homophily is a subgraph $H_{u\delta}$ of G_u for ego u with respect to a personality attribute $\delta \in P$, where the difference of the personality traits of the ego and any other member $u' \in H_u$ is less than ϕ, i.e., $|u.\delta - u'.\delta| < \phi$. We get $five$ different homophilies $H_{u\delta}$ of ego u, where $\delta=1,2,...5$.*

The goal of this study is to identify personality trait based homophilies of an egocentric

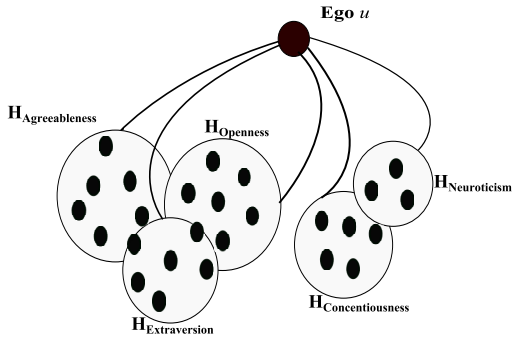

Figure 6.1: Personality trait based multiple homophilies in an egocentric network. Ego u finds different homophilies based on his five different personality traits.

social network. We make a predictive modeling of personality trait based homophily whether users are similar to an ego. For example, we find five different homophilies $H_{uOpenness}$, $H_{uConcenticiouness}$, $H_{uExtraversion}$, $H_{uAgreeableness}$, and $H_{uNeuroticism}$ of an ego u according to Figure 6.1. We also validate these homophilies in real world.

6.3 Data Collection

We have invited 865 users to collect Facebook statuses through posts on Facebook, relevant mailing lists and word of mouth technique. Since Facebook is a closed network, we collect *statuses* of a community that are known to each other. In our experiment, we use *judgmental sampling* technique [129], because we first identify most productive Facebook friends who might response in our survey actively. Later, we create a Facebook application that accesses to the users' status updates. Among the 850 Facebook users, 663 members (male=380, female=283) agreed to share their data through the application. The rest 202 Facebook users have not shown interest to share their time-line through the application. The users are members of university student and professional community, and aged between 18 and 42 years. We build a representative dataset from different age groups and professions. We collect only 96,751 Facebook English statuses as of July 25, 2016. Maximum, minimum and average word counts of the collected statuses are 6786, 145 and 854.73.

We have conducted 44 item IPIP [107] test among these 663 users to collect ground truth data on Big5 personality scores. The users are asked to fill out the survey questionnaire via an experimental web page. We have also collected a new dataset of 155 (male=97, female=58)

Facebook users who are known to each other to validate homophilies in real-life. We have collected another dataset of 123 Facebook users (male=70, female=53) for movie preference group recommendation based on similar personality traits of an ego.

We have collected 663 users' personality score by IPIP [107] test. Average scores and standard deviation of these users on personality test are shown in Table 6.1.

Table 6.1: Average Big5 IPIP personality score on a normalized 0-1 scale.

	Open.	**Consc.**	**Extra.**	**Agree.**	**Neuro.**
Average	0.6597	0.6271	0.5932	0.6515	0.4971
Std. Dev.	0.1213	0.1890	0.1257	0.1137	0.1984

6.4 Personality Building Models

We identify personality with two different strategies: **i)** with closed vocabulary (i.e., LIWC) based approach [80] and **ii)** with open vocabulary (i.e., MEH) based approach [173]. Authors in [173] show that closed vocabulary (i.e., LIWC) based approach applies fixed priori of words to analyze text. LIWC only captures dictionary words (4500 words, and word stems) and ignore a large amount of words those might be important signal to analyze one's personality accurately. For example, words such as selfie, Facebook, inbox, etc. are not analyzed by LIWC that are also useful cues to predict one's personality. Motivated by the work [173], we also apply open vocabulary based approach in our dataset of 663 Facebook users. First, we build personality identification model by using LIWC [80]. Later, we also build personality identification model using the open vocabulary approach. Then, we compare the strength between closed [80] and open [173] vocabulary based approaches to show which one better predict personality scores. Finally, we select a model that predict better personality scores between two approaches.

6.4.1 Closed Vocabulary based Approach

In this subsection, we conduct our experiment with extensively used closed vocabulary based approach. For building personality prediction model, we consider IPIP test result

of 663 users as the ground truth data of personality traits. Motivated by the prior work on $personality$ prediction from [80], we measure word uses in users status updates with LIWC. LIWC 2007 determines 74 different types of categories, each contains hundred of words [152]. We exclude the categories that are non semantic (e.g., proportion of long words, and filler).

We calculate Pearson correlation analysis between Big5 scores and each of the score of LIWC features. We conduct the correlation analysis among statuses of total 663 users who attended in the IPIP test. We analyze the association through linear regression to predict the score of a given personality score. We find that a number of LIWC features are correlated with a personality dimension. A potential problem arises when collinearity found between personality and LIWC features. When there is a perfect linear relationship exists among independent variables, the outcome for a regression model cannot be unique. We check variance inflation factor (VIF) among the independent variables to detect collinearity problem [43].

To remove collinearity among independent LIWC features, we have computed lasso penalized linear regression using *glmnet* R package [43] [93]. This technique reduces the coefficients to a low value or zero, thus the model does not get overfitted. Table 6.2 presents that openness personality trait has the strongest (24.4%) and neuroticism personality trait has the weakest (16.2%) strength among all the personality traits. We find that these models moderately fitted across all the personality traits. The result has a low relative error (MAE are ranged from 0.091 to 0.127) which indicates that the model performs better than the constant mean baseline.

Table 6.2: Adjusted R^2 scores of the linear regression models.

Big5 traits	**Open.**	**Consc.**	**Extra.**	**Agree.**	**Neuro.**
Adjusted R^2	24.4%	17.6%	23.9%	21.2%	16.2%

Motivated by the work [43], we also investigate the prediction potential using a machine learning classification study. Sumner et al. [189] suggested that computing MAE and RMSE for error measure in regression analysis are not adequate. In particular, when the majority of the individuals are around the mean of unimodal distribution, these error measures can often

mask large errors.

According to the suggestion of Sumner et al. [189] [43], we apply different supervised binary machine learning algorithms on our dataset. We classify above-median level as *high* class label and below-median as *low* class label value dimension. We have experimented with few classifiers including Logistic Regression, Naive Bayesian, Adaboost, Random Forest, support vector machine and RepTree classifiers using WEKA [87] machine learning toolkit. For each personality trait, we have applied these classifiers to understand the prediction performance of these personality building models.

Table 6.3: Best performing classifier to predict different traits of personality using LIWC.

Big5 traits	**Highest AUC achieving classifier**	**AUC**	**TPR**	**TNR**
Open.	Logistic Reg.	0.617	0.669	0.472
Consc.	Logistic Reg.	0.569	0.544	0.519
Extra.	Naive Bayes	0.605	0.631	0.402
Agree.	Logistic Reg.	0.602	0.491	0.424
Neuro.	Logistic Reg.	0.587	0.593	0.398

Table 7.7 presents the best classifier, content type, it's true positive rate (TPR), true negative rate (TNR) and Area under the ROC curve (AUC) for computing each of the personality traits [68]. Performance of the classifiers were conducted using AUC values under the 10-fold cross validation. The curve is plotted the TPR against the FPR at different threshold. The space of ROC curve is better than another if it is to the northwest (tp rate is higher, fp rate is lower, or both) of the first [68]. We observe that our classifiers achieved moderate improvement over random chances for openness, extraversion and agreeable personality traits. For the rest of personality traits (i.e., conscientiousness and neuroticism), our models achieved lower potential than random chances.

6.4.2 Open Vocabulary based Approach

We again analyze our dataset with open vocabulary based approach using MEH [33]. We analyze two categories of words: i) words ('1-gram') and ii) topics.

First, we analyze '1-gram' words with a big data analysis tool, MEH. Unlike content coding software (e.g., LIWC), the MEH is highly dynamic for extracting words and phrases from a dataset. Motivated by the work [173], we extract two different types of linguistic features: i) '1-gram' words, and ii) topics. During analysis of '1-gram' words with MEH, we use *MEH-Output_verbose* file which contains frequency of '1-gram' words, represented as percentage of each observation. We find a total of 803 unique '1-gram' words excluding stop words.

Motivated by the study [173], we also compute another type of language feature, *topics*, consists of word using Latent Dirichlet Allocation (LDA) [27]. We use R *Mallet* package implementation [101] to extract top 267 frequent topics and their percentage of frequency from our dataset of 663 Facebook users.

Table 6.4: Adjusted R^2 scores of the linear regression models using '1-gram' words and topics.

Big5 traits	**Open.**	**Consc.**	**Extra.**	**Agree.**	**Neuro.**
1-gram	64.1%	38.6%	57.1%	59%	40.3%
topic	49%	45%	42%	47.1%	46.2%

Then, we compute Pearson correlation between percentage of words and IPIP test result for both '1-gram' words and topics, independently.

Table 6.5: Best performing classifier to predict different traits of personality using *1-gram* words.

Big5 traits	**Highest AUC achieving classifier**	**AUC**	**TPR**	**TNR**
Open.	RandomTree	0.669	0.671	0.461
Consc.	SVM	0.589	0.65.3	0.451
Extra.	RepTree	0.653	0.673	0.536
Agree.	Adaboost	0.659	0.66	0.48
Neuro.	SVM	0.597	0.602	0.493

Unlike LIWC, Open vocabulary based approach does not categorize similar words into a group. Thus, similar words contribute to build the model individually each time, which might generate collinearity among independent variables. Open vocabulary based approach considers similar words (i.e., happy, cheer and joy) as an individual predictor. Thus, we identify a

large number of collinear independent variables by observing VIF scores. Later, we compute *lasso penalized linear regression* using *glmnet* R package [43] [93] to remove collinearity among independent '1-gram' words and topics. This technique reduces the coefficients to a low value or zero, thus the model does not get overfitted. We use linear regression algorithm each with a 10-fold cross-validation with 10 iterations. Table 6.4 presents the adjusted R^2 strength across all the personality traits for each type of linguistic feature.

Table 6.6: Best performing classifier to predict different traits of personality using *topics*.

Big5 traits	**Highest AUC achieving classifier**	**AUC**	**TPR**	**TNR**
Open.	SVM	0.621	0.653	0.461
Consc.	RepTree	0.625	0.601	0.443
Extra.	RepTree	0.601	0.579	0.531
Agree.	Adaboost	0.623	0.601	0.542
Neuro.	SVM	0.638	0.669	0.391

Sumner et al. [189] suggested that computing MAE and RMSE are not sufficient to check the prediction potential of a regression model. Using the similar approach that used in Subsection 6.4.1, we compute prediction potential by different supervised binary machine learning algorithms on our dataset. For each personality trait and each type of linguistic features (i.e., '1-gram' words and topics), we have applied previous classifiers to understand the prediction performance of these personality building models. Table 6.5 and 6.6 present best performing classifiers to predict trait of personality using '1-gram' words and topics, respectively. Authors [173] did not demonstrate their analysis by removing collinear variables. They also did not show the prediction potential of their model using classification technique. Since classification with cross validation is reliable metric to assess how well a model work for unseen data [189], we have investigated the potential of our models with classification techniques. We observe that these classifiers achieved significant improvement than random chances.

6.5 Model Selection

Authors [173] describe that closed vocabulary based approach (i.e., LIWC) suffers less expressiveness due to a priori fixed set of words (e.g., 4500 dictionary words, word stems and 74 categories). Apart from dictionary words, SNS users generally use diverse set of words e.g., local dialects, emoticons, buzz words (i.e., selfie). LIWC based approach doesn't consider these words while words being analyzed. Thus, we may ignore a large portion of personality cues while these texts being analyzed. In contrast authors [173] proposed data driven approach, where they consider all the words (both dictionary and non-dictionary) are written by users. Open vocabulary based approach suffers severe collinearity problem due to a large number of independent variables. We solve the collinearity problem in the subsection 6.4.2 that was ignored in previous study [173].

LIWC only analyzes word categories and it does not capture phrase level words (i.e., '2-gram' words, '3-gram' words, etc). Since, we compare between open and closed vocabulary based approaches, we ignore phrase level words in open vocabulary based approach. Thus, we use '1-gram' words and topics in our dataset to compute personality building models using open vocabulary based approach. Table 6.2 and 6.4 present the strength of personality building model between open and closed vocabulary based approaches. We observe that closed vocabulary based approach shows moderate prediction strength across all personality traits. Conscientiousness and neuroticism personality building models show low prediction strength across all the personality traits. On the other hand, open vocabulary based approach shows better prediction strength (see Table 6.4) than closed vocabulary based approach.

Based on the personality prediction strength from Table 6.2- 6.5, we finally select open vocabulary approach as our working model. We observe in open vocabulary based approach that one linguistic feature predicts better personality prediction score than other features for few personality traits. For example, openness, extraversion and agreeableness personality traits using '1-gram' based model shows better prediction potential than topics based model. Again, topical modeling linguistic feature shows better prediction strength for conscientiousness and neuroticism personality traits than '1-gram' linguistic feature. Since every linguistic feature contributes to compute the personality score based on their strength (weaker or stronger), to find out final personality score, we combine all the linguistic features obtained

from the previous steps. It is observed in previous studies [106] [158] that combining different experts, we can build better model to predict an attribute (i.e., personality).

6.5.1 Ensemble of Models

It is necessary to prioritize the features based on their importance, as we compute personality scores from two linguistic features (i.e., '1-gram' words and topics) in Facebook. For example, some may think that '1-gram' feature can reveal a personality score of a person more accurately, while other may emphasize on 'topics' to determine the personality score correctly. Ordering among linguistic features associates different weights to compute final personality score. Weight signifies the relative importance of a particular linguistic feature type. To build our ensemble/combined model, we perform following two steps: i) computing weights from neural networks, ii) combining the personality building model with a weighted linear ensemble technique.

6.5.1.1 Learning Weights from Neural Networks

In this part, we determine the *weight* of each linguistic feature type (e.g., '1-gram' words and topics) to determine personality trait. For each type of linguistic feature and each personality traits, we model a neural network with a new dataset of 198 Facebook users (30% of our total dataset). We model our network with two types of linguistic features and five types of personality traits; we build in a total of 10 (2×5) neural networks using R *caret* package implementation [117]. For a single neural network, we use nine input neurons in the input layer, five neurons in the first hidden layer, three neurons in the second hidden layer and one output neuron in the output layer. For each personality trait, we take linguistic scores (i.e., percentage of *happy* '1-gram' words) as input and gives a *personality* prediction score as output.

Consider a scenario, where we are interested in predicting personality score openness for the linguistic feature *topics*. We select the best subset of linguistic features using R *leaps* package implementation [127] by forward selection approach. Then, we normalize the linguistic feature scores (i.e., topics) in the interval [0,1] with max-min normalization technique to get better precision. We keep 90% datapoints of new dataset in the training set

and the rest are in the test set using 10-fold cross validation with 10 iterations. For each feature type and personality, we compute the strength of different models. Table 6.7 presents the strength (the adjusted R^2) of our neural network based linear regression models that will be used as *weights* of our ensemble models in the next Subsubsection, 6.5.1.2.

Table 6.7: Weights (the adjusted R^2) derived from Neural Networks.

Big5 traits	**Adjusted R^2 score of '1-gram' words**	**Adjusted R^2 score of topics**
Open.	25.2%	21.5%
Consc.	11%	17%
Extra.	20%	18%
Agree.	17.3%	14.2%
Neuro.	16.1%	21.5%

6.5.1.2 Weighted Linear Ensemble

In this part, we build a weighted linear ensemble model from different types of features of 663 Facebook users [183]. We have already built different models from '1-gram' words and *topics* linguistic features that are described in Subsection 6.4.2. Since, we train different neural networks that produce weights, we compute weighted linear ensemble score using the weights in Table 6.7.

Finally, we build our weighted linear ensemble model using the weights generated from another dataset (according to Table 6.7), thus our models do not get over-fitted. Table 6.8 presents the strength of our ensemble models and performance of the respective classifiers. We observe that our models obtain a substantial improvement with prediction potential compared with single feature based personality identification models (according to the Table 6.2).

Table 6.8: Adjusted R^2 scores of the ensemble models.

Big5 traits	**Open.**	**Consc.**	**Extra.**	**Agree.**	**Neuro.**
Adjusted R^2	75%	66.9%	70.3%	73%	65.8%

Note that, we use two different datasets for our weight learning and training. Using two different datasets is somewhat similar to cross validation where we learn from one dataset

and apply on another dataset. If we learn weights (i.e., contribution of different content type) from dataset and then again apply the ensemble on the same dataset, this would be like doing training/testing on the same dataset. Thus, we keep the training and testing dataset separate while building ensemble.

Table 6.9: Best performing classifier to predict different traits of personality using ensemble of models.

Big5 traits	**Highest AUC achieving classifier**	**AUC**	**TPR**	**TNR**
Open.	RandomTree	0.743	0.772	0.335
Consc.	Logistic Reg.	0.695	0.722	0.371
Extra.	Random Forest	0.683	0.681	0.324
Agree.	Adaboost	0.713	0.704	0.327
Neuro.	Adaboost	0.679	0.67	0.341

6.6 Homophily Identification

In this section, first we have done MEH analysis on the users' statuses. Then, we collect Big5 scores of the 663 users from 44 item IPIP test. Later, we compute lasso penalized linear regression between two linguistic features (i.e., '1-gram' words and topics) extracted from users' statuses and Big5 scores of the users using *glmnet* R package [43] [93]. Then, we build ensemble based personality prediction model that we have described in Subsection 6.5.1. After building the personality model, we predict Big5 personality scores over the new dataset of 155 users. Later we identify clusters of similar individuals with respect to an ego. Figure 6.2 shows our methodology to identify a personality trait based homophily:

The users are the members of a student community. The members are aged between 20-35 years, and have a diverse education majors. If the personality score of a user i is similar to *ego* u, then we denote the user as *alter* i. To extract similar alters of an ego, we apply agglomerative hierarchical clustering [137] technique. We extract different clusters with different stretch values. We find the best clustering result at a heuristic stretch value of +0.02 to -0.02. We claim that members within same cluster possess similar personality

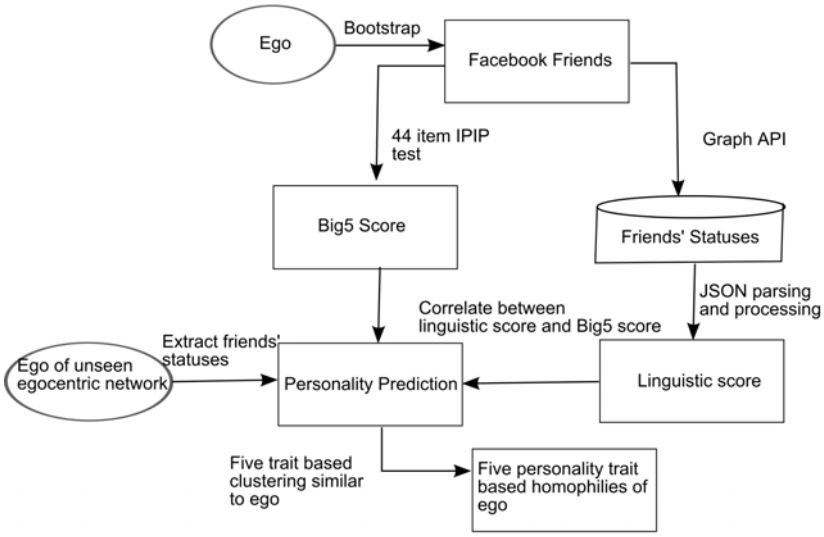

Figure 6.2: Methodology of personality trait based homophily identification.

trait. For each personality trait of an ego u, we build five separate homophily clusters. Alter v_{i} may belong to different clusters of an ego u. This indicates that alter v_{i} has close match with the ego u from multiple personality traits. Our models for five personality trait based homophilies correctly clustered users ranged from 73% to 87% among these 155 users. We find that openness and neuroticism personality traits cluster users with an accuracy highest 87% and lowest 73%, respectively.

6.7 Homophily Validation

In this section, we validate whether homophily members of an ego are similar to each other in real-life that we have extracted in Section 6.6. Every ego contains five homophilies based on five personality traits. To validate these homophilies, we ask individuals about themselves and other members who belong to same homophily about their personality using the trait based IPIP test in real-life. We divide the 44 item questionnaire into five sets which we call *trait based IPIP questionnaire*. Finally, we check the reliability of the answers reported by homophily members using statistical techniques (e.g., ICC and Cohen's Kappa). Figure 6.3 shows our methodology to validate multiple personality trait based homophilies.

