

M.SC. ENGG. THESIS

Identifying Comprehensive Personality Profile from Multiple Online Media Usage

by
Nagib Meshkat

Submitted to

Department of Computer Science and Engineering
in partial fulfilment of the requirements for the degree of
Master of Science in Computer Science and Engineering



Department of Computer Science and Engineering
Bangladesh University of Engineering and Technology (BUET)
Dhaka 1000

November 2015

Dedicated to my loving parents, wife and son

AUTHOR'S CONTACT

Nagib Meshkat

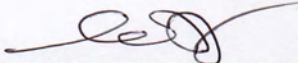
Student

Department of Computer Science & Engineering
Bangladesh University of Engineering & Technology (BUET).

Email: diko007@gmail.com

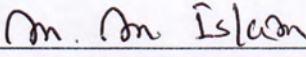
The thesis titled "Identifying Comprehensive Personality Profile from Multiple Online Media Usage", submitted by Nagib Meshkat, Roll No.0413052031, Session April 2013, to the Department of Computer Science and Engineering, Bangladesh University of Engineering and Technology, has been accepted as satisfactory in partial fulfillment of the requirements for the degree of Master of Science in Computer Science and Engineering and approved as to its style and contents. Examination held on October 28, 2015.

Board of Examiners

1.  _____

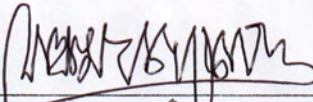
Dr. Mohammed Eunus Ali
Professor
Department of Computer Science and Engineering
Bangladesh University of Engineering and Technology, Dhaka.

Chairman
(Supervisor)

2.  _____

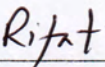
Dr. Md. Monirul Islam
Acting Head and Professor
Department of Computer Science and Engineering
Bangladesh University of Engineering and Technology, Dhaka.

Member
(Ex-Officio)

3.  _____

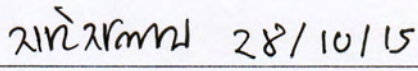
Dr. M. Kaykobad
Professor
Department of Computer Science and Engineering
Bangladesh University of Engineering and Technology, Dhaka.

Member

4.  _____

Dr. Rifat Shahriyar
Assistant Professor
Department of Computer Science and Engineering
Bangladesh University of Engineering and Technology, Dhaka.

Member

5.  _____

Dr. Salekul Islam
Associate Professor & Head of the Dept.
Department of Computer Science and Engineering
United International University, Dhaka.

Member
(External)

Candidate's Declaration

This is hereby declared that the work titled “Identifying Comprehensive Personality Profile from Multiple Online Media Usage” is the outcome of research carried out by me under the supervision of Dr. Mohammed Eunos 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.

Nagib Meshkat

Nagib Meshkat
Candidate

Acknowledgment

I express my heart-felt gratitude to my supervisor, Dr. Mohammed Eunos Ali for his constant supervision of this work. He helped me a lot in every aspect of this work and guided me with proper directions whenever I sought one. His patient hearing of my ideas, critical analysis of my observations and detecting flaws (and amending thereby) in my thinking and writing have made this thesis a success.

I would also want to thank the members of my thesis committee for their valuable suggestions. I thank Dr. Md. Monirul Islam, Dr. M. Kaykobad, Dr. Rifat Shahriyar and Dr. Salekul Islam.

In this regard, I remain ever grateful to my beloved parents, wife and family. They always exist as sources of inspiration behind every success of mine I have ever made.

Abstract

Online media platforms have become the most popular tools for social interaction and sharing information in present world. With the increasing popularity of online media and varying needs of people using it, different media platforms have emerged. Previous research has shown some relationships between users' personality and online media usage. Personality is shown to be useful in predicting social behavior, generating online feed, designing user interface, detecting virtual community, etc. Until now, personality is identified only from single social networking site by processing whole textual contents of online usage. This does not give a comprehensive picture of a user as different media platforms reveal different aspects of personality. In this thesis, we identify a person's comprehensive personality profile from two types of data independently- a person's posts in two major online media platforms (i.e., the most widely used social media platform Twitter and a major online comment posting platform Disqus) and a person's topic related data (i.e., the person's list of active topics and sentiment identified from online usage). Identifying personality from a person's topic related data allows us to predict personality without processing entirety of online posts and comments which is particularly useful in cases where a person's social media data is not directly available. In this thesis, we describe the type of data collected, our methods of analysis, our proposed novel approaches and the machine learning techniques that allow us to successfully predict personality. Our method of identifying personality outperforms existing methods which justifies the inference that single online media analysis is not enough to build a comprehensive virtual identity of a person.

Contents

<i>Board of Examiners</i>	ii
<i>Candidate’s Declaration</i>	iii
<i>Acknowledgment</i>	iv
<i>Abstract</i>	v
List of Figures	ix
List of Tables	x
1 Introduction	1
1.1 Motivation	2
1.2 Related Work and Limitations	3
1.3 Objectives and Scope of the Thesis	4
1.4 Challenges	5
1.5 Our Solution Overview	6
1.5.1 Identifying Personality from Posts	6
1.5.2 Identifying Personality from Topic Profile	7
1.6 Applications of the Thesis	8
1.7 Outline of the Thesis	8

2	Related Works	9
2.1	The Big Five Personality Trait	9
2.2	LIWC Dictionary	11
2.3	Topic Identification	11
2.4	Personality and Online Platform Usage	12
2.5	Summary	14
3	Methodology	16
3.1	Input and Output of Our Method	16
3.2	Our Proposed Method	17
3.3	Method to Identify Comprehensive Personality	17
3.3.1	Method to Identify Personality from Twitter	18
3.3.2	Method to Identify Personality from Disqus	22
3.3.3	Combine Personality Scores	27
3.4	Method to Identify Topic Profile	28
3.5	Method to Predict Personality from Topic Profile	29
3.5.1	Extracting Features from Topic Profile	31
3.5.2	Prediction Classifier	35
3.6	Summary	38
4	Experiment and Analysis	39
4.1	Data Collection and Dataset	39
4.2	Experimental Setup	40
4.3	Experimental Results and Analysis of Identifying Personality from Posts	42
4.3.1	Effects of Feature Reduction Step	43
4.3.2	Comparison of Personality Traits between Twitter and Dis- qus	44
4.4	Experimental Results and Analysis of Identifying Personality from Topic Profile	46
4.4.1	Analysis on Topic Profile	46

<i>CONTENTS</i>	viii
4.4.2 Analysis on Prediction Classifier	49
4.4.3 Analysis on Feature Extraction	54
4.5 Result Summary	61
5 Conclusion and Future Work	64

List of Figures

2.1	Big five factor personality domains	10
3.1	Correlation between neuroticism and anxiety in Twitter	19
3.2	Correlation between neuroticism and anxiety in Disqus	22
3.3	Block diagram of training and target attributes of our model	30
3.4	Condition 1 of our clustering algorithm	33
3.5	Condition 2 of our clustering algorithm	34
3.6	An illustration of some selected training data in 2-dimensional plot for conscientiousness trait of range (0.1-0.2)	37
4.1	Comparison of accuracy of our method and Golbeck's method	42
4.2	Analysis of topic coverage of Twitter and Disqus	47
4.3	Analysis of topic sentiment of Twitter and Disqus	47
4.4	Analysis of sentiment over Disqus and Twitter	48
4.5	Highest accuracy of k -NN and our clustering method	57
4.6	Accuracy of predicting neuroticism and extraversion for different d values	57
4.7	Accuracy of predicting openness, agreeableness and conscientiousness for different d values	58
4.8	Accuracy of prediction for different values of PC	60

List of Tables

3.1	Some defined symbols and their explanation	19
3.2	Some defined symbols and their explanation	26
4.1	Accuracy of our method and Golbeck’s method	43
4.2	Number of features after feature reduction for each trait	43
4.3	Accuracy (in %) of our method for each trait with and without applying feature reduction	43
4.4	Average and standard deviation of scores of inferred big five traits on a normalized 0-1 scale.	45
4.5	Percentage of people having stronger big five traits in Twitter and Disqus on a normalized 0-1 scale	45
4.6	Accuracy of zeroR, linear regression and non-linear regression . .	49
4.7	Accuracy of C4.5, naive bayesian and logistic regression	51
4.8	Accuracy of support vector machine with different parameters . .	52
4.9	Effective classifiers for each personality trait and the highest ac- curacy achieved	53
4.10	Accuracy of prediction model with and without topic removal . .	54
4.11	Comparison of accuracy of k -NN and our distance based cluster- ing	56
4.12	Accuracy (in %) for different values of d	58
4.13	Accuracy (in %) for different values of Principal Components (PC)	61

Chapter 1

Introduction

Contemporary modern world has witnessed the widespread emergence of online social media and similar technologies. People's behaviour in these platforms has become an interesting topic of research. In this thesis, we analyze people's personality in two major online platforms: the most widely used social media platform Twitter and a major online comment posting platform Disqus.

Twitter has become a valuable source for quantitative socio-researchers in the last few years. Disqus provides users the opportunity to follow different communities and to participate in discussion by commenting. As the purpose and way of usage differs in Twitter and Disqus, identity revealed from those platforms can give different aspect of a user's behaviour such as way of communication, subject of interests and sensitivity in different environment and condition.

In this chapter, we give an overview of our thesis. In Section 1.1 we discuss the motivation for our thesis. All related works and their limitations are summarized in Section 1.2. Then in Section 1.3 we provide the objectives of our thesis and define the scope of our works. In Section 1.4 we discuss the major challenges of achieving our objectives. We provide our solution overview to achieve our objectives in Section 1.5.

1.1 Motivation

Internet has a profound impact on people's online and real-life experiences. People use the Internet for a wide domain of reasons that include seeking entertainment, sharing opinions, reading news, etc. For accommodating people's different needs, different types of online platforms have emerged. With the emergence of different online platforms, people are using a wide range of platforms and applications in virtual world. People are using different social networking sites such as Facebook and Twitter to share their opinion; different comment posting platforms such as Disqus and Echoes to post comments or reviews; different image sharing sites such as Instagram and Flickr to post images. According to [14], about 42% people use more than one online platforms for communicating in the Internet.

Identifying personality from user's online usage has become a topic of interest due to its various applications in news feed generation and community detection. Earlier research [17, 24, 36] found that online usage reveals a person's personality. In identifying personality from these social sites, researchers tend to use the big five model of personality [34]. In big five model, a person's personality is defined by his scores on five major traits of personality: neuroticism, extraversion, openness, agreeableness and conscientiousness. Earlier works of personality analysis were limited to different social networking sites. For instance, personality of users were identified from different social networking sites like Twitter and Facebook only. A major shortcoming of all these works is that social networking sites alone cannot reveal all aspects of human personality [19]. Therefore, analysis on multiple types of media platforms is essential for identifying comprehensive personality profile of a user.

Another major shortcoming of the existing methods of identifying personality is that all of them rely on processing entirety of posts and comments. As a result, personality cannot be identified without processing entire statuses or posts from online platform. Due to increasing concerns of user privacy and security, identifying a user's social media profile has become extremely difficult in many cases. In numerous cases, only a user's topic related data is available. For instance, Google

has topic list and their sentiment of a user based on search history. Amazon similarly has topic related data of users available through their purchasing history. In our thesis, by “topic” we refer to predetermined meaningful category such as politics, religion and sports. With the increasing number of online users and different online platforms, topic related information for users is getting available in many cases. Even in many cases where a person has no social media account, his topic related data is available through news article browsing, search history, etc. So a method to predict personality from this type of data has become an absolute necessity in current context.

1.2 Related Work and Limitations

Connection between personality of individual and their behaviour on online platforms like Facebook [32], Twitter [28] and blogs [36] has drawn the focus of researchers for the last few years. It has been shown that extraversion and conscientiousness are positively correlated with the perceived ease of use of social networking sites [29]. Openness is positively correlated to the size of a user’s social network. Positive reply, re-post and friend density is positively correlated to agreeableness. By analysing posts, shares of links and events in Twitter, one’s mood and emotional condition can also be detected [7]. All of these studies short-fall multiple online platform analysis.

There are a handful studies on multiple online platform analysis of human behaviour. Xu *et al.* quantified the extent of user engagement of different platforms as well as their correlations in [35]. They presented the differences of usage in six major social networking sites. Their findings were only limited to usage pattern and they did not identify personality of a user. Work of Hughes *et al.* is closely related to our thesis [19]. Hughes *et al.* investigated the relationship between users’ personality and social media usage by examining the big five personality traits in Facebook and Twitter [19]. This study is limited to similar type of social media activity analysis.

With more and more individuals using social networking sites, it is important that we understand who is using the sites and for which reasons. Previous studies have begun to consider how individual differences impact upon online behaviour. Most of the studies were limited to particular social media analysis which does not provide information on individuals' behaviour in different environment. Earlier studies on cross platform analysis were mostly confined in social networking sites where people maintain social relationship through posting and sharing. In online comment posting platform like Disqus, people get the chance to share their opinion and communicate with people of different psychology and interest. Through this communication and argument, people reveal interesting characteristics of their psychological traits which are evidently deficient in the usage of social networking sites like Twitter and Facebook. To the best of our knowledge, there is hardly any research on the widespread multiple media platform analysis on human behaviour and personality traits.

All these previous works require processing the entirety of posts and statuses for identifying personality. No personality prediction method is proposed for sources like topic usage. As we discussed in the previous section, identifying personality from topic related data has become a necessity in contemporary world.

1.3 Objectives and Scope of the Thesis

The objective of our thesis is to bridge the gap of current shortcomings. Our thesis can be divided into two major parts. In first part, our objective is to overcome the limitations of existing methods of identifying personality from posts and statuses. To achieve our objective, we develop a comprehensive personality profile of a person from multiple media platforms' usage. We have discussed in the previous section that social networking site alone cannot reveal all aspects of a user's personality. To overcome this limitation, in this part of our thesis we identify personality profile of a person in big five model from his posts of one social networking site (i.e., Twitter) and one comment posting platform (i.e., Disqus).

While predicting personality from posts, we only consider the textual contents of the posts made by that user and do not consider any additional info (for example, age, time and location) of the posts.

In 2nd part of our thesis, our objective is to develop a model to identify personality of a person without analyzing entirety of posts or statuses. We propose a model to identify a person's personality from his topic related data. Our proposed model can be utilized in real-life scenario when a person's statuses or posts from online media are not available. The scope for this part of our thesis is to develop a prediction model where a person's personality is predicted by analyzing his topic profile. For a person, we define topic profile as a list of all active topics along with their sentiment value. In this part, we predict a person's all five personality traits in big five model by only analyzing his topic profile as an input.

1.4 Challenges

In this thesis, we address a number of challenges to achieve our objectives. Challenges of identifying personality from posts and challenges of predicting personality from topic profile are discussed separately in the following sections:

Personality from Posts: Analysis of a user's personality in multiple media platforms is a vast unexplored area. "Will usage of multiple platforms offer a significant improvement in identifying personality?" is a question still unanswered. How to identify a better personality profile of a person from analyzing similarities and differences between multiple online platforms is a major challenge in our work. Identifying personality and other personal measures from comment posting platform is not explored in previous work. As a result, identifying personality from Disqus is a major technical challenge in our thesis.

