INFERENCE OF PERSONALITY USING SOCIAL MEDIA PROFILES

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF INFORMATICS INSTITUTE
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

ÜMİT ATEŞ

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE
IN
THE DEPARTMENT OF INFORMATION SYSTEMS

JUNE 2014

INFERENCE OF PERSONALITY USING SOCIAL MEDIA PROFILES

Submitted by **Ümit ATEŞ** in partial fulfillment of the requirements for the degree of **Master of Science in Information Systems**, **Middle East Technical University** by,

Prof. Dr. Nazife Baykal _____
Director, **Informatics Institute**

Prof. Dr. Yasemin Yardımcı Çetin _____
Head of Department, **Information Systems**

Assist. Prof. Dr. Tuğba Taşkaya Temizel _____
Supervisor, **Information Systems, METU**

Examining Committee Members:

Assoc. Prof. Dr. Aysu Betin Can _____
Information Systems, METU

Assist. Prof. Dr. Tuğba Taşkaya Temizel _____
Information Systems, METU

Assist. Prof. Dr. Aybar Can Acar _____
Medical Informatics, METU

Assoc. Prof. Dr. Altan Koçyiğit _____
Information Systems, METU

Assoc. Prof. Dr. Ahmet Uysal _____
Psychology, METU

Date: **25/06/2014**

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and result that are not original to this work.

Name, Last Name : Ümit Ateş

Signature :

ABSTRACT

INFERENCE OF PERSONALITY USING SOCIAL MEDIA PROFILES

Ateş, Ümit

M.S., Department of Information Systems

Supervisor: Assist. Prof. Dr. Tuğba Taşkaya Temizel

June 2014, 104 pages

People have an inherent need to express themselves to other people in the community by sharing their experiences, ideas, activities, and memories. As a means, they mostly prefer to use social media such as Twitter, Facebook, personal blogs, and wikis. Many people consistently contribute to such social media platforms by writing their own experiences, sharing photos and status. The majority of shared content is personal information. There are studies in the literature which make use of shared social media content to predict users' Big 5 Personality Traits such as agreeableness, conscientiousness, extraversion, neuroticism and openness. These studies usually utilize linguistic features, social network information, and the frequency of their interaction with the platform such as number of posted status updates, photos, videos and likes. The aim of this thesis is to identify which features of the shared content in Facebook are correlated with users' Big 5 Personality Traits and develop a model

based on these features for personality prediction. The contribution of this thesis is twofold. First, we show that the existing solutions in predicting Big 5 Personality work better when there is sufficient evidence in terms of number of posts in their social media profile. Second, we show that the inclusion of information regarding users' friends such as their Big 5 Personality information improves the accuracy compared to other methods in the literature.

Keywords: Big 5 Personality Traits, Personality Prediction, Social Network, Friendship.

ÖZ

SOSYAL MEDYA PROFİLLERİ KULLANARAK
KULLANICI KİŞİLİK ÇIKARIMI

Ateş, Ümit

Yüksek Lisans, Bilişim Sistemleri Bölümü

Tez Yöneticisi: Yard. Doç. Dr. Tuğba Taşkaya Temizel

Haziran 2014, 104 sayfa

İnsanlar doğası gereği deneyimlerini, düşüncelerini, eylemlerini ve anılarını paylaşarak toplumdaki diğer bireylere kendilerini ifade etme ihtiyacı duyarlar. Bu eğilimlerini, çoğunlukla Twitter, Facebook, Blog ve Wiki sayfaları gibi sosyal medyayı araçlarını kullanarak gerçekleştirirler. Bir çok insan düzenli olarak bu tür sosyal medya ortamlarında kendi deneyimlerini, fotoğraflarını ve güncel durum bilgilerini yazarak paylaşımlarda bulunurlar. Paylaşılan içeriklerin önemli bir kısmı kişisel bilgi barındırmaktadır. Literatürde, sosyal medyada paylaşılan içerikleri kullanarak kullanıcıların Big 5 Kişilik Karakterleri'nin örneğin uyumluluk, sorumluluk, dışadönüklük, duygusal dengesizlik ve açıklık kestirimini yapan çalışmalar mevcuttur. Bu çalışmalar genellikle dilbilimsel özellikleri, kullanıcının sosyal ağ bilgilerini ve paylaşılan durum bilgisi, fotoğraf, video ve beğenilen içerik sayısı gibi kullanıcının sosyal medya platformunu kullanım alışkanlıklarını kullanarak yapılmıştır. Bu tezin amacı Facebook'ta paylaşılan içeriklerin hangi özelliklerinin kullanıcıların Big 5

Kişilik Karakterleri ile ilişkili olduklarını belirlemek ve belirlenen bu özellikleri kullanarak kullanıcıların kişilik kestirimini yapacak bir model geliştirmektir. Bu çalışmanın literatüre iki yönlü katkısı bulunmaktadır. Birincisi, eğer paylaşımı yapılan durum bilgisi sayısı anlamında yeteri kadar paylaşım varsa, kullanıcıların Big 5 kişilik kestirimini yapan var olan yöntemler daha iyi sonuçlar vererek çalışmaktadır. İkincisi ise, kişilik kestiriminde kullanılan özelliklere ek olarak kullanıcıların arkadaşlarının Big 5 kişilik karakterleri dikkate alındığında literatürde yer alan yöntemlere göre daha iyi sonuçlar alınmaktadır.

Anahtar Sözcükler: Big 5 Kişilik Karakterleri, Kişilik Kestirimi, Sosyal Ağlar, Facebook, Arkadaşlık.

To my dear family

ACKNOWLEDGMENTS

I would like to thank my supervisor Assist. Prof. Dr. Tuğba Taşkaya Temizel for her encouraging, advice and guidance in this thesis study.

I would like to thank my family, collogue and friends for their support, motivation and encouragement during the study.

I would also like to express my gratitude to the examining committee members Dr. Aybar Can Acar, Dr. Aysu Betin Can, Dr. Ahmet Uysal and Dr. Altan Koçyiğit for their valuable feedback.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

SNA	Social Network Analysis
LIWC	Linguistic Inquiry and Word Count
CCI	Correctly Classified Instances
ICI	Incorrectly Classified Instances
MAE	Mean Absolute Error
AuC	Area Under Curve
RMSE	Root Mean Square Error
SMO	Sequential Minimal Optimization

CHAPTER 1

INTRODUCTION

Internet usage has been significantly increased during last two decades. People have started to spend their time on web sites that anyone can edit and contribute to. Therefore, to fulfill this need, some web technologies, where users can interactively collaborate and contribute, are introduced. These technologies are Blogs, Wiki Pages, Portals and Social Networking Sites. These are introduced in the name of Web 2.0 Technologies. These technologies enable users to contribute and share content without requiring them to have any technical knowledge in web programming. By the help of these technologies, people are able to reach out to others with similar interests.

During the last decades, some social networking sites are introduced and have become highly popular in world wide. These networking sites are Facebook, Twitter, YouTube, LinkedIn, Instagram, Academia and etc. Each of them has different objectives to persuade people to share their experiences, ideas or moments of their life solicitously. Facebook provides users a communication network consisting of their friends, families and other people with whom they have acquaintance in their real social life. Twitter enables people to broadcast their ideas, instant critics to other people where they may probably know each other in real life. LinkedIn focuses on business life, and it provides a business networking platform to business people to communicate, follow each other and help their recruitment through enhanced searching facilities based on their professions.

These social networking sites affect our real life. Many people are well organized in such platforms. For instance Twitter has become an important alternative media to real media, it is faster to spread news and gives more freedom of speech. Even they may cause social movements as in Greece, Egypt, Turkey and lastly Ukraine.

These platforms have also attracted many researchers recently. Facebook is one of the platforms that academics are interested in because it has a large number of subscribers

worldwide and contains personal information. As reported in a research, written by Wilson et al [1], they remark that academics from different disciplines are studying on Facebook. These disciplines are ranging from law, economics, sociology, and psychology, to information technology, management, marketing, and computer science. According to them, researches about Facebook can be categorized in five categories: descriptive analysis of users with 24 % of total papers, motivations for using Facebook with 19 % of total papers, identity presentation with 12 % of total papers, role of Facebook in social interactions with 27 % of total papers, and finally privacy and information disclosure with 18 % of total papers. According to their study, between 2004 and 2008, the number of published articles about Facebook was low. But when Facebook became global in 2008 and 'Like' facility was introduced in 2009, the number of articles significantly increased [1].

There are studies and projects such as YouAreWhatYouLike [2], Five Labs [3] and research academies such as Facebook Data Science [3] working on personality inference using disclosed information in social media profiles. These researches are made to help decision driver in advertising a product, conducting a campaign, finding volunteer for social event and etc.

In this study, we aim to improve performance of existing models developed for inference of personality traits by incorporating homophily information. According to homophily, individuals tend to select person having similar characteristics in friendship. In other words, adjacent nodes in friendship network have similar interest and characteristics. Therefore, there is correlation between actions of individuals in such network [5]. In this study, we investigate whether we could use these similarities to predict individuals' personality more successfully.

1.1 Purpose of the Study

This thesis aims to investigate whether there is any relation between users' Big 5 personality traits and disclosed information in Facebook such as friendships, status updates, likes, photos and social network attributes. Several predictive models are constructed which make use of such features that have been found correlated with Big 5 Personality traits and their predictive performances are compared using myPersonality database [2 - 13]. We have also studied the effect of information volume on accurately inferring personality traits. In other words, we have showed to what extent the predictive performance of the prediction model can be improved by using different volumes of personal information. Finally, we have

explored whether there is a significant effect in the accuracy of the predictive models for personality traits when we have incorporated users' friends' personal information. The research questions we aim to answer in this thesis are;

1. What information disclosed in Facebook is correlated with users' Big 5 Personality traits?
2. How does the amount of personal information (i.e. status updates) affect the accuracy of the predictive performance of the models developed for inferring users' personality traits?
3. Is there a relation between users' and their friends' personality traits? Can a user's personality be inferred more accurately using his/her friends' personality information?
4. Is there a relation between users' and their spouses' personality traits? Can a user's personality be inferred more accurately using his/her spouse personality information?

1.2 Significance of the Study

In literature, there are studies to predict Big 5 Personality Traits using linguistic features that are extracted from written or speech text [14] [15]. But personality prediction on social media is quite popular and recent topic. The first well known research was conducted by Golbeck et al. in 2011 [16]. There are other studies that employ users' demographic information, status updates and likes in inference of personality traits [4] [14] [17]. This thesis will contribute to the existing literature on inference of personality traits domain using social media in two main ways. The first contribution is to show the effect of information volume on predictive performance of the models. The current studies give equal weights to both linguistic features and social network features while creating the models. However when the number of status updates is low, the performance of models may decrease as the extracted linguistic features on such limited information may give misleading and inaccurate information. So models should be constructed in this case by giving more importance to other features which have sufficient information such as social network information. The second main contribution is show whether information about users' spouse or friends can improve the predictive performance of the models.

1.3 Terms and Definitions

Big 5 Personality Traits: These are five main psychological traits that define individual's characteristics. These five traits are Agreeableness, Conscientiousness, Extraversion, Neuroticism, and Openness defined in Section 2.1.

Social Network: It represents relationships between individuals.

Social Network Analysis: Analyzing relationships between nodes in a social network.

Social Media: It is a platform on Web 2.0 technology where people can share, consume and exchange information between each other. So the dynamic content of web sites rely on its user updates.

Personality Prediction: Prediction of Big 5 Personality Traits of any individual by using certain attributes belonging to related individual.

CHAPTER 2

LITERATURE REVIEW

In this chapter, first, we present a brief explanation about Big 5 Personality Traits, their facet and Social Network Analysis (SNA). Most commonly used terms and measurements in SNA are described in detail. Second, we mention about existing studies, which focus on personality prediction in a chronological order.

2.1 Big 5 Personality Traits

In Psychology, there are five major characteristics known as "Big Five" that define human personality (Goldberg 1992). These characteristics are agreeableness, conscientiousness, extraversion, openness and neuroticism. These Big 5 Personality Traits can be evaluated by factor analysis of personality description questionnaires that have become a standard over the years. These personality traits are explained in detail below;

- **Agreeableness (social adaptability, likability, friendly compliance, agreeableness, and love)** [18]**:** These people are friendly, calm, peace keeper and optimistic. They easily trust others. They are nurturing people. That's why they always help others.
- **Conscientiousness (dependability, task interest, will to achieve, impulse control, and work)** [19]**:** These people are well organized therefore these people achieve most of the task they have. They care about their responsibilities. They tend to commit themselves to work that is assigned to these people. Moreover, they are hardworking and try to do their best. These people are extremely reliable.

- **Extraversion (active, energetic, assertive, forceful, outgoing, sociable, talkative, and adventurous)** [18]**:** These people are so energetic and attractive. They are outgoing people. They are also friendly people; therefore, they can make friends easily. They usually spend time with their friends. They are also peaceful people. That's why they get on well with other people. They are assertive; they believe themselves to overcome difficulties.
- **Neuroticism (emotionality, anxiety, dominant assured, satisfaction, and affect)** [18]**:** These people usually feel insecure. Therefore, they cannot easily get on well with others. They do not trust others in their social life. They are so sensitive; therefore they can be easily depressed with negative emotions. Moreover, they are generally anxious for their life. They are not happy with their current status.
- **Openness (culture, intelligence, intellect, intellectual interests, and intellectance)** [18]**:** They are curious and intelligent; therefore they tend to find out new ways to do something new. They appreciate diverse views, ideas, and experiences [16]. Moreover they are imaginative.

These personality traits are not directly opposed to each other, a person can highly present symptoms of some of these traits together. Individuals can be high in reflecting some of them and also can be low in reflecting others. Therefore, to have an idea about personality of an individual, all traits must be considered together.

In order to evaluate human personality, some inventories were published by the academics. The first inventory was published under the name of "NEO Personality Inventory" by Costa & McCrae in 1985. It only contains facets for Neuroticism, Extraversion, and Openness personality traits. In 1992, this inventory was revised by the same people. In this revised version, they added facets for Agreeableness and Conscientiousness personality traits. The revised version is called "NEO Personality Inventory – Revised (NEO PI-R)" [20]. The inventory contains a questionnaire survey having 240 items inside. Table 1 depicts facets that are used to evaluate Big 5 Personality;

NEO PI-R is generally considered as long for research purposes; subjects avoid filling such length survey. Therefore, Costa and McCrae published its short version called NEO-FFI having 60 items.

To briefly and efficiently evaluate Big 5 Personality, John, Donahue, and Kentle developed a survey having 44 items. This survey was called Big 5 Inventory (BFI) [21].

Table 1: Big 5 Personality Traits with Their Facets [20]

Big 5 Personality Trait	Facets (and correlated trait adjective)
Extraversion	Gregariousness (sociable), Assertiveness (forceful), Activity (energetic), Excitement-seeking (adventurous) , Positive emotions (enthusiastic), Warmth (outgoing)
Agreeableness	Trust (forgiving), Straightforwardness (not demanding), Altruism (warm), Compliance (not stubborn), Modesty (not show-off), Tender-mindedness (sympathetic)
Conscientiousness	Competence (efficient) , Order (organized), Dutifulness (not careless), Achievement striving (thorough), Self-discipline (not lazy), Deliberation (not impulsive)
Neuroticism	Anxiety (tense) , Angry hostility (irritable) , Depression (not contented), Self-conscientiousness (shy), Impulsiveness (moody), Vulnerability (not self-confident)
Openness	Ideas (curious), Fantasy (imaginative), Aesthetics (artistic), Actions (wide interests), Feelings (excitable), Values (unconventional)

2.2 Social Network Analysis

Individuals who interact with each other create a social network. Within this network nodes represent individuals, while edges represent type of relationship between these nodes. Analyzing such networks is called social network analysis. The aim of the social network analysis is to figure out the role of each node in a network, and find out coupling between nodes or sub networks and discover how information exchange between nodes. Therefore, analyzing social networks are important to find out a way to prevent disease separation, advertise of a product/idea, prevent a crime, organize people for some mission and etc [11].

In social network analysis, there are some measurements that give information about social network, relations between individuals and individuals within the network. These measurement types can be grouped as below;

- Measurement for Relations: Indirect link, frequency, stability, multiplicity, strength, direction and symmetry.

- Measurement for Individuals: Degree (in, out and all), closeness, brokerage, betweenness, centrality, prestige and diversity.
- Measurement for Network: Size, centralization, symmetry, connectivity, density and transitivity.