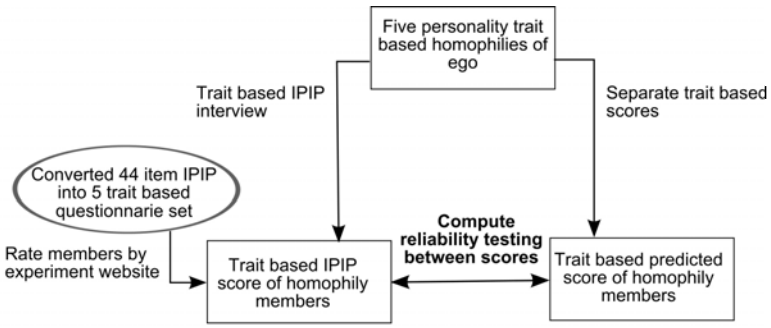

Figure 6.3: Methodology of personality trait based homophily validation.

6.7.1 Individual Trait based IPIP Scoring Method

We prepare an experimental web page where we take the trait based IPIP test. Trait based IPIP test has five sets of questionnaires for openness, conscientiousness, extraversion, agreeableness and neuroticism, respectively. Each trait has 8-10 questions based on the trait. For example, when we test extraversion trait, we ask questions 1, 6R, 11, 16, 21R, 26, 31R, and 36 from 44 item IPIP questionnaire pool following the procedure mentioned in [108]. Among the questions, some are marked as R (reverse). For each of the question that is marked as *R* for a trait, we reverse the score in a scale of 1-5, e.g., in likert scale 2 is considered as 4 for an *R* marked question. Then for each trait, we compute max-min normalization to get the final score out of [0-1]. We test our method with the proven IPIP sites [107] and find identical result on different personality traits.

6.7.2 Validation Results

In this subsection, we present experimental results of our proposed personality trait based homophily validation technique. Though people use IPIP test to self report themselves, but researchers also show that individual can check his similarity with his friends, co-workers, and others [107]. Hence, we conduct experiment where individual reports answer to questions about himself as well as about his friends. In our approach, if a user is discovered in the homophily of an ego from the homophily identification predicted score, he is directed to the experimental website to fill up the personality trait based IPIP questionnaire. For different users, homophily network might be different. For example, we randomly pick a member as an ego from our new dataset of 155 members. We then select 45 distinct members by hier-

archical clustering technique for different personality trait based predicted scores (Open.-22, Conc.-13, Extra.-15, Agree.-11, Neuro.- 10) who are similar to the ego. Within 45 selected members, 9 members fall into multiple trait based homophilies of the ego. Raters and ego need a common set of known friends inside a homophily to get actual rating based on real world relationship. Since every person rates other members in the homophily, we need a measurement of their scoring reliability.

6.7.3 Reliability among Scores of all Raters in a Homophily

We compute ICC to assess the strength of consensus among the internal raters. Inter-rater reliability describes what percent of our ratings are real [114]. We use two-way random model. We compute *consistency* and *absolute agreement* analysis of the ratings by the homophily members. For consistency, raters need not to agree perfectly among themselves. If their changing behavior of ratings are same, then they possess high consistency score. Later, we also compute absolute agreement. For absolute agreement, raters need to agree perfectly about their ratings. For both of these cases, we compute single and average measurement. Single measurement determines to what extent rating of a single person is reliable, if he rate himself. Average measurement determines reliability of raters on average. Table 6.10 compares between *ICC* scores of personality trait based and two baseline homophily identification techniques.

We also compute *Kappa* [203] by the observation of two external raters among the members of the homophily. Though these two raters are not a homophily member, they need to know all of the members of a homophily to rate them accurately. Raters judge these homophily members with particularly low (0-0.5) or high (0.51-1.0) values of a trait. Table 6.11 compares *kappa* values between personality trait based and two other baseline homophily identification techniques.

6.8 Comparison with other Baseline Techniques

In this section, we propose and justify two comparable baseline techniques with our proposed technique to identify personality trait based homophilies. A number of user activities such

as tagging items and listening music have been identified to extract homophilies [5] [23]. In this chapter, we propose two baseline techniques to identify homophilies are: i) users who like similar Facebook fan pages, and ii) users who interact frequently in Facebook. We first find out members of homophilies who are similar to an ego in terms of two different criteria (e.g., *page-likes* and *degree of interaction*). Then, we assume that these homophily members are likely to possess similar personality trait of an ego as they behave similarly such as *liking* similar pages and interacting frequently with similar members in Facebook. To validate these homophilies, we apply two baseline homophily identification techniques with our evaluation dataset of 155 Facebook users. Later, we ask these three different homophily members (personality trait based homophily with two baseline techniques) about the similarity of their personality by trait based *IPIP* questions (according to subsection 6.7.1). Finally, we show that our technique more accurately distinguish personality trait based homophilies than that of two baseline techniques.

6.8.1 Page-likes Similarity

In this baseline technique, we identify homophilies of Facebook users who *like* similar Facebook fan pages of an ego. We consider that users who *like* similar Facebook Fan pages are likely to possess similar personality traits. For example, a person with high score in openness is likely to *like* pages that present content with excitement, new experiences and strong stimulus. On the other hand, a person with strong conscientiousness personality score usually like pages with career building, health awareness, etc. Thus, we identify homophily (similar) members among the 155 users of evaluation dataset for Facebook page like similarity. We collect *page-likes* (tag: 'about') of these users using our Facebook application. We collect a total of 19,766 page-likes (tag: "about") as of July 25, 2016 through our Facebook application. The *page-likes* dataset has a maximum, minimum and average word counts 1887, 3 and 1234.12, respectively.

Authors in [122] predicted links based on similarity of co-authorship network. Following their technique, we also identify homophily of an ego based on their patterns of *liking* Facebook pages. We consider all of the *page-likes* of a single user as a single document. We compute *bag-of-words* among the documents of all users by removing stop words, and using

lemmatization technique. Given an ego u, we compute similar alters V' based on similarity of *terms* found in user *page-likes* document using Jaccard's coefficient. Let us consider that $X=(x_1, x_2,....,x_n)$ and $Y=(y_1, y_2,....,y_n)$ are two vectors of term frequency of page-likes of an ego u and an alter $v_j \in V'$, respectively. Then, we can compute the Jaccard's coefficient by the following equation:

$$J(x_i, y_i) = \frac{\sum_i |\Gamma(x) \cap \Gamma(y)|}{\sum_i |\Gamma(x) \cup \Gamma(y)|} \tag{6.1}$$

In this baseline, we compute Jaccard's coefficient between *page-likes* documents of an ego u and all the alters $u' \in V'$. After computing the Jaccard's coefficients, we apply agglomerative hierarchical clustering [137] that we used during homophily identification in Section 6.6. To make an unbiased experiment, we keep same stretch value of *+0.02* to *-0.02*. In our clustering result, we find 45 alters belong to the same cluster (homophily) of the ego u.

We assume that these 45 *alters* possess similar personality traits of ego u, since they like similar Fan pages to ego in Facebook. To prove our hypotheses, we conduct experiment where these alters report answer to IPIP test about himself as well as about his members in the same homophily. Table 6.10 and 6.11 shows that *ICC* agreement and *Kappa* coefficient in different personality traits are *poor* based on *page-likes* homophily result. Thus, we conclude that Facebook users who *like* similar pages do not possess similar personality traits in real world.

Table 6.10: ICC score for different techniques of homophily computation.

Homophily technique	**Type**	**Measure**	**Open.**	**Consc.**	**Extra.**	**Agree.**	**Neuro.**
Big5	Consistency	Single	0.619	0.253	0.589	0.539	0.231
	Absolute agreement	Average	0.872	0.501	0.807	0.788	0.457
Page-likes similarity	Consistency	Single	-0.11	0.15	0.07	0.18	0.08
	Absolute agreement	Average	0.05	0.17	0.09	0.14	-0.03
Similar k-users of interaction	Consistency	Single	0.141	0.09	0.03	0.11	-0.01
	Absolute agreement	Average	0.173	0.11	0.013	0.16	0.01

6.8.2 Maximum Degree of Interaction

In this baseline technique, we identify homophilies among users who frequently interact among themselves in Facebook. Authors in a study [57] extract *active communities* by computing the *degree of interactions* among users. Higher degree of interaction signifies that strong tie strength exists among the users and they are similar with each other than the rest of members in the network. Based on the interactions, it is possible to find out a group of users (homophily). We call this homophily as *interaction based homophily*. We assume that these users have similarity among themselves, since they have strong tie strength and they interact frequently in Facebook. Motivated by the approach [57], we identify $u' \in V'$ users (alters) who have highest degree of interactions with ego u. We call that these users (alters) belong to the same homophily of ego u.

Table 6.11: Kappa score for different techniques of homophily computation.

Homophily technique	**Kappa**	**Open.**	**Consc.**	**Extra.**	**Agree.**	**Neuro.**
Big5	Value	**0.709**	0.231	**0.632**	**0.484**	0.207
	Approx sig.	**0.001**	0.117	**0.001**	**0.002**	0.171
Page-likes similarity	Value	0.087	0.043	0.204	0.10	0.15
	Approx sig.	0.731	0.764	0.125	0.317	0.231
Similar k-users of interaction	Value	0.08	0.04	0.10	-0.01	0.10
	Approx sig.	0.734	0.79	0.343	0.831	0.31

To extract interaction graph of Facebook ego network, we use NodeXL [91], an open-source network analysis and visualization tool. We collect interaction data such as *likes, comments*, and *sharing* of object between an ego u, and each alter u'. We consider the same 155 alters those we used in previous experiments. We count total degree (sum of in-degree and out-degree) for each pair of ego u and alter connection. For each interaction (e.g., commenting or liking a post) between them, we increase the total count by one. We collect all the interaction data between ego u and each alter u' as of July 25, 2016 through NodeXL tool. We find maximum, minimum and average interaction degree counts are 183, 7 and 25.31, respectively.

Then, we build homophily (cluster) of ego u based on higher degree of interactions. We

select the cluster that contains 45 different alters, since the size of our *page-like* homophily is also 45. For the sake of uniform clustering parameters in all experiments, we keep the size of cluster same. We assume that these 45 users possess similar personality traits as they have higher degree of interactions with the ego u. To prove our hypotheses, we conduct experiment (the same experiment that we conduced in Section 6.6 and Subsection 6.8.1) where these alters report answer to *IPIP* test about himself as well as about his members in the same cluster. Table 6.10 and 6.11 show that *ICC* agreement and *Kappa* coefficient in different personality traits are *poor* based on *degree of interaction* result. Thus, we infer that users who interact frequently through the Facebook do not possess similar personality traits in real world.

6.9 Personality Trait based Group Recommendation

Group recommendation is a well studied topic in recommender systems [8] [83]. Recommending a group has a number of real scenarios that include selecting restaurants for taking lunch with friends, finding travel destination with family members, etc. In this chapter, we show that finding personality trait based homophilies can facilitate to recommend members of a group to perform certain task by an ego. In particular, we demonstrate how to find out the preferences of movies for a group of Facebook users who are similar to ego u based on their personality traits.

It is evident from previous studies [71] [44] that personality traits influence user needs and preferences. Motivated by the prior studies [4] [102], we investigate whether there is any correlation between user personality traits and group movie preference of users in Facebook. The intuition behind our movie preference scenario is as follows. Since individuals possess different personalities, it is highly likely that the preference of movie watching may differ based on the type of movie contents. In this study, we investigate whether we can accurately recommend a group of users to recommend movies who are likely to possess similar personality trait of ego u. We find strong correlation between personality score of ego and preference of movies of her homophily members.

6.9.1 Movie Preference

To evaluate our models for movie preference application, we first collect a new dataset of 123 (male=70, female=53) active Facebook users through our application. Users are from the same ethnic group, but they are from different educational and professional background, aged between 20 and 28 years. Later, we predict homophily members by our models and prepare a questionnaire to find out movie preference of these users through interviews. All the recruited participants must have watched some selected movies to attend in the interview. We recruit these participants by observing their Facebook *movie watched list*. Then, we compute correlation coefficient between the predicted personality group scores of homophily members and results of the questionnaire in real life. In particular, we will investigate the following three hypotheses that correlate personality traits with movie preferences of users.

H1. High *openness* personality homophily members is strongly associated with sci-fi/adventurous movies. An individual who possesses high score in openness personality is likely to give high importance on futuristic view and active imagination. He likes to experience challenges and excitements in his life. Since sci-fi/adventurous movies contain new ideas and stimuli in the movie contents, it is likely that an individual with high score in *openness* personality prefers to watch those movies.

H2. High *extraversion* personality is strongly associated with comedy movies. Extraversion is about enjoying life, seeking happiness, and sensuous gratification for oneself. Since comedy movies contain these attributes in their movie content, we hypothesize a positive link between extraversion personality and content of the comedy movies.

H3. High *agreeableness* personality is strongly associated with romance or drama movies. A person with high agreeable score generally trusting, generous and helpful. Since drama and romance genre of movies contain sentimental, emotional, sacrificing and compassionating content, we hypothesize a positive link between agreeableness personality score and content of romance and drama genre of movies.

6.9.2 Movie Preference Experiment

First of all, we predict homophily of ego u based on Big5 personality traits in our dataset of 123 Facebook users. In our predicted homophilies of ego u, we find 37, 22, 41, 32,

Table 6.12: Correlation coefficients (CC) between group personality score and results of movie preference interviews. Note: *p<0.05, and **p<0.01

Movie genre	**Openness**	**Extraversion**	**Agreeableness**
Sci-fi/Adventurous	0.31**	0.091	-0.03
Comedy	0.07	0.281**	0.13
Romance/Drama	0.05	0.11	0.27**

and 12 homophily members in openness, conscientiousness, extraversion, agreeableness, and neurotic personality traits, respectively. According to our hypotheses, we assume that homophily members of openness, extraversion and agreeable personality traits have association with sci-fi/adventurous, comedy, and romantic/dramatic movies, respectively. To prove our hypotheses with group recommendation, we conduct the following experiments.

We conduct a semi-structured interview during July 2016 in different locations (e.g., restaurants, university library, etc.). Most of the interviews are taken in the face-to-face settings. At the end of the interview, the homophily members are compensated by a small gift. We also take few interviews through skype for homophily members who stay in distant locations.

We initially hypothesized that the homophily members who like to watch *sci-fi/adventurous*, *comedy*, and *romance/drama* genre of movies, they tend to possess the high score in *openness*, *extraversion*, and *agreeableness* personality traits, respectively. First, we select a list of total six movies from three different genres: sci-fi/adventurous, comedy, and romance/drama. All of the movies are rated between 8.2 to 8.8 according to imdb (Internet movie database) rating. We confirm that all the homophily members have watched all of the selected movies previously. They are asked to rate all the six movies based on their preferences in a likert scale of 1-5. In likert scale, 1 is *strongly disinterested*, 2 is *disinterested*, 3 is *neither disinterested nor interested*, 4 is *interested* and 5 is *strongly interested* to recommend as worth watching a particular movie. Based on the answer of the likert scale, we normalize the score between 0 and 1. After rating these genre based movies by the homophily members, we compute mean of normalized (0-1) score. Then, we compute correlation coefficients between mean normalized ratings of the three genre based movies and predicted personality

scores from our model for each homophily member. Table 6.12 presents the correlation coefficients between movie genre preferences and personality scores of Facebook users in different homophilies.

6.9.3 Experimental Result

To evaluate our hypotheses, we observe the correlations between individuals' personality ratings with their self-reported movie preferences. Table 6.12 shows that **H1**, **H2**, and **H3** conform with the preference of movie genre selection in real-world. We find strong correlation (0.31**) between openness personality and preference of sci-fi/adventurous movie genres. We also find strong correlation (0.281**) between extraversion personality score and preference of comedy movie genre. Similarly, we find strong correlation (0.27**) between agreeableness personality score and preference of romance/drama genre of movie.

Thus, our personality trait based homophily identification technique is able to recommend a group of Facebook users who possess similar personality score to ego u. This technique can also recommend a group of similar users to ego u more accurately about their product, preference of gadget selection, and other similar types of applications in real-life.

6.10 Discussion

Our work is the first study i) to identify the correlation of word usage with personality trait based homophily in an egocentric social network using two different personality identification techniques, and ii) to validate the identified homophilies in real-world.

From Table 6.2 and 6.4, we observe that both of the techniques of open vocabulary based approach (i.e., '1-gram' words and topics) outperform closed vocabulary based approach. In open vocabulary based approach, openness and conscientiousness personality traits show the strongest and weakest strength, respectively, among all the personality traits using '1-gram' technique. From Table 6.4, we also notice that conscientiousness and neuroticism personality traits show improvement using topical modeling than '1-gram' technique. Again, we observe from Table 7.7, 6.5 and 6.6 that open vocabulary based approach shows moderate improvement over random chances for all personality traits than closed vocabulary based

approach. From Table 6.8 and 6.9, we observe that our ensemble based method shows substantial improvement over individual closed and open vocabulary based approach.

While validating personality trait based homophilies in real world, we find in Table 6.10 and 6.11 that our ensemble based technique discovered homophilies using Big5 technique show better agreement scores than other two baseline homophily identification techniques. For example, ICC and kappa values of openness homophily (Big5) show substantial (0.61-0.80) agreement among the raters [203]. Again, we observe that ICC and kappa values of agreeable homophily (Big5) show moderate (0.41-0.60) agreements. We also notice that extraversion homophily (Big5) shows moderate (0.41-0.60) and substantial (0.61-0.80) ICC and kappa values, respectively. In contrast, both conscientiousness and neurotic homophilies show fair (0.21-0.40) ICC and kappa agreements among the raters. It has also been proved in previous studies that personality prediction for neuroticism is difficult [15]. Again, conscientious people have less propensity to share information in the social media frequently [104]. We notice that ICC and kappa strength of different personality trait based homophilies are correlated with the strength of personality (according to Table 6.8) building model. Thus, we find fair agreement among the raters while discovering homophilies for conscientiousness and neuroticism traits.

In contrast, we find slight or *less than chance* (< 0.0) agreement (according to Table 6.10 and 6.11) among the raters using two baseline (e.g., page-like similarity and similar k-users of interaction) techniques. For example, we find *less than chance* ICC agreement of *page-like similarity* technique for openness and neuroticism homophilies while other three trait based homophilies show slight agreement. For kappa values, the agreements also show random scores. We also observe *less than chance* or *slight* agreement while discovering homophilies with similar *k-users of interaction* technique. We find no correlation with personality model and it's derived homophilies using these two baseline techniques. Therefore, to identify personality trait based homophilies using ensemble technique, Big5 model outperforms other two baseline techniques.

In Section 6.9, we find strong statistical correlation between personality trait based homophily with respect to personality of an ego and movie watching preference. Table 6.12 presents that sci-fi/adventurous, comedy, and romance/drama movie genres are strongly cor-

related with openness, extraversion, and agreeableness personality trait based homophilies, respectively. Since we get less ICC and Kappa values for conscientiousness and neurotic personality trait based homophilies, we do not find such association to recommend movie preference.

In a previous study, it is shown that a moderate personality prediction strength can be achieved from myPersonality [40] dataset (N=250) with minimum and average word count of 1 and 585.004 respectively. In another well cited study [80], authors successfully predict personality from Facebook with a sample size of 279 Facebook users. Therefore, the size (N=663) of our dataset is sufficient to predict personality from social media usage. There are few datasets available for modeling personality from social media [40] [173]. Since we need to validate our homophilies in real world, we build and evaluate models with our own datasets that are collected from the same demographics.

6.11 Summary

In this chapter, we have presented homophily identification and validation techniques for users in social media (i.e., Facebook). To identify homophily, we have first built a prediction model that takes a user's Facebook status as input and gives the personality scores of the user as output. We have identified personality prediction models using both closed and open vocabulary based approaches. We have also compared the strength of those models using classification and regression prediction potential techniques. We have first computed personality scores by combining two different linguistic features (i.e., '1-gram' words and topics). Then, by using hierarchical clustering technique, we have identified a community for an ego, where five different personality traits based similarities are considered for the identification of homophilies. We have also presented a novel validation technique that enables us to verify and compare our identified homophilies with other techniques in real world. Our work is the first one that identifies personality trait based homophilies from Facebook and validates those homophilies in real world. We have also demonstrated group movie recommendation, an application of personality trait based homophily identification.

We have observed that values and preferences of a user change over time. In the next two

chapters, we will investigate how a user preference changes by the influence of which psychological attributes, i.e., personality and values. First, we will extract users' eat-out (i.e., tangible) preferences and then we will extract users' movie (i.e., intangible) preferences by exploiting the fusion of multiple social media sites.

Chapter 7

Predicting Eat-out Preferences from Users' Psychological Attributes

In this chapter, we discuss the first case study pertaining to our fourth research problem. Section 7.2 describes the methodology of our work and Section 7.3 presents the data collection process. Then, Section 7.4 shows the feature selection process and Section 7.5 presents the eat-out (i.e., tangible) preference building model. Later, Section 7.6 shows the evaluation of our prediction model. Next, Section 7.7 describes the implications of our experiment and Section 7.8 concludes the chapter.