Personality from Topic Profile: In previous work, no prediction model is proposed for predicting personality from topic profile. In order to develop this prediction model, we need to identify appropriate features from topic profile which is a huge technical challenge. Number of unique topics are large and a person may

not be active in all the topics. Considering too many topics will increase missing attributes and considering too few topics will decrease overall accuracy. We need to develop methods to correctly extract features from topic profile without compromising accuracy. We also need to develop a prediction classifier to utilize the correctly extracted features. Our prediction model must be tolerant to the presence of random errors because topic profile of a person may contain random errors. All these issues make the prediction model a challenging research topic in the field of identifying personality from online usage.

1.5 Our Solution Overview

We give an overview of our proposed solution in this section. Details of our method is discussed in Chapter 3.

1.5.1 Identifying Personality from Posts

We first identify a user's personality in big five model separately from Twitter and Disqus. Previous research has shown that personality traits have correlation with linguistic features [23] [26]. We first calculate linguistic scores from a person's Twitter and Disqus posts utilizing LIWC psycho-linguistic dictionary. Details of LIWC is discussed in Section 2.2. To identify personality from LIWC scores obtained from Twitter, we implement Golbeck's method [17]. Same method cannot be used in case of Disqus due to differences in usage pattern of a person in these two platforms. We propose a novel approach to identify personality traits from Disqus which consists of non-linear regression method. Our method of identifying personality traits from Disqus is explained in detail in Section 3.3.2. At last step, we combine the personality traits obtained independently from Twitter and Disqus. The novel idea in our thesis is to overcome the limitations of social networking site utilizing a comment posting platform. As a result, our method exhibits higher accuracy in identifying openness and conscientiousness traits than

existing methods. Comparison between our method and existing method is presented in Section 4.3.

1.5.2 Identifying Personality from Topic Profile

Our proposed model to predict personality from topic profile consists of two phases: feature extraction phase and prediction classifier phase.

- We propose a feature extraction method that addresses the challenges of extracting information from topic profile.
- We create a finite number of classes from personality trait values for each personality trait and propose a prediction classifier to predict personality class from our extracted features.

In our feature extraction phase, we apply topic removal technique, topic clustering technique and principal component analysis. At first step, we consider each topic as a point in multi-dimensional graph and cluster all the unique topics in topic clusters. Our distance based clustering method to cluster similar topics is discussed in details in Section 3.5.1. Experimental results show that our clustering method performs better in our dataset than other clustering methods such as k -nearest neighbor clustering. Afterwards in the next step of feature extraction, we calculate scores of a person for all the different topic clusters. Lastly, from these scores of topic clusters of a person, we identify a limited number of features utilizing principal component analysis. This step which is discussed in Section 3.5.1 is an important step in our prediction model largely due to scalability reasons.

After extracting features from a person's topic profile, we propose a prediction model to successfully predict personality traits of a person from these features. We divide personality in 10 personality classes and then apply different prediction classifiers. We apply linear regression, decision tree based classifier, logistic regression, naive bayesian classifier and support vector machine based classifier to our dataset and analyze their accuracy for predicting different personality traits. From our analysis and experiments, we propose a prediction model

in Section 3.5.2 which is best suited for our purpose. To our best knowledge, no prediction model is proposed in previous literature to identify personality from only topic related data.

1.6 Applications of the Thesis

Personality is proved to be useful in numerous applications from improving on-line experience to predicting behaviour in social life. We propose a novel method to identify a person's personality more accurately than existing methods. Identifying personality correctly can be used in many real life scenarios that include predicting job satisfaction, predicting professional success, etc. Identifying personality correctly will improve accuracy of group targeted advertisements, online community detection, news feed generation, personalized content generation, etc. Predicting personality from topic profile is a novel approach which will help us to predict personality from only topic related data of a person. This will be specially beneficial in cases when posts and statuses of a user is unavailable or social media profile of a user is inactive.

1.7 Outline of the Thesis

The rest of the thesis is organized as follows:

In Chapter 2, we describe formal terms and several supporting technologies and algorithms that we use in our model. Also the literature review of several existing related techniques and their limitations are given in this chapter. Our proposed methods are introduced and explained in Chapter 3. Chapter 4 presents the experimental results, analysis and performance of our proposed methods. Finally, concluding remarks and suggestions for future work are provided in Chapter 5.

Chapter 2

Related Works

In this chapter we survey all the related works of identifying personality from online usage. In Section 2.1 we discuss previous works of classifying personality in big five model and in Section 2.3 we survey works related to topic and sentiment identification. We give a description of LIWC dictionary in Section 2.2. Lastly, in Section 2.4 we discuss all the existing methods for identifying personality from online usage with their limitations.

2.1 The Big Five Personality Trait

The big five model of personality has become the most widely used and the most reliable method of identifying personality [10] [33]. The big five model of personality is a hierarchical organization of personality traits in terms of five basic dimensions. These dimensions are neuroticism, extraversion, openness, agreeableness and conscientiousness. Figure 2.1 show these dimension using pie chart. These dimensions are discussed below:

- **Neuroticism:** Neuroticism is a measure of affect and emotional control [12]. A low value of neuroticism suggests a good control over emotional responses. On the other hand, people with high neuroticism value tends to



Figure 2.1: Big five factor personality domains

be nervous with a propensity to show emotional response. Anger, anxiety, worry, etc. are related to neuroticism. In previous literatures, neuroticism was sometimes referred to as *emotional stability*.

- **Extraversion:** Extraversion refers to energy, positive emotions, assertiveness, sociability, tendency to seek stimulation in the company of others, talkativeness etc. Extraverts are typically adventurous, sociable and talkative whereas introverts are typically quiet and shy [13].
- **Openness:** Openness is related to appreciation for art, adventure, unusual ideas, creative expressions and variety of experiences. It reflects the degree of intellectual curiosity and creativity of a person. Individual with high openness have broad interests and seek novelty whereas low openness is linked with familiarity and convention [11]. Openness was previously referred to as *openness-to-experience, intellect, etc.*

- **Agreeableness:** Agreeableness is related to how friendly a person is, with high values of agreeableness generally refers to being kind, sympathetic and warm [12]. It is also a measure of one's trusting and helpful nature. On the contrary, low agreeableness is linked with being suspicious and antagonistic to others.
- **Conscientiousness:** Conscientiousness refers to a person's work ethics and dutifulness. Individual with high conscientiousness scores has a tendency to be organized and dependable, show discipline [27].

2.2 LIWC Dictionary

Linguistic Inquiry and Word Count (LIWC) is a psycholinguistic dictionary which produces statistics on 70 different features of text in five categories [20]. 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).

2.3 Topic Identification

Topic refers to a predetermined meaningful category such as politics, religion and sports [8]. Identification of topic from short texts such as online posts is proposed in previous literatures [22]. A combination of algorithmic techniques such as shortest links in graph is used to classify topics [1]. In our thesis we use Texalytic tool which implements these techniques to identify topics from online contents [15]. In Texalytic, Sentiment of online posts is identified using POS tag tree which gives sentiment as positive, neutral or negative class [1].

2.4 Personality and Online Platform Usage

In the last few decades, researchers have shown broad interest in studies on human behaviour over the social media platforms and considerable research have been done on extracting psychological information of social media users. Connection between personality of individual and their behaviour on a social network like Facebook [32], Twitter [28], blogs [36] has drawn the focus of researchers for the last few years.

Previous research has linked a person's online platform usage with different personality traits. Ryan and Xenos found positive correlation with neuroticism and Facebook usage [31]. Correa *et al.* found similar positive correlation between online messenger usage and neuroticism [9]. It is also found that people high in neuroticism generally has high online platform usage [2] [6].

Ross *et al.* found that people who use more online platforms have higher extraversion traits [30]. Amichai-Hamburger and Vinitzky found that people with high extraversion have high numbers of friends and followers in online platform [3]. So, previous works suggest that extraversion has positive correlation with number of freinds and followers. It has also been shown that extraversion is positively correlated with the perceived ease of use of social network services [29].

McElroy *et al.* found that openness is related to information seeking in online usage [25]. Openness is also positively correlated with the size of a user's social network [3]. So, it is hypothesized that openness is positively correlated with informational use of online platforms [3].

Positive reply to a post, re-post and friend density is positively correlated to agreeableness. However, many studies found that it is difficult to find a strong correlation with agreeableness and online usage [30][3][9].

Relationship with conscientiousness and online usage is not well established. Butt and Phillips claimed that online usage has negative correlation with conscientiousness [6] whereas Ross *et al.* found no empirical evidence of such claims [30]. Amichai-Hamburger and Vinitzky also did not find any correlation with conscientiousness and Twitter usage [3].

All the works discussed so far establish the relationship of different personality traits with different online platforms' usage. Some later works propose method to calculate trait values from online usage [36][17].

Tal Yarkoni [36] identified traits value based on the word use among bloggers. Identifying traits from blogs are easier than from social media sites because extremely large and topically diverse writing sample are readily accessible in blogs. The actual personality of the participants was measured by a 100-item personality questionnaire. The words used by the participants in blog posts or comments were extracted and then a linguistic analysis was computed on it. Yarkoni calculated 66 psycho-linguistic scores from the blogs. For a sample size of 694, he observed strong correlation between these scores and big five traits.

Using Twitter data, Golbeck *et al.* found that big five personality traits are reflected through the Twitter posts of respected user [17]. They identified the big five personality of 167 subjects through a questionnaire. Users filled a 45-question version of the big five personality inventory from which users respective big five traits were calculated. At the same time, Golbeck *et al.* collected all Twitter data of the user using website crawler. In that process, they gathered all the public data available from their Twitter profiles that includes information about the user (user name, birthday, relationship status, etc.) and all recent twitter posts. This information were aggregated, quantified, and passed through a text analysis tool to obtain a feature set. Using these statistics describing the Twitter profile of each user, they developed a model that can predict personality on each of the five personality factors. Accuracy of the prediction differs significantly depending on which trait is being predicted. From twitter, a person's mood can also be detected along with big five personality traits [7].

All of these studies were conducted on a single social network platform. Human behaviour across multiple social media platforms was not depicted in those studies. There are a handful studies on cross platform analysis of human behaviour. In [35] Xu *et al.* quantified the extent of user engagement of different platforms as well as their correlations. They presented the differences in usage

in six major social networking sites. Their findings were only limited to usage pattern and they did not identify any personal information like personality, mood, topic sensitivity.

In [16] Davenport *et al.* investigated the reasons for narcissism in virtual world by comparing and contrasting the role of narcissism in the usage of Facebook and Twitter. They analysed posts and friend network in both sites and found that narcissism is strongly evident in tweets than Facebook status updates whereas it is a stronger predictor of Facebook friends than Twitter followers. However, their study lacks the analysis of other personality traits across multiple social networks.

Hughes *et al.* investigated the relationship between users' personality and social media use by examining the personality correlates (Big5 traits, Sociability and Need-for-Cognition) of social and informational use of Facebook and Twitter [19]. They found differential relationships between personality and Facebook and Twitter usage. This study is limited to similar type of social media activity analysis. Moreover, they did not perform any analysis on the relationship of interest and topic sensitivity and social media usage.

Most of the previous works identify personality from a person's online media posts. Relationship with other online traits such as person's discussion topics and friends' list is rarely surveyed in previous works. Bryan and Ahsley proposed some method to identify personality trait's related to success from a person's friends' list [21]. However, their work does not identify all personality traits in general. So, in general it can be said that all previous personality prediction models require processing entirety of online posts, comments etc.

2.5 Summary

We have discussed formal terminologies and existing methods for identifying personality in this chapter. Important formal terms including personality, topics, etc. are already well defined in previous literature. In our thesis we have utilized these definition as well as some established methods from previous literatures. In the

following chapter, we propose our method to overcome the limitations of current literature.

Chapter 3

Methodology

In this chapter we present our method to identify comprehensive personality profile from Twitter and Disqus usage. Before explaining our method, in Section 3.1 we describe the input and output of our method. We explain our method in Section 3.2.

3.1 Input and Output of Our Method

In our thesis, we identify a person's personality profile in big five model. More specifically, for a person our method produces output of his big five trait scores. In our thesis, we identify personality profile independently from two different types of data. At first, we identify personality from Twitter and Disqus posts of a person. Input for this part of our thesis is all the Twitter and Disqus posts of a person. We only analyze the textual contents of those posts and any additional info of the post is discarded. Formally, the problem is to identify a person's scores on each of the big five personality traits by analyzing all the Twitter and Disqus posts.

In the second part of our thesis, we identify a person's personality traits from his topic profile. So, input for this part of our thesis is topic profile of the person. We have previously defined topic profile of a person in Section 1.3. In brief, topic profile of a person is a list of all active topics for the person and sentiment value

for each topic. As topic profile of a person is not directly available in online, in our thesis we first identify topic profile of a person from his Twitter and Disqus posts. Afterwards, we develop a prediction model to identify personality trait of a person from his topic profile. We only analyze a person's topic profile as an input for our prediction model and no additional info is directly analyzed. Formally, the problem is to produce a person's personality scores from his topic profile.

3.2 Our Proposed Method

We have defined the input and output of our method in previous section. In this section we describe our proposed method. Methodology part of our thesis can be divided into following three major components:

- We propose a method to identify comprehensive personality profile from Twitter and Disqus posts which we explain in Section 3.3.
- We propose a method to create topic profile of a person from his Twitter and Disqus profile which is explained in Section 3.4.
- We propose a method to identify personality traits from topic profile in Section 3.5.

3.3 Method to Identify Comprehensive Personality

Our method to identify comprehensive personality profile analyzes a person's posts from his Twitter and Disqus accounts. Analysis of multiple media platform enables us to overcome the limitations of single platform analysis. A major technical challenge in this phase of our thesis is that personality identification from comment posting platforms is not discussed in previous literature. Our method consists of three phases:

- We identify personality trait of a person from his Twitter posts which we discuss in Section 3.3.1

- We propose and implement our method to identify personality trait of the person from his Disqus posts which is discussed in Section 3.3.2.
- We combine the traits of a person identified separately from Twitter and Disqus which we discuss in Section 3.3.3.

We collect n -persons Twitter posts, Disqus posts and personality scores and evaluate accuracy of our method using k -fold cross validation. Dataset of n persons is randomly divided into two parts- training set and validation set. Let, n_t be the number of training set and n_v be the number of validation set. Because we are using k -fold cross validation, at each fold number of training data is $n_t = \lceil \frac{(k-1) \times n}{k} \rceil$ and number of validating data is $n_v = \lfloor \frac{n}{k} \rfloor$. We perform all the analysis on training dataset which contains n_t number of persons and propose our model based on this dataset. Afterwards, we verify our model using validation set which contains n_v number of persons.