In this study we have used the below measurements in inference of personality traits;

- **Density:** This attribute can be calculated for a network in order to compute density of relation between nodes. Therefore, it is calculated as proportion of number of edges existing in the network relative to number of maximum possible edges in the same network. If a network has high density, then the information flow in the network will have high diffusion between nodes.
- **Brokerage:** Brokerage is the number of connected neighbors' pairs that the individual does not directly connect with. This attribute can be calculated for each individual in a network.
- **Betweenness:** This attribute can also be calculated for each individual in a network. It indicates the number of shortest connected paths between pairs via each individual. However, these pairs are not connected to each other directly. If an individual is high in betweenness, it shows that the individual is critical for the flow of information between others since other individuals do not know each other directly.
- **Transitivity:** Transitive triad is based on "friends of my friends are also my friends" idea. Three individuals are accepted as transitive triads if three of them are directly connected with each other or at least two of them are connected directly with each other. One of them is only accessible via another individual in the triad. The number of existing transitive triads relative to the number of possible triads in the same network indicates the transitivity of a network.

2.3 Related Works

2.3.1 Inference of Personality Traits using Linguistic Features

Mairesse et al. compare the performance of models developed for inferring personality traits using different sources such as essay corpus and transcribed recorded speech corpus with different machine learning algorithms [14]. Also in their study, individual's personality is evaluated by themselves and others. According to their results, they claim that personality

evaluated by others can be better modeled with regard to self-reporting personality. This claim proves that other people can more objectively rate them. They also assert that speech corpus is better source for predicting personality. Because they claim that individuals are more themselves while they are speaking.

Mohammad et al. study on personality prediction using essays written by individuals [15]. In their research, they focus on relations between personality and usage of different emotional word categories such as excitement, guilt, yearning, and admiration. They claim that when fine grained emotion features (calculated using NRC Hashtag Lexicon) are accounted, the prediction performance is increased. However, in their model, when coarse effect features (calculated using Turney Lexicons) and specificity features (calculated using NRC Emotion Lexicon) are considered, they have observed no significant changes in prediction performance. In addition, they also extracted most correlated emotion categories for each personality trait.

2.3.2 Personality Inference using Network Structure

Staiano et al. conduct personality prediction using social network structure [22]. In their study they populate ego-networks using an application installed on subjects' mobile phone. The application basically keeps track of calls, proximity information that is gathered by Bluetooth technology and includes a survey that evaluates personality traits. Using these attributes, they increase personality prediction accuracy up to 65% - 70% points.

Selfhout et al. study how personality traits affect social relationships between individuals [23]. The experiment is conducted with 205 participants who are freshman in a university. For each month, participants fill a survey about Big 5 Personality Traits and friendship networks. The participants fill the survey during their first semester (4 months). To analyze the effects of personality traits on friendships, they have used Simulation Investigation for Empirical Network Analyses (SIENA) tool. Their results show that individuals high in extraversion make new friends easily. They usually have large friendship networks. In addition, individuals high in agreeableness are easy to get along with others, therefore they are chosen as friends by others and they also have large friendship networks. It is also claimed that gender is another factor on friendships. Their results show that people have more friends with the same gender. Another finding is that women have more friends according to men. The most important limitation in this study is diversity of participants. They are all high educated participants and mostly women (82% of all participants).

2.3.3 Inference of Personality Traits using Social Media

Personality inference using social media is one of the new hot topics in the literature. There are also studies that observe social media usage habits according to user's personality.

In 2007, Lampe et al. study the relations between various types of user profile elements and the number of users in friendship network [24]. Their hypothesis is based on signaling theory and common ground theory. They firstly categorize profile elements in four main groups. These are control, referents (e.g. location), preference, and contact variables. Then they determine the number of usage percentages for each profile elements. After all, they measure correlation between profile elements with number of friends. They conclude their research as basic user information (age, gender etc.) that is related to number of friends but the amount of information in profile (such as about me, user posts etc.) is weakly related to number of friends. In this study, they only evaluate number of profile elements such as number of likes, number of favorite book and etc. However, they do not evaluate any content of these profile elements.

In 2008, Klemper et al. investigate the relation between users' personality (Big 5 Personality value) and acceptance/use of social network sites [25]. They prepare a questionnaire about personality traits and acceptance of social network website usage for Facebook users in Midwestern University. They find that users whose personality high in agreeableness, and openness look for usefulness of social network sites. On the other hand, users whose personality high in conscientiousness and neuroticism look for ease use of social network sites. However, users high in extraversion look for both usefulness and ease use of social network sites. The subjects used in this research are not well diverse. They are in the same education level, age group and they are studying at the same University. Therefore they may have similar preferences in acceptance of social networking sites.

In 2009, Schrammel et al. investigate relations between users' personality traits and their usage patterns and information disclosure behavior on online communities [17]. They have prepared three surveys; the first one is about personality traits, the second one is about usage patterns and the third one is about information disclosure. According to the survey results, they find that extravert and open people have more friends. They cannot find any significant evidence for effect of personality traits on information disclosure.

In 2010, Mislove et al. have inferred user profiles' attributes using his/her friends' user profile attributes in social media of universities [26]. These attributes are college,

matriculation year, department or high school. They claim that users are usually friends with other users who have similar profile attributes. They also observe that if two users share the same contents/links in a dense cluster or community, there is a tight correlation between these two users' profile attributes. Based on these inferences, they claim that, the other users' profile attributes can be predicted with 80% accuracy using 20% of users attributes. In this study, they have written a crawler to collect information from Facebook profiles. During the crawling, privacy settings are the main challenge for the study with 30% - 40% of subjects made their profile inaccessible to others.

In 2011, the first well known research that aims to predict user's personality using social media was done by Golbeck et al [16]. They show that there is a strong relation between personality traits and user's status updates. They have also found that people high in conscientiousness use less swear words compared to others and rarely use words that match in perceptual processes (seeing, hearing, feeling etc.). However, they mostly tend to talk about people they know. Moreover, people high in agreeableness, use more affective and positive feeling words while people high in neuroticism use words that impact negative feelings. People high in extraversion and openness tend to make new friends from different groups of people. Although there are many features that were disclosed in Facebook profiles, they only employed linguistic features of status updates in their study.

Sumner et al. study the correlation between user's Big 5 Personality Traits with Facebook usages [27]. For this study, they gather the usage information from 537 Facebook profiles. They also prepare 44 questions to determine users' personality traits. Users are also asked about their privacy concerns. After all, gathered information such as status updates, photo descriptions and about me statements are analyzed by Linguistic Inquiry and Word Count (LIWC) program. In evaluation, they apply zero-order Spearman's correlation on the Big 5 personality traits and Facebook usages. Their results show that there are relations between personality traits and Facebook usage. For instance; extravert people tend to share photos, and comment on other's shares. On the other hand, agreeable people tend to share photo and attract comments on others' shares. They also assert that people, who are high in openness, are more likely to share his/her profile information, and they tend to comment on others' shares. However, the strength of relations is not strong enough to infer an individual's personality.

Gosling et al. [28] study on two different research questions. The first one is to investigate relations between self-reported Facebook usage and personality traits. The second one aim to identify whether there is any relation between observable Facebook profile information and

personality traits. For the first research question, they prepare two questionnaires. The first questionnaire is Ten Item Personality Inventory to determine personality traits. The second one has questions about users' Facebook usages. For the second research question, unacquainted observers are selected for each user. They assess each user's personality versus his/her Facebook profile. In addition, each user's accuracy criterion is calculated. Accuracy criterion is obtained by combining self-reports and reports provided by four well-acquainted informants. All results are evaluated and a correlation matrix was formed between Facebook usage and personality traits. According to these results; they find that there is a significant correlation between personality traits and Facebook usages. They claim that, extraverts are more willing to use social media to socialize themselves in the society. They frequently check news feeds and likes or comments. Therefore, they have more friends, photos and etc in the Facebook. Another supporting claim is that people high in openness use social media to explore new activities. They frequently change their profile pictures compared to other users.

Chen et al. [29] work on social media to get answers for the following questions; "For what purposes do people use Facebook?" and "What is the impact of user personality on information disclosure on social network sites?". They claim that people low in extraversion and interdependent self-construal (allocentrists) disclose the least information about themselves. They are the least honest people and disclose information according to audience. They disclose information about themselves differently in social media and in real life. However, people high in extraversion and independent self construals (idiocentrists) disclose more information about themselves. They are more honest people, therefore they do not self-disclose differently in social media and in real life. Subjects' diversity is limited in this study; they are all from Psychology Department in the Southeastern University. Therefore, it is hard to generalize findings.

Bachrach et al. examine the relations between Facebook usage and personal profiles [3]. They extract the features having high level usage in Facebook profiles such as: number of published photos, events, number of joined groups and number of objects that user likes. They conclude that users' personality impacts on Facebook usage patterns. For instance, neurotic profiles do not have so many friends in social media compared to other people. Extravert profiles mostly like sharing posts while conscientious profiles share photos. According to the study, people high in extraversion can be successfully predicted by observing their Facebook usage patterns. However, people high in agreeableness are the hardest candidates to predict their personality traits by just observing their usage patterns of

social media. In this research, they have only considered the number of status updates, likes, groups while ignoring the content of what users like, share and join.

Adalı et al. study Twitter as a social media to predict user's personality [18]. In their study they focus on users' activities on Twitter. These activities are number of followers, number of favorite/retweeted messages, times that user spent on the social media and etc. According to the results of their research, they claim that behavioral features (following, retweeting) can be used in prediction like textual features. One of the limitations of this study is the number of attended users. Only 71 users have attended the experiment.

Bai et al. study on predicting user personality based on their behavior on social media. In their experiments, they use Renren as social network site. Renren is highly popular in China. Chinese people mostly prefer this social networking site instead of Facebook. They write a third party application to gather users' information from this web site. The application also enables users to submit 44 questions about personality inventory. According to their results, agreeable users spend more time in online chat. Moreover, conscientious users spend more time on questbook to help other people. Users high in extraversion trait tend to have more friends compared to others. People high in openness trait tend to keep up to date their statuses, since they are reflected in a strong intellectual curiosity and a preference for novelty and variety [30].

Farnadi et al. utilize machine learning techniques to infer users' personality traits by analyzing their status updates in Facebook [6]. They have proposed a model which uses LIWC features, social network features and temporal features. According to their results, users high in extraversion and conscientiousness have more friends compared to other users. Conscientious users mostly share status updates between 00 AM and 11 AM. In their previous works, they also point that users' status updates are more important cues for their personality. In this study, they do not consider the frequency of status updates during a specific time period.

Markovikj et al. also study on parameters that are highly correlated with user's personality and propose a predictive personality model [8]. In their study they have used linguistic features of users' status updates. These features are extracted using LIWC Tool, POS Tagger, Affin Dictionary and General Inquirer Tool (H4Lvd Dictionary). In addition, they have used demographic and egocentric network data. Using Sequential Minimal Optimization (SMO) classification algorithm, they claim that prediction performance can be improved at a 0.8 – 0.9 true positive rate. In their study, they have quantitatively accounted

likes, groups, events and etc. They do not regard what they have liked, and which group they have joined and what kind of events users have attended.

Alam et al. work on modelling users' Big 5 Personality Traits using status updates with different machine learning algorithms such as Sequential Minimal Optimization for Support Vector Machine, Bayesian Logistic Regression and Multinomial Naïve Bayes [9]. In comparison, they claim that Multinomial Naïve Bayes sparse model perform better results compared to other models. In this study, they split data set as 66% train and 34% test sets. However, the number of participants is just 250 and it is quite limited.

Appling et al. investigate the relations between users' Big 5 Personality Traits and their speech acts extracted from status updates [10]. They label status updates with the "Assertive", "Commissive", "Declarative", "Directive" and "Expressive" speech acts. According to the correlation results, people high in conscientiousness and agreeableness rarely use sentences marked as assertive speech acts. However, people high in extraversion frequently use these kinds of sentences. In addition, neurotic people rarely use commissive sentences. In their study, they do not take into consideration the content of user's speech.

2.4 Available Information on Facebook

Users disclose information about them with other users on Facebook. Sharing information is gender, interests, photos, activities, political views, religion and etc. Here is the disclosed information that is available on Facebook used in inference of personality traits in the literature.

2.4.1 Profile Information

This category contains information about users' age, birthdate, birth place, home town, relationship status and etc. Table 2 shows the list of attributes used in personality prediction.

Table 2: Profile Information used in Personality Prediction

Attribute	Literature Reference
About Me	Significant([21], [22]) – Insignificant([18], [10])
Gender	Significant([8], [18], [24], [29])
Hometown	Significant([18])
Current City	Significant([18])
Birth Information	Significant([8]) – Insignificant([21], [24])
Relationship Status	Significant([18])
Interested In	Significant([18])
Favorites	Significant ([18], [21])
Political Views	Significant([18])
Religion	No Reference

2.4.2 Facebook Activities

This category contains information about users' activities on Facebook. Their likes, sharing, attending events, joining groups are considered in this group. Table 3 shows the list of attributes used in inference of personality traits.

Table 3: Users' Activities on Facebook Used In Personality Prediction

Attribute	**Literature Reference**
Likes	Significant([21], [22], [25], [28], [29])
Status Updates	Significant([21], [28], [29])
Photos	Significant([21], [22], [25])
Tags	Significant([21], [22], [25])
Friends	Significant([10], [18], [21], [22], [23], [25], [28], [29])
Events	Significant([10], [28], [29])
Groups	Significant([21], [22], [25])
Works	No Reference
Schools	Significant ([18])

CHAPTER 3

RESEARCH METHODOLOGY

In this chapter, we discuss research methodologies used in inference of personality traits. First, we will mention about the base methodologies that are applied in the literature so far. Then we will introduce our hypotheses. For each hypothesis, we describe our motivation and the proposed methodology.

3.1 Base Methodology

The previous works show us that Facebook users express themselves on Facebook as they do in their real life. However, environmental factors affect Facebook profiles; they are generally their actual profiles not idealized version of their personality [32]. If a user is a neurotic person, he/she tends to share status updates using words including negative emotion such as hate, anger, kill, annoyed and etc. Moreover, if a user is conscientious person, he/she tends to share status updates having swear words. Table 4 shows some supplementary status updates from myPersonality database shared by people;

Table 4: Sample status updates shared by individuals with different personality traits

Personality Trait	Sample Status Updates
Neurotic	It's official: I'd rather have watched 6 depressing foreign language films than this boring 2-hour long piece of crap called Adventureland
	Needs sleep but doesn't want it to stop being tonight. Damn sleep is an awful habit.
	Today's stream of conscientiousness: Hooray for comic books, poetry and nerds in general...forgot my phone at home today, blargh (couldn't reach me, that's why)...swamped at work...pretzels suck...missing NOLA (especially the Quarter) alot today for some reason..."the holidays" are annoying me already...Wikipedia is addictive...so is Facebook, gods help me...!
Conscientiousness	Attention-Houston area ANIMAL LOVERS! Montgomery County Animal Shelter is closing tomorrow. You can adopt for free tomorrow between 12-4. They will be euthanizing all animals without a home!!! Go & save a life!! 30 cats and 80 dogs left. Please re-post
	To CTYers past and present, I urge you to do two things: 1) Dust off your Garden State soundtrack, there was a reason we listened to it non-stop for 42 days. 2) Mark July 16-18 in your calendar - we are all descending on the land of milk and honey. Carlizzle fo' shizzle.
	studying until my eyeballs fall out.
Extraversion	Is getting ready to get crunk and party like its 1999! holla at a playa! LOL
	Memo is over--- going to get something done to my hair, then cleaning the apartment and going out with *PROPNAME* for a little early birthday celebration!
	Cleveland celebrating birthdays with the King, New Years in the 'burg, cruisin' the Caribbean....gonna be a nice little weekend!
Openness	is thinking about going to graduate school? Not quite sure, just exploring the idea. There is quite a bit to consider.
	"Sit back and curiously observe the ride. You're not going to enjoy it, but it is going to take you somewhere." --Me under the inspiration of some lost muse.
	is thinking hard and excited for her new Xanga layout!
Agreeableness	Some people say pain can be controlled by the mind. I do not agree with these people. It doesn't mean I don't admire them though.
	Accept the things to which fate binds you, and love the people with whom fate brings you together, but do so with all your heart.

Based on these inferences, in the literature, there are some methodologies that use linguistic features of status updates to infer personality traits in Facebook [6] [16] [17]. They have used different linguistic tools such as LIWC Tool, NLTK and General Inquirer and etc. to extract linguistic features. These studies propose a model to classify unknown profiles using relations between users' personality and these linguistic features.