7.1 Introduction

In recent times, Twitter, a microblogging site, has become major platform of communications for users in the web. Twitter allows a user to share ideas, thoughts, and opinions with his friends and followers. The large amount of tweets generated by diversified users provide a scope for the researchers to identify different psychological attributes such as personality [80], values [43] and preferences [151] of the involved users. On the other hand, location based social networking sites such as Foursquare have become increasingly popular in recent years. By using these sites, users can share their location information of visiting different venues and places through check-ins, which provide an important set of information about user's activities and preferences. Thus, by combining the user interactions from different types of social networking sites, it is now possible to derive a new level of information about

user habits and preferences. In this chapter, we investigate how a user's psychological attributes such as personality and values affect her *eat-out* preference in different categories of restaurants. In particular, we use user's tweets in Twitter and check-ins in Foursquare, to predict user's eat-out preference in different categories of restaurants from her psychological attributes obtained from tweets.

People may visit different categories of restaurants, i.e., *cheap, moderate, expensive,* and *very expensive*, based on their interest, choice, or other socio-economic factors. For example, one may visit an expensive or very expensive restaurant, because of its attractive ambiance and decoration, and good flavor and presentation of foods. On the other hand, one may visit cheap restaurants because they cater foods quickly. Similarly, some people may prefer moderate restaurants as they provide better quality food, ambiance, and decoration than that of the cheap quality of restaurants. Several psychological studies [105, 113, 208] show that psychological attributes such as personality [113] and values [61] influence user's choice in different life-style and food consumption activities. Weaver et al. [208] find that user personality influences user selection and preferences as different choices are capable of providing different ranges of stimuli to the user and these stimuli are directly linked to her personality. Keller et al. [113] identify direct and indirect effect of Big5 personality traits on eating styles and food choices by conducting manual survey among 951 users. Similarly, *values* have profound impact on behavior and decision making [163], reading habits [9], buying products [202], etc.

Based on the above observation, we find that both *personality* and *values* influence user behavior and actions to a large extent. We are the first to present a detailed study that establishes strong correlations between different psychological attributes and eat-out preferences solely from social media interactions of users.

Few recent studies identify food habits and eating styles [1, 113, 205] from different demographic and visual attributes of social media interactions. Abbar et al. [1] identify users' dietary choices based on interest, demographics, and influence of social networks by analyzing Twitter data. Wagner et al. [205] show culinary preference (food consumption behavior) based on gender by analyzing Flickr images. However, none of these studies focuses on finding the association between user's *Big5 personality traits* and *values* with her *eat-out be-*

havior from his social media usage. Predicting eat-out preference of a user has a number of benefits. For example, by knowing the eat-out preference of Twiter users, restaurant owners can launch targeted advertisements. Similarly, a food chain service provider can decide on a location for its new outlet by knowing the eat-out preference of users in a region. One can also predict the economic state of a region to some extent by knowing the *eat-out* preference of the region.

In this chapter, we present a detailed study on how different psycholinguistic and psychological attributes of a user influence her eat-out preference from user's social media interactions. We first identify different psycholinguistic attributes such as *LIWC word categories* [153], and psychological attributes such as *Big5 personality traits* [214] and *value dimensions* [178] from user's tweets. Towards this direction, we collect tweets of 731 Twitter users who use Foursquare links in their tweets. Among these tweets, we take a total of 72,662 Foursquare links of restaurants that are categorized into four categories by Foursquare based on the food price. In the collected tweets, we find 23,986, 36,187, 10,335 and 2,154 links for *cheap, moderate, expensive* and *very expensive* categories of restaurants, respectively. Then, for each user we calculate the frequency of visits in different categories of restaurants. Next, we conduct users' linguistic features by using LIWC from her tweets. Then, we compute Big5 personality [108] traits and values of users by using IBM Watson personality insights API [1]. Later, we compute pearson correlation between linguistic features derived from LIWC, personality traits and value dimensions with the users' frequency of their visits in different categories of restaurants. Then, we build linear regression models from the correlated LIWC categories, Big5 traits and values to predict one's eat-out preference. However, strength of these independent linear regression models are not impressive. We observe that our psychological models (i.e., derived from personality and values) perform better than that of linguistic features based model. Then, we also integrate both of the psychological models by using a linear weighted ensemble technique as our eating behavior directly linked both of these attributes. Our integrated model improves significantly than that of independent personality and value models.

We find from previous studies [105,113,208] that both of the psychological attributes are

[1] https://personality-insights-livedemo.mybluemix.net/

important to persuade a person in decision making, we only ensemble these two attributes (i.e., personality and values) ignoring linguistic feature based model. Our ensemble model shows better performance than that of single models. Ensemble model has 14.31% (moderate category) and 21.4% (expensive category) R^2 strength for the weakest and the highest models, respectively. To measure the strength of our ensemble models, we again test with different classifiers. These classifiers have on average AUC scores of 63.8%. These AUC scores are significantly better than the baselines (AUC-49.3% for *ZeroR* classifier).

In summary, we have the following contributions:

- We are the first to exploit the data fusion of Twitter and Foursquare to predict users' *eat-out* preference from their linguistic features.
- We also predict users' *eat-out* preference from their two key psychological attributes: personality and values derived from social media interactions.
- We also integrate these two psychological attributes: personality and values to build a regression model of predicting *eat-out* preferences with better accuracy.
- We validate our hybrid model with a different dataset of Twitter.

A preliminary version of this work has appeared in [162], where we predict the eat-out preference of a user from her word usage patterns in tweets (Section 7.5.1). In this chapter, we have made significant advancement from our initial study. First, we build four different regression models to predict users' eat-out behavior from Big5 personality traits. Then, we build regression models where we use human values instead of personality. Then we integrate both personality traits and value dimensions to predict eat-out preferences. We also validate whether our models accurately predict users' frequencies of visiting different categories of restaurants which are similar to the original check-ins that they post through tweets.

7.2 Methodology

In this section, we first identify users' *eat-out* preferences through Foursquare check-ins from their linguistic features obtained from tweets. Later, we also predict their *eat-out* preferences

from Big5 personality traits and values dimensions. Since all users do not share Foursquare check-ins in their tweets, we predict their eat-out preferences from their psychological attributes. However, we can easily identify users' Big5 personality traits and value dimensions from their tweets by using IBM Watson personality insights API. Arnoux et al. [14] shows that IBM API for personality outperforms the state of the art techniques. To identify the eat-out preference of a Twitter user from her psychological attributes, we perform the following steps:

- *The Twitter and Foursquare crawler.* We only crawl the tweets of a user containing the Foursquare check-ins. We conduct our analysis in English words. To crawl the English tweets having Foursquare check-ins of a user, we use the Twitter advance search technique to find out the user.

- *Performing linguistic analysis.* We compute LIWC categories of words from users' tweets.

- *Computing personality and values from tweets:* We compute users' Big5 personality and value scores by using *IBM Watson personality insights API*.

- *Computation of visiting different categories of restaurants.* We use an *html* parser to get the price information (restaurant category) from the links of the foursquare check-ins. Later, we compute visiting frequencies of a user to different categories of restaurants.

- *Correlation analysis.* We compute *Pearson's correlation* analysis between users' linguistic features, Big5 personality traits and value dimensions with her visiting frequencies of different categories of restaurants based on the price.

- *Model building from LIWC categories, personality, values, and restaurant visiting frequency:* We select LIWC categories, personality and value features that are correlated with users' visiting frequency in a restaurant type. Then, we build regression models to predict *eat-out* preferences by using these correlated LIWC categories, personality traits and values dimensions independently.

- *Ensemble of models:* We build a weighted ensemble based predictive model to find users' eat-out preference from personality and values combinedly to improve the accuracy of independent personality/value based model.

- *Validation:* We validate our prediction results by using a different dataset of Twitter. We compute the statistical accuracy of our models between predicted visiting frequencies and original Foursquare check-ins that users' share through Twitter.

7.3 Data Collection

Extracting data of users' psychological attributes and restaurants' visiting frequency is one of the most challenging tasks in our work. We need psychological attributes of a user and also her restaurant visiting frequency, but there is no single source from where we can get all these information together. Therefore, to collect psychological attributes and restaurant visiting frequency, we make a fusion of both Twitter and Foursquare dataset. We collect 731 Twitter users for building model of *eat-out* preferences. Then, we also collect 220 Twitter users to validate whether our prediction model accurately predict frequency of visiting different categories of restaurants in real life.

Dataset for building eat-out model: We first search users who share Foursquare links in their tweets, since not all Twitter users use these links in their tweets. We use judgmental sampling technique [129], because we identify Twitter users who are active in Twitter and they share Foursquare check-ins through tweets. Therefore, we use Twitter advance search technique to find such user whose tweets contain Foursquare links. If a user uses Foursquare links, the check-ins of her tweets usually contain some keywords such as "4sq", and "Foursquare". We only search tweets in English words, because analyzing a single language for all users likely to produce uniform linguistic features, personality and value scores. We select users who live in different states such as "California", "Texas", "Florida", "Virginia" etc. of United States to ensure sufficient English language proficiency that reflects in their tweets. We notice that Foursquare links of these states contain available pricing information of different restaurant types. After selecting Twitter id's of users who regularly

tweets using Foursquare links, we collect their tweets by using python *tweepy* [2] implementation package.

We collect tweets of 731 users. We compute personality and value scores for the 731 users. From these 731 users, we find a total of 656,101 tweets. The users have maximum, minimum, and average of 3210, 189, and 897.54 tweets, respectively. Among the tweets, we discover a total of 72,662 Foursquare links for different restaurants. We divide the restaurant links based on prices, such as cheap, moderate, expensive and very expensive. As a ground truth data, we use frequency of visits in a restaurant type by a user.

Table 7.1: Restaurant Category with Price Sign ($).

Category	**Sign**
Cheap (C)	$
Moderate (M)	$$
Expensive (E)	$$$
Very Expensive (VE)	$$$$

Table 7.1 shows the restaurant category. We find dollar sign ($) on the Foursquare link page if the link is related to a restaurant and categorized by Foursquare. One ($) , two ($$), three ($$$) and four ($$$$) dollar sign signify cheap, moderate, expensive and very expensive categories of restaurants, respectively. Figure 7.1 shows website of a *moderate* ($$) type restaurant based on price which is directed from a Foursquare link.

We conduct analysis on each user file containing recent tweets and count the total number of Foursquare links. For our study, we perform our experiment with substantial tweets and Foursquare check-in data. We discard users who have less than 50 Foursquare check-ins. Each of the Foursquare links point to a web page in the Foursquare website. In some cases, the type and other information of the location are directly shown in the page. Otherwise, clicking the name of the place takes to the location page having restaurant's pricing information.

We perform html parsing to obtain information about the locations pointed by the Foursquare links. Not all links are related to restaurants, or food related services. Therefore, we discard

[2] http://www.tweepy.org

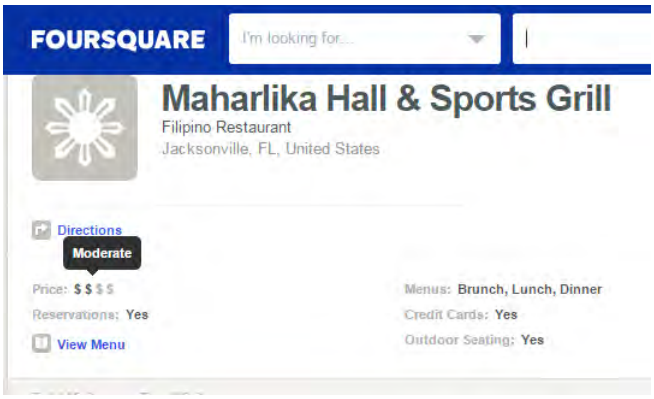

Figure 7.1: Moderate category of restaurant based on price ($).

the links that are not related to restaurants, or similar places. If a link is related to restaurant, Foursquare usually categorizes the restaurants based on the price of the available foods. We parse the Foursquare links of a Twitter user and collected information about the place or restaurant mentioned in the link. Thus, we calculate how many times a user visits each type of restaurant. We calculate the total number of categorized links. Then, we calculate the percentage of each type of restaurant that a user visited.

We find 33.01%, 48.80%, 14.22% and 2.96% links related to cheap, moderate, expensive and very expensive category of restaurants, respectively. Later, we conduct linguistic analysis of users' tweets by using LIWC. We also compute Big5 personality traits and value dimensions by using IBM Watson personality insights API from words in English language.

Dataset for validation in real life: For validating our eat-out model with the frequency of Foursquare check-ins, we collect another Twitter dataset. We again use judgmental sampling technique [129] to select 220 users who share frequently Foursquare check-ins through tweets. We select these users by using Twitter advance search technique. Then, we collect users' tweets by using python *tweepy* [3] implementation package as of November 20, 2017.

7.4 Feature Selection

In this section, we identify best subset of features to predict users' *eat-out* preferences from LIWC categories, personality and values derived from their tweets. We find few correlations between LIWC categories and eat-out preferences. Users' preference may also depend

[3] http://www.tweepy.org/

on personality, values, or both. According to Big5 model [108], personality has five different traits: *openness-to-experience, conscientiousness, extraversion, agreeableness, and neuroticism*. Values [178] have also five broader dimensions: *openness-to-change, hedonism, self-transcendence, conservation, and self-enhancement*.

First, we conduct LIWC based analysis of users' tweets. For this purpose, we use LIWClite7- a student version of LIWC tool. LIWC analyzes 70 different features of text in different categories. The categories are *linguistic processes* (word count, words longer than 6 letters, total pronouns, common verbs etc.), *psychological processes*, *personal concerns* (work, leisure etc. related words) and *spoken categories* (assent, noninfluencies etc.). The *psychological processes* is divided into five categories. It includes *social process* (words related to family, friends etc.), *affective process* (words related to positive emotion, negative emotion, anger etc.), *cognitive process* (insight, discrepancy, inhibition etc. related words), *perceptual process* (see, hear etc. related words) and *biological process* (body, health etc. related words) [153].

Since both of our independent variables (LIWC categories) and dependent variable (visiting frequency of different categories of restaurants) are continuous values, we apply *Pearson's correlation* (ρ) analysis to find significant correlation between these two variables. Table 7.2 shows only few important Pearson's correlations between LIWC category of words and restaurant type. We see that *cheap* restaurant is correlated with family, friend, leisure, and money LIWC categories of words. *Moderate* restaurant is positively correlated with leisure LIWC category of words. *Expensive* restaurant has correlation with family, friend, and money category of words. *Very expensive* restaurant has positive correlation with money category of words.

Then, we find association between personality and values with different categories of restaurants. One may think that personality traits may influence a person to choose a specific category of restaurant, while others may think that value dimensions may persuade a person to select a specific category of restaurant. It is difficult to state which traits of personality and values influence a person having an eat-out in which category of restaurants. We first identify features from both personality traits and value dimensions that are suitable to predict users' eat-out preference. Therefore, we also use Pearson's correlation coefficient to find out

Table 7.2: Pearson's Correlations between LIWC category of words and visiting frequency in different types of restaurants. [for significance level: *p<0.05, **p<0.01]

category	**Cheap**	**Moderate**	**Expensive**	**V. exp.**
family	**0.116****	-0.029	**-0.112****	-0.035
friend	**-0.145****	-0.002	**0.211****	-0.0006
leisure	**-0.114****	**0.107****	0.023	-0.0008
money	**-0.075***	-0.014	**0.086***	**0.093***

Table 7.3: Pearsons correlations between Big5 personality traits and visiting frequencies of different categories of restaurants. [for significance level: *p<0.05,**p<0.01

	Cheap	**Moderate**	**Expensive**	**V. Expensive**
Open.	-0.086	**0.103****	0.002	-0.033
Consc.	**-0.108****	0.019	**0.097***	**0.074***
Extraver.	-0.059	-0.006	**0.074***	0.054
Agreeable	**-0.108****	0.011	**0.106****	**0.079***
Neurotic	-0.015	0.017	0.030	0.038

correlation between users psychological attributes, i.e., personality and values, and visiting frequency of different restaurants variables, since both of our independent (personality and values) and dependent (visiting frequency of different categories of restaurants) variables are continuous. Tables 7.3 and 7.4 show the pearson correlations between personality traits and values, respectively, with different categories of restaurants where N=731. These tables show that specific personality traits and value dimensions are correlated with particular categories of restaurants.

7.5 Building Eat-out Models

In this section, we first build regression models from LIWC categories of words, personality traits, and value dimensions independently. Then, we investigate prediction potential of these regression models. Finally, we build a weighted linear ensemble of regression models by combining both personality traits and value dimensions.

Table 7.4: Pearsons correlations between value dimensions and visiting frequencies of different categories of restaurants. [for significance level: *p<0.05,**p<0.01

	Cheap	**Moderate**	**Expensive**	**V. Expensive**
Self-trans.	**0.078***	**0.081***	**-0.0771***	0.029
Open.	0.004	0.034	-0.001	**0.073***
Hedonism.	0.052	0.019	**0.096****	**0.073***
Self-enhan.	**-0.086***	0.049	0.007	**0.0742***
Conserv.	-0.013	-0.011	0.020	0.033

7.5.1 Eat-out Prediction from Linguistics

We perform linear regression analysis using WEKA machine learning toolkit to predict the eat-out preference of a person from her word use pattern in tweets [87]. In our linear regression model, we consider visiting frequency of a restaurant category is the dependent variable and LIWC category of words in her tweets are the independent variables. Table 7.6 presents the strength of the regression model. The R^2 of the linear regression models are moderate across all four restaurant categories based on price. Cheap and expensive type restaurant have better prediction strength than moderate and very expensive type of restaurant. Very expensive type restaurant has the lowest R^2 value among them.

We find that a number of LIWC category of words are correlated with the visiting frequency of a restaurant type. A potential problem arises when collinearity found between LIWC category and restaurant type. When there is a perfect linear relationship exists among independent variables, the outcome for a regression model can not be unique [43]. To eliminate collinearity among independent LIWC categories, we compute lasso penalized linear regression using *glmnet* R package [43] [93]. This technique reduces the coefficients to a low value or zero. Finally, we perform the linear regression analysis with a 10-fold cross-validation with 10 iterations. Motivated by the work, we also conduct the prediction potential using machine learning classification techniques [43]. Sumner et al. suggested that computing MAE and RMSE for error measure in regression analysis are not adequate [189]. In particular, when the majority of the individuals are around the mean for unimodal distribution, these error measures can often mask large errors.

Table 7.5: Best performing classifier to predict different restaurant categories from users' linguistic features.

Restaurant category	**Highest AUC achieving classifier**	**AUC**	**TPR**	**TNR**
Cheap	SVM	0.624	0.624	0.376
Moderate	Logistic Reg.	0.608	0.59	0.41
Expensive	Logistic Reg.	0.670	0.624	0.376
V. expensive	Logistic Reg.	0.564	0.551	0.449

Table 7.6: Strength of linear regression models by using personality traits and value dimensions.

Restaurant Category	**R^2 of linguistic**	**R^2 of Big5**	**R^2 of values**
Cheap	14.0%	15.8%	15.21%
Moderate	10.8%	12.4%	11.91%
Expensive	17.0%	19.1%	17.21%
V. Expensive	9.12%	13.7%	14.91%

We experiment with major classifiers that include logistic regression, Support vector machine (SVM), naive Bayesian, Adaboost, Random Forest and RepTree classifier using WEKA machine learning toolkit. We use these six classifiers to compute the accuracy of our models. Following the suggestion of Sumner et al., we perform supervised binary machine learning classification algorithms [189] [43]. We classify above-median level as high class label and below-median as low class label restaurant category. Table 7.5 presents the best classifier, it's true positive rate (TPR), false positive rate (FPR) and Area under the ROC curve (AUC) for computing visiting frequency of each of the restaurant category [68].

7.5.2 Eat-out Prediction from Personality

We also perform linear regression analysis by using WEKA machine learning toolkit to predict the eat-out preference of a person from her personality traits. In our linear regression models, visiting frequencies of restaurant categories are dependent variable and Big5 personality traits are independent variables. Table 7.6 presents the strength of the regression model.

Table 7.7: Best performing classifier to predict different restaurant categories from Big5 personality traits.

Restaurant Category	**Highest AUC achieving classifier**	**AUC**	**TPR**	**FPR**
Cheap	Naive Bayes	0.621	0.617	0.389
Moderate	RepTree	0.594	0.453	0.331
Expensive	RepTree	0.652	0.657	0.342
V. Expensive	Naive Bayes	0.603	0.584	0.602

The R^2 of the linear regression models are moderate across all four restaurant categories based on price.