3.3.1 Method to Identify Personality from Twitter

We begin by analyzing the data pattern of personality trait scores and psycholinguistic scores of tweets. First we remove any unnecessary parts from tweets and tokenize each tweet. Subsequently, utilizing LIWC we calculate psycho-linguistic scores of a person from all his tweets. Details of LIWC dictionary is explained in Section 2.2.

Some selected persons neuroticism value and anxiety scores from twitter posts are plotted in Figure 3.1 which clearly shows a positive correlation between neuroticism scores and anxiety scores of twitter posts. There are a number of significant correlations between various LIWC scores and personality traits. However, none of the relations are strong enough to independently identify personality scores. Many scores are not correlated with a particular trait at all. To identify these correlations, we calculate Pearson correlation coefficients between subjects' personality scores and each of the feature scores obtained from all twitter posts.

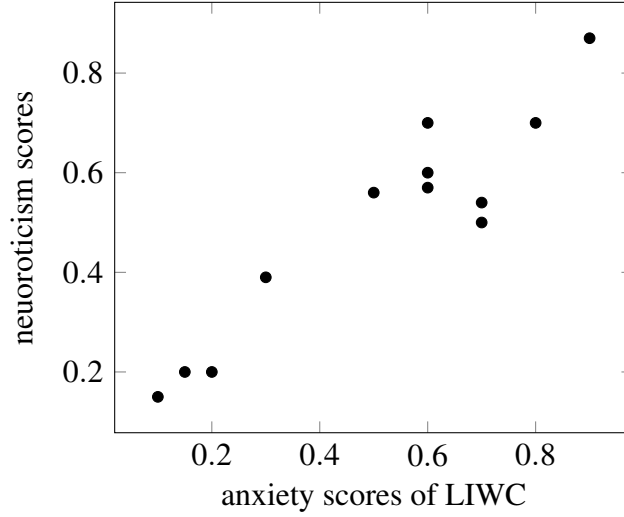


Figure 3.1: Correlation between neuroticism and anxiety in Twitter

We have n_t persons each having l feature scores. For each of the l feature and each personality trait, we calculate the feature's Pearson correlation co-efficient with that personality trait. We have $M = 5$ personality traits. For convenience, we define some notations which is shown in Table 3.1:

Table 3.1: Some defined symbols and their explanation

symbol	explanation
r_{jm}	Pearson correlation coefficient for j -th feature and m -th personality trait
f_{ij}	i -th person's j -th feature score
\bar{f}_j	average of all persons' j -th feature scores
p_{im}	i -th person's m -th personality trait score
\bar{p}_m	average of all m -th personality trait scores

Using the notations defined previously, the formula for r_{jm} is shown in Equation 3.1.

$$r_{jm} = \frac{\sum_{i=1}^{n_t} (f_{ij} - \bar{f}_j)(p_{im} - \bar{p}_m)}{\sqrt{\sum_{i=1}^{n_t} (f_{ij} - \bar{f}_j)^2} \sqrt{\sum_{i=1}^{n_t} (p_{im} - \bar{p}_m)^2}} \quad (3.1)$$

Algorithm 1 Algorithm to identify Pearson Coefficient

```

1: procedure CALCPEARS ( $r$ )
2:   for each personality trait  $m$  do
3:     for each feature  $j$  do
4:       calculate average of  $j$ -th feature  $\overline{f_j}$ 
5:       initialize three variables  $x, y$  &  $z$ 
6:       for each person  $i$  from training set do
7:          $x = x + (f_{ij} - \overline{f_j})(p_{im} - \overline{p_m})$ 
8:          $y = y + (f_{ij} - \overline{f_j})^2$ 
9:          $z = z + (p_{im} - \overline{p_m})^2$ 
10:      end for
11:      calculate  $r_{jm}$  ( $r_{jm} = \frac{x}{\sqrt{y}*\sqrt{z}}$ )
12:    end for
13:  end for
14:  return calculated pearson coefficient  $r$ 
15: end procedure

```

By calculating r we identify which psycho-linguistic scores to consider while identifying personality traits. We disregard j -th psycho-linguistic score for identifying m -th personality trait if $|r_{jm}| \leq \beta_t$, where β_t denotes significance threshold. Algorithm to identify Pearson Coefficient is shown in Algorithm 1 and algorithm for reducing feature set to identify personality from Twitter is shown in Algorithm 2.

We utilize a linear regression model for each personality trait to identify trait value from these significant psycho-linguistic scores. Let, $l^{(m)}$ be the number of total reduced psycho-linguistic features (only features having significant Pearson coefficients) for personality trait m and $f^{(m)}$ refers to the feature scores in the reduced feature set where the feature set is reduced for identifying m -th personality trait. We also assume that each person's 0-th feature score is 1, so $f_{i_0}^{(m)} = 1$ for all

Algorithm 2 Algorithm to identify reduced feature set

```

1: procedure CALCFEATSET
2:   calculate Pearson coefficients( calculate  $r$  )
3:   for each personality trait  $m$  do
4:     initialize reduced feature set for  $m$ -th trait  $f^{(m)}$ 
5:     initialize variable  $x$ 
6:     for each feature  $j$  do
7:       if  $|r_{jm}| \geq \beta$  then
8:         for each person  $i$  from training set do
9:           assign  $f'_{ix} = f_{ij}$ 
10:        end for
11:        increment  $x$ 
12:       else
13:         disregard this feature
14:       end if
15:     end for
16:     save reduced feature set for  $m$ -th personality  $f^{(m)}$ 
17:   end for
18: end procedure

```

i. A person's personality trait score is calculated using the following formula,

$$p'_{im} = \sum_{j=0}^{l^{(m)}} \theta_{jm} f'_{ij} \quad (3.2)$$

p'_{im} = i -th person's m -th predicted personality trait which is a function of $f^{(m)}$

θ_{jm} = correlation co-efficient between j -th feature score and m -th personality trait

$f'_{ij} = f_{ij}^{(m)}$ = i -th person's j -th feature score where feature set is reduced for m -th trait

$l^{(m)}$ = total number of reduced features for m -th personality

From this model and our reduced feature set, we predict personality of a person by using co-efficients (values of θ_{jm} for different j and m) identified by Golbeck

[17].

3.3.2 Method to Identify Personality from Disqus

Methods to identify personality from social media sites like Twitter are not applicable for identifying personality from comment posting platforms like Disqus. People use Disqus differently and as a consequence, a person's usage pattern in Disqus is different from Twitter. We begin by analyzing the relationship between a person's personality trait scores and different psycho-linguistic scores of Disqus posts. To identify psycho-linguistic scores of Disqus posts, we utilize LIWC dictionary which have been explained in Section 2.2. Personality scores and psycho-linguistic scores are not directly correlated in Disqus. Some selected persons neuroticism trait scores and anxiety scores are shown in Figure 3.2. This graph plot shows that the relationship between neuroticism scores and anxiety scores is not linear in Disqus.

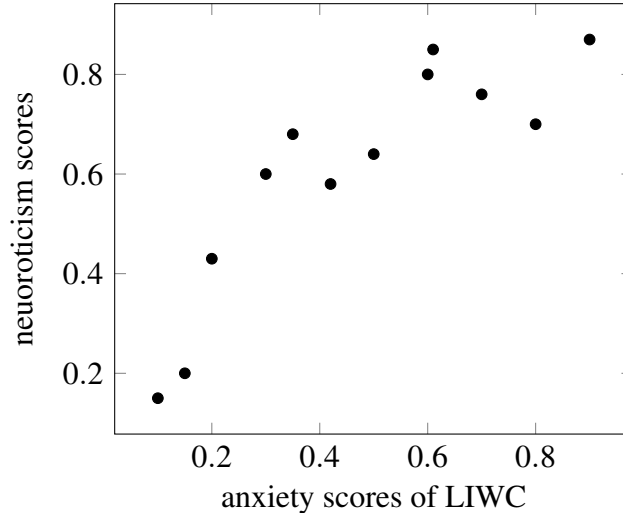


Figure 3.2: Correlation between neuroticism and anxiety in Disqus

On the contrary, Figure 3.1 shows a linear relationship between neuroticism scores and anxiety scores in Twitter. Among all the psycho-linguistic features,

not all are useful for identifying personality scores. But in case of Disqus, we cannot use Pearson correlation coefficient to remove redundant features. Pearson correlation coefficient calculates the direct linear correlation-ship and as the graph plot of some psycho-linguistic scores and personality trait scores can be best fit using non-linear curve, direct Pearson correlation coefficient may wrongly disregard some important psycho-linguistic scores.

We evaluate Spearman's correlation coefficients (ρ) to identify significant features. As the relationship between psycho-linguistic scores and personality trait scores can be explained using monotonic function, we can interpret the relationship using ρ . We propose a method to identify reduced feature set utilizing ρ and null hypothesis in the following paragraphs.

For each personality trait m , we calculate a new rank function (R^m) of trait scores of every person from training set. Similarly, for a particular feature j , we calculate rank function (R^j) of feature scores of every person from training set. Rank of a personality score of m -th trait is the numeric position of that trait score among all the m -th personality scores and rank of a feature score of j -th feature is the numeric position of that feature score among all the j -th feature scores.

$P_{i_m} = i$ -th person's m -th personality trait score

$R^m(P_{i_m}) =$ numeric position of P_{i_m} among all m -th personality scores

$f_{i_j} = i$ -th person's j -th feature score

$R^j(f_{i_j}) =$ numeric position of f_{i_j} among all j -th feature scores

Spearman's correlation coefficient between j -th feature score and m -th personality trait (ρ_{jm}) is shown in Equation 3.3.

$$\rho_{jm} = 1 - \frac{6 \times \sum_{i=1}^{n_t} (R^m(P_{i_m}) - R^j(f_{i_j}))^2}{n_t(n_t^2 - 1)} \quad (3.3)$$

For each personality trait m and each feature score j , we identify ρ_{j_m} and utilize null hypothesis to disregard any redundant feature. We initially assume that j -th feature and m -th personality is unrelated to each other. Then using ρ_{j_m} and number of training set (n_t), we identify the probability p which is the probability that our observed data is inconsistent with the hypothesis. So, when $p \leq \beta_d$, (where β_d is the significance level) we reject the hypothesis and use the j -th feature score for identifying m -th personality. Algorithm for identifying ρ is shown in Algorithm 3.

Algorithm 3 Algorithm to identify Spearman's coefficient

```

1: procedure CALCRANK
2:   sort all  $n_t$  persons'  $m$ -th trait scores from low to high
3:   for  $i = 1$  to  $n_t$  do
4:     assign  $R(p_{i_m}) = \text{index of } p_{i_m}$  in sorted order
5:   end for
6:   for  $j = 1$  to  $n_t$  do
7:     sort all  $n_t$  persons'  $j$ -th feature scores from low to high
8:     for  $i = 1$  to  $n_t$  do
9:       assign  $R(f_{i_j}) = \text{index of } f_{i_j}$  in sorted order
10:    end for
11:  end for
12: end procedure
13: procedure CALCSPEAR
14:   call procedure CalcRank to calculate rank of all traits and feature scores
15:   for each personality trait  $m$  do
16:     for each feature  $j$  do
17:       initialize variable  $x$ 
18:       for each person  $i$  from training set do
19:          $x = x + (R^m(P_{i_m}) - R^j(f_{i_j}))^2$ 
20:       end for
21:       calculate  $\rho_{j_m} = 1 - \frac{6 \times x}{n_t(n_t^2 - 1)}$ 
22:     end for
23:   end for
24:   return  $\rho$ 
25: end procedure

```

Let, $f^{(m)}$ be the reduced feature set for m -th trait and $l^{(m)}$ be the number of reduced features for m -th trait. For calculating convenience, we also assume that each person's zero-th feature score is 1 ($f_{i_0}^{(m)} = 0$, for all i). Linear regression model cannot predict personality with significant accuracy. We analyze our data pattern with various non-linear model and identify the best fit for Disqus data. We propose Equation 3.4 for identifying personality trait score.

$$p'_{i_m} = \sum_{j=0}^{l^{(m)}} \theta_{j_m} f_{i_j}^{(m)} + \sum_{j=1}^{l^{(m)}} \theta'_{j_m} \times \sqrt{f_{i_j}^{(m)}} \quad (3.4)$$

p'_{i_m} = i -th person's m -th predicted personality trait which is a function of $f^{(m)}$

θ_{j_m} = correlation co-efficient of j -th feature score and m -th personality trait

θ'_{j_m} = correlation coefficient of square root of j -th feature score and m -th personality trait

$f_{i_j}^{(m)}$ = i -th person's j -th feature score where feature set is reduced for m -th trait

$l^{(m)}$ = total number of reduced features for m -th personality

We utilize our training dataset (number of trained dataset = n_t) to identify the values of θ and θ' . E^m refers to error in predicted and actual personality trait scores when identifying m -th trait which is a function of related θ and θ' . So,

$$E^m(\theta, \theta') = E(\theta_{0_m}, \theta_{1_m}, \dots, \theta_{l^{(m)}_m}, \theta'_{1_m}, \theta'_{2_m}, \dots, \theta'_{l^{(m)}_m}) \quad (3.5)$$

In our model, error is calculated as the square of the difference of our predicted and real trait values. Formula for error function is shown in Equation 3.6.

$$E^m(\theta, \theta') = \frac{1}{2n_t} \sum_{i=1}^{n_t} (p'_{i_m} - p_{i_m})^2 \quad (3.6)$$

Algorithm 4 Algorithm to identify coefficient values (θ and θ') of features

```

1: procedure EVALTHETA
2: Input:  $m$ ,  $m$ -th personality trait of  $n_t$  persons, reduced feature scores  $f^{(m)}$ ,  $\alpha_l$ 
3: Output: coefficient of correlation ( $\theta$  and  $\theta'$ ) for  $l^{(m)}$  features
4:   initialize  $\theta_{0_m}, \theta_{1_m}, \dots, \theta_{l_m^{(m)}}$ 
5:   initialize  $\theta'_{0_m}, \theta'_{1_m}, \dots, \theta'_{l_m^{(m)}}$ 
6:   repeat
7:     for  $j = 0$  to  $l^{(m)}$  do
8:       initialize variable  $x$  and  $y$ 
9:       for  $i = 1$  to  $n_t$  do
10:         $x = x + (p'_{i_m} - p_{i_m}) * f_{i_j}^{(m)}$ 
11:         $y = y + (p'_{i_m} - p_{i_m}) * \sqrt{f_{i_j}^{(m)}}$ 
12:      end for
13:      update  $\theta_{j_m}$  ( $\theta_{j_m} = \theta_{j_m} - \frac{\alpha_l * x}{n_t}$ )
14:      update  $\theta'_{j_m}$  ( $\theta'_{j_m} = \theta'_{j_m} - \frac{\alpha_l * y}{n_t}$ )
15:    end for
16:  until ( $\theta$  is not converged) OR ( $\theta'$  is not converged)
17:  return output ( $\theta_{0_m}, \theta_{1_m}, \dots, \theta_{l_m^{(m)}}$ ) and ( $\theta'_{0_m}, \theta'_{1_m}, \dots, \theta'_{l_m^{(m)}}$ )
18: end procedure

```

We implement gradient descent algorithm to identify θ and θ' . In each step of our algorithm, we update the coefficient value and repeat until the result is converged. We provide the definition of some symbols in Table 3.2.