There are also predictive models that make use of demographic information obtained by Facebook [4] [14] [17]. These studies use basic profile information such as gender, location, age and etc. As in linguistic features, these studies also calculate the correlation between personality traits and profile information. Considering correlation values, they have proposed prediction models.

Moreover, there are some methodologies that predict user's personality by taking into account their Facebook usage habits and friendship networks [3] [27]. These studies focus on number of sharing, like, comment, joined groups and attended events. They also study their social networks by using friendship networks.

The previous works show that there are tight relations between user status updates, profile information, facility usage statistics and their personality. Based on these studies, we will use different data sets. We aim to obtain a predictive model to satisfy hypothesizes described in next sections.

3.2 Preprocessing

In this chapter of thesis, proposed improvement based on existing studies is discussed. This improvement is going to be tested and verified in the next "Experiments" chapter of the thesis.

Improvement: The personality traits of a user whose status updates are few cannot be identified accurately as there is no or very limited information. Incorporation of such cases in a supervised model degrades the predictive performance of the model.

LIWC Tool processes a given text and produces an output by denoting usage percentages of word categories [33]. If the length of text is limited, the usage percentage of word categories does not make any sense. However, there are many users where they have only a few posts shared in social media. That's why for these users, it is difficult to infer their personality by looking at their status updates. In addition, the performance of the predictive model will be affected negatively if these users information is incorporated to the model. Here are some status updates shared by users who have shared only one post;

"*it will be a long week....*"

"I'm just sippin on chamomile..."

For these cases, it is difficult to say anything about their personality just by looking at their limited status updates. There are too many users that do not like sharing status updates in Facebook. To clarify, we have plotted which shows the number of status updates posted by users (Y axis shows the number of total users while X axis shows the number of shared status updates.). As it can be seen from the histogram, most of the users share less than 200 status updates. There are noticeable amount of users who have less than 10 status updates.

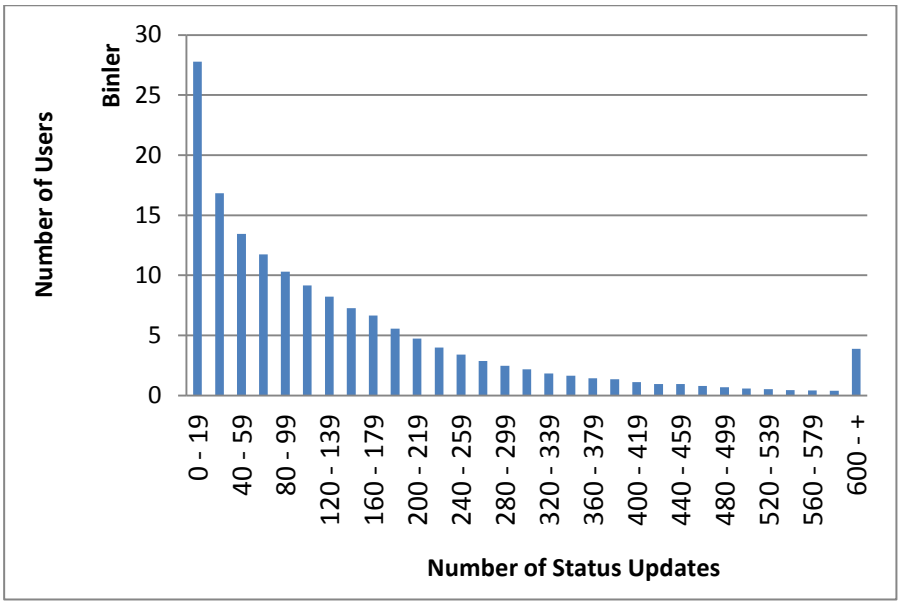


Figure 1: Number of users by their number of shared status updates

After discarding people who do not have sufficient amount of status updates, the performance of the model can be increased. Because, users for whom there are more evidence to infer their personality will remain.

3.3 Hypotheses

In this chapter of thesis, hypotheses, which our master thesis is based on, are introduced. These hypotheses are going to be tested and verified in the next "Experiments" chapter of the thesis.

Hypothesis 1: There is a relation between users' personality traits and their spouses' personality traits. Incorporation of spouses' personality in a supervised model improves the predictive performance of the model.

Individuals' personality traits have an important role in starting a new relationship according to homophily. Therefore, there might be a relation between users' personality and their

spouses' personality. To evaluate this relation, we calculated Pearson Correlation between couple's personality using data provided by myPersonality database. According to the results shown in Table 5, individuals prefer to have a relationship with people who are similar to themselves. However, neurotic and agreeable individuals usually start a relationship with conscientious individuals.

Table 5: Pearson Correlation Matrix of Big 5 Personality Traits for Couples. Statistically significant correlations ($p < 0.05$) are bolded. (r: pearson correlation coefficient, t: t-distribution value)

		Second User									
		OPE		**CON**		**EXT**		**AGR**		**NEU**	
Mean±StdDev		3.95±0.66		3.49±0.73		3.49±0.82		3.53±0.7		2.72±0.83	
Correlation		**r**	**t**	**r**	**t**	**r**	**t**	**r**	**t**	**r**	**t**
First User	**OPE**	**0.117**	5.566	-0.037	1.748	-0.037	1.753	0.008	0.393	**0.043**	2.058
	CON	**-0.037**	1.748	**0.148**	7.075	**0.051**	2.402	**0.064**	3.053	**-0.113**	5.373
	EXT	-0.037	1.753	**0.051**	2.402	**0.115**	5.477	0.015	0.689	**-0.066**	3.119
	AGR	0.0080	0.393	**0.064**	3.053	**0.015**	0.689	0.038	1.803	-0.012	0.555
	NEU	**0.043**	2.058	**-0.113**	5.373	**-0.066**	3.119	-0.012	0.555	-0.04	1.911

As it can be seen in correlation matrix, there is significant correlation between users' personality and their spouses' personality traits. In personality inference, incorporating spouses' personality traits may increase prediction performances of the models.

Hypothesis 2: If two friends have too many common friends, their personality should be same or close to each other. Incorporation of such friends' personality in a supervised model improves the predictive performance of the model.

When the number of common friends increases, it may indicate that two users are close to each other. Since an individual might prefer to selects friends having similar preferences and interests, we claim that the similarity between two friends increases as they have maximum number of common friends in Facebook. They may do the same job, study in the same school, have similar interests and live in the same neighborhood (thus same social background). This information shows us that there are so many common things shared by

these friends. Therefore, these two users may have similar personality traits, which may improve the performance of predictive models.

Hypothesis 3: If two friends have too many common likes, their personality should be same or close to each other. Incorporation of such friends' personality in a supervised model may improve the predictive performance of the model.

Kosinki et al. researched on predicting personality of users by studying their likes in Facebook. According to results, they observed that there are tight relations between users' likes and dichotomous variables such as gender, race, relation status, religion etc [4].

Cantador et al. researched on relation between personality traits and user preferences in Facebook such as music, book, movies, activity groups, fun pages etc. Their results showed that users who have the same personality traits tend to like similar type of contents [13].

Based on the previous researches, we assert that, if two friends have similar interests and likes same contents on Facebook, their personality traits should be close to each other. In other words, friends who have similar personality trait scores, likes similar contents and vice versa.

3.4 Evaluation Criteria

The proposed models in Experiments section will be tested with below described parameters.

- ***Correctly Classified Instances:*** It indicates that the percentage of the number of correctly classified instances over the number of all instances that are predicted by the proposed model [34]. To formulize it;

$$CCI = 100 * \frac{number\ of\ correctly\ classified\ instances}{number\ of\ all\ classified\ instances}$$

- ***Incorrectly Classified Instances:*** It implies the percentage of the number of incorrectly classified instances over the number of all instances that are predicted by the proposed model [34]. To formulize it;

$$ICI = 100 * \frac{number\ of\ incorrecly\ classified\ instances}{number\ of\ all\ classified\ instances}$$

- ***Mean Absolute Error:*** It is another parameter to evaluate prediction performances. In prediction, the distance between actual value and predicted value is called

absolute error. Mean Absolute Error is the average of summation of these absolute errors calculated for each instance [35]. If the parameter converges to zero, it denotes prediction is performed in best performances. However, if the parameter converges to one, it denotes prediction performance is worst. To formulize it;

$$e_i = |P_i - O_i|,$$

$$e_i: Absolute\ Error,\ P_i: Predicted\ Value, O_i: Actual\ Value$$

$$MAE = n^{-1} * \sum_{i=1}^{n} e_i\ ,\ n: Number\ of\ predicted\ instances$$

- ***Root Mean Square Error:*** This parameter also evaluates the prediction performance by finding the difference between predicted and actual values. However, the parameter increases with proportion of square of absolute error. Therefore, if absolute error increases for predicted instances, the parameter quickly converges to one compared to Mean Absolute Error [35]. To formularize it;

$$RMSE = \sqrt[2]{n^{-1} * \sum_{i=1}^{n} e_i^2}$$

CHAPTER 4

EXPERIMENTS & RESULTS

In this chapter, we explain the datasets, experiments and their results.

4.1 Data Set

In this thesis, we have used the database provided by myPersonality Project [36]. This project was created and proceeded by David Stillwell and Michal Kosinski. They developed a Facebook application named myPersonality. This application mainly collects information from user profiles and enables users to take psychometric tests (more than 25 different questionnaires) to calculate Big 5 Personality Values, IQ scores, Satisfaction with Life Scale and etc. After collecting the raw data, they have processed this data to create new data sets while concerning privacy issues.

Currently the application has 4,282,857 individual Facebook profiles from various age groups, background and cultures. Almost 40% of these users (1,674,259 users) give the application access their information on Facebook such as status updates, likes, friends, groups, photos and etc. There are 1,048,575 individual users who have completed comprehensive 366 question Big 5 Personality Trait survey. For 74,521 users, their friendship graph is generated and egocentric networks are defined. On this friendship graph, social network analysis has been applied to calculate transitivity, brokerage, betweenness and etc.

There are 2,720,324 female users while there are 1,482,036 male users. And 80,497 users' gender could not be obtained because of their privacy settings. Figure 2 shows the number of users and their gender information.

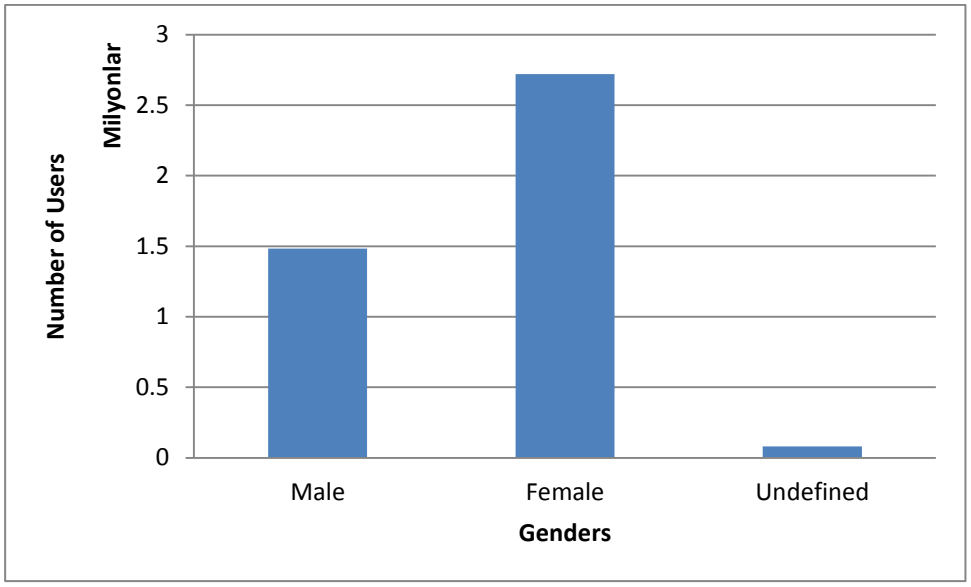


Figure 2: The number of users and their gender

If we consider user's demographic details, the users are generally in between 11 - 21 age years old meanwhile there are people who are more than 60 years old. The average age of users is 26. Figure 3 shows the age distribution of the participated users.

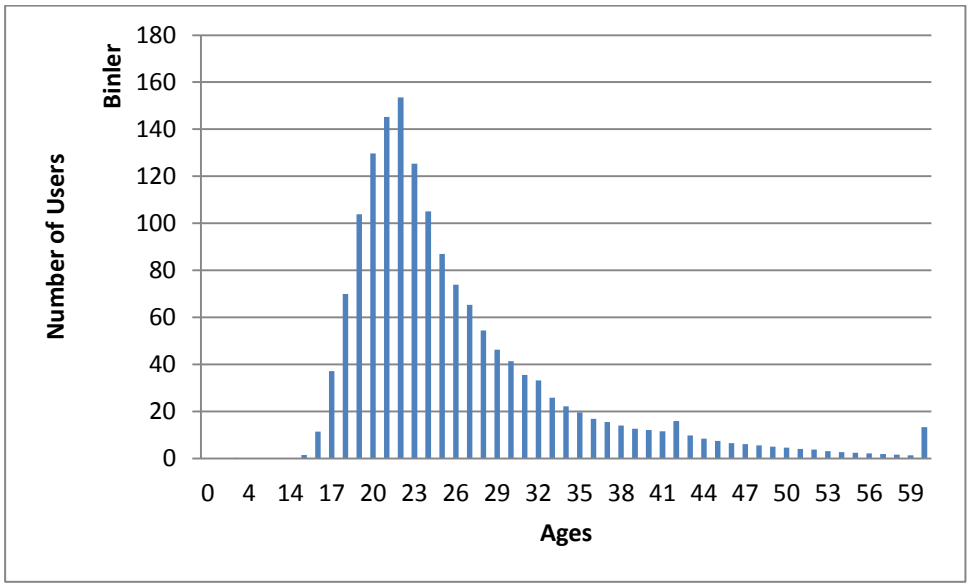


Figure 3: The number of users and the distribution of their respective ages

The majority of the users who utilized myPersonality application are from the United States. However, there are users from the United Kingdom, Canada, India and etc. Figure 4 shows the country distribution of users.

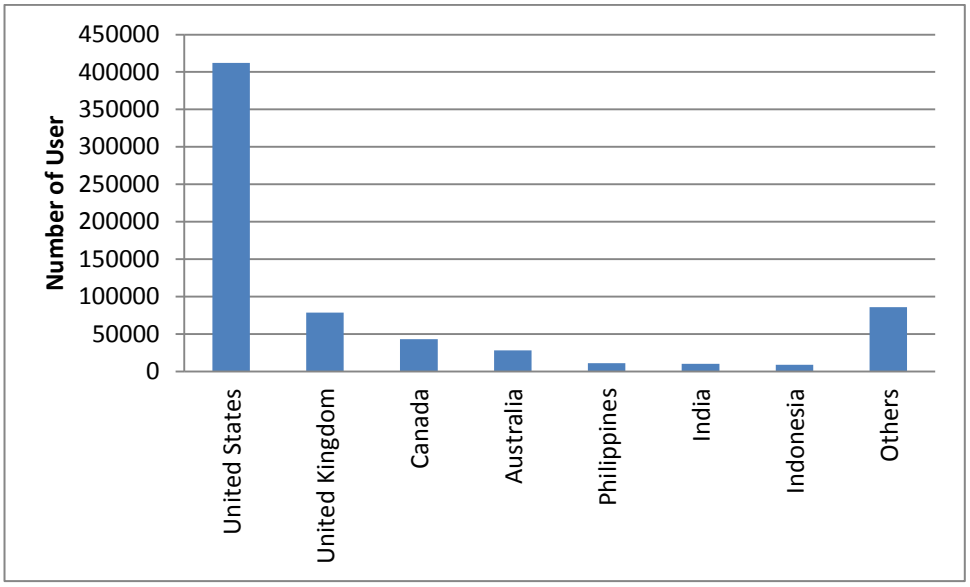


Figure 4: The country distribution of users

The users' status updates were also gathered by myPersonality application. The average number of user's status updates per user is approximately 120. The majority of the participated users has less than 20 status updates. Figure 5 shows the status updates' distribution of users.