Motivated by the work, we also conduct the prediction potential using machine learning classification techniques [43]. Following the suggestion of Sumner et al., we perform supervised binary machine learning classification algorithms [189] [43]. We classify above-median level of visiting restaurant frequencies as high class label and below-median as low class label restaurant category.

Table 7.7 presents the best classifier, it's TPR, FPR and AUC for computing visiting frequency of each restaurant category [68]. We use *ZeroR* classifier as baseline. The average AUC score of our baseline classifier is 0.493. We find the weakest AUC score (0.594) for moderate category of restaurants and the strongest AUC score (0.652) for expensive category of restaurants by using our classifiers. We also find 0.621 and 0.603 AUC scores by using our classifiers for cheap and very expensive categories of restaurants, respectively.

7.5.3 Eat-out Prediction from Values

We perform linear regression analysis by using WEKA machine learning toolkit to predict the eat-out preference of a person from her values. In this regression models, visiting frequencies of each restaurant category is also dependent variable and value dimensions are independent variables. Table 7.6 presents the strength of the regression model. The R^2 of the linear regression models are moderate across all four restaurant categories based on price. We also find out prediction potential using machine learning classification techniques [43].

Table 7.8: Best performing classifier to predict different restaurant categories from values.

Restaurant Category	**Highest AUC achieving classifier**	**AUC**	**TPR**	**FPR**
Cheap	Random Forest	0.620	0.608	0.443
Moderate	Random Tree	0.571	0.553	0.385
Expensive	Random Forest	0.633	0.631	0.382
V. Expensive	Random Forest	0.619	0.609	0.431

Following the suggestion of Sumner et al., we again perform supervised binary machine learning classification algorithms [189] [43].

Table 7.8 presents the best classifier, it's TPR, FPR, and AUC for computing visiting frequency of each of the restaurant category [68]. We use *ZeroR* classifier as baseline. The average AUC score of our baseline classifier is 0.493. We find the weakest AUC score (0.571) for moderate category of restaurants and the strongest AUC score (0.633) for expensive category of restaurants by using our classifiers. We also find 0.620 and 0.619 AUC scores by using our classifiers for cheap and very expensive categories of restaurants, respectively.

7.5.4 Eat-out Prediction using Personality and Values

We find associations in users' eat-out preferences with their personality traits and value dimensions (according to Section 3.4.1). Users' different psychological attributes (i.e., personality and values) may contribute differently while identifying users' eat-out preferences. Among our built regression models, one model may predict a certain restaurant category better than the other. For example, *cheap* category of restaurants can be predicted better with personality traits (R^2- 15.8%) than value dimensions (R^2-15.21%); on the other hand, *very expensive* category of restaurants can be predicted better with value dimensions (R^2-14.9%) than personality traits (R^2-13.7%). Since every psychological attribute contributes to predict users' eat-out preferences based on their strength, we combine all the regression models obtained from the previous independent personality and value models.

Thus, it is necessary to prioritize the psychological attributes based on their importance, as we compute eat-out preferences from personality and values. Performance of individual

Table 7.9: Weights (R^2 strength of regression models) derived from 30% of the dataset.

Psychological attributes	**R^2 of regression models**			
	Cheap	**Moder.**	**Exp.**	**V. Exp**
Personality	13.4%	11.2%	18.1%	11.5%
Values	13.4%	10.8%	16.4%	15.0%

Table 7.10: R^2 strength of ensemble model by integrating both Personality and Values derived from 70% of the dataset.

	Cheap	**Moderate**	**Expensive**	**V. Expensive**
Ensemble Model	18.47%	14.31%	21.4%	16.83%

regression model represents weights of different psychological attributes. To combine personality and value dimensions, we perform the following two steps: i) computing weights (the R^2 strength) for personality traits and value dimensions, and ii) combining the models with an weighted linear ensemble technique.

Learning weights: We determine the weight of personality and value scores to determine preference of eat-out behavior. To this end, we build regression models by using the data of 219 (30% of the total dataset) eat-out preference instances. To compute weights, we take both personality and value scores as input and the R^2 strength of the models as output. Table 8.3 presents the R^2 strength that are computed with different regression models to predict eat-out preferences by using 30% of the dataset.

A weighted linear ensemble: We build a weighted linear ensemble model from personality and values of 512 (70% of the total dataset) instances. We have built different regression models from personality and values to predict users' eat-out preferences that are described in previous sections. Since we train regression models that produce weights, we compute weighted linear ensemble score using the weights in Table 8.3. Finally, we build our weighted linear ensemble model using the weights generated from another dataset, so that our models do not get over-fitted. Table 7.10 presents the R^2 strength of our final ensemble based regression model. We also compute classification techniques to investigate the prediction potential of our ensemble model. Table 8.4 presents the eat-out preference classification result by using the ensemble of personality and values. We observe that the average AUC

Table 7.11: Classification result to identify eat-out preferences by using both Personality and Value scores

Eat-out Category	**Best classifier**	**AUC**	**TPR**	**FPR**
Cheap	Naive Bayes	0.654	0.643	0.397
Moderate	RepTree	0.595	0.599	0.591
Expen.	RandomForest	0.681	0.692	0.281
V. Exp.	RandomForest	0.631	0.628	0.411

of our classifier is 64.0%, and the baseline (*ZeroR*) accuracy is 49.3% among the ensemble classifiers.

We also observe that our classifiers largely outperform the random baseline. The accuracy of our random baseline is 50%. We observe that our ensemble of classifiers achieves higher accuracy than the independent personality and value based classifiers. Majority of the independent personality and value based classifiers have AUC value less or slight above than 60% but in the ensemble model, we see that majority of AUC values are significantly greater than 60% which indicates that when personality and values are combined, the model produces higher accuracy.

7.6 Evaluation

In this section, we validate whether our prediction model for *eat-out* preferences by using our hybrid model is similar to the original check-ins that they post through tweets. We investigate the accuracy of our hybrid model with a different Twitter dataset of 220 users. Then, we compute statistical significance test ($\tilde{\chi}^2$) to investigate the prediction accuracy with real check-ins of different categories of restaurants. Finally, we compare the accuracy of our hybrid model with other independent models.

7.6.1 Ground Truth Extraction

To measure the accuracy of our technique, we first extract ground truth of the eat-out preferences for the validation dataset of 220 Twitter users. We count user's visiting frequency of

Table 7.12: Chi-square test result between predicted eat-out preferences and real restaurant check-ins. Comparison among different models.

Restaurant category	**Linguistic model**		**Personality model**		**Value model**		**Hybrid model**	
	Chi-sq.	Sig.	Chi-sq.	Sig	Chi-sq.	Sig.	Chi-sq.	Sig.
Cheap	15.49	0.078	**16.94**	0.049	**17.01**	0.047	**19.01**	0.026
Moderate	7.92	0.551	13.23	0.21	11.33	0.261	16.83	0.053
Expensive	**16.95**	0.049	**17.53**	0.051	**16.99**	0.049	**21.17**	0.013
V. expensive	9.79	0.382	15.23	0.084	16.09	0.061	**16.96**	0.049

220 users to each type of restaurant. We find 29.07%, 49.10%, 10.51% and 4.32% links related to *cheap, moderate, expensive* and *very expensive* category of restaurants respectively.

7.6.2 Validation

In this subsection, we investigate the similarity between the predicted eat-out preferences by using our hybrid model and real frequency of Foursquare check-ins in our validation dataset. We also predict the eat-out preferences of the validation datasets by using our linguistic models. Then we normalize the predicted visiting frequency and original check-ins of the validation dataset into 0 to 1 scale. Based on the normalized score, we define three different labels: *low* (less than 0.4), *medium* (between 0.4 to 0.7), and *high* (greater than 0.7). Then, we compute *chi-square* ($\tilde{\chi}^2$) test to find the correlation between predicted eat-out preferences obtained from our models and Foursquare visiting frequency of restaurants' check-ins. We get two categorical variables (predicted eat-out preferences and original restaurant check-ins) each with four different types, which form a 4×4 contingency table. Table 7.12 shows the significant correlation between predicted eat-out preferences and collected restaurant check-ins. According to Lancaster et al. [119], we obtain significance level *p<0.05 for $\tilde{\chi}^2$>= 16.92 with degrees of freedom (df) equal to 9 (based on contingency table), where *df= (#rows-1)×(#col-1).*

Table 7.12 presents that our hybrid model largely predicts users' real frequencies of vis-

iting different categories of restaurants from our validation dataset accurately. We find that linguistic based model predicts accurately *expensive* categories of restaurants only. In contrast, both of our psychological models (i.e., personality and values) perform identical performance to predict frequency of visiting different categories of restaurants on our validation dataset. These two models correctly predict *cheap* and *expensive* categories of restaurants. In contrast, our hybrid model derived from both of the psychological models predicts *cheap, expensive,* and *very expensive* categories of restaurants accurately. However, prediction accuracy of our hybrid model is close to significant (16.83) for moderate category of restaurants. In our experiment, we find that our hybrid model performs better in predicting frequency of check-ins of different categories of restaurants than that of independent models.

7.7 Discussion

Our work is the first study to predict users' eat-out preference from their psychological attributes, i.e., Big5 personality traits and value dimensions, derived from social media usage.

Insights from linguistics: We observe from Table 7.2 that *very expensive* category has positive correlation with *money* LIWC category of words. For instance, it is likely that the users who usually write about money in their tweets, they tend to visit very expensive restaurants. Again, the users who visit *expensive* category restaurants likely to use friend and money category of words in their tweets. We observe that users who tend to use money category of words frequently, usually visit both of the very expensive, and expensive categories of restaurants. The result of the analysis also focuses that visiting a *moderate* type restaurant correlates with the usage of *leisure* LIWC category of words. It can be assumed that users are more prone to visit restaurants at leisure time. We also find that users who visit cheap restaurants frequently, they generally use family category of words. We observe that these users have negative correlation with friend, leisure, and money categories of words.

However, we also find few correlations that are not intuitively explainable. For example, cheap category have strong negative correlations with friend, and leisure LIWC category of words. Similarly, expensive category of words have negative correlation with family LIWC category of words. Therefore, we provide more focus on psychological attributes,

i.e., personality traits and value dimensions that have intuitive relations with users' eat-out behavior.

Findings from psychological attributes: Our experiment results show few intuitive findings from Tables 7.3 and 7.4. We find that majority of the personality traits and value dimensions are correlated with users' eat-out behavior. We observe from Table 7.3 that openness personality trait is strongly correlated with moderate categories of restaurants. People with high openness score tend to visit moderate category of restaurants, because they likely not to spend money by visiting expensive and very expensive restaurants, but they perhaps explore different dishes in moderate restaurants. High conscientiousness personality trait is strongly negatively correlated with cheap category of restaurants. In contrast, conscientiousness personality trait is positively correlated with visiting expensive and very expensive categories of restaurants. For instance, it is likely that the users who are conscious about their friends and family members, they prone to visit these restaurants for better hygiene, decor, quality of food (i.e., flavor, odor, texture) and ambiance. We find that extraversion personality trait has correlation with expensive category of restaurants. We also find that agreeable people are strongly correlated with their visiting frequencies of expensive and very expensive categories of restaurants. We observe that agreeable personality trait shows strongly negatively correlation with cheap category of restaurants. Similar to contentiousness personality trait, agreeable people are concern about the hygiene, and health of their friends and family members during visit to a restaurant, therefore they tend to visit expensive and very expensive categories of restaurants. They prone not to visit cheap category of restaurants. On the other hand, we do not find any association between neurotic personality traits and visiting any categories of restaurants. According to this study [94], eating behavior largely influence by conscientiousness and neurotic personality traits. However, neurotic people may not interested in sharing their check-ins about visiting restaurants via Foursquare, therefore we perhaps could not find any association between neuroticism personality traits and eat-out behavior. Another study [15] from Facebook reveals that neurotic people likely to share less information with their friends, therefore in our study, we also find less check-ins of neurotic people to predict eat-out preference.

We also observe from Table 7.4 that users' eat-out preferences have association with few

value dimensions. We find that self-transcendence value has correlation with visiting cheap category of restaurant. Since people with high self-transcendence scores likely to think about the welfare of others, therefore they do not tend to relish their life alone eating in expensive restaurants. We find that hedonism value dimensions have correlation with expensive and very expensive restaurants. People with high hedonism score likely to enjoy their lives visiting excellent categories of restaurants. We observe that self-enhancement value dimension are negative correlated with cheap categories of restaurants, because people with high self-enhancement score prone to eat in better quality of restaurants. However, we do not find association between conservative value dimension and any category of restaurants. We find an usual findings too. For example, we observe that openness value dimension has association with very expensive restaurants.

We also observe from Tables 7.3 and 7.4 that openness-to-experience (personality trait) and openness-to-change (value dimension) show different correlations with users' eat-out preferences. Since personality is largely an endogenous human characteristics while values are learned adaptation strongly influenced by the environment [147]. Social psychologists address personality traits and value dimensions as nature, and nurture, respectively. Therefore, openness-to-experience and openness-to-change attributes of personality and values likely to show different correlations with users' eat-out preference.

Strength of combined model: Table 7.10 presents the strength of our ensemble model derived combinedly from personality and values. We observe that expensive and moderate categories of restaurants show the strongest (21.4%) and the weakest (14.31%) R^2 scores, respectively, to predict from personality traits and value dimensions combinedly. We observe in Table 8.4 that RandomForest is the best classifier to predict expensive category of restaurants. The classifier has AUC, TPR, and FPR scores of 68.1%, 69.2%, and 28.1%, respectively. The baseline classifier, ZeroR, has AUC, TPR, and FPR scores of 49.3%, 61.5%, and 61.0%, respectively. From the table, we can observe that the RandomForest classifier of expensive category of restaurants performs largely better than the random baseline and other classifiers. We also observe that our ensemble based classifiers perform better than that of independent models of personality and values to predict users' eat-out preferences.

We observe that prediction of moderate category of restaurants shows weak potential than

that of other restaurant categories. We find in our dataset that people of high scores in all personalities and values prone to go to moderate categories of restaurants. Therefore, we do not find sufficient distinct features to predict moderate categories of restaurants accurately. Though we obtain moderate strength in our regression models, we can rely on the prediction result. From the value study of Chen et al. [43], we find that we can trust models that have moderate R^2 strength. We observe that different value dimensions show the R^2 strength between 13.8% to 18.2%. However, we improve the R^2 prediction strength in our ensemble model than that of independent models, therefore we can trust our ensemble model to predict eat-out preference accurately. In a previous well cited study [80] related to psycholinguistic research from social media, we observe that the size of the datasets is ranged from 250 to 300. Therefore, the size of our dataset (N=731) is sufficiently large to predict users' eat-out preferences by using psychological attributes.

7.8 Summary

In this chapter, we have predicted users' eat-out preferences from their psychological attributes, i.e., personality traits and value dimensions, derived from social media interactions. We have exploited the data fusion of Twitter and Foursquare to find out how psychological attributes affect users' eat-out preferences in real life. We have demonstrated which types of personality traits and value dimensions better predict which type of restaurant categories by computing correlations between them. Then, we have also built model to predict users eat-out from her personality and values, independently. We have improved the model by integrating both personality traits and value dimensions to predict users' eat-out preferences. We have also validated how our models perform in predicting users' real restaurant check-ins. The main advantage of our approach is that we can successfully predict eat-out preference of a person in spite of his not using Foursquare check-ins in her tweets.

In this chapter, we have showed that we can predict users' preferences of a tangible product, i.e., *eat-out*, from their psychological attributes. In the next chapter, we will show another case study to predict preference of an intangible product, i.e., *movie genre* and *rating*, from users' psychological attributes. We will conduct the experiment by making a fusion of an-

other two social media sites: *Twitter*, and *IMDb* to predict users' preferences.

Chapter 8

Predicting Movie Preferences from Users' Psychological Attributes

In this chapter, we investigate the second case study of our fourth research problem: prediction of movie preferences from users' psychological attributes. Section 8.2 describes the methodology of our work and Section 8.3 presents the data collection process. Then, Section 8.4 shows the technique to predict movie genre from psychological attribute and Section 8.5 describes the approach of predicting movie rating from users' psychological attributes. Later, Section 8.6 discusses the findings of this case study and Section 8.7 concludes the chapter.

8.1 Introduction

In recent years, social media sites such as Facebook and Twitter have become popular platforms for users to express feelings, and share happiness and sorrows. Thus, user tweets and statuses have become a major source of information for predicting user behavior. Many useful human attributes such as user sentiment [116], preferences [113], personality [80] and values [43] can be extracted from these user tweets and statuses.

A number of studies have been conducted to predict one's behavior or actions from her personality traits. For example, personality influences on profile picture choice [124], eating style, food choices [113], and artistic and scientific creativity [70]. Several studies also show the influence of personality on consumer buying behavior [200], career choice [167], and

political attitudes [25]. Weaver et al. [208] find that personality influences movie preference of the user since different movies are capable of giving rise to different ranges of stimuli, which are directly connected to human psychological state. On the other hand, basic human values are the factors that determine who will value which thing the most in her life. Values have profound impact on behavior and decision making [163], reading habits [9], buying products [202], etc. In the light of above discussion, we find that both personality and values influence user behavior and actions to a large extent. In this study, we investigate how these psychological attributes influence user movie genre preference and rating behavior in real life.

First, we develop a novel technique to predict user's movie genre preference from her psycholinguistic attributes i.e., personality and values obtained from social media. Since user tastes vary depending on her personality and values, movie genre choices might also vary depending on these psychological attributes. We collect users' tweets from Twitter and their movie genre from IMDb. Then we build a model that takes user tweets as input, computes her personality and values and outputs her genre preference. This model can recommend movies to a user by only knowing her personality and values from social media usage.

We further extend our work by predicting users' movie rating behavior from her personality, values, movie genre and storyline. In particular, after computing personality and value dimensions from tweets, we use these attributes along with movie genre and storyline to predict user's movie rating behavior. The key intuition of our approach is that since people are influenced by their psycholinguistic attributes like personality and values in both social and real life, they might be influenced by these factors in movie choices also. The main benefit of our approach is that, by using our proposed model, one can determine how a user will rate a movie of a specific genre and story only by knowing her personality and values from tweets and decide whether to recommend the movie to her or not.

Various attributes of a movie have impact on users such as actors, directors, awards, box office, year of release, movie type or genre, storyline, etc. Movie genre refers to the type of the movie, i.e., action, comedy, science fiction (sci-fi), biographies, horror, etc. People with different tastes would generally like different genres of movies. Some like movies which make them feel thrilled while others like movies which make them laugh. Again, it can also

happen that a person may like a movie of one genre while he may dislike another movie of the same genre. The key factor behind this scenario might be the story of the movie. In this chapter, we first investigate how user's personality and values persuade her in selecting movie genre. Later, we also explore how movie genre and story along with personality and values influence her movie rating behavior.

In our study, we have collected data of 330 users who have both Twitter and IMDb profiles. We collected a total of 4,70,899 tweets of these users to compute their personality and values. First, we compute Big5 personality [108] traits and Schwartz value [178] dimensions of users from tweets by using IBM Watson personality insights API[1]. Next, we collected the movie reviews, ratings of these users in IMDb and also collected the corresponding movie genres and stories. In total, we have collected ratings of 7,168 distinct movies of 28 different movie genres. To predict genre preference from personality and values, we take ratings of 6725 distinct movies of 5 different movie genres which have sufficient number of ratings. After computing personality and values scores, we select personality and value features that are highly correlated with users' movie genre preference. Then we build a classification model to predict movie genre preferences from users' personality, and values independently. Next, we build an ensemble model that combines both personality and values of a user to predict her preference in choosing different movies based on their genre. Personality and values based classifiers have average AUC scores of 58.0%, and 60.0%, respectively. Ensemble based classifiers have average AUC score of 63.0%. These AUC scores are significantly better than the baseline (AUC-56.2% for ZeroR classifier) for movie genre prediction.

To predict user's movie rating, we conduct a psycholinguistic based analysis on the textual content of the movie storylines by using Linguistic Inquiry and Word Count (LIWC) [154]. We find several significant correlations between movie rating and psycholinguistic scores. Then we build four different regression models to predict user rating behavior by incrementally adding reasonable features such as users' personality, values, genre, storyline, and reviews. First, we build a model from personality, movie genre and storyline that has an adjusted R-squared value of 68.2%. Then, we build another model where we use human values instead of personality. The model has an adjusted R-squared value of 68.5%. We obtain lit-

[1]https://personality-insights-livedemo.mybluemix.net/

tle improvement (0.3%) for adjusted R-squared value compared with personality trait based model. Then we further find that the model improves when we use both personality and value dimensions with genre and storyline. The model achieves an adjusted R-squared value of 72.3%. Later, we assume that the topics about which a user writes most in her reviews are her interest topics. Therefore, we consider that if user's interest topics have a similarity with movie topics, then she is likely to rate the movie high. Thus, we compute similarity coefficient between user's review and storyline of the movies. After adding the similarity coefficient with existing feature set, we obtain the adjusted R-squared value of 73% which improves than that of previous models. In this way, we can recommend movies to the users more efficiently who write reviews in IMDb. We also build the corresponding binary classifier of the regression models to understand the prediction potential and find the AUC value of our RepTree based classifier is 66.9% which is better than the baseline (AUC-50% for ZeroR classifier) of movie rating prediction.