Table 3.2: Some defined symbols and their explanation

symbol	explanation
α_l	learning rate of algorithm
θ_{j_m}	coefficient of correlation of j -th feature and m -th trait
θ'_{j_m}	coefficient of correlation of square root of j -th feature and m -th trait
n_t	total number of training data
p'_{i_m}	i -th person's predicted score of m -th personality trait
p_{i_m}	i -th person's actual score of m -th personality trait
$f_{i_j}^{(m)}$	i -th person's j -th feature score in reduced feature set

Using our defined notation, Equation 3.7 provides the update formula for θ . This equation can be rewritten as Equation 3.8 which we implement in our algorithm.

$$\theta_{j_m} = \theta_{j_m} - \alpha_l \frac{\partial}{\partial \theta_{j_m}} E^m(\theta, \theta') \quad (3.7)$$

$$\theta_{j_m} = \theta_{j_m} - \frac{\alpha_l}{n_t} \sum_{i=1}^{n_t} (p'_{i_m} - p_{i_m}) \times f_{i_j}^{(m)} \quad (3.8)$$

(simultaneously update for $j = 0, 1, 2, \dots, l^{(m)}$)

Similarly, Equation 3.9 provides the update formula for θ' which is rewritten as Equation 3.10.

$$\theta'_{j_m} = \theta'_{j_m} - \alpha_l \frac{\partial}{\partial \theta'_{j_m}} E^m(\theta, \theta') \quad (3.9)$$

$$\theta'_{j_m} = \theta'_{j_m} - \frac{\alpha_l}{n_t} \sum_{i=1}^{n_t} (p'_{i_m} - p_{i_m}) \times \sqrt{f_{i_j}^{(m)}} \quad (3.10)$$

(simultaneously update for $j = 0, 1, 2, \dots, l^{(m)}$)

We identify values of coefficient from our training dataset. Algorithm to identify the coefficient values (θ and θ') is shown in Algorithm 4. Algorithm to identify personality trait scores from these values and reduced feature set is shown in Algorithm 5. Lastly, we identify accuracy of our model utilizing the validation set.

3.3.3 Combine Personality Scores

We analyze the differences of trait values of a person obtained independently from Twitter and Disqus. Using co-efficient analysis, we identify whether a trait is strong in Twitter or Disqus. From this analysis, we identify all five trait values of

Algorithm 5 Algorithm to identify personality scores from Disqus

```

1: procedure IDPERS
2:   for each personality trait  $m$  do
3:     call procedure EvalTheta( $m$ ) to calculate related  $\theta$  and  $\theta'$ 
4:     for each person  $i$  in validation set do
5:       initialize  $i$ -th person's  $m$ -th trait score  $p'_{i_m}$ 
6:       for  $j = 0$  to  $l^{(m)}$  do
7:          $p'_{i_m} = p'_{i_m} + \theta_{j_m} f_{i_j}^{(m)}$ 
8:          $p'_{i_m} = p'_{i_m} + \theta'_{j_m} \sqrt{f_{i_j}^{(m)}}$ 
9:       end for
10:      print  $p'_{i_m}$  as predicted score of  $i$ -th person's  $m$ -th trait
11:    end for
12:  end for
13: end procedure

```

a person.

3.4 Method to Identify Topic Profile

We propose a method to identify topic profile from a person's Twitter and Disqus usage which is pretty straightforward. Steps of our method is given below:

- We identify topics from each Twitter and Disqus post of a person utilizing Texalytic tools.
- We also identify sentiment of each post using Texalytic tools
- Finally we create a list of all active topics for a person. We calculate sentiment of each topic by averaging the sentiment of all posts related to that topic.

3.5 Method to Predict Personality from Topic Profile

In this section we discuss our proposed method to predict a person's personality trait values from his topic profile. Our prediction model can be explained using two phases- a feature extraction phase and a prediction classifier phase. We give a brief overview of the model in this section.

In feature extraction phase, we identify n'_c scores of a person from his topic profile. These n'_c scores are regarded as training attributes for our prediction model. Now we discuss how we identify training attributes from topic profile. Let, v^i be topic profile of i -th person which contains his sentiment value on x^i topics ($v^i = \{v_1^i, v_2^i, \dots, v_{x^i}^i\}$). Number of active topics is different for different persons and as a result, x^i differs for different values of i . From i -th person's scores on x^i topics we calculate his scores on n_c topic clusters where $n_c \leq x^i$. Let, v'^i be the set of these scores of i -th person which contains i -th person's scores on n_c topic clusters ($v'^i = \{v_1'^i, v_2'^i, \dots, v_{n_c}'^i\}$). Unlike x^i , n_c is fixed for all persons. Identifying n_c topic clusters and calculating a person's score on a particular topic cluster are discussed in Section 3.5.1. Using principal component analysis, we identify n'_c scores from a person's scores on n_c topic clusters where $n'_c \leq n_c$. Let, v''^i be i -th person's scores of these n'_c attributes. Using this notation, scores of training attributes for i -th person are $v_1''^i, v_2''^i, \dots, v_{n'_c}''^i$. Using these attributes our objective is to identify each five personality trait of the person. We divide continuous personality trait scores in b discrete ranges. Instead of directly predicting the numeric score of personality traits, we predict one of the b class values for each personality trait. So, our prediction classifier has 5 target attributes where each attribute can take any of the b class values. Using our notation, training and target attributes of our model is shown using block diagram in Figure 3.3. In this figure, P_j^i is i -th person's j -th target attribute.

From the implementation of previous part of our thesis, we have n persons' topic profile and personality trait scores. From each person's topic profile we

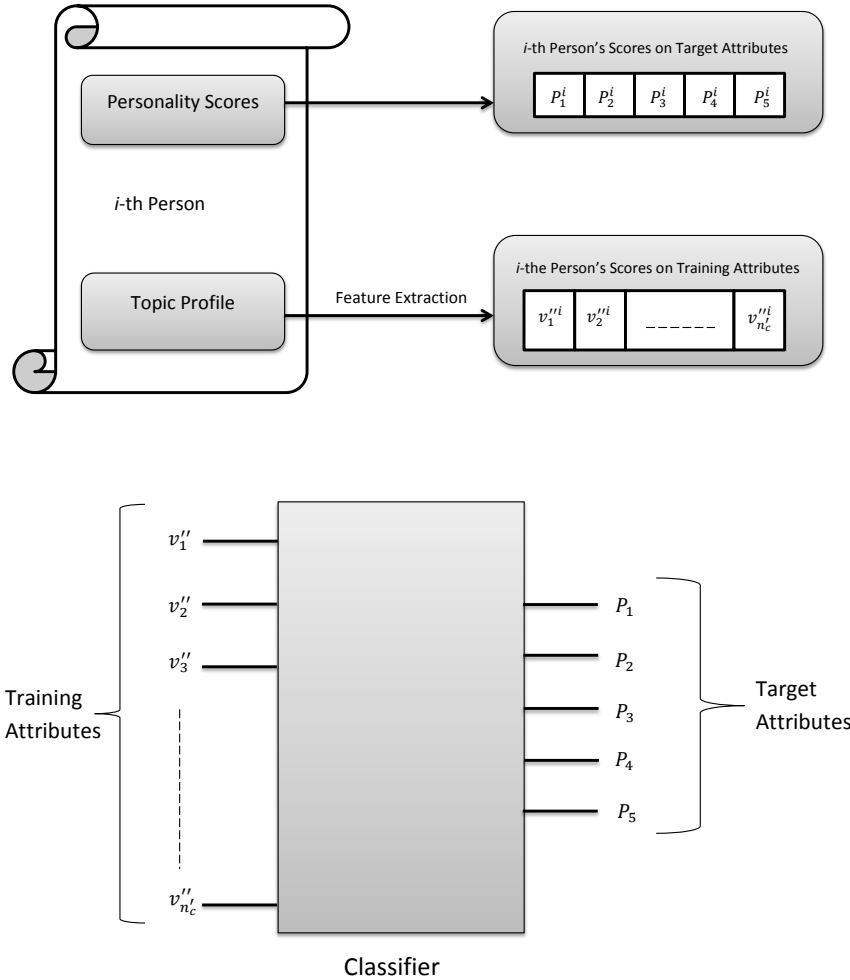


Figure 3.3: Block diagram of training and target attributes of our model

identify n'_c training attributes and from personality scores we identify 5 target attributes. We divide this dataset into training set and test set. Training set is utilized to train our model and test set is utilized to validate our model. In our experiment, we identify accuracy of our proposed model using k -fold cross validation.

3.5.1 Extracting Features from Topic Profile

At the very beginning of our feature extraction phase, we create a list of all unique topics from the topic profile of all the persons. Then we remove all the topics that are present in only a few number of persons. This will remove all the less occurring topics from our topic profile. The intuition behind this step is that topics occurring for only a limited number of people can only be utilized to predict a small group of people. As we want our method to be scalable in larger dataset, we discard all these topics.

A major technical challenge of identifying features is that there are many missing attributes- not all person is active in all topics. A straightforward approach to address missing features is selecting only the most frequently used subset of all topics. But this approach disregards many information regarding many topic usage which consequently decreases the accuracy of prediction model. We address this challenge by proposing a topic clustering method. In this section, we explain our clustering method.

After topic removal phase, let there be x unique topics. For each topic we calculate its m different scores including scores on psychological processes, scores on time, scores on personal concerns, etc. We regard each topic as a point in m -dimensional Euclidean space. So, topic list $T = \{t_1, t_2, \dots, t_x\}$ is a collection of x points in \mathbb{R}^m (m -dimensional Euclidean space). t_i is the i -th topic in topic list and has m dimensions, where $t_{i,j}$ refers to the j -th dimension of t_i .

Our clustering algorithm is based on a distance threshold d . We define T_u as the set of unprocessed points which is initialized to T in the beginning of our algorithm. We define “cluster” as a set of topics and denote C to indicate the set of all clusters. We also initialize the set of all clusters C as an empty set. Each execution of our algorithm starts by randomly selecting s points from T_u where distance between any two points is at least d . These s points are removed from the set T_u . For calculating the distance between any two points t_i and t_j , we utilize the

Euclidean distance formula for m dimensions:

$$\text{distance between } t_i \text{ and } t_j = \sqrt{\sum_{k=1}^m (t_{i_k} - t_{j_k})^2}$$

In this initialization phase, each of the s points create a new cluster which is added to C . From empty set, after this step C becomes, $\{C_1, C_2, \dots, C_s\}$ where each cluster contains a single point.

In each of the following steps, our algorithm selects a point from the unprocessed points' set T_u and removes it from T_u . Let the point be t_p . Subsequently, distance between t_p and every cluster of C is calculated. We define distance between a point and a cluster of points as Euclidean distance of m -dimension between the point and cluster centroid. A cluster centroid has m -dimensions where i -th dimension is the average of all the points' i -th dimension in the cluster. Our algorithm executes one of the following depending on the condition:

Condition 1: We identify the closest cluster from t_p based on the distance from cluster centroid. Let, the closest cluster be C_a . When distance between centroid of C_a and t_p is less than d , we add the point t_p to the cluster set C_a and update the centroid of C_a .

This condition is shown in Figure 3.4. For demonstration purpose, only two arbitrary features are shown and as a result, the graph is two-dimensional. In this figure, yellow colored point is t_p and C contains three clusters, points belonging to these three clusters are shown as red, green and blue colors respectively. All remaining black colored points belong to unprocessed points' set T_u . Our algorithm calculates distance between t_p and cluster centroid of red, green and blue colored clusters respectively. Among these distances, minimum distance is between t_p and red colored cluster. This minimum distance is less than d . Therefore in the next step, yellow colored point is added to the red cluster and cluster centroid is updated.

Condition 2: When the distance between t_p and the closest cluster C_a is more than d , a new cluster C_{new} is created which initially contains the point t_p . If any

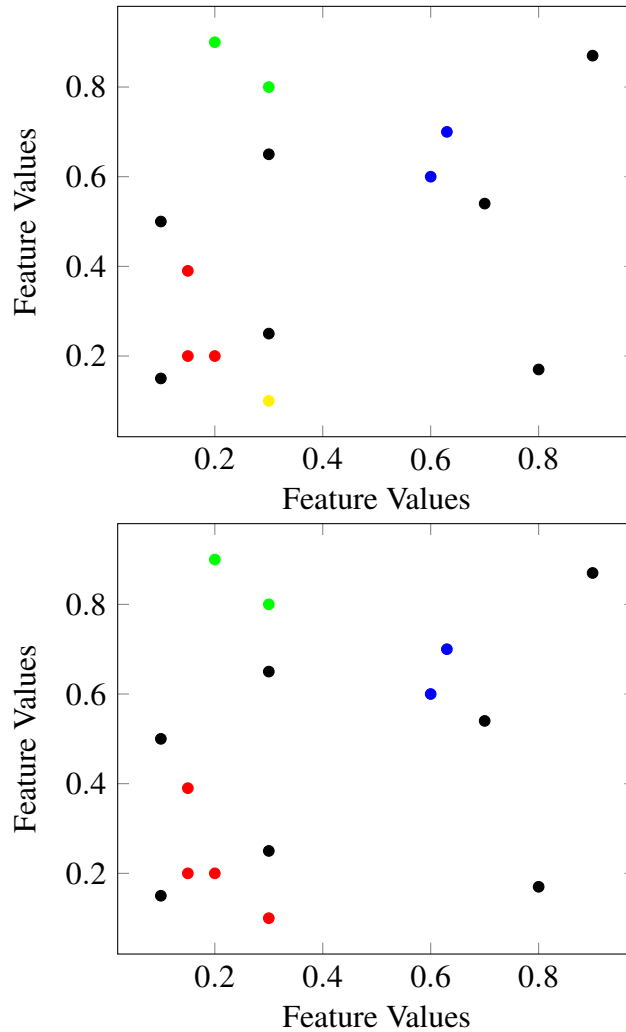


Figure 3.4: Condition 1 of our clustering algorithm

previously processed point's distance from the new cluster C_{new} is less than its own cluster C_{prev} , then this point needs to be updated. In such case, we remove the point from C_{prev} , add it to C_{new} and update the centroid of both C_{prev} and C_{new} .