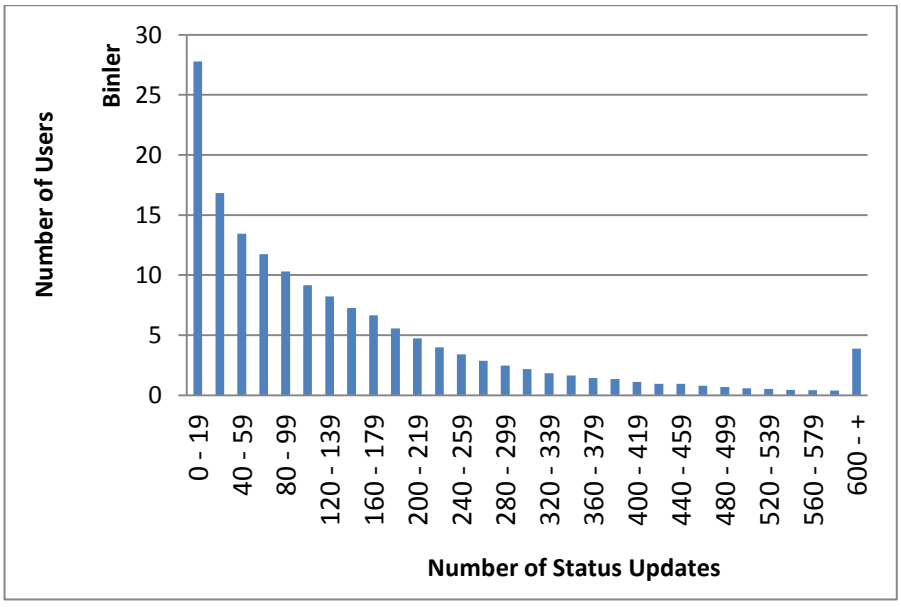


Figure 5: The distribution of users and their shared number of status updates

Facebook profiles provided by myPersonality includes different political views, most of which are liberal, moderate, democrat, and conservative. Figure 6 shows the political views distribution of the users. As can be seen, the majority declared themselves as liberal.

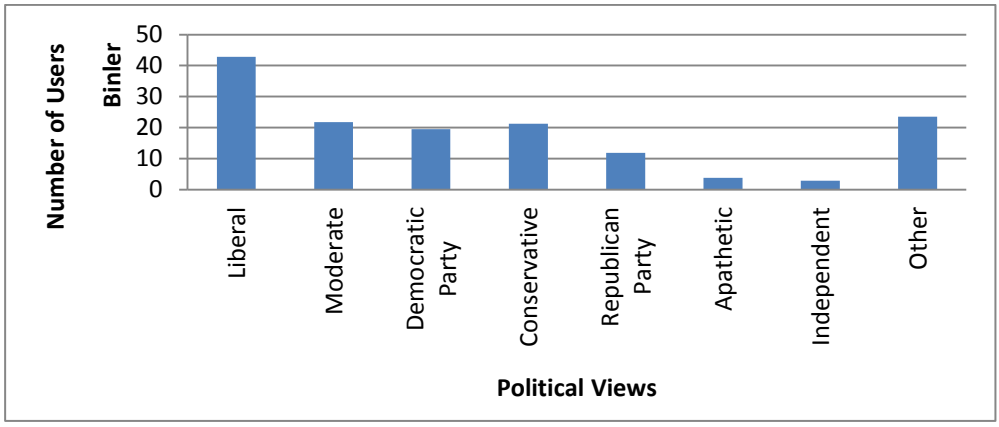


Figure 6: The distribution of users and their political views

During our research we used the following data tables. For some experiments, only one data table was used whereas for some other experiments we merged two or more data tables (by user id) to verify the proposed hypotheses.

- ***User's demographic details (demog.csv):*** This table contains basic attributes about Facebook users such as their gender, birthday, age, relationship status, interests, locale, network size, time zone and user's motivation for participating to Facebook. In this table 4,282,857 users' basic information resides.
- ***Facebook activity (freq.csv):*** This table retains summary information about user activities. These activities are tagging, liking, posting, joining a group, attending an event and etc. This table contains the number of these activities such as the number of like, status updates, attending an event, joining groups, work and education background, tagged items and friends. In this table, 1,674,261 users' Facebook activity summaries reside.
- ***BIG 5 Personality Scores (big5.csv):*** This table comprises Big 5 Personality Test results that users have taken. It contains the scores of main traits such as conscientiousness, agreeableness, openness, neuroticism and extraversion. The scoring values are between 0 and 5. The higher score on traits implies that the trait describes the individual better. The table also contains the date the user has taken the survey. In this table, there are test results for 1,048,575 unique Facebook users.
- ***LIWC tags for the status updates (liwc.csv):*** For each user, user's all posts are merged into a single text file. Then the file is processed by LIWC Tool to extract word usage frequencies in 64 linguistic and psychological processes, personal concerns, and spoken categories. The results are concatenated in this table. The table

contains usage frequencies in LIWC word categories for each user. It has results for 153,717 Facebook profiles.

- ***Couples (couples.csv):*** Facebook users may share their relationship status as a profile attribute. Relationship requests are sent from one user to another, if accepted, it is shown on both user profiles like whether they are "In a Relationship", "Engaged", "Married" and etc. This table contains relationship status between two Facebook users. In addition, it also contains some Facebook usage statistics like number of common friends, likes, groups and tags. 8,165 relationship records reside in this table.
- ***Egocentric network stats (sna.csv):*** Social network parameters for each user are present in this table. These parameters are network size, ego betweenness, normalized ego betweenness, density, brokerage, normalized brokerage and transitivity. These parameters are calculated for 74,521 users.
- **Facebook friendship DIADS (fb_friendship.csv):** This table comprises friendship information between users. It shows users who have sent a friend request and keeps track to whom this request has been sent. There are 137,224,401 friendship records.

In our study, to be able to affirm our hypotheses, we have used aforementioned tables. For some hypotheses, we need to use data from different tables. That's why during the experiments we have merged these tables based on "user id" attributes. Here are tables we created by manipulating above tables.

BIG5_LIWC: While big5.csv dataset contains 1,048,575 unique Facebook users' Big 5 Personality trait information, liwc.csv dataset contains only 153,717 unique Facebook users' status update tags. These two tables were merged in a single table according to user id fields. While merging these tables, the users who did not exist in both tables were removed. Also there were some duplicate records in LIWC tags database and these records were also removed. As a result, we obtained a merged dataset with ***115.863 records***. Each record in this table has 76 different attributes such as user id, Big 5 Personality Scores and LIWC Tags Frequencies.

- **BIG5_LIWC_FREQ_DEMOG_SNA:** User's demographic details, Facebook activities, social network attributes, Big 5 Personality scores and LIWC tags for the status updates were merged in this dataset. After merging these datasets, users who have less than 50 posts were removed. At the final stage, there were 20,931 unique user records. For each user, there were 104 attributes.

- **BIG5_LIWC_FREQ_DEMOG_SNA_COUPLE:** In this dataset, we have merged Couples table (*couples.csv*) with Big 5 Personality Trait score, LIWC tags for the status updates, User's demographic details, Facebook activity, Egocentric network stats tables. The tuples about couples who do not exist in the Big 5 Personality Trait table were discarded. After merge operation finished, we calculated the number of common friends, likes, groups, events, schools and works between two users for each friendship. Moreover, for each couple, we calculated Euclidean distance between two users' personality. Finally, we had 1,126 unique couple records with 117 attributes in this merged dataset.
- **BIG5_LIWC_FREQ_DEMOG_SNA_FRIENDSHIP:** In this dataset, we have merged Facebook fr*iendship ta*ble with Big 5 Personality Trait Score, LIWC tags for the status updates, User's demographic details, Facebook activity, Egocentric network stats tables. If one of the users in this friendship do not exist in Big 5 Personality Trait Score table, the related tuple is removed from the dataset. After merging, we calculated the number of common friends, likes, groups, events, schools and works between two users for each friendship. Moreover, for each friendship, we calculated Euclidean distance between two users' personality trait scores. Finally, if one of the users in a friendship has less than 30 friends, that tuple was eliminated from the dataset. This number was determined by an ad-hoc inspection of the base table since in lower friendship numbers, users seemed to be inactive in Facebook so there will not be sufficient information to run the experiments. Finally, we had 34,291 unique friendship records with 132 attributes.

Table 6 shows the details of the data sets used in each experiment. In the table, FREQ, DEMOG and SNA imply Facebook usage statistics, demographic information and social network analysis attributes respectively. And also it summarizes the number of records and the number of attributes.

Table 6: myPersonality data tables used in the experiments

Exp No	Tables							Number of Records	Number of Attributes
	BIG5	LIWC	FREQ	DEMOG	SNA	COUPLE	FRIENDSHIP		
1	√	√	-	-	-	-	-	115,863	76
2	√	√	-	-	-	-	-	115,863	76
3	√	√	√	√	√	-	-	20,931	104
4	√	√	√	√	√	-	-	20,931	104
5	√	√	√	√	√	√	-	1,126	117
6	√	√	√	√	√	-	√	34,291	132

4.1.1 Preprocessing

In order to state whether the user is agreeable, extravert, conscientious, neurotic or open in line with the literature [11] [37] [38], we calculated the mean values for each Big 5 Personality Trait scores. For agreeableness score the mean value is 3.55. If user agreeableness score is equal or greater than this value, the user has been accepted as an agreeable person. Secondly, for conscientiousness score, the mean value is 3.46. If user's conscientiousness score is equal or greater than this value, the user has been accepted as a conscientious person. Thirdly, for extraversion score, the mean value is 3.51. If user's extraversion score is equal or greater than this value, the user has been accepted as extravert person. Fourthly, for neuroticism score, we obtained the mean value as 2.74. If user's neuroticism score is equal or greater than this value, the user has been accepted as a neurotic person. Lastly, for openness score, the mean value is 3.80. If user's openness score is equal or greater than this value, user has been accepted as an open person.

4.2 Tools Used in Experiments

While performing experiments we have used Weka Data Mining Software for modeling and LIWC Tool for textual analysis.

4.2.1 Weka Data Mining Software Tool

Weka Software is a collection of machine learning algorithms for data mining tasks [39]. The Weka Project was proposed and developed by Machine Learning Group at the University of Waikato. The aim of the group is developing an open source machine learning algorithm library that can be used by universities and industries. That's why they start to develop this framework under GNU General Public License [40].

Weka is fully implemented with JAVA programming language. The tool has its own graphical user as shown in Figure 7 and command line interfaces that help users to access functionalities such as filtering, classifying, visualizing and etc. It can also be used as a library in any other projects.

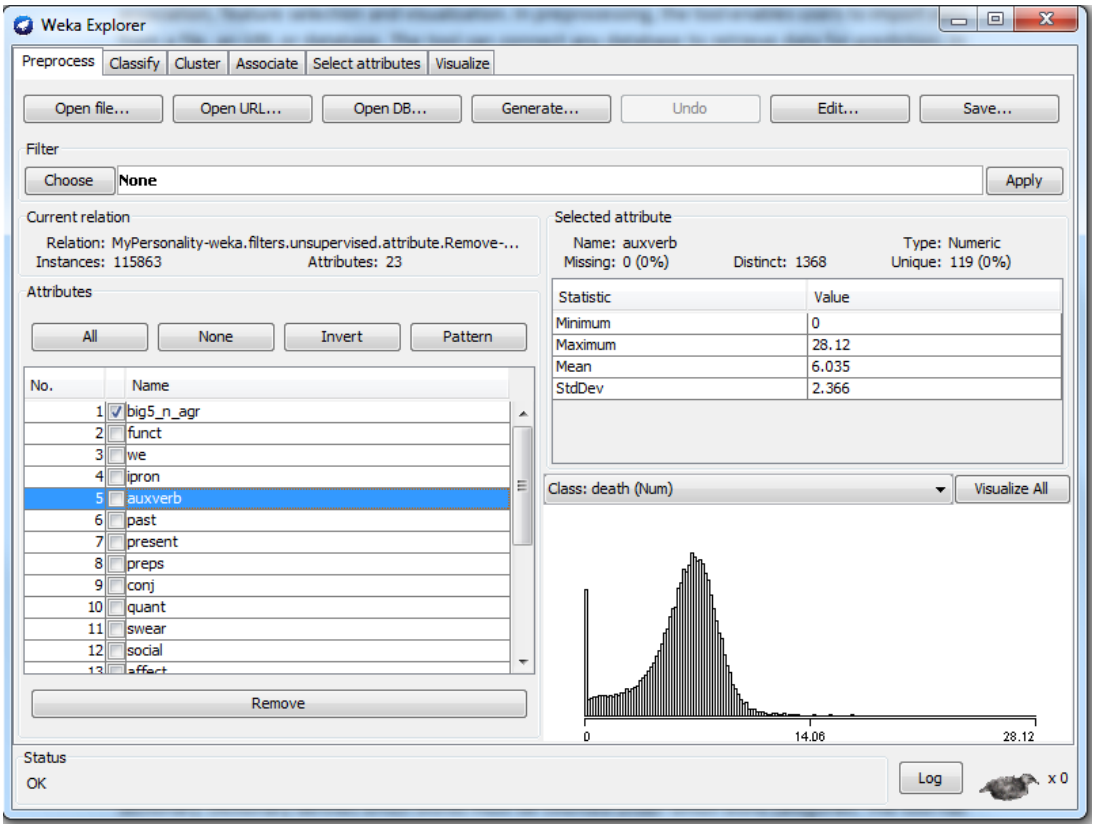


Figure 7: Weka Data Mining Software Tool Graphical User Interface

Weka Data Mining Software Tool supports several standard data mining tasks such as preprocessing, classifying, clustering, association, feature selection and visualization.

In preprocessing, the tool enables users to import data from different sources such as a file, a URL or database. The tool can connect to any database to retrieve data for prediction. In addition, it enables users to apply filtering algorithms on data sets. Weka Data Mining Software Tool also calculates maximum, minimum, average and standard deviation values for each given attribute. It plots histograms for selected attributes.

In classification, it enables to configure training and testing options. In clustering, the tool provides algorithms to cluster instances.

4.2.2 Linguistic Inquiry and Word Count Dictionary & Tool

LIWC is a text analysis software program developed by James W. Pennebaker, Roger J. Booth, and Martha E. Francis [33]. They developed it to provide an efficient and effective method for emotional, cognitive and structural analysis in individuals' verbal and written speech samples.

LIWC Tool calculates usage percentages of each word category on any given text according to its dictionary. This dictionary defines which words must be counted under which word categories.

The tool has its own default dictionary in English named with LIWC2007. In their own default dictionary, there are approximately 4,500 words and word stems in 82 different word categories given in Appendix F. For instance; "*Social Processes*" word category contains words implying an action or phenomenon in social event such as mate, talk, they and child. The category has also three sub categories named as "*Family*", "*Friends*" and "*Humans*". It also enables users to define their own dictionaries. There are some available dictionaries in French, German, Turkish languages and etc.

The tool reads a given text in any file type which is created by any word processing software. Then, the tool tokenizes all words used in a text. For each tokenized word, it seeks a match in all dictionaries. If the tool finds a match for the word in a dictionary, it increases its usage percentage under the word's associated dictionary category. During the process, the tool also counts various structural statistics such as number of words, sentences, punctuations and etc. The tool gives an output to show usage percentages of word categories with some structural statistics [41].