In summary, we have the following contributions:

- We are the first to exploit the data fusion of Twitter and IMDb to predict users' movie genre preferences from their tweets and movie rating behavior from their tweets, movie genre, storylines and reviews.
- We integrate two psychological attributes: personality and values to build a classification model to predict users' movie genre preference.
- We also integrate personality and values to build a regression model that can accurately predict users' movie rating behavior from their personality, values, movie genre and storyline.
- We further improve the model by using similarity coefficient between user review and movie storylines for existing users in IMDb who write reviews for movies.
- Our experimental study on 330 users shows the efficacy of our approach.

A preliminary version of this work has appeared in [136] where we predict movie genre preference from personality and values. However, a person may not always like all the movies of a specific genre. Therefore, in this chapter, we further build a model which predicts

how a particular user will like a specific movie considering few additional attributes such as storylines, reviews, etc. We use two user attributes: personality and values, and two movie attributes: genre and storyline. Then we build regression models to predict how the user will rate the movie on a scale of 1 to 10. Our model also works for users who do not have IMDb profile or have not written any reviews. However, we further improve the model for the users who write reviews by adding similarity coefficient between the story topic and user interest topic collected from her reviews. In our previous chapter, we used a dataset of 232 users. In this work, we collected data from 98 more users, and currently the size of our dataset is 330 users.

8.2 Methodology

In our chapter, we predict users' i) movie genre preferences, and ii) movie rating behavior, from their psychological attributes. First, we extract users' tweets by using Twitter API. Then we extract IMDb movie genre, rating, movie story lines, and reviews of the users who have same Twitter *username* that we extracted. Later, we identify users' psychological attributes from their tweets. Then, we compute correlations between users' psychological attributes and their preferences in IMDb. Next, we build classification and regression models that predict movie genre preferences and rating behavior, respectively. To improve our models of movie rating, we add different features from users' IMDb data. We present our methodologies in two different modules: a) movie genre preferences, and b) movie rating behavior.
a) To build classification model to predict genre preference, we perform the following steps:

- *Computing personality and values from tweets:* We compute users' personality and value scores by using *IBM Watson personality insights API*.

- *Model building from personality, values, and movie genre:* We select personality and value features that might be correlated with users' movie genre preference. Then we build classification model to predict movie genre preferences from users' personality and values independently.

- *Ensemble of models:* We build a weighted ensemble based predictive model to find

users' movie genre preference from personality and values combinedly to improve the accuracy of independent personality/value model.

b) To build regression model to predict user rating, we perform the following steps:

- *Model building from personality and values individually*. We select the features from personality, value dimensions, LIWC attributes of movie storylines, genres which might be correlated with the user rating behavior. Then we build individual regression models from personality or values along with movie genre and storyline to predict user's movie rating.

- *Model building from both personality and values*. We build a model using both psycholinguistic attributes, i.e. personality traits and value dimensions along with movie genre and storyline. The model can predict the rating behavior of a user for a movie of a specific genre and story from her tweets where she does not need to have an IMDb profile.

- *Final model building with similarity coefficients*. For existing users in IMDb, we enhance our model by adding similarity coefficient between storyline and user review. Towards this direction, we collect users' movie reviews and find their interest topics in reviews. Then we compute the jaccard coefficient for user interest topics and story topics. We find that there exists a correlation between this jaccard coefficient and the corresponding user rating. Therefore, we add the jaccard coefficient attribute with the other attributes of our model and build the final regression model to improve the accuracy of movie rating prediction.

8.3 Data Collection

Getting data with user psychological attributes, genre preference and rating behavior is one of the most challenging tasks in our work. We need psycholinguistic attributes of a user and also her movie ratings, and reviews but there is no single source from where we can get all these information together. Since users only write reviews and rate movies in the movie rating websites (i.e., IMDb), we cannot extract their personality/values from such

sites. On the other hand, in social media, users express their feelings, and thoughts and these sites can be a great source of analyzing their psycholinguistic attributes but we cannot collect users' movie rating behavior from social media. Therefore we collect users' both psycholinguistic and movie preference attributes by manually linking profiles of IMDb and Twitter users. Though we can compute users' personality/values by analyzing their reviews in IMDb, the content of reviews largely address to the subject matter of the movie story only. In contrast, majority of the users' express their opinions, thoughts, and feelings in social media on regular basis. Thus social media data is more appropriate to analyze one's personality/values effectively.

We found that users in IMDb with the same username can be found in Twitter also. We manually check the user name from movie reviews and search for the reviewer in Twitter with the same username, profile picture and location. We check the latest movies that are released in the year between 2015 to 2017, so that reviewers information (i.e., current location) are likely to be updated. We manually search same users in Twitter by using the same name of IMDb. We confirm a user by her location, and profile photo that are same in IMDb. First, we randomly select both high (i.e., *The Martian*-8.0) and low (i.e., *Ghost in the Shell*-6.5) rated movies from the yearly searching list [2] of IMDb. Then, we start finding users by manually checking in the review section of the previously selected movies from the yearly list.

We extract a total of 349 IMDb users whose Twitter profile can be found with the same username. Then we collect users' tweets by using python *tweepy* [3] implementation package. Among the collected 349 users, we discard 19 users as they have few tweets to analyze personality and values by using IBM Watson personality insights API. The API requires at least 100 words to start analyzing. We observe that we can compute personality/values by using the API for the users who have on an average 12-15 tweets. We compute personality and values scores for the rest of the 330 (349-19) users. From these 330 users, we find a total of 4,70,899 tweets. The users have maximum, minimum, and average of 3221, 172, and 1427 tweets, respectively. When we search and find cross-linked users, we observe that few users are highly active in IMDb but they have few tweets (i.e., 5-8) in Twitter. We do not consider these users as we cannot compute their personality/values by using IBM Watson

[2] http://www.imdb.com/year/

[3] http://www.tweepy.org/

Table 8.1: Personality based classification of movie genres.

Genres	**Best classifier**	**AUC**	**TPR**	**FPR**
Drama	Random Tree	0.59	0.81	0.71
Thriller	Rep Tree	0.54	0.17	0.09
Comedy	Random Tree	0.55	0.01	0.01
Action	Rep Tree	0.61	0.09	0.05
Adventure	Random Tree	0.61	0.09	0.04

personality insights API. In contrast, we find a large number of users who are highly active in Twitter and have a decent amount of reviews in IMDb. We use highly active users in Twitter since we get accurate personality/value score from IBM Watson Personality API.

We fetch users' movie reviews, and ratings from IMDb by using Python *BeautifulSoup* [4] implementation package. The corresponding storylines and genres of movies are extracted using *IMDbPY* [5] python package. In total, we have 45,212 instances of user rating. Users rate a total of 7,421 distinct movies from their account creation date. We find a total of 28 different genres of movies. To build model for genre preference, among these 28 genre of movies, we consider only 5 genre of movies as these are most prominent in the dataset. The genres' names and percentages are: *drama (17.52%), thriller (12.48%), comedy (11.01%), action (12.39%), and adventure (9.71%)*, respectively.

8.4 Genre Preference from Tweets

In this section, we predict user's movie genre preferences from her personality and values. First, we identify best subset of personality and value dimensions as independent variables to predict user's movie genre preferences. Later, we build classification models from both personality and values, independently. Then, we combine both personality and values based classifiers models to increase the accuracy of the independent models.

[4]https://pypi.python.org/pypi/beautifulsoup4

[5]https://pypi.python.org/pypi/IMDbPY

8.4.1 Feature Selection

Users' movie genre preference may depend on personality, values, or both. According to Big5 model [108], personality has five different traits: *openness-to-experience, conscientiousness, extraversion, agreeableness, and neuroticism*. Values [178] have also five broader dimensions: *openness-to-change, hedonism, self-transcendence, conservation, and self-enhancement*. One may think that personality traits may influence a person to choose a movie of a specific genre, while others may think that value dimensions may persuade a person to select a certain movie genre. It is difficult to state which traits of personality and values influence a person to watch which genre of movies. Thus, we first identify features from both personality traits and value dimensions that are suitable to predict movie genre preference.

We use genre of movies as ground truth data (dependent variable), and personality and value scores as independent variables. We select the best subset of personality and value traits (predictors) using *forward selection* approach of *leaps* [6] R package implementation. Since the forward selection approach starts with no predictors, we can identify which predictors are more significant than the other when these are added stepwise to the list. Leaps package performs an exhaustive search to find out best subset of personality and value dimensions by using an efficient branch-and-bound algorithm. We first select the best subset of 3 features both from the 5 broad personality and values dimensions, independently. Then, we also identify the best subset of 5 features among the 10 features of personality and values, combinedly. For example, we select *extraversion, conscientiousness, and neuroticism* personality traits and *hedonism, conservation, and openness-to-change* values independently as features for computing movie genre preference by using the best subset selection approach. Then we identify the best subset of features for computing movie genre preferences from both personality and values are: *neuroticism, extraversion, agreeableness, hedonism, and self-transcendence.*

8.4.2 Genre Classification

In this subsection, we first build classification models from personality traits and value dimensions independently. Then, we investigate prediction potential of these classification

[6] https://cran.r-project.org/web/packages/leaps/

Table 8.2: Value based classification of movie genres.

Genres	**Best classifier**	**AUC**	**TPR**	**FPR**
Drama	Random Tree	0.61	0.81	0.71
Thriller	Random Tree	0.58	0.21	0.11
Comedy	Random Tree	0.61	0.01	0.01
Action	Rep Tree	0.60	0.09	0.06
Adventure	Rep Tree	0.61	0.07	0.03

Table 8.3: Weights (AUC of best classifiers) derived from 30% of the dataset.

Psychological attributes	**AUC of best classifier**				
	Action	Adven.	Thril.	Dram.	Com.
Personality	0.56	0.60	0.61	0.61	0.65
Values	0.60	0.62	0.58	0.60	0.63

models by using machine learning techniques. Finally, we build a weighted ensemble of classification model by combining both classifiers of personality and value.

We observe that each movie may fall in the intersection of more than one genre names. For example, *Gone Girl* movie has three different genres: *crime, drama, and thriller*. Since we build classifier for movie genre preference, we need only one class label for each instance. Therefore we distribute the genre names for a single movie name into multiple rows. In IMDb, users may rate a movie on a scale of 1 to 10. Since we identify users' movie genre preferences, we discard instances that have low rating, i.e., rating <7. When a user gives a low score to a movie genre, it signifies that she has less interest on that movie genre. Finally we have a total 23,410 instances that have a single movie genre name.

8.4.2.1 Genre Classification using Personality

We apply Naive Bayes, Support Vector Machine (SVM), Random Forest, Random Tree and RepTree classifiers in our dataset by using WEKA [87] machine learning toolkit. Table 8.1 presents the best classifier, its true positive rate (TPR), false positive rate (FPR) and area under the ROC curve (AUC) for predicting different genre of movies from users' personality traits. TPR defines how many samples are correctly classified as positive among all positive samples and FPR defines how many samples are incorrectly classified as positive among all

Table 8.4: Classification result to identify genre by using both personality and value scores.

Genres	**Best classifier**	**AUC**	**TPR**	**FPR**
Drama	RandomTree	0.648	0.843	0.654
Thriller	RandomForest	0.601	0.185	0.106
Comedy	RandomForest	0.604	0.315	0.202
Action	RepTree	0.648	0.584	0.321
Adventure	RandomTree	0.651	0.653	0.301

negative samples available during the test. We conduct the performance of the classifiers by using AUC values under the 10-fold cross validation. We find that on an average the AUC of our classifier is 58%. We use *ZeroR* classifier as baseline method, which has an AUC score of 50%. We observe that our classifiers always outperform the baseline. We also find that mean absolute error (MAE) of our classification model is 0.27.

8.4.2.2 Genre Classification using Values

In this subsection, we again apply Naive Bayes, SVM, Random Forest, Random Tree and RepTree classifiers in our dataset to predict different genre of movies from users' value dimensions. We conduct the performance of the classifiers by using AUC values under the 10-fold cross validation. Table 8.2 presents the best classifier, its TPR, FPR, and AUC for predicting different genre of movies from users' value dimensions. We find that the average AUC of our classifier is 60.2%. The AUC of our baseline is 50%. We observe that our classifiers largely perform better than the baseline. We also find that MAE of our classification model is 0.30.

8.4.2.3 Genre Classification using Personality and Values

We find associations in users' movie genre preferences with their personality traits and value dimensions (according to Section: Feature Selection). Users' different psychological attributes (i.e., personality and values) may contribute differently while identifying movie genre preferences. Among our built classifiers, one classifier may predict a movie genre better than the other. For example, *action* genre of movies can be predicted better with

Table 8.5: Pearson Correlation between psycholinguistic attributes and movie rating. [for significance level: *p <0.05,**p <0.01]

Personality traits	**Rating**	**Value dimensions**	**Rating**
Conscientiousness	0.140^{**}	Openness to change	0.020^{**}
Neuroticism	0.137^{**}	Self transcendence	0.0218^{**}
Openness to experience	-0.075^{**}	Conservation	0.193^{**}
Extraversion	-0.052^{**}	Self enhancement	0.082^{**}
Agreeableness	0.044^{**}	Hedonism	0.037^{**}

personality traits (61%) than value dimensions (60%); on the other hand, *thriller* genre of movies can be predicted better with value dimensions (58%) than personality traits (54%). Since every psychological attribute contributes to identify users' movie genre preferences based on their strength, we combine all the classification models obtained from the previous independent personality and value models. Thus, it is necessary to prioritize the psychological attributes based on their importance, as we compute movie genre preferences from personality and values. Performance of individual classifier represents weights of different psychological attributes. To combine personality and value dimensions, we perform the following two steps: i) computing weights for personality traits and value dimensions, and ii) combining the models with an weighted linear ensemble technique.

Learning weights: We determine the weight of personality and value scores to determine preference of movie genres. To this end, we build classification models by using the data of 7,023 (30% of the total dataset) movie genre preference instances. To compute weights, we take both personality and value scores as input and the best AUC score of movie genre classifier as output. Table 8.3 presents the best AUC scores that are computed with different classifiers to predict genre of movie preferences by using 30% of the dataset.

A weighted linear ensemble: We build an weighted linear ensemble model from personality and values of 16,387 (70% of the total dataset) instances. We have built different classifiers from personality and values to predict users' movie genre preferences that are described in previous sections. Since we train different classifiers that produce weights, we compute weighted linear ensemble score using the weights in Table 8.3. Finally, we build our

Table 8.6: Pearson's correlations between LIWC categories of storyline and movie rating [for significance level: *p <0.05,**p <0.01]. LIWC categories with no significant correlations are omitted from the table.

LIWC Category	**Abbreviation**	**Examples**	**Movie Rating**
Word Count	WC		0.029^{**}
Question Marks	Qmarks		-0.031^{**}
Unique	Unique		-0.034^{**}
First-person plural	We	We, us, our	0.019^{**}
Past Tense	Past	Went, ran, had	0.032^{**}
Present Tense	Present	Is, does, hear	-0.054^{**}
Future Tense	Future	Will, gonna	-0.014^{**}
Numbers	Number	Second, thousand	0.014^{**}
Swear words	Swear	Damn, piss, fuck	-0.018^{**}
Social processes	Social	Mate, talk, they, child	0.038^{**}
Friends	Friend	Buddy, friend, neighbor	0.018^{**}
Humans	Humans	Adult, baby, boy	-0.019^{**}
Affective Processes	Affect	Happy, cried, abandon	-0.022^{**}
Negative Emotion	Negemo	Hurt, ugly, nasty	-0.038^{**}
Anxiety	Anx	Worried, nervous, fearful	-0.049^{**}
Anger	Anger	Hate, kill, annoyed	-0.016^{**}
Sadness	Sad	Crying, grief, sad	-0.013^{*}
Cognitive mechanisms	Cogmech	Cause, know, ought	-0.012^{*}
Insight	Insight	Think, know, consider	-0.011^{*}
Causation	Cause	Because, effect, hence	0.040^{**}
Tentative	Tentat	Maybe, perhaps, guess	-0.017^{**}
Perceptual processes	Percept	Observing, heard, feeling	0.018^{**}
See	See	View, saw, seen	0.015^{**}
Hear	Hear	Listen, hearing	0.016^{**}
Body	Body	Cheek, hands, spit	-0.022^{**}
Health	Health	Clinic, flu, pill	0.023^{**}
Sexual	Sexual	Horny, love, incest	-0.032^{**}
Relativity	Relativ	Area, bend, go	0.017^{**}
Motion	Motion	Arrive, car, go	0.017^{**}
Space	Space	Down, in, thin	0.013^{*}
Work	Work	Job, majors, xerox	0.029^{**}
Achievement	Achieve	Earn, hero, win	0.044^{**}
Home	Home	Apartment, kitchen	-0.025^{**}
Money	Money	Audit, cash, owe	-0.014^{**}
Religion	Relig	Altar, church, mosque	-0.013^{*}
Death	Death	Bury, coffin, kill	-0.028^{**}

weighted linear ensemble model using the weights generated from another dataset, so that our models do not get over-fitted. Table 8.4 presents the movie genre preference classification result by using the ensemble of personality and values. We observe that the average AUC of our classifier is 63.4%, and the baseline (*ZeroR*) accuracy is 56.2% among the ensemble classifiers. We also observe that our classifiers largely outperform the random baseline. The accuracy of our random baseline is 20%. We also find that MAE of our model is 0.18. We observe that our ensemble of classifiers achieves higher accuracy than the independent personality and value based classifiers. Some of the independent personality and value based classifiers have AUC value less than 60% but in the ensemble model, we see that all AUC values are greater than 60% which indicates that when personality and values are combined, the model gives us higher accuracy.

8.5 Rating Behavior for a Movie

In this section, we predict user's movie rating behavior from her personality, values, movie genre, review, and story lines. First, we compute psycholinguistic categories (i.e., LIWC categories) of movie story lines and find out correlation with user's movie rating. Then we find out best subset of features from LIWC categories of movie story lines to compute user's movie rating. Later we build regression model for computing user's movie rating from users' personality and values. Next, we improve the above models by adding features such as movie genre, reviews,story lines, etc. incrementally.

8.5.1 Feature Selection

In this subsection, we compute correlation between users' different attributes (e.g., psychological attributes, movie story lines, and genre) and their movie rating behavior to identify important features for building rating prediction models. First, we identify important features from users' psychological attributes, i.e., personality traits and value dimensions, to predict movie rating behavior.

We use *Pearson correlation coefficient* to find out correlation between users' psychological attributes and movie rating variable, since both of our independent (personality and

Table 8.7: One Way ANOVA test results for association between genre and movie rating.

	Sum of Sq.	**df**	**Mean Sq.**	**F**	**Sig.**
Between Groups	4266.8	4	158.03	37.06	.000
Within Groups	160702.82	37689	4.26		
Total	164969.6	37693			

values) and dependent (rating) variables are continuous. Table 8.5 shows the pearson correlation (ρ) between different psycholinguistic attributes and movie rating where N=45,212. The test shows that all the personality and value traits are strongly correlated with movie rating.

Later, we also compute pearson correlation between linguistic features in movie storylines and users' rating. We use LIWC to extract linguistic features from movie storylines. LIWC produces statistics on 70 different features of text in five categories. These categories include Standard Counts (word count, words longer than six letters, number of prepositions, conjunctions, pronouns), Psychological Processes (emotional, cognitive, sensory, and social processes), Relativity (words about time, the past, the future), Personal Concerns (occupation, financial issues, health), and Other dimensions (counts of various types of punctuation, swear words). Among 70 LIWC categories, we find a total of 36 categories which are significantly correlated with movie ratings. Table 8.6 shows significant correlations between LIWC categories and movie rating (N=45,212).

We also compute correlation between movie genre and movie rating. Since movie genre is categorical variable and movie rating is continuous variable, we perform one way Analysis of Variance (ANOVA) test using SPSS [48] to find out associations between movie genre and movie ratings. Table 8.7 presents that movie genre and movie ratings have strong correlation with each other where we find $p < 0.001$. We also find the *F-statistic* from our test. In our experiment, F-statistic is the ratio of variation of movie rating between the group of genres divided by the variation of movie rating within the group of genres [75]. We observe that our F-statistic score has a reasonably large value (37.06). Therefore, we infer that movie genre categories have also a greater effect on users' movie rating behavior.