This condition is shown in Figure 3.5. Similar to previous figure, this figure also shows yellow colored point as t_p . C contains three clusters- points belonging to these three clusters are shown as red, green and blue colors respectively. All re-

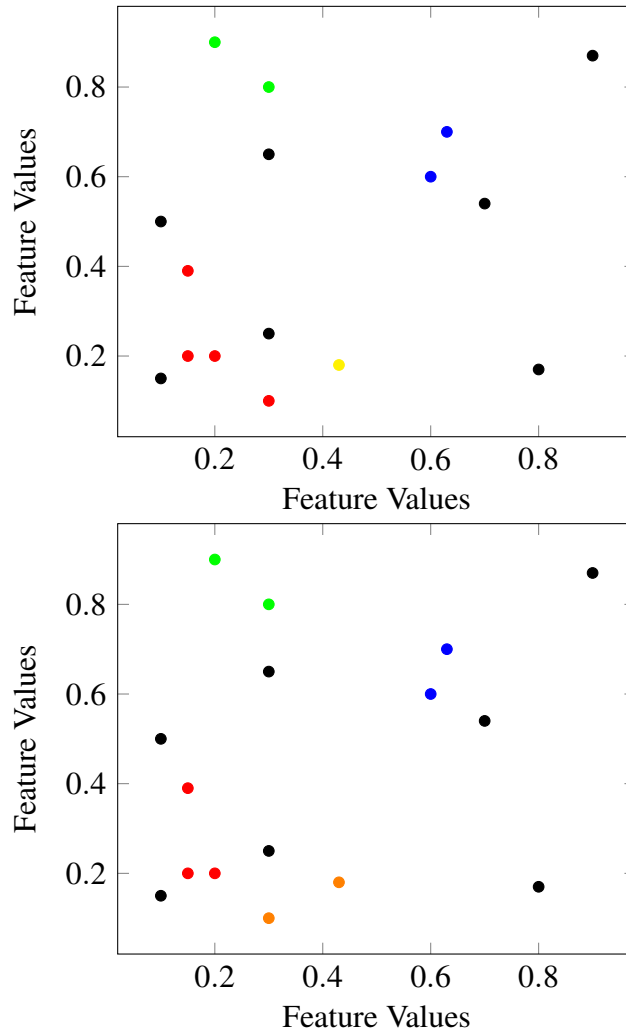


Figure 3.5: Condition 2 of our clustering algorithm

maining black colored points belong to unprocessed points' set T_u . Our algorithm calculates distance between t_p and cluster centroid of red, green and blue colored clusters separately. Among these distances, minimum distance is between t_p and red colored cluster. This minimum distance is greater than d . Therefore, yellow colored point is not added to the red cluster. A new cluster with the point t_p is created which is shown using orange color. One previously processed points from

red cluster is also removed from red cluster and added to this new cluster.

Our algorithm starts with a set of unclustered points T and produces a set of clusters C , where the total number of clusters is not fixed before execution. Our algorithm is shown in Algorithm 6. Each execution of our algorithm may produce different clustering depending on the initially selected points. We execute our algorithm r times from beginning and select the clustering where the maximum standard deviation of all the clusters is minimum.

Our topic clustering algorithm produces distinct number of topic clusters. Let, the number of topic clusters be n_c and the set of clusters $C = \{C_1, C_2, \dots, C_{n_c}\}$. To extract features from a person's topic profile, we identify the person's sentiment value of n_c clusters (C_1, C_2, \dots, C_{n_c}). For a person, sentiment of a cluster is the average of sentiment of all the topics within the cluster. If a person is not active on any topic, that topic is disregarded while calculating average. Sentiment of a cluster gives us a real value between -1 and +1. Afterwards, we plot all the training dataset (n_t persons' dataset where each person has sentiment value on n_c clusters) and identify principal component vector for n'_c principal components. For each person in our validation set, we first identify n_c clusters' sentiment value from his topic profile. Afterwards, we evaluate n'_c principal components using previously identified principal component vectors. For a person, these values of n'_c training attributes are utilized as features for our prediction classifier.

3.5.2 Prediction Classifier

We propose a prediction model to predict personality class from extracted features. First, we identify class from a discrete value of personality trait. We divide the personality trait value to create b personality classes each with equal range ($\frac{1}{b}$). So personality prediction problem transforms into predicting class from already extracted features.

For each personality trait, we propose an independent prediction model. So one personality trait has no influence on predicting other personality traits. The intuition behind our model is that big five personality traits are independent of

Algorithm 6 Algorithm to cluster a set of points

```

1: procedure CLUSTER
2: input: a set of points  $T = \{t_1, t_2, \dots, t_x\}$ , distance threshold  $d$ 
3: output: a set of clusters  $C$  (each cluster is a set of points)
4:   initialize  $C$  to empty set
5:   initialize  $T_u = T$ 
6:   randomly select  $s$  points from  $T_u$ 
7:   for each point  $t_i$  of selected  $s$  points do
8:     remove  $t_i$  from  $T_u$ 
9:     create a new cluster  $C_d$ 
10:    push  $t_i$  to  $C_d$  and update centroid of  $C_d$ 
11:    push  $C_d$  to set of clusters  $C$ 
12:   end for
13:   while  $T_u$  is not empty do
14:     initialize  $t_p$  to the top point in  $T_u$ 
15:     remove  $t_p$  from  $T_u$ 
16:     initialize cluster  $C_a$  to first cluster of  $C$  ( $C_1$ )
17:     initialize variable mindist to distance between  $t_p$  and  $C_1$ 
18:     for each cluster  $C_i$  from  $C$  do
19:       calculate variable curdist = distance between  $t_p$  and  $C_i$ 
20:       if curdist  $\geq$  mindist then
21:         update  $C_a$  to  $C_i$ 
22:         update mindist to curdist
23:       end if
24:     end for
25:     if mindist  $\leq d$  then
26:       push  $t_p$  to  $C_a$ 
27:       update centroid of  $C_a$ 
28:     else
29:       create a new cluster  $C_d$ 
30:       push  $t_p$  to  $C_d$ 
31:       for each point  $t_i$  in  $C$  do
32:         calculate  $d_1$  = distance of  $t_i$  from current cluster
33:         calculate  $d_2$  = distance of  $t_i$  from  $C_d$ 
34:         if  $d_1 \geq d_2$  then
35:           remove  $t_i$  from its current cluster
36:           push  $t_i$  to  $C_d$ 
37:           update centroid of both clusters
38:         end if
39:       end for
40:     end if
41:   end while
42:   return  $C$  (set of clusters)
43: end procedure

```

each other. Our model applies one-vs-all technique to predict b classes. So for each class, we apply our prediction model and predict the class with highest probability.

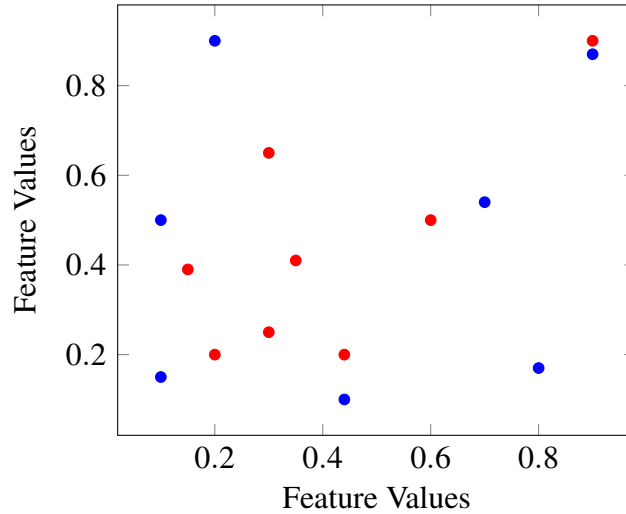


Figure 3.6: An illustration of some selected training data in 2-dimensional plot for conscientiousness trait of range (0.1-0.2)

For a particular class of a particular trait, each trained dataset with n'_c -features is plotted as an n'_c -dimensional point. This point is labeled as either -1 or a +1, where +1 represents that the point belongs to that particular class. For illustration purpose we have shown few trained data points of conscientiousness trait of range (0.1-0.2) in Figure 3.6. Blue points represents the points of this class and red points represents points outside this class. Our model implements a support vector machine based classifier to predict class from our extracted features. There are few intuitions behind our proposed model. Extracted features from our proposed methods collectively make up a complex class boundary in n'_c -dimensional graph. However, due to errors in topic extraction and sentiment extraction, some points lie outside the boundary. Our support vector machine based model is robust in presence of random errors. As support vector machine is a large margin classifier our proposed model identifies the boundary with considerable margin which

improves the prediction accuracy. Our model uses Gaussian kernel because the dataset is not linearly separable in a hyper plane. As we need our proposed model to work in real life scenario, it needs to be computationally feasible. Gaussian kernel with support vector machine allows us to predict a complex non-linear hyperplane without heavy computational cost.

3.6 Summary

In this chapter, we propose our method to identify comprehensive personality profile from Twitter and Disqus usage. To achieve this objective, we also propose method to identify personality from Disqus and identify topic profile from online usage. In the following part, we propose our method to predict personality from topic profile. For this purpose, we also propose a variety of techniques both in feature extraction and prediction phase. In summary, our proposed methods consist of some improvised existing methods as well as some novel approaches to achieve the objectives of our thesis.

Chapter 4

Experiment and Analysis

We implement our proposed methods and perform experiments to measure the accuracy of our methods. This chapter contains the data collection methods, experimental settings and experimental results of our thesis. In Section 4.1, we explain our data collection method. Afterwards, in Section 4.2, we explain our experimental environment and settings for our implementation. Rest of this chapter contains the experimental results and analysis of the results.

4.1 Data Collection and Dataset

For this research we develop an application that crawls Disqus network. Initially we select 20 random Disqus users of different geography and age. Therefore we list all the people who have affiliation with them (i.e. replied to their comments or have comments on same topics). In this process we collect information about 2530 Disqus users. From these Disqus users, we identify 173 users who have linked their public Twitter profile to Disqus profile as of March 11, 2015 and populate our Dataset.

From Disqus profiles we collect basic profile info of a user (i.e., name, location, date of profile creation, about me) using Disqus API. Besides that we gather headline of the news articles on which the user reacted, comments on those ar-

ticles, number of likes and dislikes on those comments. From Twitter profile we collect latest tweets of the user along with basic profile info. We use Twitter API to collect the tweets posted by these users. Due to the constraint of Twitter API, we could collect up to 3200 recent tweets per user. We omit users who have less than 50 tweets. For Disqus profiles we consider those who have at least 30 comments.

After the data collection phase, we have 105 users to analyze. On average there is around 1300 tweets per user. In Disqus on average a user discussed on 424 distinct articles through 1170 comments - around 3 comments per news article. Our collected dataset is available in online through this URL link "sites.google.com/nagibteaching/dataset".

4.2 Experimental Setup

From our dataset of 105 persons' Twitter and Disqus posts, we identify personality profile of a person by implementing our proposed method. To identify personality from Twitter, for each person we first calculate 70 psycho-linguistic scores using LIWClite7 from his Twitter posts. LIWClite7 is a software tool to identify LIWC scores from texts which is freely available for research purposes [20]. From among these scores, for each trait we identify significant scores by implementing the method discussed in Section 3.3.1. On average 18 scores per personality trait remain after this step. Utilizing these scores and Golbeck's coefficients [17], we identify trait scores of a person by implementing Equation 3.2. To identify trait from Disqus, for each person we calculate 70 psycho-linguistic scores from all his Disqus posts utilizing LIWClite7. Then we identify significant scores from these 70 scores by implementing the method discussed in Section 3.3.2. After this step, on average we have 21 feature scores per personality trait. Afterwards, we identify trait scores of a person from these scores by implementing Equation 3.4. To create a person's combined personality profile, we consider scores of openness and conscientiousness from his Disqus profile and score of extraversion from his Twitter profile. To identify scores of his neuroticism and agreeableness, we

consider online platform for which his feature scores are higher. All our implementations in this part are done in C++ programming language and are available in online through this URL link "sites.google.com/nagibteaching/imp1".

Second part of our experiment is related to topic profile. First, we identify topic profile from a person's Disqus and Twitter dataset using Texalytic tools [15]. A person's personality trait values are available from previous part of our experiment. We evaluate the accuracy of our prediction model based on the dataset of 105 persons' personality profiles and topic profiles.

We implement our prediction model discussed in Section 3.4 which consists of two phases- feature extraction phase and prediction classifier phase. After applying our topic removal method, we have 226 unique topics ($x = 226$). We apply our topic clustering algorithm with distance threshold $d = 0.05$ to these 226 topics and identify 93 unique clusters ($n_c = 93$). From a person's topic profile, we calculate his scores on these 93 clusters. Afterwards, we apply our principal component analysis to identify scores on 41 principal components from these 93 scores ($n'_c = 41$). These 41 attributes are training attributes and scores of these attributes are used as features for each person.

In prediction phase, we divide personality scores in 10 equal ranges where each range represents a class ($b = 10$). Using previously mentioned 41 features, we predict a person's personality range using zeroR, linear regression, decision tree based classifier, logistic regression, naive bayesian and support vector machine based classifier. We utilize WEKA data mining tool [18] to implement all these classifiers in JAVA programming language. All our implementations with various parameter values are available online through this URL link "sites.google.com/nagibteaching/imp2". We apply support vector machine using linear kernel and Gaussian kernel with three different σ values (0.01, 0.03, 0.05). In Section 4.4.2 we analyze our experimental results.

All the experimental implementations of this thesis are done on a personal computer equipped with Intel Core 2 Duo CPU T6600 at 2.2 GHz processor and 4GB RAM.

4.3 Experimental Results and Analysis of Identifying Personality from Posts

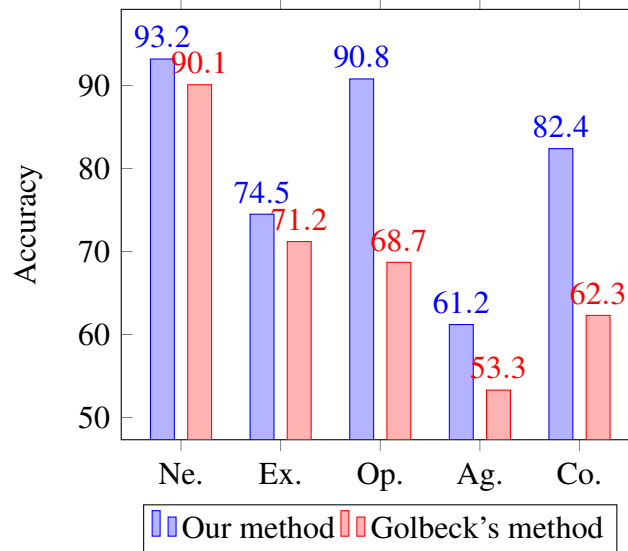


Figure 4.1: Comparison of accuracy of our method and Golbeck's method

In this section we analyze our experimental result of identifying personality from posts. In this part, we identify comprehensive personality profile of a person from his Twitter and Disqus posts. Our method offers significant improvement in identifying big five personality traits over existing methods. Accuracy of our method and Golbeck's method is shown in Table 4.1. In this table, a person's identified trait value is considered to be accurate when difference in the predicted and actual normalized trait value is less than 0.1. Same accuracy is shown using histogram in Figure 4.1.