Here are the sample status updates from myPersonality database;

"*I saw HP6... funny, lots of awesome awkward silences, but the scriptwriter needs to have his fingers broken. Not only did they achieve new and improved levels of editing important plot elements out, but they are getting increasingly more creative with the n*"

When LIWC Tool is executed on the above paragraph, it produces the following output depicted in Table 7;

Table 7: Linguistic analysis result produced by LIWC2007 Tool

Category	Usage	Category	Usage	Category	Usage	Category	Usage
WC	44.00	conj	6.82	inhib	0.00	relig	0.00
WPS	44.00	negate	2.27	incl	6.82	death	0.00
Sixltr	29.55	quant	6.82	excl	6.82	assent	2.27
Dic	81.82	number	0.00	percept	6.82	nonfl	0.00
funct	47.73	swear	0.00	see	2.27	filler	0.00
pronoun	9.09	social	6.82	hear	2.27	Period	9.09
ppron	9.09	family	0.00	feel	2.27	Comma	6.82
i	2.27	friend	0.00	bio	2.27	Colon	0.00
We	0.00	humans	0.00	body	2.27	SemiC	0.00
you	0.00	affect	13.64	health	0.00	QMark	0.00
shehe	2.27	posemo	11.36	sexual	0.00	Exclam	0.00
They	4.55	negemo	2.27	ingest	0.00	Dash	0.00
ipron	0.00	anx	2.27	relativ	9.09	Quote	0.00
article	4.55	anger	0.00	motion	2.27	Apostro	0.00
verb	11.36	sad	0.00	space	4.55	Parenth	0.00
auxverb	6.82	cogmech	20.45	time	2.27	OtherP	0.00
past	4.55	insight	0.00	work	4.55	AllPct	15.91
present	6.82	cause	2.27	achieve	9.09		
future	0.00	discrep	2.27	leisure	0.00		
adverb	2.27	tentat	2.27	home	0.00		
preps	11.36	certain	0.00	money	0.00		

4.3 Experiments

In this section, the experiments which are used to verify the hypothesis in Section 3.2 are discussed. We have performed seven experiments in total. Here is the table showing which hypothesis / objective is verified under which experiment/s;

Exp No	Hypothesis / Objective
1	To verify base methodologies proposed in the literature.
2	***Improvement:*** *The personality traits of a user whose status updates are few cannot be identified accurately as there is no or very limited information. Incorporation of such cases in a supervised model degrades the predictive performance of the model.*
3	To verify base methodologies proposed in the literature
4	To test performances of different machine learning algorithms
5	***Hypothesis 1:*** *There is a relation between users' personality traits and their spouses' personality traits. Incorporation of spouses' personality in a supervised model improves the predictive performance of the model.*
6	***Hypothesis 2:*** *If two friends have too many common friends, their personality should be same or close to each other. Incorporation of such friends' personality in a supervised model improves the predictive performance of the model.* ***Hypothesis 3:*** *If two friends have too many common likes, their personality should be same or close to each other. Incorporation of such friends' personality in a supervised model improves the predictive performance of the model.*

4.3.1 Experiment 1

Objective: In this experiment we aim to verify the base methodologies proposed in the literature claiming that users' personality traits can be predicted by using their status updates [8] [9] [11] [12] [16]. We would like to ensure that we obtain similar correlation values between user's personality and LIWC categories as in the literature.

Dataset: In this experiment we have used BIG5_LIWC dataset.

Methodology: In this experiment, for each personality, we have calculated Pearson Correlations between LIWC tags and the corresponding personality.

In the preprocessing phase, for the feature selection, initially we have calculated Pearson Correlation between the scores of personality traits and the linguistic features and reported in Appendix A: Pearson Correlation Table for Experiment 1. Then we have used Information Gain attribute evaluator algorithm to assess the worth of each attribute in classification. The results are reported in Appendix E.

When we compared Pearson Correlation table and Information Gain results, we observed that they give similar results in terms of weight of each attribute. Therefore, we have only considered Pearson Correlation table in attribute elimination for this and following experiments.

We selected the correlated attributes where their r values are higher than 0.05 and p values are 0. Then the remaining attributes that are under the determined correlation values are filtered out from the proposed model.

In the classification phase, we have used SMO machine learning algorithm with the suggested configuration settings [44] specified in Table 8 to classify instances. Moreover, 10 fold cross validation method is applied.

Table 8: Suggested configuration for classifying instances by SMO

Parameters	The complexity parameter C: 1.0 The epsilon for round-off error: 1.0E-12 Data transformation: Normalize training data Kernel: PolyKernel The tolerance parameter: 0.0010

Results: As can be seen from the correlation table in Appendix A, agreeable people most frequently use words in *anger, swear, negemo, relativ, time, incl, preps, funct, motion, affect, cogmech, relig, conj, quant, we, auxverb, death, ipron, present* word categories and least frequently use words in *posemo, social and, past* word categories. Users' agreeableness personality trait is predicted using these attributes. As a result, Correctly Classified Instance (CCI) score is obtained as 56.947%. The prediction results are summarized for each fold in Table 9;

Table 9: Classification results in predicting user's agreeableness.

Fold	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
CCI	56.891	56.926	57.078	57.117	56.93	56.904	56.891	56.887	56.887	56.947
ICI	43.109	43.074	42.922	42.883	43.07	43.096	43.109	43.113	43.113	43.052
MAE	0.431	0.431	0.429	0.429	0.431	0.431	0.431	0.431	0.431	0.43
RMSE	0.656	0.656	0.655	0.655	0.656	0.656	0.656	0.656	0.656	0.656
AuC	0.56	0.561	0.562	0.563	0.561	0.561	0.56	0.56	0.56	0.562

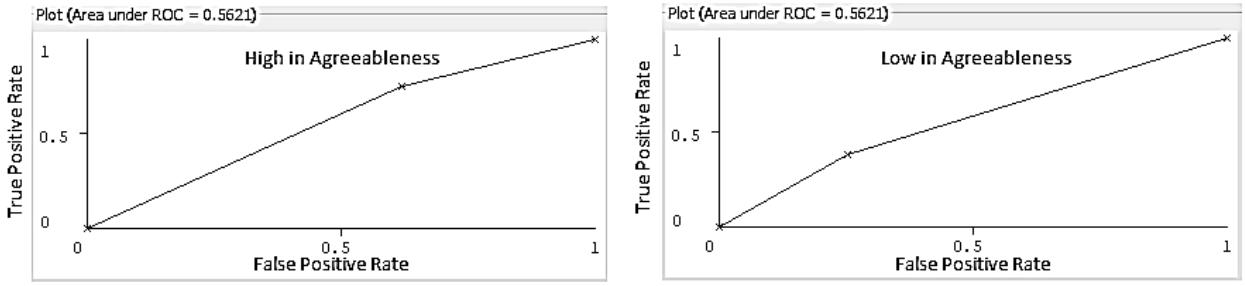


Figure 8: Roc curve for Agreeableness

For conscientiousness people, they most frequently use words in *posemo, relativ, preps, time, incl, achieve, funct, motion, article, social, quant, work, we, family, space, cogmech, home, conj, relig, affect, certain, ipron, auxverb, present and they* word categories and they least frequently use words in *anger, negemo, swear, verb, body and death* word categories. Users' conscientiousness personality trait is predicted using these attributes. As a result, CCI score is obtained as 58.232%. The prediction results are summarized for each fold in Table 10;

Table 10: Classification results in predicting user's conscientiousness

Fold	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
CCI	58.609	58.665	58.646	58.554	58.49	58.44	58.271	58.202	58.18	58.232
ICI	41.391	41.335	41.354	41.446	41.51	41.56	41.729	41.798	41.82	41.768
MAE	0.414	0.413	0.414	0.415	0.415	0.416	0.417	0.418	0.418	0.418
RMSE	0.643	0.643	0.643	0.644	0.644	0.645	0.646	0.647	0.647	0.646
AuC	0.578	0.579	0.579	0.578	0.577	0.576	0.575	0.574	0.574	0.574

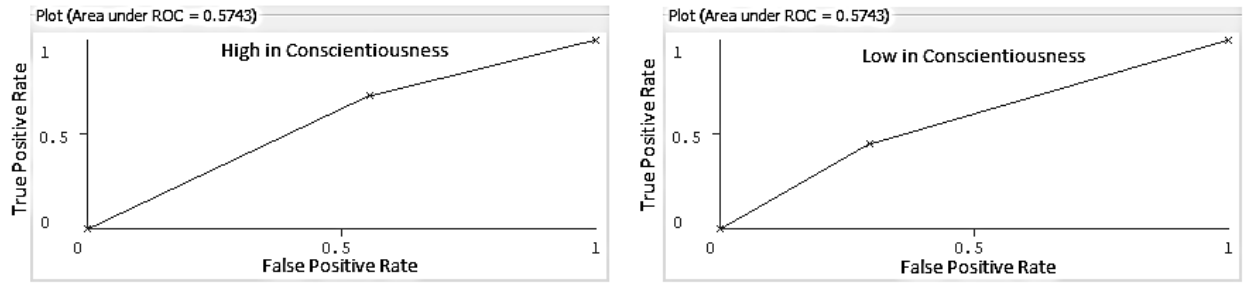


Figure 9: Roc curve for Conscientiousness

For extraversion people, they most frequently use words in *posemo, sexual, affect, incl, social, humans and bio* word categories. Users' extraversion personality trait is predicted

using these attributes. As a result, CCI score is obtained as 52.273%. The prediction results are summarized for each fold in Table 11;

Table 11: Classification results in predicting user's extraversion

Fold	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
CCI	52.318	52.184	52.208	52.219	52.276	52.253	52.242	52.247	52.237	52.273
ICI	47.683	47.816	47.792	47.781	47.724	47.747	47.758	47.753	47.762	47.727
MAE	0.477	0.478	0.478	0.477	0.477	0.478	0.478	0.478	0.478	0.477
RMSE	0.691	0.692	0.691	0.691	0.691	0.691	0.691	0.691	0.691	0.691
AuC	0.506	0.504	0.505	0.505	0.506	0.505	0.505	0.506	0.505	0.506

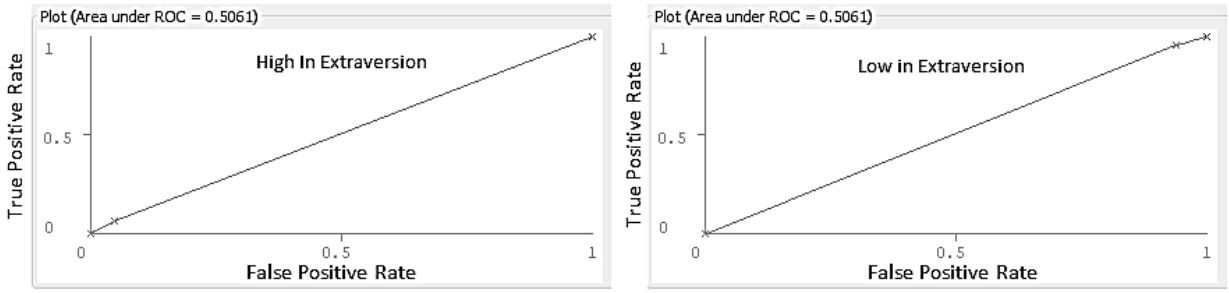


Figure 10: Roc curve for Extraversion

For neurotic people, they most frequently use words in *negemo, sad, i and anger* word categories and they least frequently use words in *leisure* category. Users' neuroticism personality trait is predicted using these attributes. As a result, CCI score is obtained as 56.207%. The prediction results are summarized for each fold in Table 12;

Table 12: Classification results in predicting user's neuroticism

Fold	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
CCI	56.21	56.21	56.21	56.209	56.208	56.208	56.207	56.207	56.207	56.207
ICI	43.79	43.79	43.79	43.791	43.792	43.792	43.793	43.793	43.793	43.793
MAE	0.438	0.438	0.438	0.438	0.438	0.438	0.438	0.438	0.438	0.438
RMSE	0.662	0.662	0.662	0.662	0.662	0.662	0.662	0.662	0.662	0.662
AuC	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5

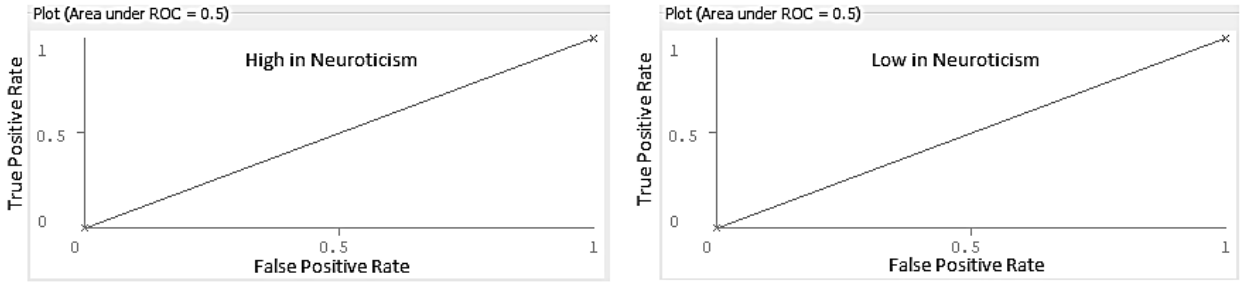


Figure 11: Roc curve for Neuroticism

Lastly, people having openness personality trait most frequently use words in *article, insight, cogmech, funct, percept, death, ipron, pronoun, auxverb, space, conj, hear, cause and tentat* word categories. They least frequently use words in *verb, family and posemo* word categories. Users' openness personality trait is predicted using these attributes. As a result, CCI score is obtained as 56.209%. The prediction results are summarized for each fold in Table 13;

Table 13: Classification results in predicting user's openness

Fold	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
CCI	56.279	56.339	56.302	56.358	56.22	56.154	56.211	56.223	56.219	56.209
ICI	43.721	43.661	43.698	43.642	43.78	43.846	43.789	43.777	43.781	43.791
MAE	0.437	0.437	0.437	0.436	0.438	0.439	0.438	0.438	0.438	0.438
RMSE	0.661	0.661	0.661	0.661	0.662	0.662	0.662	0.662	0.662	0.662
AuC	0.531	0.532	0.532	0.533	0.531	0.528	0.530	0.531	0.530	0.529

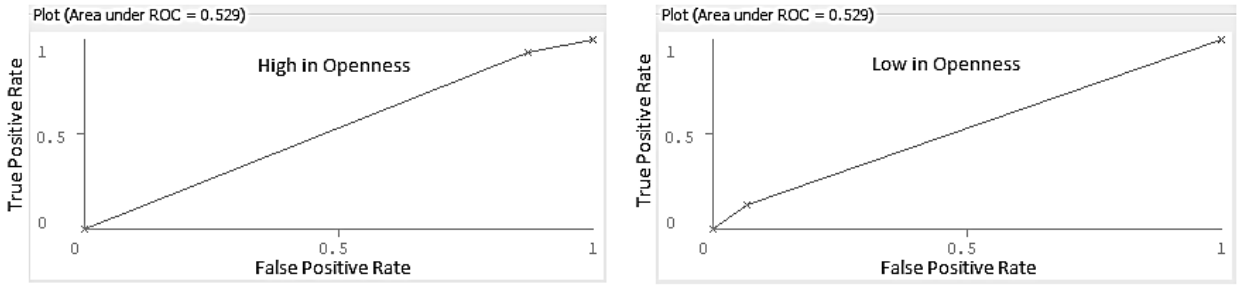


Figure 12: Roc curve for Neuroticism

To sum up, people having openness personality can be predicted more successfully compared to other personality traits. The prediction results are summarized for each fold as in Table 14;

Table 14: Personality prediction performances in each trait for Experiment 1

	MAE	RMSE	TP Rate	FP Rate	Precision	Recall	F-Measure	AuC
Agreeableness	0.43	0.66	0.569	0.447	0.572	0.569	0.552	0.562
Conscientiousness	0.42	0.64	0.582	0.433	0.58	0.582	0.575	0.574
Extraversion	0.48	0.69	0.523	0.511	0.528	0.523	0.402	0.506
Neuroticism	0.44	0.66	0.562	0.562	0.316	0.562	0.404	0.5
Openness	0.44	0.66	0.562	0.5	0.575	0.562	0.483	0.529

Discussion: In this experiment we observed that there were significant correlations between LIWC tags and personality traits as in base methodologies claim. However, when we used these correlated tags in prediction, the results were not so successful. There were too many users whose personality could not be predicted correctly.

As declared in the objectives, there are existing studies in personality prediction using linguistic features of social media with different or same dataset. Table 15 shows comparison of our study with existing studies on personality prediction performances.

Table 15: Comparison of personality prediction performances with existing studies

	Our Study TP Rate	[16] **TP Rate**	[11] **TP Rate**	[9] **TP Rate**	[8] **TP Rate**
	10-fold cross-validation	**10-fold cross-validation**	**10-fold cross-validation**	**10-fold cross-validation**	**66% training 33% test**
Agreeableness	0.569	0.482	0.528	0.584	0.86
Conscientiousness	0.582	0.595	0.524	0.58	0.92
Extraversion	0.523	0.553	0.576	0.575	0.928
Neuroticism	0.562	0.531	0.448	0.569	0.864
Openness	0.562	0.653	0.548	0.575	0.948

If we compare the prediction performances, the prediction performances are similar with existing studies except the study of Markovikj et al. In their study, they have used different linguistic tools and dictionaries such as General Inquirer Tool [42] and AFINN Words [43]. These tools and dictionaries may increase the performance of prediction model. Since we do

not have any content of status updates, we cannot verify how prediction performances change considering these tools and dictionaries.

Since we have used k-fold cross validation in our prediction model, it reduces the variance while increasing the bias [44] [45]. This would increase our confidence dramatically in the reliability of the model performance, because we have multiple tests, which are at least slightly different.