Next, we use a lexical analyzer tool, Empath [67], a tool that can generate and validate new lexical categories on demand, to extract topics from all reviews of a user. We assume

these topics are her interest topics since she writes about the topics in most of her reviews. We also extract the movie story topics using the same tool. Then we find the similarity between her interest topics and story topics using Jaccard coefficient [142] and observe if there is any correlation between her interest topics and the corresponding rating given by her for the movie. We find a Pearson correlation coefficient of 0.017 which is significant at level 5% (N=45,212).

8.5.2 Building Models to Movie Rating

In this subsection, we build different models to investigate the impact of personality and values independently on user rating behavior using linear regression models. Next we combine the models to observe the impact of these psychological attributes together. Later, we increase the strength of the model by adding a new attribute, Jaccard coefficient of users' review and movie story lines, for the users who write reviews in IMDb. Then, we investigate the potential of different regression models and we also build classification models to verify the strength of our models.

We observe that each movie may fall in the intersection of more than one genre names. For example, *Gone Girl* movie has three different genres: crime, drama, and thriller. Since movie genre is an attribute for rating prediction, we distribute the genre names for a single movie name into multiple rows. In this way, we have a total of 45,212 different movie rating instances. In IMDb, users may rate a movie on a scale of 1 to 10. Since the target variable is continuous, we use different independent variables to build regression models that fit well with our dataset.

Predicting user's movie rating behavior refers how accurately our model predicts a user's rating for a certain movie. For building a movie rating prediction model, we use some user attributes and also some movie attributes. We use personality traits and values dimensions independently and combinedly as user attributes. We also use genre and LIWC categories of storyline as movie attributes. In this chapter, we use different combinations of user and movie attributes to predict rating behavior. For movie attributes, we use genre and storyline together because a movie is not uniquely identifiable by its genre only. If we use only genre with personality and values, it will not be movie specific, because a user may not rate equally

Table 8.8: Strength of different regression models for movie rating prediction.

Features	**Best Model**	**Adj. R^2**	**R^2**	**RMSE**
Personality, Genre, Story	Rand. Forests	0.682	0.77	1.183
Values, Genre, Story	Rand. Forests	0.685	0.787	1.178
Personality, Values, Genre, Story	Rand. Forests	0.723	0.815	1.104
Personality, Values, Jaccard Coeff., Genre, Story	Rand. Forests	0.73	0.823	1.095

all the movies of same genre. Therefore, we build different regression models by using following reasonable combination of feature sets to predict movie rating accurately.

8.5.2.1 Building Model using Personality, Genre and Story

We compute linear regression, Multivariate Adaptive Regression Spline (MARS) model, Classification And Regression Trees (CART), Random Forest Tree ensembles with our dataset using Salford Predictive Modeler[7] (SPM 8.0) and SPSS [48]. In this model, we use user personality traits, movie genre, and correlated LIWC categories of movie storylines to predict user rating. Table 8.8 presents the adjusted R-squared, R-squared scores and root mean squared error (RMSE). Among all the models, we observe that Random Forests Tree Ensembles best fit the data with an adjusted R-squared value of 0.682 and RMSE of 1.183.

8.5.2.2 Building Model using Values, Genre and Story

We also investigate the impact of Basic Human Values on movie rating. For this, we use value dimensions, movie genre, correlated LIWC categories of movie storyline as features and build several regression models like Subsubsection 8.5.2.1 and observe the strength of the models. From Table 8.8, we find that the best model is Random Forests Tree Ensembles. The model obtains an adjusted R-squared value of 0.685 and RMSE of 1.178 which is slightly better than that of the previous model using personality traits.

[7] http://www.salford-systems.com

8.5.2.3 Building Models using Personality, Values, Genre and Story

We find that both personality and values have substantial impact on user rating behavior. In this combination of feature subset to build regression model, we use both personality and values along with movie genre and stories. Table 8.8 shows that the model improves moderately (0.038) than that of previous models where the best regression model is Random Forests Tree Ensembles. The model achieves an adjusted R-squared value of 0.723 and RMSE of 1.104.

8.5.2.4 Building Model using Personality, Values, Genre, Story and Reviews

In our previous models, we observe that our model can predict user rating for a movie of a specific genre and story from only her tweets where a user may not have an IMDb profile. Furthermore, we can improve the model for existing users in IMDb by using their reviews. In IMDb, a user's profile contains all the reviews that she writes for different movies. In these reviews, users largely write about their interest topics and we extract users' interest topics from their reviews using *Empath* [67]. Then we find the Jaccard coefficient between the topics and story topics. We observe that the Jaccard coefficient has a correlation with movie rating. We build regression models using the attribute along with previous attributes and find that the best model is Random Forests Tree Ensembles. The model improves slightly (0.007) than that of previous model and obtains an adjusted R-squared value of 0.73 and RMSE of 1.095.

8.5.3 Classification Analysis

According to the studies [43, 189], computing few error measures such as MAE, RMSE are not sufficient to check the prediction potential of a regression model. Since rating scores have unimodal distributions, RMSE suffers in lack of investigating strength of a prediction model. To overcome this limitation, we apply different binary classification algorithms on our dataset. For this, we label the target variable 'rating' as *high* if its value is greater than or equal to the median 7, otherwise we label it *low*. We apply several classifiers, i.e. Naive Bayes, Adaboost, Rep Tree, Support Vector Machine by using WEKA [87] machine learning toolkit.

Table 8.9: Strength of different classifiers for movie rating prediction.

Classifier	**AUC**	**TPR**	**FPR**	**MAE**	**RMSE**
Naive Bayes	0.625	0.612	0.450	0.4105	0.524
Adaboost	0.651	0.655	0.513	0.444	0.467
Rep Tree	0.669	0.664	0.427	0.396	0.481
SVM	0.556	0.644	0.533	0.356	0.596

Table 8.9 shows, the best classifier Rep Tree has AUC value of 66.9% which is better than the baseline AUC value of 50% for ZeroR classifier. Our model largely outperforms the random baseline.

8.6 Discussion

Our work is the first study to predict i) movie genre preference, and ii) rating behavior from psycholinguistic attributes, i.e., personality and values, derived from social media usage. We observe in Table 8.1 that personality trait based genre prediction classifiers perform better than the baseline (AUC-50%) ZeroR classifier. We find that drama and thriller movie genres show weak classification accuracy. In contrast, action and adventure genre of movies show moderately strong classification accuracy. We observe in Table 8.2 that value based classifiers show identical performance to personality trait based classifiers for adventure genre of movies. However, value based classifier shows significant better result for comedy genre of movies and weaker result for action genre of movies than that of personality trait based classifiers. From Table 8.4, we see that our ensemble based classifiers shows higher AUC scores than that of personality/value based independent classifiers. We find that adventure genre classifier shows the strongest (65.1%) and thriller genre classifier shows the weakest (60.1%) AUC scores. We also find that classifiers for thriller, and comedy movie genres show low TPR scores. These classifiers also show substantial improvement for TPR and FPR scores for movie genre classifiers than personality/values based independent models.

In this chapter, we also investigate the impact of personality and value dimensions on the movie rating behavior of users. We find few interesting results that are quite intuitive. From

Table 8.5, we see that all personality traits and value dimensions are strongly correlated with users' movie rating behavior. We also observe that openness to experience and extraversion personality traits are strongly negatively correlated with movie rating. People with high openness to experience personality score may evaluate a movie with low rating, because they tend to expect newness in a movie content than that of a movie usually presents. Table 8.6 presents the Pearson correlations between LIWC categories of movie storylines and movie rating behavior. Among the correlated 36 LIWC categories of movie storylines, we find that 20 (55%) LIWC categories are negatively correlated while the rest 16 (45%) LIWC categories are positively correlated with the users' movie rating behavior. We observe that people usually give low rating to a movie having emotional negativity in the LIWC categories of its storyline. For example, users likely to give low rating for the movies that have *negemo*, *anger*, *sad*, and *anxiety* LIWC categories of words in the storylines. We also find that people tend to give low rating having LIWC categories of words such *death*, and *religion* in the storyline of a movie. Moreover, we find that people also give low rating to a movie that has mundane content such as *sexual*, *money*, and *home* in the storyline. In contrast, we observe that people usually give high rating in a movie that has *social*, *friend*, *human*, *motion*, *work*, and *achievement* LIWC categories of words in the storyline. Since people watch movies to eliminate boredom from their daily lives and rejuvenate themselves to work further, they have high propensity to rate the movies high that have frequent content related to social, emotional, fun, friend, love, success and challenge in the storylines.

We use regression based machine learning techniques to predict user rating behavior from personality, values, movie genre, and storylines. Among the models, we find tree based models perform well over the dataset. Table 8.8 shows the best regression model is Random Forests Tree Ensembles because tree based regression model works well when the number of features is large. Our movie rating prediction models achieve the adjusted R-squared score of 68.2% and 68.5% by using personality and values, respectively. When we combine both personality traits and value dimensions, we find that the adjusted R-squared score of our model is 72.3%. The model moderately improves over the previous models. To further improve the model for the existing users in IMDb who write reviews for movies, after adding the Jaccard coefficient between user interest topics and storyline topics with existing feature

set, our model improves slightly. We observe that the adjusted R-squared score of the model is 73%.

To understand the prediction potential of movie rating models, we also apply supervised binary classification techniques following the suggestions of [43, 189]. We observe in Table 8.9 that Rep Tree is the best classifier to predict movie rating. The classifier has AUC, TPR, and FPR scores of 66.9%, 66.4%, and 42.7%, respectively. The MAE and RMSE scores of the model are 39.6% and 48.1%, respectively. The baseline classifier, ZeroR, has AUC, TPR, and FPR scores of 50%, 61.5%, and 61.0%, respectively. The baseline classifier has MAE and RMSE scores of 47.35% and 48.65%, respectively. From the table, we can observe that the Rep Tree performs largely better than the random baseline and other classifiers.

Our approach has several limitations. According to Tables 8.1, we find that drama, thriller, and comedy genres have low accuracy for personality trait based classifiers. We also find from Table 8.2 that thriller genre has low accuracy for value dimensions based classifiers. Though the classifiers have weaker accuracy, they balance each other while we combine both personality traits and values dimensions (according to Table 8.4). In our study, we use users' tweets, and retweets, but we do not consider users' likes/favorites, photo captions, etc. while building regression and classification models. These features might be important attributes to identify a user's personality/values accurately. In previous studies, authors predict user's personality with Facebook *Likes* [115] and value of an individual from different interaction features such as *page-likes, shared-links*, and *statuses* [134]. Therefore, we may get higher MAE for building classification models (according to Table 8.9). Since we use LIWC to analyze movie storylines, this approach usually correlates text corpus with a fixed set of words whereas a lexicon based (open vocabulary) [34] approach analyzes all the texts of user data. For example, in the storyline of *The Amazing Spider-Man 2* movie, we find few important words such as *Osborn*, *Oscorp*, *Spiderman*, etc. are not recognized by LIWC whereas these words have greater impact on rating the movie accurately.

8.7 Summary

In this chapter, we have predicted movie genre preferences and movie rating behavior from users' personality and values derived from their social media interactions. We have exploited the data fusion of Twitter and IMDb to find out how psycholinguistic attributes affect users' movie choices in real life. We have demonstrated which types of personality traits and value dimensions better predict which type of movie genre by using classification techniques. Then, we have also built model to predict user's movie rating from her personality and values, movie genre and storyline. The model does not require information from user's IMDb profile such as reviews, so the model works even if the user has no IMDb profile or has not written reviews in IMDb. Finally, we have improved the model for the users who have written reviews in IMDb by adding similarity coefficient between storyline topics and user interest topics obtained from her written reviews.

Though the average AUC score of the genre classifiers is below 65%, we believe that the accuracy will improve more when the models will be trained on a larger dataset. Matching Twitter profiles with IMDb profiles was a crucial task in collecting our dataset. Since we have collected data manually, we have worked with only a few hundreds. Increasing the size of the training dataset will improve the performance.

Predicting users' preferences is a well studied topic in social media literature, but incorporating psychological attributes with the previous approaches is largely ignored. We have showed that people such as business analyst, entrepreneur, and policy maker can take into account users' psychological attribute while making a complex decision for prospective audience. In the next chapter, we will conclude our dissertation by recapitulating all the research problems, techniques, results, and implications of these problems.

Chapter 9

Conclusions

In this thesis, we have introduced and developed techniques to identify values from multiple interaction features, capturing the change of values, psychological group identification, and users' preference prediction from their psychological attributes derived from social media interactions. We have four objectives that we have discussed throughout the thesis:

First, we have identified values from *user generated* content (i.e., statuses) and *user supported* content (i.e., page-likes, and shared-links) from Facebook interactions. We have built three different linear regression models from each content type. Since different models show different accuracy for different value dimensions, we have built an ensemble model to build a unified value scores. We have also built value models for *silent* and *active* users.

Value scores computed by using our ensemble model improve the accuracy of independent *self-transcendence, openness-to-change,* and *conservation* value models by 32.22%, 17.62%, and 22.22%, respectively. Our study reveals that only *user generated* contents are not sufficient to compute the value scores accurately for all types of users, *user supported* contents also play important role to identify values of both *active* and *silent* users. We find that *self-transcendence, hedonism,* and *self-enhancement* value dimensions of *silent* users can be predicted more accurately for *silent* users by using the *user supported* content than the *user generated* content.

Second, we have captured the *change of values* from social media usage. We have also validated *the change of values* in real life. In this research, we have demonstrated that the change of values are correlated to the users' decision making process by using a movie

experiment. Towards this direction, we have built regression model to predict the value scores of Facebook users. Then, we have segmented the Facebook statuses of new dataset by *6-months* interval. We have predicted the value scores from each of the segmented statuses by using our linear regression model. Then, we have predicted the change values for each of the 6-months interval (on an average 14 different intervals) by using the *ARIMA*, the *LSTM*, and the *HMM* models, independently. We have also developed a hybrid model by using both the *ARIMA* model and the *LSTM* model. Our hybrid model reduces the average RMSE for the *openness-to-change* value dimension by 58.35%, 41%, and 32.32% than that of *HMM, ARIMA,* and *LSTM* models, respectively. We have also observed that our hybrid model captures the value changes accurately for *self-transcendence, openness-to-change,* and *conservation* value dimensions in real life.

We have observed an interesting insight from our experiment, i.e., when the priority of a certain value goes high, the other value may go down. For example, an individual with high *openness-to-change* score is likely to possess a low *conservation* score. Our study finds that when the values change among individuals, their social media interaction patterns also reflect the change in their SNS usage. Since values change over time, users' changing preferences influenced by the values can be captured accurately by using our hybrid model.

Third, we have identified group of users in an egocentric network who exchange similar psychological signals among themselves. Following two important socio-psychological notions: i) *social identity theory*, and ii) *homophily*, we have identified following two types of groups:

- We have classified group of users (*role identities*) who exchange similar psychological cues among themselves during interaction in an egocentric network. We have analyzed users' interaction pattern by using both *closed* and *open vocabulary* based approaches. We have built an integrated model that can classify five different role identities from users' interaction. Our combined model has achieved greater AUC scores than that of independent vocabulary based approaches. We have also validated our hybrid model with different datasets of different networks. Since Facebook is a closed egocentric network, we have predicted all the role identities correctly. In contrast, Twitter is by default an open network, therefore we could not predict the *family, and friend role*

identities effectively.

We have observed that *friend* role identity achieves the best classification strength (AUC-0.723). *Friends* are likely to interact with each other frequently for personal, emotional, study, career, foods, social network, and movie related topics by using the *non-dictionary* words. On the other hand, *family members* build the weakest (AUC-0.635) classifier since they interact less with others through social network.

- We have predicted *Big5 personality* traits from users Facebook statuses by analyzing both *closed* and *open vocabulary* based approaches. Then, we have clustered the similar users based on their predicted personality scores. We have also validated whether members inside the cluster are similar to each other in real life by computing *intraclass correlation coefficient* and *Cohen's Kappa* scores. We have also demonstrated group movie recommendation, an application of personality trait based homophily identification. Our experiment results have showed that our discovered clusters can correctly group members ranged from 73% to 87%.

  Since Facebook reveals the actual human personality traits, we accurately identify *homophily* (group) of Facebook users who possess similar personality traits from their social media interactions.

Fourth, we have examined that users psychological attributes, i.e., *personality* and *values*, are important attributes to identify users preferences in real life. We have conducted our experiment on two case studies: i) *eat-out preferences*, and ii) *movie preferences.*

- We have investigated users' eat-out preferences by exploiting two important social media sites, *Twitter* and *Foursquare*. We have extracted users psychological attributes from tweets, and computed *eat-out* behavior from the pricing information of the restaurants derived from the Foursquare links. Then, we have built regression models to predict users *eat-out* behavior from their combined psychological attributes. We have also validated our result in real life. Our classification results show moderate AUC scores which are ranged from 59.5% to 68.1%.

  Our study finds that *openness* personality trait is strongly correlated with moderate categories of restaurants. People with high *openness* score tend to visit moderate cat-

egory of restaurants, because they likely not to spend money by visiting expensive and very expensive restaurants, but they perhaps explore different dishes in moderate restaurants. In contrast, *conscientiousness* personality trait is positively correlated with visiting expensive and very expensive categories of restaurants. People with high *hedonism* value score likely to enjoy their lives visiting excellent categories of restaurants. We observe that *self-enhancement* value dimension are negatively correlated with cheap categories of restaurants, because people with high *self-enhancement* score prone to eat in better quality of restaurants.

- We have identified users *movie genre* and *rating preferences* from their social media interactions. We have cross-linked the users' Twitter and IMDb profiles who write review, and give rating on movies of different genres. We have built an integrated classification model to predict movie genre from users psychological attributes derived from their social media usage. We have also built a combined regression model from both users' psychological attributes and movie attributes (i.e., storyline, genre, etc.) to predict their movie rating preferences.

  Our genre classification models have achieved moderate performance to predict important movie genres. The AUC scores are ranged from 60.1% to 64.8%. On the other hand, we find that our linear regression models to predict movie rating also achieve higher accuracy by combining psychological attributes, and different movie attributes. Adjusted R^2 of these models are ranged from 68.2% to 73.0%.

  Our study shows that we can improve the prediction potential by integrating psychological attributes with other movie features. People with high *openness-to-experience* personality score may evaluate a movie with low rating, because they tend to expect newness in a movie content than that of a movie usually presents. We observe that people usually give low rating to a movie having an emotional negativity in its storyline. In contrast, we observe that people usually give high rating in a movie that has *social, friend, human, motion, work,* and *achievement* related words in the storyline.

9.1 Future Works

There are many avenues to extend our works of this dissertation. Chapter 3 has described the technique of computing values from multiple interaction features. In this work, we are interested to augment photo related activities and interactions of different social media platforms, so that we can build a comprehensive *value* score from users' different virtual profiles. Chapter 4 has demonstrated how we capture the change of values from social media usage. We also plan to identify the factors that might have influence on users' change of values. Chapter 5 has described a framework to identify users' social role identities from their social media interactions. We are interested to integrate a concept hierarchy in our role identification framework that might help in finding more complex roles in real-life. Chapter 6 has investigated a novel technique to build and validate personality trait based homophily identification from SNS usage. In this work, we plan to use a dynamic lexical analyzer which is more flexible and robust. Chapters 7 and 8 have presented two case studies: eat-out, and movie rating preferences, respectively, from social media usage. We plan to investigate the business opportunity from the *eat-out* preference model and integrate the psychological attributes with other baseline models for movie preferences. In summary, we briefly describe the future directions as follows:

- **Value computation from different interaction features:** There are many aspects of this work that were beyond the scope of this dissertation. Some of the most immediate future works can be investigated in these extensions.

    1. **Incorporating photos and photo related activities:** We observe that people predominantly perform photo related activities such as uploading photo album, and writing photo caption in Facebook. Motivated by the study of Eftekhar et al. [65], we plan to incorporate photo related attributes to compute values accurately with other interaction features such as statuses, and page-likes.

    2. **Inferring values from different social medias:** From the study of Al Maruf et al. [7], we observe that people use different social networks for different purposes. Single social media may not be sufficient to represent completely a user's psychological attribute, i.e., personality and values. In our thesis, we have only

considered the closed network, i.e., Facebook, to compute the value scores from different interaction features. We would like to incorporate the value scores derived from an open network, i.e., Twitter, with the result of the closed network to obtain comprehensive value scores.

- **Change of values:** In future, we want to identify the factors such as actors (e.g., friends), social structure (e.g., school and workplace), change of emotions (e.g., anger, fear, and happiness), and important events (e.g., demise of a family member) that may influence the change of values of an individual.

- **Group identification:** We can extend our group identification, i.e., *social role identity* and *homophily*, problems in different promising directions.

  1. **Integrating concept hierarchy with the social role identities:** We plan to integrate a concept hierarchy with social media interactions that can extract role identities for more *general* to more *specific*, and vice versa.