Our experiment shows that Twitter expresses a person's extraversion profoundly whereas Disqus expresses a person's openness and conscientiousness profoundly. We can identify personality of a person with higher accuracy than existing methods by combining the two online media platforms' usage.

Table 4.1: Accuracy of our method and Golbeck’s method

Personality	Our method	Golbeck’s method
Neuroticism	93.2	90.1
Extraversion	74.5	71.2
Openness	90.8	68.7
Agreeableness	61.2	53.3
Conscientiousness	82.4	62.3

4.3.1 Effects of Feature Reduction Step

In our experiment, we reduce the feature set before applying our personality prediction method. Initially for a person, we have 70 features from his Twitter and 70 features from his Disqus profile. Our feature reduction method is applied independently for each trait. Number of features after reduction step for each personality trait is given in Table 4.2.

Table 4.2: Number of features after feature reduction for each trait

Platform	Features	Reduced features				
		Ne.	Ex.	Op.	Ag.	Co.
Twitter	70	23	17	22	8	20
Disqus	70	22	18	27	11	24

Table 4.3: Accuracy (in %) of our method for each trait with and without applying feature reduction

	Ne.	Ex.	Op.	Ag.	Co.
All features	82.2	30.1	50.6	44.7	48.7
Reduced features	93.2	73.5	90.8	61.2	82.4

Accuracy of our method with and without applying feature reduction is shown in Table 4.3. Our experiment exhibits that extraversion, openness and conscientiousness traits cannot be identified correctly for most of the persons without

applying feature reduction method. Especially when the training dataset is small, excluding feature reduction step may result in very poor accuracy.

4.3.2 Comparison of Personality Traits between Twitter and Disqus

In this section we compare a person's personality traits identified from Twitter and Disqus. Table 4.4 shows the average of inferred big five traits and standard deviation of inferred big five traits obtained from both networks. We also compare inferred big five trait scores obtained from both networks at different threshold level (α). When the absolute difference of the normalized trait values of a person in two platforms is less than α , the two trait values are considered equal. Otherwise, the person's normalized trait value of one platform is considered stronger than the other platform when the value is greater than the other platform. Table 4.5 depicts the percentage of people having stronger trait values at different α levels.

As the way of expression of thoughts and opinion differs from Disqus to Twitter, LIWC scores obtained from these two social networks have varieties in some contents. For example, *Negative emotion, Anxiety, Anger, Sadness, Discrepancy, Swear Words* - are associated to neuroticism [36]. Score of these features are found higher in Disqus. In Disqus, discussion often turns into argument through comments and replies. People of high level of neuroticism express emotions during arguments and hence Disqus reveals neuroticism more evidently than Twitter. In our study, 77.14% users express stronger neuroticism (at $\alpha = 0.01$) in Disqus than Twitter.

Disqus being attached to millions of websites and online forums, provides users the opportunity to wander around varieties of articles and demonstrate their broad interest and novelty seeking attitude. We find that scores of *Cognitive Process, Social Process, Inclusive, Work and Space*, etc. LIWC features are higher in Disqus than Twitter. As these features are associated to openness, in Disqus people show more openness than Twitter through interacting with others more

Table 4.4: Average and standard deviation of scores of inferred big five traits on a normalized 0-1 scale.

		Neuro.	Extra.	Open.	Agree.	Cons.
Avg.	Twitter	0.39	0.62	0.47	0.46	0.32
	Disqus	0.55	0.42	0.67	0.41	0.45
	Difference	0.29	0.15	0.26	0.16	0.27
Stdev.	Twitter	0.18	0.21	0.19	0.13	0.17
	Disqus	0.23	0.16	0.23	0.17	0.15
	Difference	0.21	0.14	0.19	0.19	0.23

Table 4.5: Percentage of people having stronger big five traits in Twitter and Disqus on a normalized 0-1 scale

Threshold (α)		Neuro.	Extra.	Open.	Agree.	Cons.
0.01	Disqus	77.1429	33.5714	90.4762	54.2857	89.5238
	Twitter	20.001	64.9048	6.66667	40.003	6.66667
0.03	Disqus	73.3333	27.8095	88.5714	45.7143	88.5714
	Twitter	16.1905	59.1905	6.66667	36.1905	6.66667
0.05	Disqus	62.8571	22.2857	77.1429	40.003	84.7619
	Twitter	12.381	55.381	3.80952	31.4286	6.66667
0.07	Disqus	51.4286	18.5714	71.4286	32.381	80.9524
	Twitter	7.61905	48.6667	3.80952	23.8095	4.7619
0.10	Disqus	40.9524	12.8095	50.4762	22.8571	68.5714
	Twitter	5.71429	41.1429	2.85714	20.9524	3.80952

thoughtfully and perceptually. In our study, 90.48% users have distinguishably higher openness score (at $\alpha = 0.01$) in Disqus.

People who have high level of conscientiousness tend to discuss on human rights, crime, law and justice. Due to character limitation, Twitter posts reveal less information than Disqus posts on conscientiousness. Deliberative comments on political and social issues result in high score in *Achievement*, *Humans*, *Perceptual Processes*, *Tentative*, etc. LIWC features in Disqus. In our study, around 89.52% people have higher score in conscientiousness (at $\alpha = 0.01$) in Disqus.

In Disqus people interact with users who are not necessarily acquainted to

them. On the other hand in Twitter people generally follow familiar people. This influences people to express more friendliness in Twitter. LIWC score of *Family*, *Friend*, *Humans*, *Positive emotions* are higher in Twitter, which means extraversion trait of human personality is more identifiable in Twitter than Disqus. In our analysis, 64.9% user show stronger Extraversion (at $\alpha = 0.01$) in Twitter than Disqus. However, in both networks Agreeableness trait is consistent. People who are sympathetic, cooperative and kind, express similar attitude in both networks. Almost 54.29% users have analogous agreeableness values (at $\alpha = 0.01$) in Disqus and Twitter.

4.4 Experimental Results and Analysis of Identifying Personality from Topic Profile

We develop a prediction model to identify a person's personality from his topic profile. To evaluate accuracy of our model, we also develop a method to identify topic profile from a person's Twitter and Disqus usage. In Section 4.4.1 we analyze the differences of a person's topic profile in Twitter and Disqus. Then in the following sections we analyze the prediction classifiers and feature extraction module of our prediction model.

4.4.1 Analysis on Topic Profile

Twitter limits posts to 140 character. On the other hand Disqus posts have no character restriction. This helps Disqus to reveal more information on users' topic profile. To justify this proposition we have performed topic profile analysis on Disqus posts and tweets.

Although people discuss on different topics in these two platforms, common topics of discussion in both social network platforms are fewer in number. More than 80% of people post on less than 35% of their topics of interest in both networks. Frequency of posts on these topics are generally high. Beside these topics,

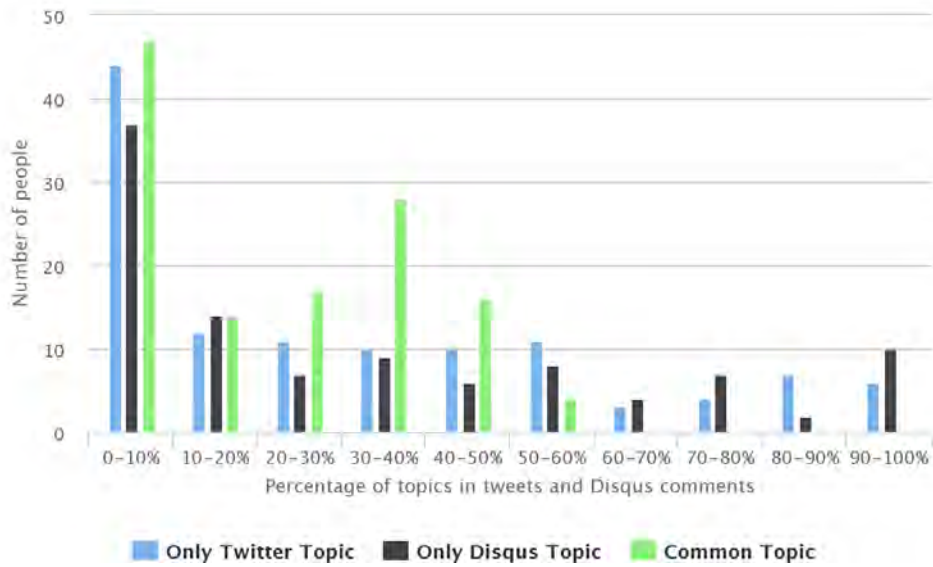


Figure 4.2: Analysis of topic coverage of Twitter and Disqus

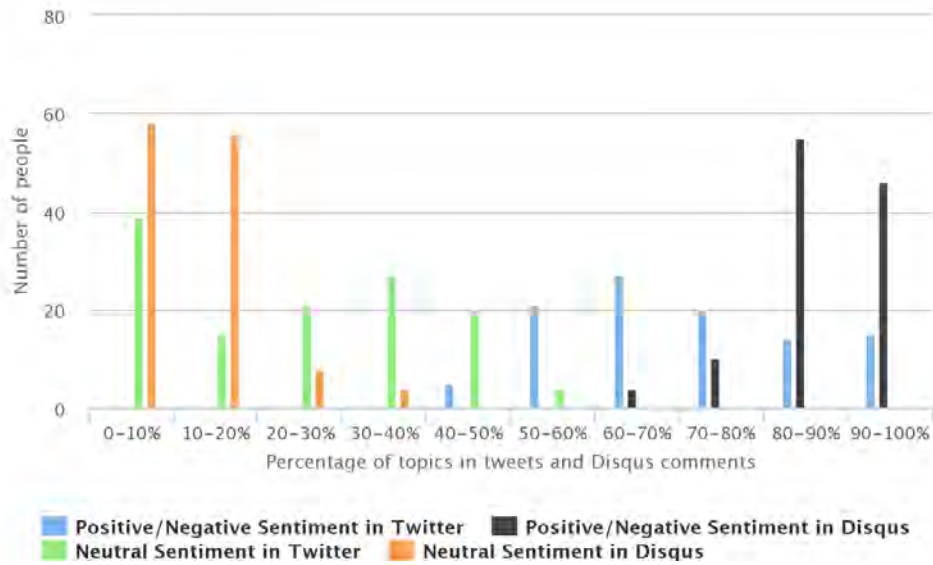


Figure 4.3: Analysis of topic sentiment of Twitter and Disqus

people share their views on other topics exclusively in Twitter or Disqus. 25% people tweets on 60-90% different topics that are not used in Disqus as a topic of

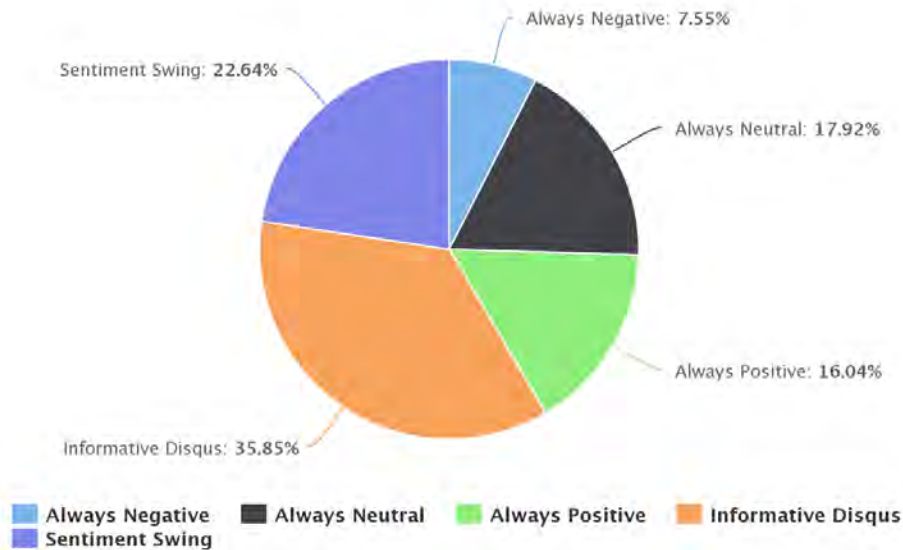


Figure 4.4: Analysis of sentiment over Disqus and Twitter

discussion. In Disqus this phenomenon also prevails. 18% users post and comment on 60-90% different topics which are not tweeted in twitter. Figure 4.2 depicts peoples' topic coverage pattern in Twitter and Disqus and Figure 4.3 depicts sentiment pattern on these topics.

In sentiment analysis, Disqus offers comprehensive result than twitter. As tweets are of limited character, it is often hard to get strong result on sentiment over a topic from Twitter. On the other hand Disqus comments being more descriptive, provides better result on a persons' sentiment over different topics. In our study, Disqus posts provide clear polarity of sentiment towards topics shared through posts and comments in more than 90% cases. On the contrary, a person's twitter posts on same topic show neutral sentiment in most of the cases. For a particular topic, it is easier to extract emotion of an user from his Disqus account. Figure 4.4 depicts sentiment on topics shared by users in Twitter and Disqus.

4.4.2 Analysis on Prediction Classifier

In this section we evaluate the accuracy of different classification models and different feature extraction methods for our dataset. All accuracy is averaged using 10-fold cross validation.

ZeroR, Linear and Non-linear Regression

ZeroR is the simplest classification method which predicts the majority occurring class for all instances. This is a very naive approach and we calculate zeroR only for determining baseline accuracy. Afterwards we apply linear regression for predicting personality from our modified feature set. Linear regression is the most frequently used methods for predicting big five personality traits. However, linear regression does not offer any significant improvement for predicting personality from topic profile. Accuracy differs for five different traits and on average, linear regression is able to predict personality with only 41.2% accuracy. Some other non-linear functions also do not fit well and cannot predict with significant accuracy. We state the accuracy of zeroR, linear regression and non-linear regression (square root fit) in Table 4.6.

Table 4.6: Accuracy of zeroR, linear regression and non-linear regression

Pers. Traits	ZeroR	Linear regression	Non-linear regression
Neu.	31.2	69.1	61.3
Ext.	37.3	43.0	40.1
Opn.	29.8	23.7	30.1
Agg.	33.2	30.1	29.1
Con.	30.3	40.1	33.1
Avg.	32.4	41.2	38.7

From our experimental results, we identify that linear regression can only be used to predict neuroticism. Neuroticism is related to people’s showing emotions during arguments on certain topics. So sentiment level in some topics is strongly

correlated to a person's neuroticism values. This is why linear regression shows high level of accuracy than baseline at predicting neuroticism. This phenomenon is not true for other four traits and hence the accuracy is low for predicting other four traits. Prediction model using square root function (non-linear regression) yields slightly poor result to linear regression.