4.3.2 Experiment 2

Objective: When we inspected the data set, we realized that there were 12% of users who had shared a maximum of ten posts. In this experiment, we investigate whether there is an effect of the amount of personal information on accurately predicting users' personality. In this experiment, we would like to test our improvement which claims that discarding users having a few status updates increases prediction performances of the proposed model.

Dataset: In this experiment we have used BIG5_LIWC dataset.

Methodology: In this experiment, we firstly calculated some basic statistical information for number of status updates to determine the threshold values. In BIG5_LIWC dataset, the minimum and maximum number of status updates is 0 and 2450 respectively. Moreover, the average number of status updates is approximately 142 and standard deviation value is 163.

Figure 13 shows the number of users versus the number of status updates.

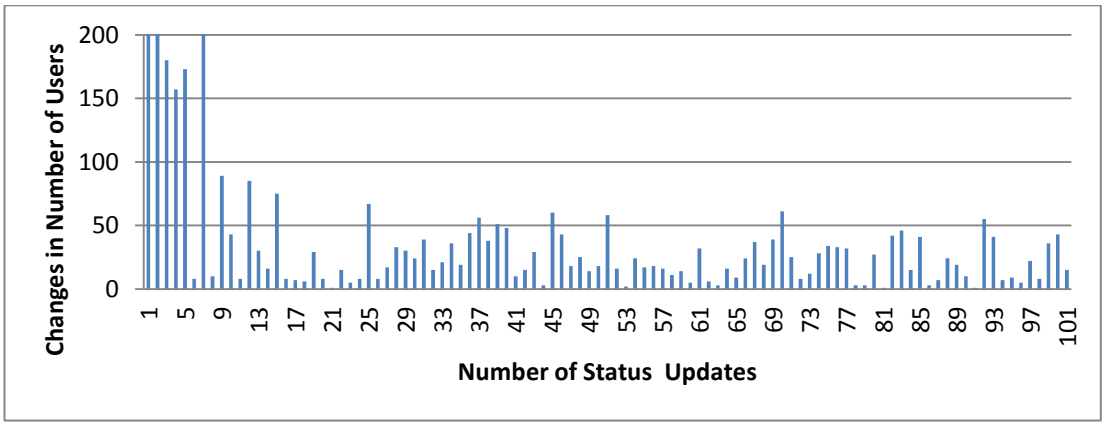


Figure 13: Changes in number of users while number of status updates increases

By observing the cut-off points, 10, 25, 50 and 100 are selected as threshold values. Table 16 shows the threshold values and the number of users that satisfy the corresponding criteria;

Table 16: Threshold values for number of status updates

Number of Status Updates	Number of Users
≥ 0	115,863
≥ 10	104,514
≥ 25	93,313
≥ 50	78,659
≥ 100	56,213

During the experiment we have used the same methodology described in Section 4.3.1. In each one of experiment iterations, we have discarded users whose number of status updates is less than a threshold value and constructed the SMO using the remaining users.

Results: Based on the data set which includes users having 10 or more status updates, the proposed model produced the results depicted in Table 17;

Table 17: Personality prediction in each trait when users have 10 or more status updates

	MAE	RMSE	TP Rate	FP Rate	Precision	Recall	F-Measure	AuC
Agreeableness	0.43	0.65	0.575	0.441	0.576	0.575	0.562	0.567
Conscientiousness	0.41	0.64	0.595	0.411	0.594	0.595	0.594	0.592
Extraversion	0.44	0.66	0.566	0.438	0.567	0.566	0.561	0.564
Neuroticism	0.44	0.66	0.56	0.56	0.314	0.56	0.402	0.5
Openness	0.41	0.64	0.587	0.462	0.588	0.587	0.554	0.563

Table 18: Personality prediction in each trait when users have 25 or more status updates

	MAE	RMSE	TP Rate	FP Rate	Precision	Recall	F-Measure	AuC
Agreeableness	0.42	0.65	0.577	0.439	0.578	0.577	0.566	0.569
Conscientiousness	0.4	0.64	0.599	0.404	0.599	0.599	0.599	0.597
Extraversion	0.43	0.65	0.57	0.43	0.571	0.57	0.569	0.57
Neuroticism	0.44	0.67	0.557	0.557	0.31	0.557	0.398	0.5
Openness	0.4	0.63	0.6	0.448	0.598	0.6	0.576	0.576

Based on the data set which includes users having 25 or more status updates, the proposed model produced the results depicted in Table 18;

Based on the data set which includes users having 50 or more status updates, the proposed model produced the results depicted in Table 19;

Table 19: Personality prediction in each trait when users have 50 or more status updates

	MAE	RMSE	TP Rate	FP Rate	Precision	Recall	F-Measure	AuC
Agreeableness	0.42	0.65	0.581	0.433	0.581	0.581	0.571	0.574
Conscientiousness	0.4	0.63	0.602	0.4	0.602	0.602	0.602	0.601
Extraversion	0.42	0.65	0.575	0.425	0.575	0.575	0.574	0.575
Neuroticism	0.44	0.66	0.561	0.531	0.57	0.561	0.449	0.515
Openness	0.4	0.63	0.604	0.447	0.601	0.604	0.581	0.578

Based on the data set which includes users having 100 or more status updates, the proposed model produced the results depicted in Table 20;

Table 20: Personality prediction in each trait when users have 100 or more status updates

	MAE	RMSE	TP Rate	FP Rate	Precision	Recall	F-Measure	AuC
Agreeableness	0.42	0.65	0.583	0.429	0.583	0.583	0.575	0.577
Conscientiousness	0.4	0.63	0.604	0.396	0.604	0.604	0.604	0.604
Extraversion	0.42	0.64	0.583	0.421	0.582	0.583	0.583	0.581
Neuroticism	0.42	0.65	0.584	0.46	0.583	0.584	0.555	0.562
Openness	0.39	0.62	0.611	0.446	0.607	0.611	0.587	0.582

Discussion: If we compare the prediction results from the tables, it can be seen that while increasing evidences (number of status updates), the prediction results are getting better although the improvement is small.

Table 21 summarizes correctly classified instances of each one of iterations during the experiment;

Table 21: Summary of personality prediction in TP Rate by number of shared status updates thresholds

	≥ 0	≥ 10	≥ 25	≥ 50	≥ 100
Agreeableness	0.569	0.575	0.577	0.581	0.583
Conscientiousness	0.582	0.595	0.599	0.602	0.604
Extraversion	0.523	0.566	0.57	0.575	0.583
Neuroticism	0.562	0.56	0.557	0.561	0.584
Openness	0.562	0.587	0.6	0.604	0.611

How prediction performance is improved in proportion to the number of status updates depicted in Figure 14.

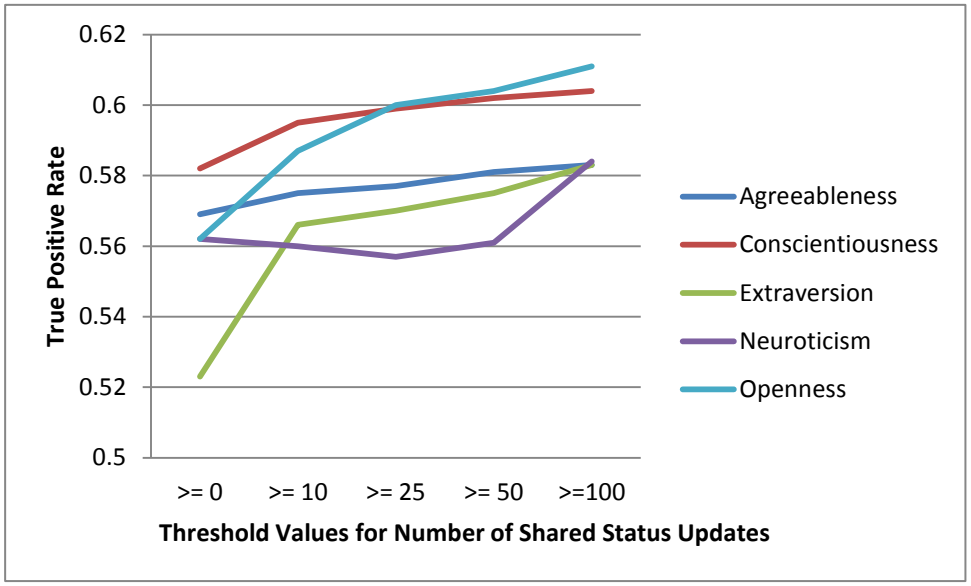


Figure 14: Changes in prediction performances for each personality trait

In addition to prediction results, t-test for Paired Two Sample is applied using MAE and RMSE of each folds of iterations to decide whether changes in prediction performance is significant or not when the number of shared status updates increase.

Table 22 shows the p-values (two-tailed) of t-test applied using RMSEs. Where p-values equal or less than 0.05 denotes that changes in RMSE is significant.

Table 22: Iteration comparison based on RMSEs applying t-test for Paired Two Sample (df = 9, r: pearson correlation coefficient)

	≥0 vs. ≥10		≥10 vs. ≥25		≥25 vs. ≥50		≥50 vs. ≥100	
	r	**Increment**	**r**	**Increment**	**r**	**Increment**	**r**	**Increment**
Agreeableness	**-0.166**	%0.3	**0.603**	%0.3	**0.485**	%0.1	**0.427**	%0.2
Conscientiousness	**-0.665**	%0.7	**0.922**	%0.3	**-0.499**	%0.5	-0.159	-%0.1
Extraversion	**0.538**	%3	**-0.301**	%0.7	**0.0426**	%0.3	**-0.762**	%0.7
Neuroticism	**1**	-%0.1	0	-%0.2	**-1.3E-13**	%0.3	**0.507**	%1.7
Openness	**-0.226**	%1.8	**0.261**	%1	**0.778**	%0.2	**0.364**	%0.7

As can be seen in Table 22, the larger the number of posts is, the higher the accuracy of the model is.

Therefore the obtained results satisfy our improvement "*The personality traits of a user whose status updates are few cannot be identified accurately as there is no or very limited information. Incorporation of such cases in a supervised model degrades the predictive performance of the model*".

4.3.3 Experiment 3

Objective: In this experiment we aim to evaluate how prediction accuracy improves if we use demographic information, Facebook activities and social network attributes [4] [14] [17]. With the results of this experiment, we would like to verify the base methodologies in the literature which have used these attributes in prediction.

Dataset: In this experiment we have used BIG5_LIWC_FREQ_DEMOG_SNA dataset.

Methodology: In this experiment, we have calculated Pearson Correlations between all attributes in the dataset and their corresponding personality trait scores.

In the preprocessing phase, for each personality, we obtain correlated attributes (r values are higher than 0.05, p values are 0) using Pearson Correlation table given in

Appendix B: Pearson Correlation Table for Experiment 3. Then remaining attributes that are under the determined correlation values are filtered out from the proposed model.

In the classification phase, we have used SMO machine learning algorithm to classify instances. Moreover, 10 fold cross validation method is applied.

Results: As correlation table shows, agreeableness is highly correlated with gender of user. There is no significant correlation between agreeableness and social network attributes. These people most frequently use words in *funct, we, ipron, article, auxverb, past, present, adverb, preps, conj, quant, social, family, friend, affect, posemo, cogmech, incl, percept, see, relativ, motion, space, time, achieve, leisure, home and relig* word categories. They least frequently use words in *swear, negemo, anger, body and death* word categories. As a result, CCI score is obtained as 59.128%. The prediction results are summarized for each fold in Table 23;

Table 23: Classification results in predicting user's agreeableness scores

Fold	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
CCI	58.835	59.183	59.124	58.951	59.1	59.193	59.405	59.236	59.157	59.128
ICI	41.165	40.814	40.876	41.049	40.9	40.807	40.595	40.764	40.843	40.872
MAE	0.412	0.408	0.409	0.411	0.41	0.408	0.406	0.408	0.408	0.409
RMSE	0.642	0.639	0.639	0.641	0.641	0.639	0.637	0.639	0.639	0.639
AuC	0.57	0.574	0.572	0.571	0.572	0.574	0.576	0.574	0.574	0.572

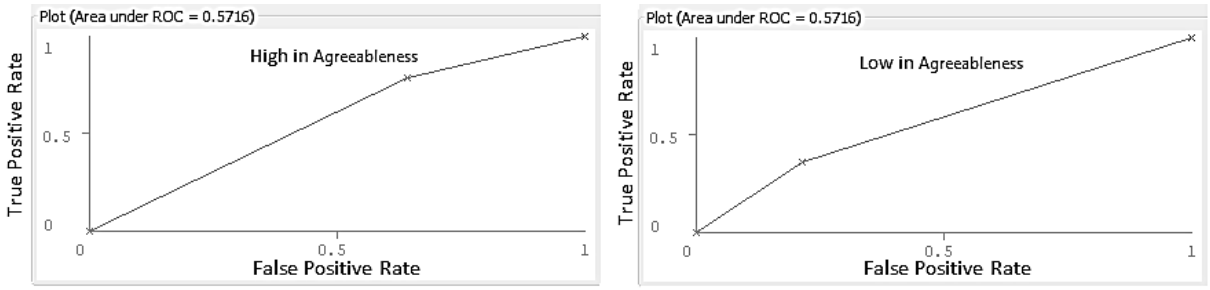


Figure 15: Roc curve for Agreeableness

Secondly, people having conscientiousness personality are highly correlated with number of like, concentration, group and education. These people most frequently use words in *funct, we, they, ipron, article, auxverb, present, preps, conj, quant, number, social, family, friend, affect, posemo, cogmech, certain, inhib, incl, motion, space, time, work, achieve, leisure,*

home, money and relig word categories. They least frequently use words in *verb, swear, negemo, anger, sad, body, sexual, relativ and death* word categories. As a result, CCI score is obtained as 61.263%. The prediction results are summarized for each fold in Table 24;

Table 24: Classification results in predicting user's conscientiousness scores

Fold	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
CCI	61.509	61.834	61.513	61.495	61.456	61.55	61.493	61.302	61.318	61.263
ICI	8.491	38.166	38.487	38.505	38.544	38.45	38.507	38.698	38.682	38.737
MAE	0.385	0.382	0.385	0.385	0.385	0.385	0.385	0.387	0.387	0.387
RMSE	0.62	0.618	0.62	0.621	0.621	0.62	0.621	0.622	0.622	0.622
AuC	0.614	0.618	0.614	0.614	0.614	0.615	0.614	0.612	0.612	0.612

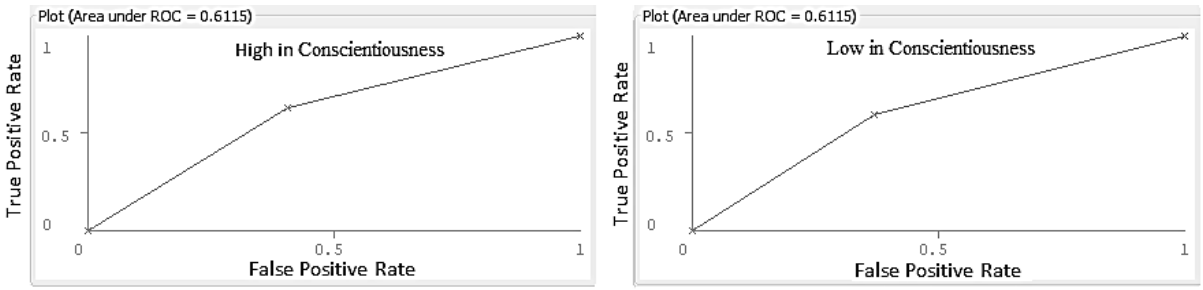


Figure 16: Roc curve for Conscientiousness

Thirdly, extraversion people are highly correlated with betweenness, brokerage and transitivity features in social network. These people have larger network size compared to others. They can be friends with other people easily. They are commonly tagged in photos and statuses by other people. They are quite popular. They most frequently use words in *social, friend, humans, affect, posemo, incl, bio, sexual and leisure* word categories. They least frequently use words in *negemo, insight, cause, tentat and death* word categories. As a result, CCI score is obtained as 61.144%. The prediction results are summarized for each fold in Table 25;

Table 25: Classification results in predicting user's extraversion

Fold	1	2	3	4	5	6	7	8	9	10
CCI	60.841	61.046	60.844	60.982	61.16	61.335	61.009	60.884	61.015	61.144
ICI	39.159	38.954	39.156	39.018	38.84	38.665	38.991	39.116	38.985	38.856
MAE	0.392	0.39	0.392	0.39	0.388	0.387	0.39	0.391	0.39	0.389
RMSE	0.626	0.624	0.626	0.625	0.623	0.622	0.624	0.625	0.624	0.623
AuC	0.608	0.61	0.608	0.609	0.611	0.613	0.609	0.608	0.609	0.611