  2. **Dynamic lexical analyzer for personality identification:** In future, we are interested in using *Empath* for deriving *Big5 personality scores* from multiple interaction features that generates on demand new lexical categories [67]. The technique uses deep learning embedding across more than 1.8 billion words of modern fiction.

- **Eat-out preference:**: From our eat-out preference study, we can explore following future aspects:

  1. **Change of eat-out preference:** We plan to explore how eat-out preferences vary among the members of same age group. Furthermore, we want to investigate how users' eat-out preferences change over time depending on the changing nature of profession, age, income, and education level.

  2. **Prediction of business sustainability:** We would like to predict the sustainability of a new restaurant service in a particular area based on the local users' tweets. Similarly, we plan to predict the financial conditions of a person from her tweets.

Furthermore, we can research whether our study can predict the success of any start-up business in the long-run.

- **Movie preference:** We could improve our movie preference prediction if we would incorporate few important movie and network features with the existing features. We plan to add more attributes such as directors, actors, awards, year of release, box office gross, etc. to build our models. We also want to conduct the experiment with the datasets of other social networks such as Facebook, Reddit, etc. to investigate in what extent our results vary with those datasets. In future, we would like to integrate our model with a real movie recommendation application for social media users.

Next, we present the list of published papers from our thesis in different conferences and journals. We also present the list of our papers with the status of "under review".

Publications

1. Rahman MM, Majumder MT, **Mukta Md. Saddam Hossain**, Ali Mohammed Eunus, Mahmud Jalal. *Can we predict eat-out preference of a person from tweets?* In Proceedings of the 8th **ACM Conference on Web Science 2016 May 22** (pp. 350-351). *ACM*.

2. **Mukta Md. Saddam Hossain**, Ali Mohammed Eunus, Mahmud Jalal. *User Generated vs. Supported Contents: Which One Can Better Predict Basic Human Values?* In International Conference on Social Informatics 2016 Nov 11 (pp. 454-470). **Springer International Publishing**.

3. **Mukta Md. Saddam Hossain**, Ali Mohammed Eunus, Mahmud Jalal. *Identifying and validating personality traits-based homophilies for an egocentric network.* Social Network Analysis and Mining. 2016 Dec 1:6(1):74. **Springer International Publishing**.

4. **Mukta Md. Saddam Hossain**, Khan EM, Ali Mohammed Eunus, Mahmud J. *Predicting Movie Genre Preferences from Personality and Values of Social Media Users.* International AAAI Conference on Web and Social Media. 2017 May (pp. 624-627). **AAAI International Publishing**.

5. **Mukta Md. Saddam Hossain**, Ali Mohammed Eunus, Mahmud Jalal. *Identifying the Change of Basic Human Value Priorities from Social Network Usage.* ACM/IEEE conference on Advances in Social Network Analysis and Mining, 2017 (pp. 619-620).

6. **Mukta Md. Saddam Hossain**, Ali Mohammed Eunus, Mahmud Jalal. *Temporal Modeling of Basic Human Values from Social Network Usage.* Journal of **the Association for Information Science and Technology (JASIST)**, Wiley, 2017. **[Accepted]**

7. Khan, EM, **Mukta Md. Saddam Hossain**, Ali Mohammed Eunus, Mahmud Jalal. *Predicting Users Movie Preference and Rating Behavior from Personality and Values.* **ACM Transactions on Social Computing (TSC)**, 2017. **[Under review]**

8. **Mukta Md. Saddam Hossain**, Ali Mohammed Eunus, Mahmud Jalal. *Predicting Eat-out Preference from Users' Psychological attributes.* Journal of EPJ Data Science, **Springer International Publishing**, 2018. **[Under review]**

9. **Mukta Md. Saddam Hossain**, Ali Mohammed Eunus, Mahmud Jalal. *Deriving Interpersonal Role Identities from Social Network Interactions.* Journal of **Applied Intelligence**, Springer, 2018. **[Under review]**

Bibliography

[1] Sofiane Abbar, Yelena Mejova, and Ingmar Weber. You tweet what you eat: Studying food consumption through twitter. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 3197–3206, 2015.

[2] James Abello, Mauricio GC Resende, and Sandra Sudarsky. Massive quasi-clique detection. In *Latin American symposium on theoretical informatics*, pages 598–612. Springer, 2002.

[3] Sibel Adali and Jennifer Golbeck. Predicting personality with social behavior. In *ASONAM*. IEEE, 2012.

[4] Panagiotis Adamopoulos and Vilma Todri. Personality-based recommendations: Evidence from amazon. com. In *Proceedings of the 9th ACM International Conference on Recommender Systems*, 2015.

[5] Luca Maria Aiello, Alain Barrat, Rossano Schifanella, Ciro Cattuto, Benjamin Markines, and Filippo Menczer. Friendship prediction and homophily in social media. *TWEB*, 6(2):9, 2012.

[6] Luca Maria Aiello, Rossano Schifanella, et al. Reading the source code of social ties. In *Websci*, pages 139–148. ACM, 2014.

[7] Hasan Al Maruf, Nagib Meshkat, Mohammed Eunus Ali, and Jalal Mahmud. Human behaviour in different social medias: A case study of twitter and disqus. In *Advances in Social Networks Analysis and Mining (ASONAM), 2015 IEEE/ACM International Conference on*, pages 270–273. IEEE, 2015.

[8] Sihem Amer-Yahia, Senjuti Basu Roy, Ashish Chawlat, Gautam Das, and Cong Yu. Group recommendation: Semantics and efficiency. *Proceedings of the VLDB Endowment*, 2(1):754–765, 2009.

[9] Anthony J Applegate and Mary Dekonty Applegate. The peter effect: Reading habits and attitudes of preservice teachers. *The Reading Teacher*, 57(6):554–563, 2004.

[10] Olatz Arbelaitz, José María Martínez-Otzeta, and Javier Muguerza. User modeling in a social network for cognitively disabled people. *Journal of the Association for Information Science and Technology*, 67(2):305–317, 2016.

[11] Nick Armstrong and Kevin Yoon. Movie rating prediction. Technical report, Technical Report, Carnegie Mellon University, 1995.

[12] Valerio Arnaboldi, Marco Conti, Andrea Passarella, and Robin Dunbar. Dynamics of personal social relationships in online social networks: A study on twitter. In *Proceedings of the first ACM conference on Online social networks*, pages 15–26. ACM, 2013.

[13] Valerio Arnaboldi, Andrea Guazzini, and Andrea Passarella. Egocentric online social networks: Analysis of key features and prediction of tie strength in facebook. *Computer Communications*, 36(10):1130–1144, 2013.

[14] Pierre-Hadrien Arnoux, Anbang Xu, Neil Boyette, Jalal Mahmud, Rama Akkiraju, and Vibha Sinha. 25 tweets to know you: A new model to predict personality with social media. *arXiv preprint arXiv:1704.05513*, 2017.

[15] Mitja D Back, Juliane M Stopfer, Simine Vazire, Sam Gaddis, Stefan C Schmukle, Boris Egloff, and Samuel D Gosling. Facebook profiles reflect actual personality, not self-idealization. *Psychological science*, 21(3):372–374, 2010.

[16] Kallol K Bagchi, Godwin J Udo, Peeter J Kirs, and Khendum Choden. Internet use and human values: Analyses of developing and developed countries. *Computers in Human Behavior*, 50:76–90, 2015.

[17] Anat Bardi, Julie Anne Lee, Nadi Hofmann-Towfigh, and Geoffrey Soutar. The structure of intraindividual value change. *Journal of personality and social psychology*, 97(5):913, 2009.

[18] Murray R Barrick and Michael K Mount. The big five personality dimensions and job performance: a meta-analysis. *Personnel psychology*, 44(1):1–26, 1991.

[19] Anne E Becker. Television, disordered eating, and young women in fiji: Negotiating body image and identity during rapid social change. *Culture, Medicine and Psychiatry*, 28(4):533–559, 2004.

[20] Thomas J Berndt. Friendship quality and social development. *Current directions in psychological science*, 11(1):7–10, 2002.

[21] Lynn Bible and Hani Tadros. Do values change over time? an exploratary study of business majors. *Journal of Legal, Ethical and Regulatory Issues*, 17(1):97, 2014.

[22] Bruce J Biddle. Recent development in role theory. *Annual review of sociology*, pages 67–92, 1986.

[23] Halil Bisgin, Nitin Agarwal, and Xiaowei Xu. Investigating homophily in online social networks. In *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, volume 1, pages 533–536, 2010.

[24] Mabel Blades. Factors affecting what we eat. *Nutrition & Food Science*, 31(2):71–74, 2001.

[25] André Blais and SIMON LABBÉ ST-VINCENT. Personality traits, political attitudes and the propensity to vote. *European Journal of Political Research*, 50(3):395–417, 2011.

[26] Peter Michael Blau. *Exchange and power in social life*. Transaction Publishers, 1964.

[27] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[28] Daniel M Blonigen, Marie D Carlson, Brian M Hicks, Robert F Krueger, and William G Iacono. Stability and change in personality traits from late adolescence to early adulthood: A longitudinal twin study. *Journal of personality*, 76(2):229–266, 2008.

[29] Johan Bollen, Huina Mao, and Alberto Pepe. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. *ICWSM*, 11:450–453, 2011.

[30] Johan Bollen, Huina Mao, and Xiaojun Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8, 2011.

[31] Johan Bollen, Alberto Pepe, and Huina Mao. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. *arXiv:0911.1583*, 2009.

[32] George EP Box, Gwilym M Jenkins, Gregory C Reinsel, and Greta M Ljung. *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

[33] RL Boyd. Meh: Meaning extraction helper (version 1.0. 6)[software], 2014.

[34] Ryan L Boyd, Steven R Wilson, James W Pennebaker, Michal Kosinski, David J Stillwell, and Rada Mihalcea. Values in words: Using language to evaluate and understand personal values. In *international conference on web and social media (ICWSM)*, pages 31–40, 2015.

[35] Erin M Bryant and Jennifer Marmo. The rules of facebook friendship: A two-stage examination of interaction rules in close, casual, and acquaintance friendships. *Journal of SPR*, 29(8):1013–1035, 2012.

[36] John D Burger, John Henderson, George Kim, and Guido Zarrella. Discriminating gender on twitter. In *Proc. of EMNLP*, pages 1301–1309. ACL, 2011.

[37] Moira Burke, Lada A Adamic, and Karyn Marciniak. Families on facebook. In *ICWSM*, 2013.

[38] Robin Burke. Hybrid recommender systems: Survey and experiments. *User modeling and user-adapted interaction*, 12(4):331–370, 2002.

[39] Fabio Celli, Bruno Lepri, Joan-Isaac Biel, Daniel Gatica-Perez, Giuseppe Riccardi, and Fabio Pianesi. The workshop on computational personality recognition 2014. In *Acmmm*, pages 1245–1246. ACM, 2014.

[40] Fabio Celli, Fabio Pianesi, David Stillwell, and Michal Kosinski. Workshop on computational personality recognition (shared task). In *Proceedings of the Workshop on Computational Personality Recognition*, 2013.

[41] Tomas Chamorro-Premuzic. *Personality and individual differences.* Blackwell publishing, 2007.

[42] Keittima Chapphannarungsri and Saranya Maneeroj. Combining multiple criteria and multidimension for movie recommender system. In *Proceedings of the International MultiConference of Engineers and Computer Scientists*, volume 1. Citeseer, 2009.

[43] Jilin Chen, Gary Hsieh, Jalal U Mahmud, and Jeffrey Nichols. Understanding individuals' personal values from social media word use. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 405–414, 2014.

[44] Li Chen, Wen Wu, and Liang He. How personality influences users' needs for recommendation diversity? In *CHI'13 Extended Abstracts on Human Factors in Computing Systems*, pages 829–834. ACM, 2013.

[45] Eunjoon Cho, Seth A Myers, and Jure Leskovec. Friendship and mobility: user movement in location-based social networks. In *In Proc. of SIGKDD*, pages 1082–1090. ACM, 2011.

[46] François Chollet. Keras: Deep learning library for theano and tensorflow, 2015.

[47] Yan Piaw Chua and Yee Pei Chua. Do computer-mediated communication skill, knowledge and motivation mediate the relationships between personality traits and attitude toward facebook? *Computers in Human Behavior*, 70:51–59, 2016.

[48] Sheridan J Coakes and Lyndall Steed. *SPSS: Analysis without anguish using SPSS version 14.0 for Windows*. John Wiley & Sons, Inc., 2009.

[49] Raviv Cohen and Derek Ruths. Classifying political orientation on twitter: It's not easy! In *ICWSM*, 2013.

[50] Jerome T Connor, R Douglas Martin, and Les E Atlas. Recurrent neural networks and robust time series prediction. *IEEE transactions on neural networks*, 5(2):240–254, 1994.

[51] OriginLab Corporation. Interpreting results of discriminant analysis. `http://www.originlab.com/doc/Origin-Help/DiscAnalysis-Result`, 2016.

[52] David Crandall, Dan Cosley, Daniel Huttenlocher, Jon Kleinberg, and Siddharth Suri. Feedback effects between similarity and social influence in online communities. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 160–168. ACM, 2008.

[53] Lee J Cronbach. Coefficient alpha and the internal structure of tests. *psychometrika*, pages 297–334, 1951.

[54] Antonio Damasio. *Descartes' error: Emotion, reason and the human brain*. Random House, 2008.

[55] Wade M Danis, Leigh Anne Liu, and Jiri Vacek. Values and upward influence strategies in transition: Evidence from the czech republic. *Journal of Cross-Cultural Psychology*, 42(2):288–306, 2011.

[56] Shelley Derksen and HJ Keselman. Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables. *British Journal of Mathematical and Statistical Psychology*, pages 265–282, 1992.

[57] Himel Dev, Mohammed Eunus Ali, and Tanzima Hashem. User interaction based community detection in online social networks. In *DASFAA*, pages 296–310. Springer, 2014.

[58] Joan Morris DiMicco, Werner Geyer, David R Millen, Casey Dugan, and Beth Brownholtz. People sensemaking and relationship building on an enterprise social network site. In *HICSS*, pages 1–10. IEEE, 2009.

[59] Simon Dooms, Toon De Pessemier, and Luc Martens. Movietweetings: a movie rating dataset collected from twitter. In *Workshop on CrowdRec at RecSys*, volume 2013, page 43, 2013.

[60] Simon Dooms and Luc Martens. Harvesting movie ratings from structured data in social media. *ACM SIGWEB Newsletter*, (Winter):4, 2014.

[61] Ellen Dreezens, Carolien Martijn, Petra Tenbült, Gerjo Kok, and Nanne K De Vries. Food and values: an examination of values underlying attitudes toward genetically modified-and organically grown food products. *Appetite*, 44(1):115–122, 2005.

[62] Richard O Duda, Peter E Hart, and David G Stork. *Pattern classification*. John Wiley & Sons, 2012.

[63] David Dunbar, Michael Proeve, and Rachel Roberts. Problematic internet usage self-regulation dilemmas: effects of presentation format on perceived value of behavior. *Computers in Human Behavior*, 70:453–459, 2017.

[64] RI Dunbar. The social brain hypothesis. *brain*, 9(10):178–190, 1998.

[65] Azar Eftekhar, Chris Fullwood, and Neil Morris. *Capturing personality from Facebook photos and photo-related activities: How much exposure do you need?*, volume 37. Elsevier, 2014.

[66] Nicole B Ellison. Introduction: Reshaping campus communication and community through social network sites. *The EDUCAUSE*, 8:19–32, 2008.

[67] Ethan Fast, Binbin Chen, and Michael Bernstein. Empath: Understanding topic signals in large-scale text. *arXiv preprint arXiv:1602.06979*, 2016.

[68] Tom Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874, 2006.

[69] Norman T Feather. Values, valences, and choice: The influences of values on the perceived attractiveness and choice of alternatives. *Journal of personality and social psychology*, 68(6):1135, 1995.

[70] Gregory J Feist. 14 the influence of personality on artistic and scientific creativity. *Handbook of creativity*, page 273, 1999.

[71] He Feng and Xueming Qian. Recommendation via user's personality and social contextual. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, pages 1521–1524. ACM, 2013.

[72] Danyel Fisher. Using egocentric networks to understand communication. *Internet Computing, IEEE*, 9(5):20–28, 2005.

[73] Noah E Friedkin and Eugene C Johnsen. Social influence networks and opinion change. *Advances in group processes*, pages 1–29, 1999.

[74] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. glmnet: Lasso and elastic-net regularized generalized linear models. *R package version*, 1(4), 2009.

[75] Andrew Gelman et al. Analysis of variancewhy it is more important than ever. *The annals of statistics*, 33(1):1–53, 2005.

[76] Zoubin Ghahramani. An introduction to hidden markov models and bayesian networks. *International Journal of Pattern Recognition and Artificial Intelligence*, 15(01):9–42, 2001.

[77] Eric Gilbert and Karrie Karahalios. Predicting tie strength with social media. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 211–220. ACM, 2009.

[78] Erving Goffman. The presentation of self. *Life as theater: A dramaturgical sourcebook*, 2006.

[79] Jennifer Golbeck and Derek Hansen. Computing political preference among twitter followers. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1105–1108, 2011.

[80] Jennifer Golbeck, Cristina Robles, Michon Edmondson, and Karen Turner. Predicting personality from twitter. In *SocialCom*, pages 149–156, 2011.

[81] Bruno Gonçalves, Nicola Perra, and Alessandro Vespignani. Modeling users' activity on twitter networks: Validation of dunbar's number. *PloS one*, 6(8):e22656, 2011.

[82] Wei Gong, Ee-Peng Lim, and Feida Zhu. Characterizing silent users in social media communities. In *ICWSM*, 2015.

[83] Jagadeesh Gorla, Neal Lathia, Stephen Robertson, and Jun Wang. Probabilistic group recommendation via information matching. In *Proceedings of the 22nd international conference on World Wide Web*, pages 495–504. ACM, 2013.

[84] Valdiney V Gouveia, Kátia C Vione, Taciano L Milfont, and Ronald Fischer. Patterns of value change during the life span: Some evidence from a functional approach to values. *Personality and Social Psychology Bulletin*, 41(9):1276–1290, 2015.

[85] Aditya Gupta and Bhuwan Dhingra. Stock market prediction using hidden markov models. In *Engineering and Systems (SCES), 2012 Students Conference on*, pages 1–4. IEEE, 2012.

[86] Jooyeon Ha and SooCheong Jang. Attributes, consequences, and consumer values: A means-end chain approach across restaurant segments. *International Journal of Contemporary Hospitality Management*, 25(3):383–409, 2013.

[87] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18, 2009.

[88] Md Nafiz Hamid, Md Abu Naser, Md Kamrul Hasan, and Hasan Mahmud. A cohesion-based friend-recommendation system. *Social Network Analysis and Mining*, 4(1):1–11, 2014.

[89] Jiawei Han, Jian Pei, and Micheline Kamber. *Data mining: concepts and techniques*. Elsevier, 2011.

[90] David J Hand. Measuring classifier performance: a coherent alternative to the area under the roc curve. *Machine learning*, 77(1):103–123, 2009.

[91] Derek Hansen, Ben Shneiderman, and Marc A Smith. *Analyzing social media networks with NodeXL: Insights from a connected world*. Morgan Kaufmann, 2010.

[92] Trevor Hastie and Junyang Qian. Glmnet vignette. Technical report, Technical report, Stanford, 2014.

[93] Trevor Hastie and Junyang Qian. Glmnet vignette., 2014.

[94] Patrick CL Heaven, Kathryn Mulligan, Robyn Merrilees, Teneille Woods, and Yasmeen Fairooz. Neuroticism and conscientiousness as predictors of emotional, external, and restrained eating behaviors. *International Journal of Eating Disorders*, 30(2):161–166, 2001.

[95] Jonathan L Herlocker, Joseph A Konstan, Loren G Terveen, and John T Riedl. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems (TOIS)*, 22(1):5–53, 2004.

[96] Anne Hewitt and Andrea Forte. Crossing boundaries: Identity management and student/faculty relationships on the facebook. *Poster presented at CSCW, Banff, Alberta*, pages 1–2, 2006.

[97] Ka-Wing Ho. Movies genres classification by synopsis.

[98] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[99] Morris B Holbrook. *Consumer value: a framework for analysis and research*. Psychology Press, 1999.

[100] Pamela M Homer and Lynn R Kahle. A structural equation test of the value-attitude-behavior hierarchy. *Journal of Personality and social Psychology*, 54(4):638, 1988.

[101] Kurt Hornik and Bettina Grün. topicmodels: An r package for fitting topic models. *Journal of Statistical Software*, 40(13):1–30, 2011.