Decision Tree Based C4.5

Classification based on C4.5 algorithm can predict neuroticism and openness with notable accuracy. C4.5 generates decision tree based on statistical features to classify a dataset. As discussed previously, a person's neuroticism is directly related to his sentiment values of certain topics. As a result C4.5 algorithm works fairly well for predicting neuroticism trait. In case of openness, a person's openness trait can be evaluated by utilizing his sentiment range on various topics. For instance, a person with high negative value in controversial topics usually represents very low openness value, whereas a person with high negative value on some controversial topics and neutral value on some controversial topics usually represents mild openness value. As C4.5 algorithm is a statistical classifier based on decision tree, it works fairly well for predicting openness. For predicting a person's other three trait values, C4.5 algorithm offers no significant improvement over baseline accuracy.

Naive Bayesian Classifier

We apply naive bayesian classifier assuming normality and modeling each conditional distribution with a single Gaussian. Naive bayesian classifier can predict neuroticism with 65.1% and conscientiousness with 73.6% accuracy. A person's conscientiousness is directly related to the presence of some topics. For example, very positive sentiment in "world hunger" usually indicates that the person has high conscientiousness value. Naive bayesian classifier predicts a person's traits by probabilistic calculation on the sentiment of selected topics. As a result it reveals a person's neuroticism and conscientiousness with good accuracy. However,

Table 4.7: Accuracy of C4.5, naive bayesian and logistic regression

Pers. Traits	C4.5	Naive bayes.	Logistic regr.
Neu.	31.2	69.1	61.3
Ext.	37.3	43.0	40.1
Opn.	29.8	23.7	30.1
Agg.	33.2	30.1	29.1
Con.	30.3	40.1	33.1
Avg.	32.4	41.2	38.7

for predicting other traits, Naive bayesian classifier is not very effective.

Logistic Regression

We identify that for predicting personality from topic profile, classifiers based on logistic regression remains inferior to support vector machine based classifiers. Logistic Regression cannot identify complex boundary. As a result, it cannot identify personality traits with significant accuracy. We explain the reasoning of complex boundary further in the following section. In Table 4.7, we present accuracy of C4.5 classifier, naive bayesian classifier and logistic regression based classifier for each trait.

Our Prediction Classifier

Among all the classifier methods we experiment, our support vector machine based classifier yields most accurate result for predicting personality of different traits. We find that support vector machine is a powerful predictor for predicting neuroticism, extraversion, openness and conscientiousness. The effectiveness of support vector machine can be explained by our “complex boundary” assumption which we introduced in Section 3.5.2. In summary, for a particular trait, people with similar values show sentiment on certain topics. Sentiment values of different topics create a complex boundary inside which boundary, people have similar

Table 4.8: Accuracy of support vector machine with different parameters

	Support Vector Machine			
Per. Tra.	Lin. Kernel.	RBF $\sigma(0.01)$	RBF $\sigma(0.03)$	RBF $\sigma(0.05)$
Ne.	78.4	78.4	79.2	79.0
Ex.	51.1	72.1	80.2	75.7
Op.	42.4	62.6	69.2	66.8
Ag.	43.2	44.1	41.1	44.0
Co.	56.4	73.0	77.5	76.1

personality trait values. The goal of support vector machine is to increase prediction accuracy and as well as to increase the margin between the boundary. For this reason, support vector machine exhibits most accuracy for predicting personality traits.

We implement sequential minimal optimization for training a support vector machine. We apply support vector machine with linear kernel and Gaussian kernel with three different values of σ and state our result in Table 4.8. We identify that Gaussian kernel performs better because complex boundary is not linearly separable using hyperplane. When the value of σ is very low, effects of using Gaussian kernel is very little and predicted boundary is similar to boundary predicted by linear Kernel. When we use very high sigma value, it smooths out the complex boundary and data points that are close to boundary are predicted erroneously. From our experiments, we identify that $\sigma = 0.03$ exhibits the highest accuracy in our dataset for personality prediction.

Summary of Classifiers

We have analyzed the accuracy of different prediction classifiers in the previous section. In this section we compare between different classifiers and identify which classifiers can be utilized for each trait. Table 4.9 shows the effective classifiers and the highest prediction accuracy achieved for each of the five different personality traits.

Table 4.9: Effective classifiers for each personality trait and the highest accuracy achieved

Personality Traits	All Effective Classifiers	Highest Accuracy
Neuroticism	Linear regression, C4.5, Naive bayesian, SVM	79.2%(SVM)
Extraversion	SVM	80.2%(SVM)
Openness	C4.5, SVM	81.1%(C4.5)
Agreeableness	None	44.1% (SVM)
Conscientiousness	Naive bayesian, SVM	77.5%(SVM)

Among the five traits of personality, neuroticism is the easiest trait to predict from topic profile. A person's neuroticism is directly related to his expressing emotions and sentiments. By analyzing sentiment values on certain topics, neuroticism value can be easily predicted. For this reason linear regression, C4.5 algorithm, naive bayesian classifier and support vector machine -all perform well for predicting neuroticism. Among all the classifiers we applied, support vector machine with Gaussian kernel ($\sigma = 0.03$) achieved the best accuracy for neuroticism.

Predicting extraversion, openness and conscientiousness is not as straightforward as predicting neuroticism. We achieved significant accuracy with support vector machine based classifier. In Section 4.4.2 we explained our "complex boundary" assumption to give insight into the accuracy of support vector machine based classifier. Apart from support vector machine, we identify that C4.5 decision tree based classifier is effective for predicting openness and naive bayesian classifier is effective for predicting conscientiousness.

No classifier can predict agreeableness with accuracy more than 44.1%. Hughes *et al.* argued that agreeableness is not revealed in an online platform [19]. It is also reasonable that agreeableness trait cannot be predicted accurately from topic profile.

4.4.3 Analysis on Feature Extraction

In the previous section we have analyzed the accuracy of different classifier models. In this section we analyze the impact of our feature extraction step. We have applied some methods to extract features from topic profile. As we discussed in Section 3.5.1, we have applied topic removal, topic clustering and principal component analysis to our dataset. Now we analyze the impact of applying these methods.

Analysis on Topic Removal

Table 4.10: Accuracy of prediction model with and without topic removal

Personality Traits	Without Topic removal	With Topic removal
Neu.	65.2	79.2
Ext.	71.1	80.2
Opn.	47.8	69.2
Agg.	41.0	41.1
Con.	61.0	77.5
Avg.	57.2	69.4

In the topic removal phase, we remove all the topics which are present only for a very few number of persons. This phase is straightforward and offers significant improvement to our prediction model. We evaluate accuracy of a support vector machine with a Gaussian kernel with and without applying topic removal phase. All other steps remain unchanged. Experimental result is tabulated in Table 4.10. Without topic removal phase, feature set contains a lot of topics which are absent for most of the persons. This increases number of missing features which result in low accuracy.

Analysis on Topic Clustering

In the second phase of our feature extraction step, we cluster similar topics together. By clustering similar topics together, we decrease the number of missing attributes in a person's topic profile. Without clustering, our dataset contains 662 unique topics where every person is not active in all 662 topics. Even after topic removing phase, we have 240 unique topics. As a result, a large number of topics' sentiment is missing for a person which reduces the accuracy of prediction classifier. Our experiment exhibits another major advantage of topic clustering. After topic clustering, for each person we identify his sentiment on a topic cluster by averaging his sentiment of all topics inside the topic cluster. This makes our prediction model resistant to random errors which occur during sentiment extraction phase. As we are averaging the sentiment of some topics, effects of few incorrect sentiments are averaged out. In this section we analyze our experimental result to compare our clustering method with another clustering method.

Our goal is to cluster similar topics together. Centroid based clustering such as k -nearest neighbor (k -NN) is the most frequently used algorithm for cluster analysis [4]. This type of clustering method generally clusters based on some predetermined numbers of clusters. So in this method, if two topics are similar they may still belong to two different clusters if there are many other topics within close range. Another issue with this algorithms is that two very different topics may belong to same cluster if there are few topics within the range of these two topics [5]. These two cases are contrary to what we want to achieve by applying clustering method to our dataset. For these reasons, k -NN does not perform well in our dataset. Highest accuracy of k -NN clustering algorithm along with their k -value for different personality traits is shown in Table 4.11. For evaluating accuracy, we applied support vector machine with Gaussian kernel as our prediction model. Figure 4.5 shows highest accuracy achieved by k -NN and our distance based clustering algorithm side by side.

For each personality trait, we apply our clustering method with different values of distance threshold (d) and calculate prediction accuracy of support vector

Table 4.11: Comparison of accuracy of k -NN and our distance based clustering

Pers. Traits	value of k	k -NN clustering	Our distance based clustering
Neu.	6	68.1%	79.2%
Ext.	4	56.1%	80.2%
Opn.	9	50.8%	69.2%
Agg.	3	40.4%	41.1%
Con.	6	61.2%	77.5%

machine with Gaussian kernel which is enlisted in Table 4.12. Same data is represented as a two dimensional line in Figure 4.6 and Figure 4.7. In these figures, X-axis represents d value, Y-axis represents the accuracy of our prediction model and each line corresponds to one of five personality traits. We find that $d = 0.05$ is the most optimum value in our experiment where prediction model achieves the highest accuracy. In the following paragraph, we explain the reason for this optimality.

When the value of d is reduced, topic clusters reduce and a topic cluster contains only few number of topics. Major disadvantage of this arises because a person is not active in all the topics. As a result, for a large number of people, we cannot identify their emotion on many topic clusters. These missing attributes result in lower accuracy. We can observe this phenomenon in the figures where accuracy is increasing with increasing d up to $d = 0.05$. After this optimum value, accuracy of our prediction model again decreases which is evident from the figures as the lines starts to contort downwards. With greater values of d , number of topic clusters is decreased which results in less number of features for our prediction model. This decreases the accuracy of prediction. Additionally, with higher values of d , in some cases two or more completely different topics belong to same topic cluster. This also reduces prediction accuracy of our model. For these two conflicting reasons, prediction model achieves highest accuracy near $d = 0.05$ for all personality traits. However, we also identify that the curves of figures are not similar for all the personality traits. In the following section, we identify and

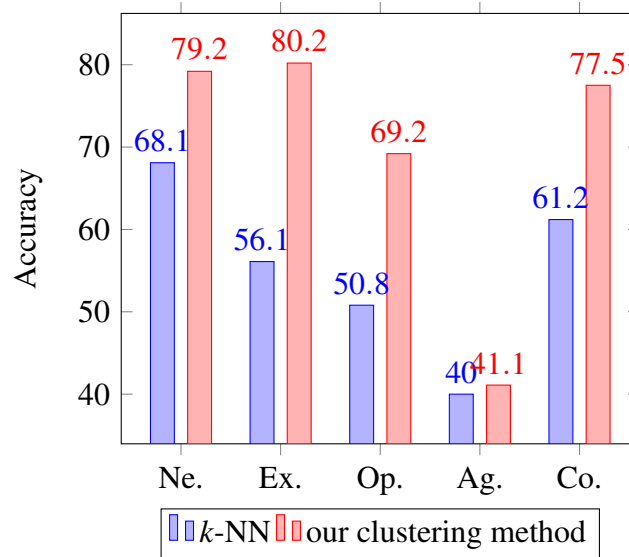


Figure 4.5: Highest accuracy of *k*-NN and our clustering method

analyze the differences between personality traits.

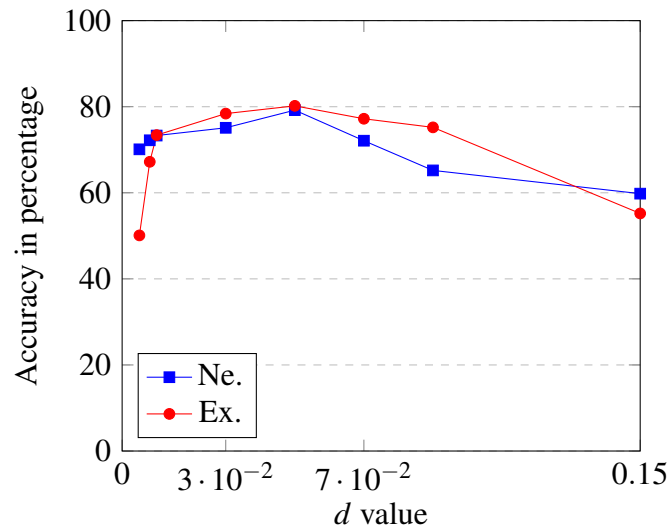
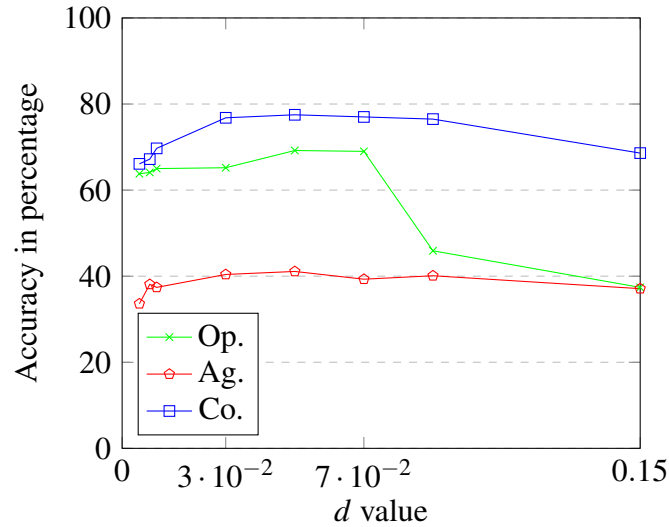


Figure 4.6: Accuracy of predicting neuroticism and extraversion for different *d* values

Neuroticism: Neuroticism trait has the lowest impact on accuracy with chang-

Table 4.12: Accuracy (in %) for different values of d

d	Ne.	Ex.	Op.	Ag.	Co.
0.005	70.1	50.1	63.8	33.6	66.1
0.008	72.2	67.2	64.1	38.1	67.2
0.01	73.3	73.4	65.0	37.4	69.7
0.03	75.1	78.4	65.2	40.4	76.8
0.05	79.2	80.2	69.2	41.1	77.5
0.07	72.1	77.2	69.0	39.3	77.0
0.09	65.2	75.2	45.9	40.1	76.5
0.15	59.8	55.2	37.4	37.1	68.6

Figure 4.7: Accuracy of predicting openness, agreeableness and conscientiousness for different d values

ing values of d . This can be explained with the fact that neuroticism is easily revealed in online usage. When the values of d is smaller, there are many missing attributes in the prediction model. However, the remaining attributes can achieve a significant accuracy in case of neuroticism. When the values of d is higher, number of topic clusters is fewer. However, in case of neuroticism, fewer topic clusters can achieve a significant accuracy.

Extraversion: Value of d has a very high impact on the accuracy of identifying extraversion. Extraversion is affected both by missing attributes and fewer number of topic clusters. So accuracy of predicting extraversion is low when d is less than the optimum value (0.05) as well as when d is greater than the optimum values. For this reason, we observe a steep curve downwards before and after 0.05 in extraversion trait's curve.