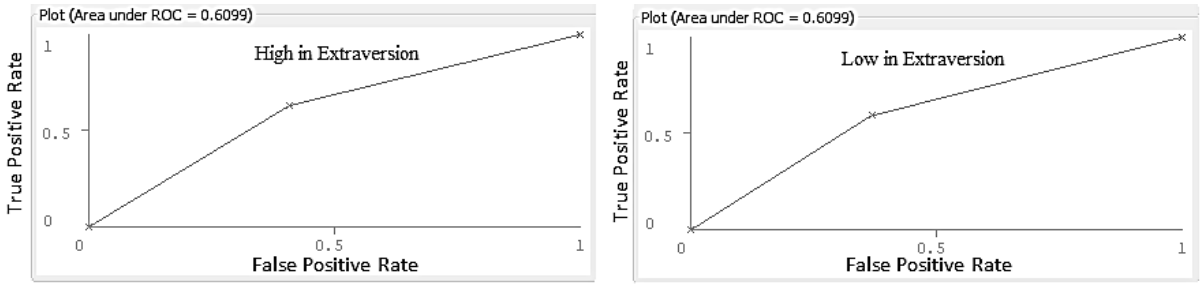
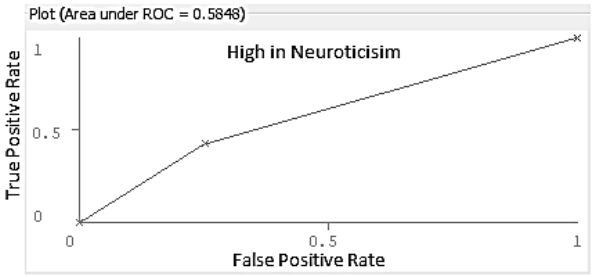


Figure 17: Roc curve for Extraversion

Fourthly, neurotic people are highly correlated with betweenness and brokerage role in social network. These people have larger network size compared to others. They can be friends with other people easily. They incline to like other people posts and share their status updates. "Gender" and "Interested in" attributes are other indicators for neurotic people. They most frequently use words in *i, negate, negemo, anx, anger, sad, discrep, bio, body and health* word categories. They least frequently use words in *article, relativ, space, work, achieve, leisure and relig* word categories. As a result, CCI score is obtained as 60.093%. The prediction results are summarized for each fold in Table 26;

Table 26: Classification results in predicting user's neuroticism scores

Fold	1	2	3	4	5	6	7	8	9	10
CCI	61.366	60.783	60.557	60.349	59.975	60.076	60.135	60.12	60.176	60.093
ICI	38.634	39.217	39.443	39.651	40.025	39.924	39.865	39.88	39.824	39.907
MAE	0.386	0.392	0.394	0.397	0.4	0.399	0.399	0.399	0.398	0.399
RMSE	0.622	0.626	0.628	0.63	0.633	0.632	0.631	0.632	0.631	0.632
AuC	0.596	0.591	0.589	0.587	0.583	0.585	0.585	0.585	0.586	0.585



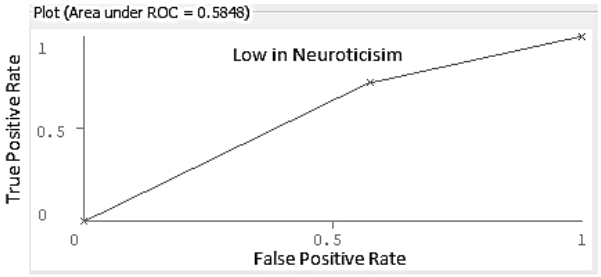
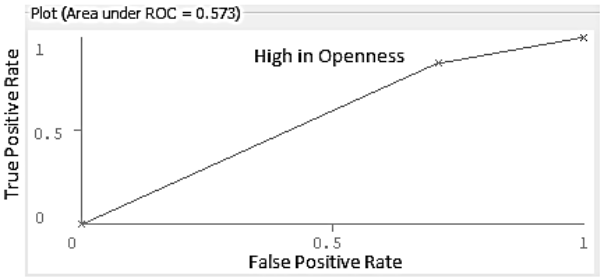


Figure 18: Roc curve for Neuroticism

Lastly, people high in openness are highly correlated with number of interests and groups. There is no significant correlation between openness and social network attributes such as betweenness, transitivity and brokerage. They most frequently use words in *funct, pronoun, ppron, you, they, ipron, article, auxverb, future, conj, number, anx, cogmech, insight, cause, tentat, certain, incl, excl, percept, see, hear, body, space and death* word categories. They least frequently use words in *verb, family, posemo, time and home* word categories. As a result, CCI score is obtained as 61.746%. The prediction results are summarized for each fold in Table 27;

Table 27: Classification results in predicting user's openness scores

Fold	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
CCI	62.799	61.333	61.513	61.758	61.647	61.637	61.746	61.678	61.705	61.746
ICI	37.201	38.667	38.487	38.242	38.353	38.363	38.254	38.322	38.295	38.254
MAE	0.372	0.387	0.385	0.382	0.384	0.384	0.383	0.383	0.383	0.383
RMSE	0.61	0.622	0.62	0.618	0.619	0.619	0.619	0.619	0.619	0.619
AuC	0.588	0.57	0.571	0.573	0.572	0.573	0.573	0.573	0.573	0.573



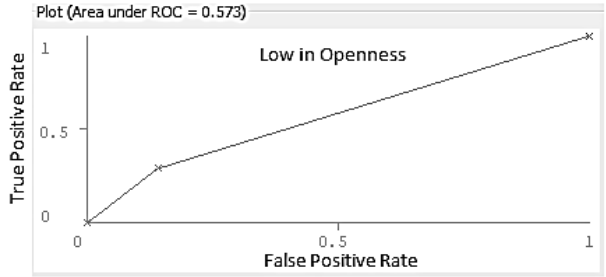


Figure 19: Roc curve for Openness

To sum up, in this experiment we observed that the trait scores of openness people can be predicted more successfully compared to other personality traits. People high in agreeableness trait are difficult to infer according to the results. Table 28 depicts the classification results of each prediction model for each personality trait;

Table 28: Personality prediction performances in each trait for Experiment 3

	MAE	RMSE	TP Rate	FP Rate	Precision	Recall	F-Measure	AuC
Agreeableness	0.41	0.64	0.591	0.445	0.59	0.591	0.571	0.573
Conscientiousness	0.39	0.62	0.613	0.389	0.613	0.613	0.613	0.612
Extraversion	0.39	0.62	0.611	0.39	0.612	0.611	0.611	0.611
Neuroticism	0.4	0.63	0.601	0.431	0.596	0.601	0.591	0.585
Openness	0.38	0.62	0.617	0.47	0.612	0.617	0.582	0.574

Discussion: If we compare these results with the previous experiment results, the predictive performance in all Big 5 Personality Traits are more successful than the previous experiment. That shows us that demographic information, Facebook activities, and social network attributes help to increase the prediction performance of models.

4.3.4 Experiment 4

Objective: In this experiment we would like to compare the performance of different machine learning algorithms using the same dataset. These machine learning algorithms are J48, SMO and Random Forest.

Dataset: In this experiment we have used BIG5_LIWC_FREQ_DEMOG_SNA dataset.

Method: In this experiment we have calculated Pearson Correlation values and eliminated uncorrelated attributes in the preprocessing section as we did in the previous experiment. However, we have used three different machine learning algorithms for classification. These algorithms are J48, SMO and Random Forest machine learning algorithm with default settings [44] [49] [50] depicted in Table 29. The experiment was repeated for each algorithm. In each one of iterations, 10 fold cross validation method is applied.

Table 29: Suggested configurations for classification algorithms

Algorithm	Settings
J48	The confidence factor used for pruning: 0.25 The minimum number of instances per leaf: 2 Amount of data used for reduced-error pruning: 3 Whether reduced-error pruning is used instead of C.4.5 pruning: false Whether to consider the subtree raising operation when pruning: true Whether pruning is performed: false Whether counts at leaves are smoothed based on Laplace: false
Random Forest	The maximum depth of the trees: unlimited The number of trees to be generated: 10
SMO	The complexity parameter C: 1.0 The epsilon for round-off error: 1.0E-12 Data transformation: Normalize training data Kernel: PolyKernel The tolerance parameter: 0.0010

Results: Using different machine learning algorithms, prediction performances are shown in Table 30;

Table 30: Percentage of correctly classified instance by machine learning algorithms

	Agreeableness	Conscientiousness	Extraversion	Neuroticism	Openness
Random Forest	54.436 %	56.318 %	58.072 %	56.395 %	57.427 %
J48	56.371 %	54.551 %	59.419 %	57.608 %	57.413 %
SMO	59.128 %	61.263 %	61.144 %	60.093 %	61.746 %

T-test for Paired Two Sample is applied using MAE and RMSE of each folds of iterations to decide whether changes in prediction performance is significant or not.

Table 31 shows the p-values (two-tailed) of t-test on RMSEs. Where p-values equal or less than 0.05 denotes that changes in RMSE is significant.

Table 31: Machine learning comparison using t-test for Paired Two Sample using RMSEs of each run using different algorithms. (df = 9, r: pearson correlation coefficient)

	Random Forest vs. J48		J48 vs. SMO		Random Forest vs. SMO	
	r	**Increment**	**r**	**Increment**	**r**	**Increment**
Agreeableness	**-0.844**	%10	0.433	%.28	**-0.453**	%12.8
Conscientiousness	**-0.132**	-%1.8	**0.73**	%13.2	**-0.107**	%11.4
Extraversion	-0.8	%12.3	**0.525**	%0.1	**-0.524**	%12.4
Neuroticism	**-0.782**	%11.4	**0.753**	%0.4	**-0.98**	%11.8
Openness	**-0.58**	%6.1	**-0.603**	%5.7	**0.261**	%11.8

As can be seen from Table 31, switching algorithm from Random Forest to J48 significantly affects MAE and RMSE values and increases prediction performances most of the time. Moreover, switching algorithm from J48 to SMO always changes MAE and RMSE significantly and increases prediction performances.

Discussion: There are too many machine learning algorithms which can be used to create models. We have used three popular algorithms to compare prediction performances. Each machine learning algorithm performed different results. We received the best outcomes from SMO Algorithm.

4.3.5 Experiment 5

Objective: In this experiment, we aim to test our first hypothesis which asserts that there is a relation between users' and couples' Big 5 Personality Trait scores and the use of this relation in personality inference increases the performance of the proposed model.

Dataset: In this experiment we have used BIG5_LIWC_FREQ_DEMOG_SNA_COUPLE dataset.

Methodology: For each personality trait, Pearson Correlation values (p) are calculated between couple's personality. Table 32 shows the correlation between the scores of the personality traits of couples;

Table 32: Pearson Correlation Matrix of Big 5 Personality Traits for Couples. Statistically significant correlations (p < 0.05) are bolded. (r: pearson correlation coefficient, t: t-distribution value)

		Second User									
		OPE		CON		EXT		AGR		NEU	
Mean±StdDev		3.95±0.66		3.49±0.73		3.49±0.82		3.53±0.7		2.72±0.83	
Correlation		R	t	r	t	r	t	r	t	r	t
First User	OPE	**0.117**	5.566	-0.037	1.748	-0.037	1.753	0.008	0.393	**0.043**	2.058
	CON	**-0.037**	1.748	**0.148**	7.075	**0.051**	2.402	**0.064**	3.053	**-0.113**	5.373
	EXT	-0.037	1.753	**0.051**	2.402	**0.115**	5.477	0.015	0.689	**-0.066**	3.119
	AGR	0.0080	0.393	**0.064**	3.053	**0.015**	0.689	0.038	1.803	-0.012	0.555
	NEU	**0.043**	2.058	**-0.113**	5.373	**-0.066**	3.119	-0.012	0.555	-0.04	1.911

Individuals usually prefer people exhibiting similar Big 5 Personality Traits in couple selection except neurotic and agreeable people. Conscientious individuals are the most preferable individuals in a relationship. In spite of that, agreeable individuals are the least preferable individuals in a relationship. Conscientious and neurotic individuals mostly select each other as couple.

In the preprocessing phase, for each personality trait, we select the highly correlated attributes (r values are higher than 0.05, p values are 0) using Pearson Correlation table given in Appendix C: Pearson Correlation Table for Experiment 5. Then the remaining attributes that are under the determined correlation values are filtered out from the proposed model.

In the classification phase, we have used SMO machine learning algorithm with 10 fold cross validation.

Results: Prediction results for each personality traits are listed in Table 33.

Table 33: Personality prediction performances in each trait for Experiment 6

	MAE	RMSE	TP Rate	FP Rate	Precision	Recall	F-Measure	AuC
Agreeableness	0.44	0.66	0.568	0.481	0.569	0.568	0.524	0.544
Conscientiousness	0.39	0.63	0.617	0.429	0.613	0.617	0.602	0.594
Extraversion	0.43	0.66	0.563	0.444	0.566	0.563	0.552	0.56
Neuroticism	0.39	0.63	0.625	0.375	0.625	0.625	0.625	0.625
Openness	0.36	0.6	0.636	0.636	0.404	0.636	0.494	0.5

To compare the prediction results with Experiment 3;

Table 34: Iteration comparison based on RMSEs applying t-test for Paired Two Sample (df = 9, r: pearson correlation coefficient)

	Exp 3	Exp 5	Exp 3 vs. Exp 5	
	TP Rate	TP Rate	t-test for RMSE	
			r	Increment
Agreeableness	0.591	0.568	**0.084**	-%2.5
Conscientiousness	0.613	0.617	-0.316	-%0.7
Extraversion	0.611	0.563	**-0.376**	-%3.6
Neuroticism	0.601	0.625	-0.557	%0.1
Openness	0.617	0.636	**0.073**	%1.6

Discussion: As it can be seen from the tables, people generally tend to select their spouses who have similar Big 5 Personality Traits with them. However, people high in agreeable and neuroticism personality traits usually select open people to start a relationship.

Moreover, as it can be seen in Table 34, the prediction performances are significantly decreased in agreeableness and extraversion personality traits. The personality prediction in conscientiousness and neuroticism are increased but the changes are not significant. However, the personality prediction performance is significantly increased in openness personality trait. Therefore; our first hypothesis "*There is a relation between users' personality traits and their spouses' personality traits. Incorporation of spouses' personality in a supervised model improves the predictive performance of the model*" is not satisfied for all the Big 5 Personality Traits.

4.3.6 Experiment 6

Objective: In this experiment we study the similarities between the personality traits of friends. We aim to improve our prediction model using these similarities if they exist. We also aim to evaluate our second and third hypothesis about personality similarities between friends.

Dataset: In this experiment we have used BIG5_LIWC_FREQ_DEMOG_SNA_FRIENDSHIP dataset.

Methodology: For each personality trait, Pearson Correlation values (p) between each user personality and his/her friends' personality are calculated. Table 35 shows the correlation between the scores of the personality traits of two users;

Table 35: Pearson Correlation Matrix of the Scores of Big 5 Personality Traits for Friendships. The statistically significant correlations (p < 0.05) are bolded. (r: pearson correlation coefficient, t: t-distribution value)

		Second User									
		OPE		**CON**		**EXT**		**AGR**		**NEU**	
Mean±StdDev		3.88±0.65		3.41±0.71		3.71±0.77		3.6±0.68		2.73±0.79	
Correlation		**r**	**t**	**r**	**t**	**r**	**t**	**r**	**t**	**r**	**t**
First User	**OPE**	**0.102**	26.823	**0.01**	2.511	0.0020	0.628	0.004	0.927	-0.005	1.184
	CON	**0.01**	2.511	**0.065**	17.042	0.015	3.839	**0.029**	7.547	**-0.04**	10.583
	EXT	0.0020	0.628	**0.015**	3.839	**0.06**	15.783	0.006	1.627	-0.023	6.079
	AGR	0.004	0.927	**0.029**	7.547	0.006	1.627	**0.043**	11.324	**-0.019**	4.963
	NEU	-0.005	1.184	**-0.04**	10.583	**-0.023**	6.079	**-0.019**	4.963	**0.047**	12.326

Individuals usually prefer people having similar Big 5 Personality Traits in friend selection. Open people mostly have friends who are high in openness and conscientiousness personality traits. Conscientious individuals are the most preferable individuals in friend selection. It denotes that they are the most appropriate people to get on well with.