[102] Gary Hsieh, Jilin Chen, Jalal U Mahmud, and Jeffrey Nichols. You read what you value: understanding personal values and reading interests. In *Proceedings of the*

32nd annual ACM conference on Human factors in computing systems, pages 983–986, 2014.

[103] Daniel Hsu, Sham M Kakade, and Tong Zhang. A spectral algorithm for learning hidden markov models. *Journal of Computer and System Sciences*, 78(5):1460–1480, 2012.

[104] David John Hughes, Moss Rowe, Mark Batey, and Andrew Lee. A tale of two sites: Twitter vs. facebook and the personality predictors of social media usage. *Computers in Human Behavior*, 28(2):561–569, 2012.

[105] Øystein Jensen and Kai Victor Hansen. Consumer values among restaurant customers. *International Journal of Hospitality Management*, 26(3):603–622, 2007.

[106] Daniel Jiménez. Dynamically weighted ensemble neural networks for classification. In *Neural Networks Proceedings, 1998. IEEE World Congress on Computational Intelligence. The 1998 IEEE International Joint Conference on*, volume 1, pages 753–756. IEEE, 1998.

[107] Oliver P. John. The big five personality test., 2000.

[108] Oliver P John, Laura P Naumann, and Christopher J Soto. Paradigm shift to the integrative big five trait taxonomy. *Handbook of personality: Theory and research*, 3:114–158, 2008.

[109] Oliver P John and Sanjay Srivastava. The big five trait taxonomy: History, measurement, and theoretical perspectives. *Handbook of personality: Theory and research*, 2(1999):102–138, 1999.

[110] Eleanna Kafeza, Andreas Kanavos, Christos Makris, and Dickson Chiu. Identifying personality-based communities in social networks. In *Advances in Conceptual Modeling*, pages 7–13. Springer, 2013.

[111] Eleanna Kafeza, Andreas Kanavos, Christos Makris, and Pantelis Vikatos. T-pice: Twitter personality based influential communities extraction system. In *BigData Congress*, pages 212–219. IEEE, 2014.

[112] Daniel Kahneman and Amos Tversky. Prospect theory: An analysis of decision under risk. *Econometrica: Journal of the Econometric Society*, pages 263–291, 1979.

[113] Carmen Keller and Michael Siegrist. Does personality influence eating styles and food choices? direct and indirect effects. *Appetite*, 84:128–138, 2015.

[114] Gary G Koch. Intraclass correlation coefficient. *Encyclopedia of statistical sciences*, 1982.

[115] Michal Kosinski, David Stillwell, and Thore Graepel. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, 110(15):5802–5805, 2013.

[116] Efthymios Kouloumpis, Theresa Wilson, and Johanna D Moore. Twitter sentiment analysis: The good the bad and the omg! *Icwsm*, 11(538-541):164, 2011.

[117] Max Kuhn. Caret package. *Journal of Statistical Software*, 28(5), 2008.

[118] Henry Oliver Lancaster and Eugene Seneta. *Chi-Square Distribution*. Wiley Online Library, 1969.

[119] Henry Oliver Lancaster and Eugene Seneta. *Chi-square distribution*. Wiley Online Library, 1969.

[120] Hyun-Joo Lee and Jewon Lyu. Personal values as determinants of intentions to use self-service technology in retailing. *Computers in Human Behavior*, 60:322–332, 2016.

[121] Anton Leuski. Email is a stage: discovering people roles from email archives. In *In SIGIR*, pages 502–503. ACM, 2004.

[122] David Liben-Nowell and Jon Kleinberg. The link-prediction problem for social networks. *Journal of the American society for information science and technology*, 58(7):1019–1031, 2007.

[123] Marjaana Lindeman and Markku Verkasalo. Measuring values with the short schwartz's value survey. *Journal of personality assessment*, 85(2):170–178, 2005.

[124] Leqi Liu, Daniel Preotiuc-Pietro, Zahra Riahi Samani, Mohsen E Moghaddam, and Lyle H Ungar. Analyzing personality through social media profile picture choice. In *ICWSM*, pages 211–220, 2016.

[125] Wendy Liu, Faiyaz Al Zamal, and Derek Ruths. Using social media to infer gender composition of commuter populations. In *Proc. of the when the city meets the citizen workshop, ICWSM*, 2012.

[126] Wendy Liu and Derek Ruths. What's in a name? using first names as features for gender inference in twitter. In *AAAI Spring Symposium: Analyzing Microtext*, 2013.

[127] Thomas Lumley and A Miller. Leaps: regression subset selection. r package version 2.9. *See http://CRAN. R-project. org/package= leaps*, 2009.

[128] François Mairesse, Marilyn A Walker, Matthias R Mehl, and Roger K Moore. Using linguistic cues for the automatic recognition of personality in conversation and text. *J. Artif. Intell. Res.(JAIR)*, 30:457–500, 2007.

[129] Martin N Marshall. Sampling for qualitative research. *Family practice*, 13(6):522–526, 1996.

[130] Julian J McAuley and Jure Leskovec. Learning to discover social circles in ego networks. In *NIPS*, volume 2012, pages 548–56, 2012.

[131] Miller McPherson, Lynn Smith-Lovin, and James M Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, pages 415–444, 2001.

[132] Robert H Moorman and Gerald L Blakely. Individualism-collectivism as an individual difference predictor of organizational citizenship behavior. *Journal of organizational behavior*, 16(2):127–142, 1995.

[133] Rajatish Mukherjee, P Dutta, Sandip Sen, and Ip Sen. Movies2go-a new approach to online movie recommendation. In *Proceedings of the IJCAI Workshop on intelligent Techniques for Web Personalization*, 2001.

[134] Md Saddam Hossain Mukta, Mohammed Eunus Ali, and Jalal Mahmud. User generated vs. supported contents: Which one can better predict basic human values? In *SocInfo*, pages 454–470. Springer, 2016.

[135] Md Saddam Hossain Mukta, Mohammed Eunus Ali, and Jalal Mahmud. Identifying and predicting temporal change of basic human values from social network usage. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, pages 619–620. ACM, 2017.

[136] Md. Saddam Hossain Mukta, Euna Mehnaz Khan, Mohammed Eunus Ali, and Jalal Mahmud. Predicting movie genre preferences from personality and values of social media users. In *Proceedings ICWSM 2017, Montréal, Québec, Canada, May 15-18, 2017.*, pages 624–627, 2017.

[137] Fionn Murtagh and Pedro Contreras. Algorithms for hierarchical clustering: an overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2(1):86–97, 2012.

[138] Liisa Myyry, Soile Juujärvi, and Kaija Pesso. Change in values and moral reasoning during higher education. *European Journal of Developmental Psychology*, 10(2):269–284, 2013.

[139] Tormod Næs and Bjørn-Helge Mevik. Understanding the collinearity problem in regression and discriminant analysis. *Journal of Chemometrics*, 15(4):413–426, 2001.

[140] George A Neuman, Stephen H Wagner, and Neil D Christiansen. The relationship between work-team personality composition and the job performance of teams. *Group & Organization Management*, 24(1):28–45, 1999.

[141] Dong Nguyen, Rilana Gravel, Dolf Trieschnigg, and Theo Meder. "how old do you think i am?"; a study of language and age in twitter. In *ICWSM*. AAAI Press, 2013.

[142] Suphakit Niwattanakul, Jatsada Singthongchai, Ekkachai Naenudorn, and Supachanun Wanapu. Using of jaccard coefficient for keywords similarity. In *Proceedings of the International MultiConference of Engineers and Computer Scientists*, volume 1, 2013.

[143] John R Nofsinger. Social mood and financial economics. *The Journal of Behavioral Finance*, 6(3):144–160, 2005.

[144] Warren T Norman. Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality ratings. *The Journal of Abnormal and Social Psychology*, 66(6):574, 1963.

[145] Anastasios Noulas, Salvatore Scellato, Cecilia Mascolo, and Massimiliano Pontil. An empirical study of geographic user activity patterns in foursquare. *ICWSM*, 11:70–573, 2011.

[146] Richard J Oentaryo, Ee-Peng Lim, David Lo, Feida Zhu, and Philips K Prasetyo. Collective churn prediction in social network. In *Proc. of ASONAM 2012*, pages 210–214. IEEE Computer Society, 2012.

[147] James M Olver and Todd A Mooradian. Personality traits and personal values: a conceptual and empirical integration. *Personality and individual differences*, 35(1):109–125, 2003.

[148] Georgios Paltoglou. Sentiment-based event detection in twitter. *Journal of the Association for Information Science and Technology*, 67(7):1576–1587, 2016.

[149] Jaram Park, Vladimir Barash, Clay Fink, and Meeyoung Cha. Emoticon style: Interpreting differences in emoticons across cultures. In *ICWSM*, 2013.

[150] J Payne, JR Bettman, and EJ Johnson. Consumer decision making. *Handbook of consumer behaviour*, pages 50–84, 1991.

[151] Marco Pennacchiotti and Ana-Maria Popescu. Democrats, republicans and starbucks afficionados: user classification in twitter. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 430–438. ACM, 2011.

[152] James W Pennebaker, Roger J Booth, and Martha E Francis. Linguistic inquiry and word count: Liwc [computer software]. *Austin, TX: liwc. net*, 2007.

[153] James W Pennebaker, Martha E Francis, and Roger J Booth. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71:2001, 2001.

[154] James W Pennebaker, Martha E Francis, and Roger J Booth. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71:2001, 2001.

[155] James W Pennebaker and Anna Graybeal. Patterns of natural language use: Disclosure, personality, and social integration. *Current Directions in Psychological Science*, 10(3):90–93, 2001.

[156] people.duke.edu. Arima models for time series forecasting. `http://people.duke.edu/~rnau/411arim.htm#arima010`, 2016.

[157] Tom Petrocelli. Closed vs open social networks., 2014.

[158] Robi Polikar. Ensemble based systems in decision making. *Circuits and systems magazine, IEEE*, 6(3):21–45, 2006.

[159] Rachel Povey, Benjamin Wellens, and Mark Conner. Attitudes towards following meat, vegetarian and vegan diets: an examination of the role of ambivalence. *Appetite*, 37(1):15–26, 2001.

[160] Jenny Preece, Blair Nonnecke, and Dorine Andrews. The top five reasons for lurking: improving community experiences for everyone. *Computers in human behavior*, 20(2):201–223, 2004.

[161] Usha Nandini Raghavan, Réka Albert, and Soundar Kumara. Near linear time algorithm to detect community structures in large-scale networks. *Physical review E*, 76(3):036106, 2007.

[162] Md Mahabur Rahman, Md Taksir Hasan Majumder, Md Saddam Hossain Mukta, Mohammed Eunus Ali, and Jalal Mahmud. Can we predict eat-out preference of a person from tweets? In *WebSci*, pages 350–351. ACM, 2016.

[163] Antonio Rangel, Colin Camerer, and P Read Montague. A framework for studying the neurobiology of value-based decision making. *Nature Reviews Neuroscience*, 9(7):545–556, 2008.

[164] Peter J Rentfrow and Samuel D Gosling. The do re mi's of everyday life: the structure and personality correlates of music preferences. *Journal of personality and social psychology*, 84(6):1236, 2003.

[165] George Athanasopoulos Rob J Hyndman. 6.1 time series components., 2016.

[166] Brent W Roberts, Kate E Walton, and Wolfgang Viechtbauer. Patterns of mean-level change in personality traits across the life course: a meta-analysis of longitudinal studies. *Psychological bulletin*, 132(1):1, 2006.

[167] Mary E Rogers, Peter A Creed, and A Ian Glendon. The role of personality in adolescent career planning and exploration: A social cognitive perspective. *Journal of Vocational Behavior*, 73(1):132–142, 2008.

[168] Jonathan P. Rowe and Erica L. Snow, editors. *Workshop and Tutorial Proceedings of the 9th International Conference on Educational Data Mining, Raleigh, NC, USA, June 29, 2016*, volume 1633 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2016.

[169] Stuart J Russell and Peter Norvig. *Artificial intelligence: a modern approach*. Malaysia; Pearson Education Limited,, 2016.

[170] Alan Said and Alejandro Bellogín. You are what you eat! tracking health through recipe interactions. In *Rsweb@ recsys*, 2014.

[171] Jesus F Salgado. The big five personality dimensions and counterproductive behaviors. *International Journal of Selection and Assessment*, 10(1-2):117–125, 2002.

[172] Bodo B Schlegelmilch, Greg M Bohlen, and Adamantios Diamantopoulos. The link between green purchasing decisions and measures of environmental consciousness. *European journal of marketing*, 30(5):35–55, 1996.

[173] H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin EP Seligman, et al. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):e73791, 2013.

[174] Shalom H Schwartz. Are there universal aspects in the structure and contents of human values? *Journal of social issues*, 50(4):19–45, 1994.

[175] Shalom H Schwartz. A proposal for measuring value orientations across nations. *Questionnaire Package of ESS*, pages 259–290, 2003.

[176] Shalom H Schwartz. Basic human values: Their content and structure across countries. *Valores e Trablho*, pages 21–55, 2005.

[177] Shalom H Schwartz. Robustness and fruitfulness of a theory of universals in individual values. *Valores e trabalho*, pages 56–85, 2005.

[178] Shalom H Schwartz, Gila Melech, Arielle Lehmann, Steven Burgess, Mari Harris, and Vicki Owens. Extending the cross-cultural validity of the theory of basic human values with a different method of measurement. *Journal of cross-cultural psychology*, pages 519–542, 2001.

[179] Skipper Seabold and Josef Perktold. Statsmodels: Econometric and statistical modeling with python. In *Proceedings of the 9th Python in Science Conference*, pages 57–61, 2010.

[180] Maarten Selfhout, William Burk, Susan Branje, Jaap Denissen, Marcel Van Aken, and Wim Meeus. Emerging late adolescent friendship networks and big five personality traits: A social network approach. *Journal of personality*, 78(2):509–538, 2010.

[181] Shai Shalev-Shwartz and Shai Ben-David. Understanding machine learning: From theory to algorithms. 2014.

[182] Phillip R Shaver and Kelly A Brennan. Attachment styles and the" big five" personality traits: Their connections with each other and with romantic relationship outcomes. *Personality and Social Psychology Bulletin*, 18(5):536–545, 1992.

[183] Joseph Sill, Gábor Takács, Lester Mackey, and David Lin. Feature-weighted linear stacking. *arXiv preprint arXiv:0911.0460*, 2009.

[184] Thiago H Silva, Pedro OS de Melo, Jussara Almeida, Mirco Musolesi, and Antonio Loureiro. You are what you eat (and drink): Identifying cultural boundaries by analyzing food & drink habits in foursquare. *arXiv preprint arXiv:1404.1009*, 2014.

[185] Aaron Smith. 6 new facts about Facebook. `http://www.pewresearch.org/fact-tank/2014/02/03/6-new-facts-about-facebook/`, 2014.

[186] Laurence Steinberg and Amanda Sheffield Morris. Adolescent development. *Annual review of psychology*, 52(1):83–110, 2001.

[187] Andrew Steptoe, Tessa M Pollard, and Jane Wardle. Development of a measure of the motives underlying the selection of food: the food choice questionnaire. *Appetite*, 25(3):267–284, 1995.

[188] Robert J Sternberg. Implicit theories of intelligence, creativity, and wisdom. *Journal of personality and social psychology*, 49(3):607, 1985.

[189] Chris Sumner, Alison Byers, Rachel Boochever, and Gregory J Park. Predicting dark triad personality traits from twitter usage and a linguistic analysis of tweets. In *11th IEEE international conference on Machine learning and applications (icmla)*, volume 2, pages 386–393, 2012.

[190] Henri Tajfel and John C Turner. An integrative theory of intergroup conflict. *The social psychology of intergroup relations*, 33(47):74, 1979.

[191] Henri Tajfel and John C Turner. The social identity theory of intergroup behavior. 2004.

[192] Jie Tang, Tiancheng Lou, and Jon Kleinberg. Inferring social ties across heterogenous networks. In *Proceedings of WSDM*, pages 743–752. ACM, 2012.

[193] Wenbin Tang, Honglei Zhuang, and Jie Tang. Learning to infer social ties in large networks. In *ECML-PKDD*, pages 381–397. Springer, 2011.

[194] Yla R Tausczik and James W Pennebaker. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54, 2010.

[195] Berhane Teclehaimanot and Torey Hickman. Student-teacher interaction on facebook: What students find appropriate. *TechTrends*, 55(3):19–30, 2011.

[196] Mike Thelwall. The heart and soul of the web? sentiment strength detection in the social web with sentistrength. pages 119–134, 2017.

[197] Ben Tufft. Adulthood begins at 25, says new research. `https://goo.gl/dNM7mJ`, 2015.

[198] John C Turner and Penelope J Oakes. The significance of the social identity concept for social psychology with reference to individualism, interactionism and social influence. *British Journal of Social Psychology*, 25(3):237–252, 1986.

[199] UCSC. Strengths and weaknesses of hidden markov models, 1996.

[200] Philip Thomas Udo-Imeh. Influence of personality on the buying behaviour of undergraduate students in universities in cross river state, nigeria. *International Journal of Marketing Studies*, 7(4):64, 2015.

[201] George Veletsianos. Higher education scholars' participation and practices on twitter. *Journal of Computer Assisted Learning*, 28(4):336–349, 2012.

[202] Bas Verplanken and Rob W Holland. Motivated decision making: effects of activation and self-centrality of values on choices and behavior. *Journal of personality and social psychology*, 82(3):434, 2002.

[203] Anthony J Viera, Joanne M Garrett, et al. Understanding interobserver agreement: the kappa statistic. *Fam Med*, 37(5):360–363, 2005.

[204] Ingmar Visser, Maarten Speekenbrink, et al. depmixs4: an r package for hidden markov models. *Journal of Stat. Software*, 36(7):1–21, 2010.

[205] Claudia Wagner and Luca Maria Aiello. Men eat on mars, women on venus?: An empirical study of food-images. In *Proceedings of the ACM Web Science Conference*, page 63. ACM, 2015.

[206] Chi Wang, Jiawei Han, Yuntao Jia, Jie Tang, Duo Zhang, Yintao Yu, and Jingyi Guo. Mining advisor-advisee relationships from research publication networks. In *KDD*, pages 203–212. ACM, 2010.

[207] Linlin Wang, Zhu Cao, Yu Xia, and Gerard de Melo. Morphological segmentation with window lstm neural networks. In *AAAI*, pages 2842–2848.

[208] James B Weaver, Hans-Bernd Brosius, and Norbert Mundorf. Personality and movie preferences: A comparison of american and german audiences. *Personality and Individual Differences*, 14(2):307–315, 1993.

[209] Susan Whelan and Gary Davies. Profiling consumers of own brands and national brands using human personality. *Journal of Retailing and Consumer Services*, 13(6):393–402, 2006.

[210] Jill Wiltfong. Global Sharers on Social Media Sites. `http://www.ipsos-na.com/news-polls/pressrelease.aspx?id=6239`, 2013.

[211] Felix Ming Fai Wong, Chee Wei Tan, Soumya Sen, and Mung Chiang. Quantifying political leaning from tweets and retweets. In *ICWSM*, 2013.

[212] Joshua Wortman. *Film classification using subtitles and automatically generated language factors*. Technion-Israel Institute of Technology, Faculty of Industrial and Management Engineering, 2010.

[213] Changhua Yang, Kevin Hsin-Yih Lin, and Hsin-Hsi Chen. Building emotion lexicon from weblog corpora. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 133–136. Association for Computational Linguistics, 2007.

[214] Tal Yarkoni. Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers. *Journal of research in personality*, 44(3):363–373, 2010.

[215] Tal Yarkoni and Jacob Westfall. Choosing prediction over explanation in psychology: Lessons from machine learning. *Unpublished manuscript. Retrieved from http://jakewestfall. org/publications/Yarkoni_Westfall_choosing_prediction. pdf*, 2016.

[216] Hongzhi Yin, Bin Cui, Ling Chen, Zhiting Hu, and Zi Huang. A temporal context-aware model for user behavior modeling in social media systems. In *Proceedings of the 2014 ACM SIGMOD international conference on Management of data*, pages 1543–1554. ACM, 2014.

[217] G Peter Zhang. Time series forecasting using a hybrid arima and neural network model. *Neurocomputing*, 50:159–175, 2003.

[218] Hao Zhao and Scott E Seibert. The big five personality dimensions and entrepreneurial status: a meta-analytical review., 2006.

[219] Yuchen Zhao, Guan Wang, Philip S Yu, Shaobo Liu, and Simon Zhang. Inferring social roles and statuses in social networks. In *KDD*, pages 695–703. ACM, 2013.

[220] W-X Zhou, Dider Sornette, Russell A Hill, and Robin IM Dunbar. Discrete hierarchical organization of social group sizes. *Proceedings of the Royal Society of London B: Biological Sciences*, 272(1561):439–444, 2005.