Openness: When the value of d is less than the optimum level (0.05), openness trait can still be predicted with high accuracy. Even though number of missing attributes is higher when d is low, remaining attributes can reveal a person's openness. We identify that some frequently occurring topics (i.e., politics) are present in most of the persons which can reveal a person's openness even when other attributes are missing. On the other hand, when d is higher than the optimum level, accuracy of prediction model starts to fall drastically. When d is too high, a topic cluster may contain two distant topics. When we average sentiment levels of all the topics inside a topic cluster, we may lose vital information. For this reason, accuracy of predicting openness is decreased with the increasing value of d after the optimum value (0.05).

Agreeableness: We identify that the value of d has very little effect on the accuracy of predicting agreeableness. As we identified earlier, agreeableness is not usually revealed in online usage. Accuracy of predicting agreeableness remains within the range 30-40% which is not very significant. With different values of d , accuracy does not change much and remains within this mentioned range.

Conscientiousness: For predicting conscientiousness, smaller d values show less accuracy. The reason is previously explained as smaller d means higher missing attributes. However, conscientiousness can be predicted using small number of topic clusters. As a result, for values of d greater than the optimum value, our prediction model still performs with significant accuracy for predicting conscientiousness.

Our analysis on topic clustering shows that our distance based clustering method performs better than the most widely used clustering method k -NN for our clus-

tering purposes. We also identify an optimum value of d for our prediction model. In the following section, we analyze the effects of different principal components.

Analysis on Principal Components

We apply Principal Component Analysis (PCA) to our feature set after the topic removal and topic clustering steps. We identify the accuracy of our prediction model on varying principal component numbers which is shown in Table 4.13. We plot the graph of accuracy against principal components in Figure 4.8.

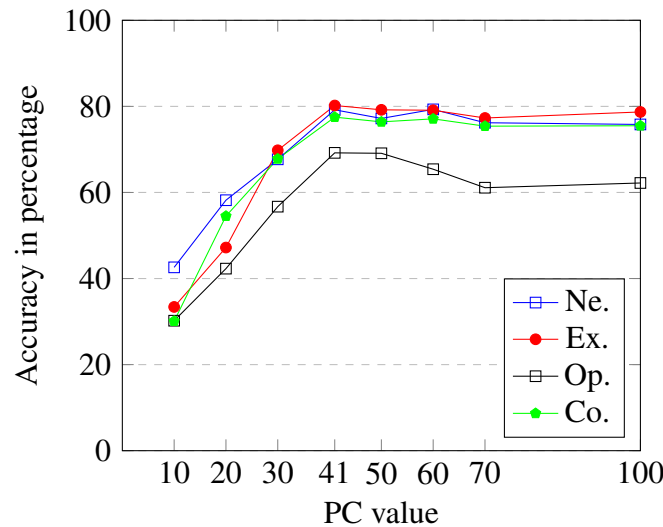


Figure 4.8: Accuracy of prediction for different values of PC

We identify that around 41 principal components are sufficient for achieving high prediction accuracy. This step of identifying principal components is important because of scalability reasons. We want our prediction model to work in real life scenario where number of unique topic clusters may be very high. In those cases, prediction model based on huge number of inputs may not be feasible. We identify that our prediction model works accurately with limited number of principal components. This step makes our prediction model scalable in real life scenario.

Table 4.13: Accuracy (in %) for different values of Principal Components (PC)

PC	Ne.	Ex.	Op.	Ag.	Co.
10	42.6	33.4	30.2	29.1	30.1
20	58.2	47.2	42.3	38.9	54.5
30	67.7	69.8	56.7	41.0	67.8
41	79.2	80.2	69.2	41.1	77.5
50	77.2	79.2	69.1	40.9	76.4
60	78.3	79.1	65.4	33.2	77.1
70	76.2	77.3	61.1	34.2	75.4
100	75.8	78.7	62.2	31.7	75.5

4.5 Result Summary

We have implemented our method to identify a person's comprehensive personality profile from his Twitter and Disqus usage and our prediction model to predict user personality from his topic profile. Key findings of our experiments are listed below:

- Our experiment with Twitter and Disqus proves that for a person, personality is revealed differently in different online platforms. We find that 60.9% persons have stronger extraversion trait in Twitter. We also find that 90.5% persons' openness trait and 89.5% persons' conscientiousness trait is stronger in Disqus.
- Our method identifies personality more accurately because different online platform reveals different aspects of human behaviour and combining them creates a more accurate profile of a person. As a result, our method identifies openness with 22.1% higher accuracy and conscientiousness with 20.1% higher accuracy than Golbeck's method.
- Methods to identify personality from social media sites is not applicable for comment posting platform because usage pattern in these two different types of media platforms is different. There is a linear correlation between

a person's psycholinguistic scores of Twitter posts and personality trait. On the contrary, in case of Disqus posts, this relationship is non-linear and can be best described using a square root function.

- Our non-linear regression based method can identify personality trait from comment posting platform (Disqus) with significant accuracy. Our method can identify neuroticism, openness and conscientiousness of a person from his Disqus posts with 77.14%, 90.48% and 89.52% accuracy. The other two personality traits (i.e., extraversion and agreeableness) are not revealed in a person's Disqus profile.
- Analysis of topic profile between Twitter and Disqus reveals that a person's active topic of discussion is different in these two platforms. More than 80% of people posts on less than 35% of their topic of interest in both platforms. Our experiment supports the assumption that online media platform has influence on a person's discussion topics.
- Our prediction model identifies personality of a person from his topic profile with significant accuracy. It predicts personality traits (except agreeableness) with 76.5% accuracy on average. So it is possible to predict personality of a person by analyzing his topic profile only.
- Our proposed support vector machine based prediction model achieves high prediction accuracy because our extracted features create a complex boundary. For this reason, our proposed model can predict neuroticism, extraversion, openness and conscientiousness with 79.2%, 80.2%, 69.2% and 77.5% accuracy respectively.
- Utilizing the topic profile directly increases a lot of missing features and utilizing only selected topic profile cannot predict personality accurately. We need a combination of intelligent techniques that include clustering, principal component analysis, etc. to extract meaningful features from topic

profile. From our experiment, we identify that applying our clustering algorithm with distance threshold (d) equal to 0.05 and identifying 41 principal components yield the most accurate prediction.

- Identifying all personality traits are not equally difficult. A person's neuroticism can be easily identified from online usage. Linear regression, decision tree based classifier, naive bayesian and our proposed model -all classification methods can predict neuroticism with more than 65% accuracy. On the contrary, a person's agreeableness is not revealed in online usage. As a result, no prediction model can achieve more than 42% accuracy while predicting agreeableness.

In summary, we can conclude from our experiment that our model identifies personality more accurately than existing methods. Our model to predict personality from topic profile is a novel approach to identify personality from topic related data. Accuracy of our prediction model makes us confident that it can be implemented in many real life scenarios.

Chapter 5

Conclusion and Future Work

In this thesis we present a new model to identify comprehensive personality profile of a user from his multiple online media usage. Our thesis can be broadly divided into two parts: in first part we identify personality from posts and in second part we identify personality from topic related data. To achieve our objective, in each part we develop new techniques and approaches. In the first part of our thesis, we propose a new method to identify personality from comment posting platform and a novel approach to combine different types of online media contents. Our experimental results clearly indicate that this model outperforms existing methods of identifying personality. In the last part of our thesis, we develop a prediction model to predict personality from topic related data. Predicting personality from such dataset is not addressed in previous literatures. We develop this model by improving some existing methods as well as proposing some novel techniques.

Our experiment on a person's Disqus and Twitter profile reveals some interesting properties. Our study reveals that online platform has influence on a person's topics of discussion in which he is interested in. Different aspect of personality is revealed in different media platforms. In future, such findings can be tested across other online platforms. We identify that for a large number of topics, a person's Twitter posts contain mostly neutral emotion whereas his Disqus posts reveal polarity (positive or negative sentiment). By extending these works, it might be

possible to predict how a person would react to a topic in general. In second part of our thesis, we develop a model to identify personality from topic profile. Our success in developing such model indicates that prediction model can also be proposed to identify personality from other types of data (for instance, search query and purchasing history).

This study is a comprehensive effort on analyzing a person's usage in Twitter and Disqus. We compare 105 person's Disqus and Twitter usage and find that some traits are better revealed in one platform. So using the two platforms, we can predict personality of a person more accurately. Moreover, through topic and sentiment analysis we prove that a person's activities is highly influenced by online platforms and single platform analysis is not sufficient to predict a person's sentiment over different topics. Our prediction model to identify personality from topic profile establishes that a person's personality is directly related to his topics and sentiment. It proves that personality can be predicted from a more restricted media usage dataset. Our experiment in two different types of online platforms makes us confident that similar approaches can be made to other social platforms. This can lead to future work in developing an individual's comprehensive virtual profile to predict any personal behaviour.

Bibliography

- [1] Saima Aman and Stan Szpakowicz. Identifying expressions of emotion in text. In *Lecture Notes in Computer Science*. Springer Publication, 2007.
- [2] Yair Amichai-Hamburger and Elisheva Ben-Artzi. Loneliness and internet use. *Computers in Human Behavior*, 19(1):71–80, 2003.
- [3] Yair Amichai-Hamburger and Gideon Vinitzky. Social network use and personality. *Computers in Human Behavior*, 26(6):1289–1295, 2010.
- [4] Kevin Beyer, Jonathan Goldstein, Raghu Ramakrishnan, and Uri Shaft. When is nearest neighbor meaningful? In *Database TheoryICDT99*, pages 217–235. Springer, 1999.
- [5] Nitin Bhatia. Survey of nearest neighbor techniques. *arXiv preprint arXiv:1007.0085*, 2010.
- [6] Sarah Butt and James Philips. Personality and self reported mobile phone use. *Computers in Human Behavior*, 24(2):346–360, 2008.
- [7] M. De Choudhury and S. Counts. The nature of emotional expression in social media: Measurement, inference, and utility. In *Human Computer Interaction Consortium (HCIC)*, 2012.
- [8] Gobinda Chowdhury. Natural language processing. *Information science and technology*, 37(1):51–89, 2003.

- [9] Teresa Correa, Amber Willard Hinsley, and Homero Gil De Zuniga. Who interacts on the web?: The intersection of users personality and social media use. *Computers in Human Behavior*, 26(2):247–253, 2010.
- [10] P Costa and R McCrae. The five factor model of personality: theoretical perspectives. In *Human Behavior*, pages 51–87, 1996.
- [11] Paul Costa and Robert McCrae. Validation of the five-factor model of personality across instruments and observers. *Journal of personality and social psychology*, 52(1):81–85, 1987.
- [12] Paul Costa and Robert McCrae. Four ways five factors are basic. *Personality and individual differences*, 13(6):635–665, 1992.
- [13] Paul Costa and Robert McCrae. A contemplated revision of the neo five-factor inventory. *Personality and individual differences*, 36(3):587–596, 2004.
- [14] Michael Cyger. Infographic: Todays social media user has multiple accounts, 2013.
- [15] Daedalus. Textalytics text analytics and semantic processing, 2013.
- [16] Shaun W. Davenport, Shawn M. Bergman, Jacqueline Z. Bergman, and Matthew E. Fearington. Twitter versus facebook: Exploring the role of narcissism in the motives and usage of different social media platforms. *Computers in Human Behavior*, 32:212–220, 2013.
- [17] J. Golbeck, C. Robles, M. Edmondson, and K. Turner. Predicting personality from twitter. In *IEEE Third International Conference on Social Computing (SocialCom)*, pages 149–156, 2011.
- [18] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian Witten. The weka data mining software, 2015.

- [19] David John Hughes, Moss Rowea, Mark Bateya, and Andrew Leea. A tale of two sites: Twitter vs. facebook and the personality predictors of social media usage. *Computers in Human Behavior*, 28:561–569, 2012.
- [20] M. Francis J. Pennebaker and R. Booth. Linguistic inquiry and word count: Liwc 2001. In *Mahway Lawrence Erlbaum Associates*, 2001.
- [21] Bryan Kennedy and Ashley Kennedy. Using the myers-briggs type indicator in career counseling. *Journal of Employment Counseling*, 41(1):38–43, 2004.
- [22] Chin-Yew Lin and Eduard Hovy. The automated acquisition of topic signatures for text summarization. In *Conference on Computational linguistics*, pages 495–501, 2000.
- [23] F. Mairesse, M. Walker, M. Mehl, and R. Moore. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of Artificial Intelligence Research*, 30(1):457–500, 2007.
- [24] Hasan Al Maruf, Jalal Mahmud, and Mohammed Eunus Ali. Can hashtags bear the testimony of personality? predicting personality from hashtag use. In *SocialCom*, 2014.
- [25] James Mcelroy, Anthony Hendrickson, Anthony Townsend, and Samuel De-Marie. Dispositional factors in internet use: personality versus cognitive style. In *MIS quarterly*, pages 809–820, 2007.
- [26] C. Nass and K. M. Lee. Does computer-generated speech manifest personality? an experimental test of similarity-attraction. In *the SIGCHI conference on Human factors in computing systems*, page 329336, 2000.
- [27] Giuseppe Pilia, Wei-Min Chen, Angelo Scuteri, Marco Orru, Giuseppe Albai, Mariano Dei, and Sandra Lai. Heritability of cardiovascular and personality traits in 6,148 sardinians. *PLoS Genet*, 2(8):132–135, 2006.

- [28] D. Quercia, M. Kosinski, D. Stillwell, and J. Crowcroft. Our twitter profiles, our selves: Predicting personality with twitter. In *IEEE SocialCom*, 2011.
- [29] P. Rosen and D. Kluemper. The impact of the big five personality: Traits on the acceptance of social networking website. In *AMCIS*, 2008.
- [30] Craig Ross, Emily Orr, Susic Mia, James Arseneault, Mary Simmering, and Rober Orr. Personality and motivations associated with facebook use. *Computers in Human Behavior*, 25(2):578–586, 2009.
- [31] Tracii Ryan and Sophia Xenos. Who uses facebook? an investigation into the relationship between the big five, shyness, narcissism, loneliness, and facebook usage. *Computers in Human Behavior*, 27(5):1658–1664, 2011.
- [32] M. Stopfer, J. Vazire, S. S. Gaddis, S. Schmukle, B. Egloff, and S. Gosling. Facebook profiles reflect actual personality, not self-idealization. In *Psychological Science*, 2010.
- [33] David Watson and Auk Tellegen. Toward a consensual structure of mood. *Psychological Bulletin*, 98:219–235, 1985.
- [34] Jerry S. Wiggins. *The Five-factor Model of Personality: Theoretical Perspectives*. The Guilford Press, 1996.
- [35] Jiejun Xu, Tsai-Ching, Ryan Compton, and David Allen. Quantifying cross-platform engagement through large-scale user alignment. *Information Systems Management*, 31:225–239, 2014.
- [36] Tal Yarkoni. Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers. *Journal of Research in Personality*, 44:363–373, 2010.