In the preprocessing phase, for each personality trait, we select the highly correlated attributes (r values are higher than 0.05, p values are 0) using Pearson Correlation table given

in Appendix D: Pearson Correlation Table for Experiment 6. Then the remaining attributes that are under determined correlation values are filtered out from the proposed model.

In the classification phase, we have used SMO machine learning algorithm with 10 fold cross validation.

This experiment' settings are as follows:

A. Using Friends' Personality Traits without Any Filtering

1. Pre-process the social network so as to have the following settings:
   - ✓ ***Setting 1:*** Friends where both of them have 30 friends or more.
2. Find two friends who have the most common friends with each other according to below indices.

$$Common\ Friends\ Indices = \frac{\begin{matrix}Number\\of\\Common\ Friend\end{matrix}}{\begin{matrix}Number\\of\\First\ User's\ Friends\end{matrix}} + \frac{\begin{matrix}Number\\of\\Common\ Friend\end{matrix}}{\begin{matrix}Number\\of\\Second\ User's\ Friends\end{matrix}}$$

3. Give the selected friend's real personality trait as an input to SMO together with the other correlated attributes.

B. Using Friends' Personality Traits with Filtering by Number of Shared Friends:

1. Pre-process the social network so as to have the following settings:
   - ✓ ***Setting 2:*** Friends where both of them have 30 friends or more. They have 10 friends in common or more.
   - ✓ ***Setting 3:*** Friends where both of them have 30 friends or more. They have 25 friends in common or more.
   - ✓ ***Setting 4:*** Friends where both of them have 30 friends or more. They have 50 friends in common or more.
2. Find two friends who have the most common friends with each other according to Common Friend Indices.
3. Give the selected friend's real personality trait as an input to SMO together with the other correlated attributes.

C. Using Friends' Personality Traits with Filtering by Number of Shared Likes:

1. Pre-process the social network so as to have the following settings:

- ✓ ***Setting 5:*** Friends where both of them have 30 friends or more. They have 10 likes in common or more.
- ✓ ***Setting 6:*** Friends where both of them have 30 friends or more. They have 25 likes in common or more.
- ✓ ***Setting 7:*** Friends where both of them have 30 friends or more. They have 50 likes in common or more.

2. Find two friends who have the most likes friends with each other according to below formula:

$$Common\ Likes\ Indices = \frac{\begin{matrix}Number\\of\\Common\ Like\end{matrix}}{\begin{matrix}Number\\of\\First\ User's\ Likes\end{matrix}} + \frac{\begin{matrix}Number\\of\\Common\ Like\end{matrix}}{\begin{matrix}Number\\of\\Second\ User's\ Likes\end{matrix}}$$

3. Give the selected friend's real personality trait as an input to SMO together with the other correlated attributes.

Results: As a result of this experiment 35 different prediction models (7 experiment settings for each Big 5 Personality Trait) are created and evaluated.

Firstly, the prediction model using the setting 1 is constructed and tested to see how the prediction performances change with using friends' Big 5 Personality Trait scores. In accordance with this purpose,

Table 36 shows the results of prediction models for each personality using Setting 1;

Table 36: Summary of personality prediction using all friends' information

	MAE	RMSE	TP Rate	FP Rate	Precision	Recall	F-Measure	AuC
Agreeableness	0.38	0.62	0.618	0.411	0.616	0.618	0.608	0.603
Conscientiousness	0.39	0.63	0.613	0.388	0.613	0.613	0.612	0.612
Extraversion	0.36	0.6	0.639	0.396	0.636	0.639	0.631	0.622
Neuroticism	0.42	0.65	0.582	0.421	0.582	0.582	0.581	0.581
Openness	0.36	0.6	0.639	0.431	0.635	0.639	0.617	0.604

Secondly, the prediction model using setting 2,3,4 is constructed and tested to see how prediction performances changes while eliminating friendships where there is not enough number of common friends according to the threshold value.

Table 37 shows the results of prediction models for each personality using Setting 2;

Table 37: The summary of the models' performances which utilize the information of friends whose number of common friends are greater and equal to 10.

	MAE	RMSE	TP Rate	FP Rate	Precision	Recall	F-Measure	AuC
Agreeableness	0.39	0.62	0.617	0.418	0.615	0.617	0.604	0.599
Conscientiousness	0.39	0.63	0.608	0.393	0.608	0.608	0.607	0.607
Extraversion	0.36	0.6	0.64	0.404	0.635	0.64	0.629	0.618
Neuroticism	0.42	0.65	0.581	0.422	0.581	0.581	0.579	0.58
Openness	0.36	0.6	0.637	0.415	0.635	0.637	0.621	0.611

Table 38 shows the results of prediction models for each personality using Setting 3.

Table 38: The summary of the models' performances which utilize the information of friends whose number of common friends are greater and equal to 25.

	MAE	RMSE	TP Rate	FP Rate	Precision	Recall	F-Measure	AuC
Agreeableness	0.37	0.61	0.624	0.414	0.622	0.624	0.611	0.605
Conscientiousness	0.4	0.63	0.604	0.396	0.604	0.604	0.604	0.604
Extraversion	0.35	0.59	0.654	0.409	0.649	0.654	0.639	0.623
Neuroticism	0.4	0.64	0.599	0.402	0.6	0.599	0.597	0.598
Openness	0.36	0.6	0.639	0.398	0.637	0.639	0.628	0.621

Table 39 shows the results of prediction models for each personality traits using Setting 4.

Table 39: The summary of the models' performances which utilize the information of friends whose number of common friends are greater and equal to 50.

	MAE	RMSE	TP Rate	FP Rate	Precision	Recall	F-Measure	AuC
Agreeableness	0.38	0.62	0.62	0.424	0.618	0.62	0.604	0.598
Conscientiousness	0.39	0.63	0.609	0.395	0.61	0.609	0.607	0.607
Extraversion	0.34	0.59	0.659	0.433	0.65	0.659	0.639	0.613
Neuroticism	0.39	0.63	0.612	0.387	0.613	0.612	0.612	0.612
Openness	0.36	0.6	0.637	0.397	0.635	0.637	0.627	0.62

To sum up the predictions results by the number of shared friends in Table 40;

Table 40: Summary of personality prediction results in TP Rate by the number of common friends thresholds (df = 9)

	Experiment 3	Experiment 6			
Number of Common Friends	**-**	**≥ 0**	**≥ 10**	**≥ 25**	**≥ 50**
Agreeableness	0.591	0.618	0.617	0.624	0.62
Conscientiousness	0.613	0.613	0.608	0.604	0.609
Extraversion	0.611	0.639	0.64	0.654	0.659
Neuroticism	0.601	0.582	0.581	0.599	0.612
Openness	0.617	0.639	0.637	0.639	0.637

Figure 20 shows how the prediction performance is changed in proportion to the number of common friends.

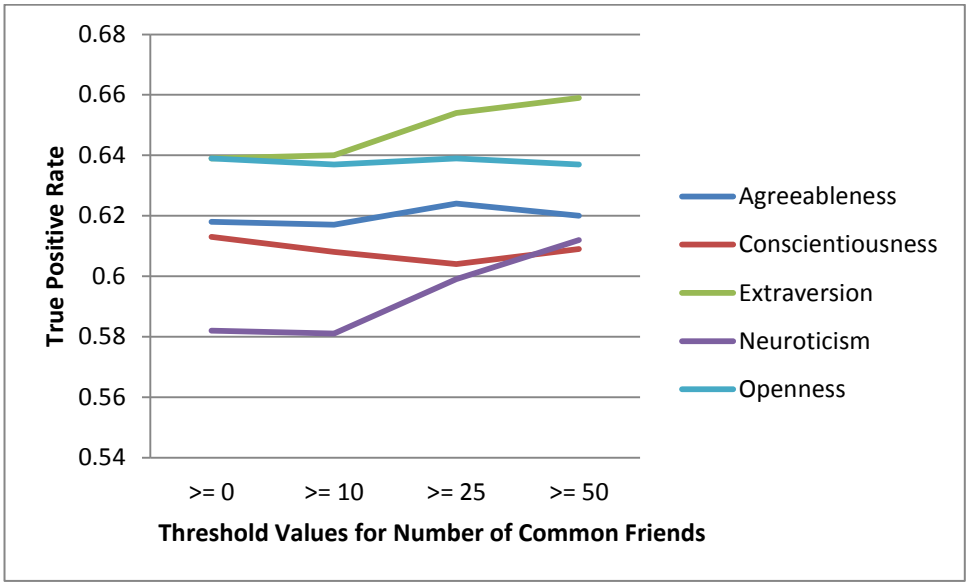


Figure 20: Changes in prediction performances by number of threshold values

T-test for Paired Two Sample is applied on MAE and RMSE of each fold of iterations to decide whether the changes in prediction performance are significant or not when the number of common friends increase.

Table 41 shows the p-values (two-tailed) of t-test on RMSEs. Where p-values are equal or less than 0.05 denotes that changes in RMSE are significant.

Table 41: Iteration comparison based on RMSEs applying t-test for Paired Two Sample (df = 9, r: pearson correlation coefficient)

	Exp 3 vs. ≥0		≥0 vs. ≥10		≥10 vs. ≥25		≥25 vs. ≥50	
	r	**Increment**	**r**	**Increment**	**r**	**Increment**	**r**	**Increment**
Agreeableness	**0.815**	%2.1	**0.647**	-%0.2	**-0.891**	%1.2	**0.96**	-%0.7
Conscientiousness	**-0.579**	-%0.6	-0.447	%0.1	**-0.171**	-%0.5	**0.256**	%0.5
Extraversion	**0.51**	%2.2	-0.483	%0.3	**-0.315**	%0.7	**-0.18**	%0.7
Neuroticism	**-0.481**	-%1.8	**0.783**	-%0.2	**0.261**	%1.3	**0.186**	%1
Openness	**0.234**	%1.8	**0.772**	-%0.2	0.722	%0.2	0.338	-%0.2

It is seen from the comparison table (Table 41) that the changes in RMSE is significant. Therefore, in regard to the prediction performances, we can assert that the larger the number of common friends is, the higher the accuracy of the prediction model is.

Lastly, the prediction model using setting 5,6,7 are constructed and tested to see how the prediction performance changes while discarding friendships which do not share common likes according to the given threshold values.

Table 42 shows the results of the prediction models for each personality trait using Setting 5.

Table 42: The summary of the models' performances which utilize the information of friends whose number of common likes are greater and equal to 10.

	MAE	**RMSE**	**TP Rate**	**FP Rate**	**Precision**	**Recall**	**F-Measure**	**AuC**
Agreeableness	0.43	0.66	0.583	0.427	0.584	0.583	0.575	0.578
Conscientiousness	0.42	0.65	0.585	0.471	0.592	0.585	0.54	0.557
Extraversion	0.41	0.64	0.598	0.413	0.598	0.598	0.593	0.593
Neuroticism	0.41	0.64	0.581	0.436	0.578	0.581	0.578	0.572
Openness	0.36	0.6	0.635	0.445	0.63	0.635	0.609	0.595

Table 43 shows the results of prediction models for each personality using Setting 6.

Table 43: The summary of the models' performances which utilize the information of friends whose number of common likes are greater and equal to 25.

	MAE	RMSE	TP Rate	FP Rate	Precision	Recall	F-Measure	AuC
Agreeableness	0.39	0.63	0.603	0.398	0.61	0.603	0.596	0.603
Conscientiousness	0.36	0.6	0.624	0.444	0.622	0.624	0.597	0.59
Extraversion	0.44	0.67	0.553	0.553	0.306	0.553	0.394	0.5
Neuroticism	0.39	0.62	0.601	0.42	0.597	0.601	0.596	0.59
Openness	0.36	0.6	0.635	0.448	0.63	0.635	0.609	0.594

Table 44 shows the results of prediction models for each personality using Setting 7.

Table 44: The summary of the models' performances which utilize the information of friends whose number of common likes are greater and equal to 50.

	MAE	RMSE	TP Rate	FP Rate	Precision	Recall	F-Measure	AuC
Agreeableness	0.45	0.67	0.561	0.446	0.572	0.561	0.537	0.558
Conscientiousness	0.42	0.65	0.596	0.594	0.558	0.596	0.452	0.501
Extraversion	0.46	0.68	0.564	0.491	0.592	0.564	0.483	0.536
Neuroticism	0.42	0.65	0.571	0.563	0.756	0.571	0.419	0.504
Openness	0.33	0.58	0.657	0.377	0.654	0.657	0.651	0.64

To sum up predictions results by number of shared likes in Table 45;

Table 45: Summary of the models in TP Rate by the number of common like thresholds (df = 9)

	Experiment 3	Experiment 6			
Number of Common Likes	-	**≥0**	**≥ 10**	**≥ 25**	**≥ 50**
Agreeableness	0.591	0.618	0.583	0.603	0.561
Conscientiousness	0.613	0.613	0.585	0.624	0.596
Extraversion	0.611	0.639	0.598	0.553	0.564
Neuroticism	0.601	0.582	0.581	0.601	0.571
Openness	0.617	0.639	0.635	0.635	0.657

Figure 21 shows how prediction performance is changed in proportion to the number of common friends.

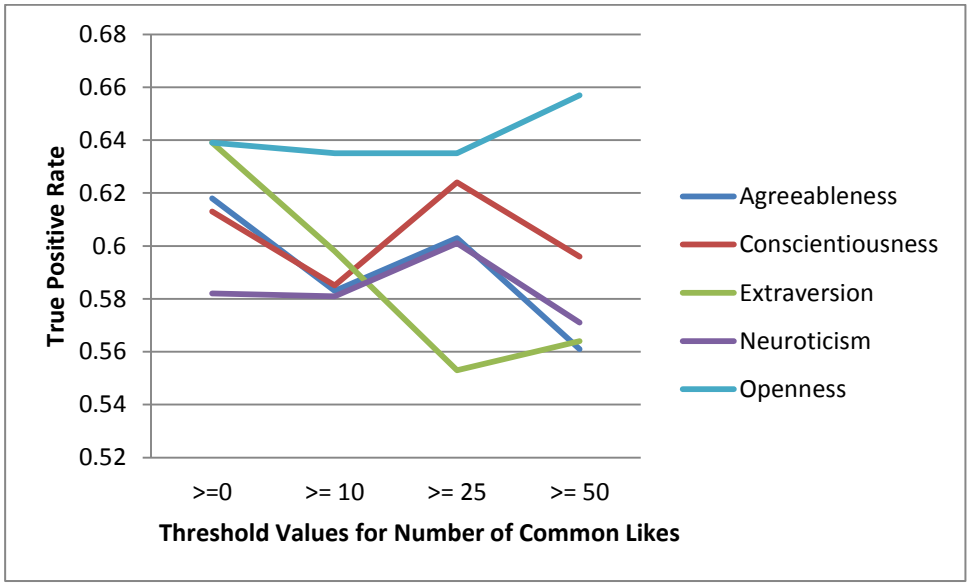


Figure 21: Changes in prediction performances by number of threshold values

T-test for Paired Two Sample is applied on MAE and RMSE of each folds of iterations to decide whether changes in prediction performance is significant or not when number of common likes increases.

Table 46 shows p-values (two-tailed) of t-test applied using RMSEs. Where p-values equal or less than 0.05 denotes that the changes in RMSE is significant.

Table 46: Iteration comparison based on RMSEs applying t-test for Paired Two Sample (df = 9, r: pearson correlation coefficient)

	Exp 3 vs. ≥0		≥0 vs. ≥10		≥10 vs. ≥25		≥25 vs. ≥50	
	r	**Increment**	**r**	**Increment**	**r**	**Increment**	**r**	**Increment**
Agreeableness	**0.815**	%2.1	**0.549**	-%0.1	**-0.825**	-%3.6	**-0.786**	%2.9
Conscientiousness	**-0.579**	-%0.6	**0.62**	-%3.3	**-0.525**	%1.1	**-0.754**	%4.6
Extraversion	**0.51**	%2.2	**0.488**	-%1.3	**-0.271**	-%2.2	-0.619	-%2.8
Neuroticism	**-0.481**	-%1.8	**-0.015**	%2.1	**0.643**	-%1.7	**0.841**	%2.3
Openness	**0.234**	%1.8	-0.33	%1.2	0.63	-%1.2	**-0.118**	-%0.3

As can be seen from the comparison tables, the changes in RMSE is significant. However, when we compare prediction performances; they do not increase in direct proportion to the

number of common likes between two users. Even, for some personality traits such as conscientiousness, neuroticism and openness, personality prediction performance decreases while the number of shared likes is increasing.

Discussion: In the experiment, we calculated the similarities between friends' personality traits. Additionally, we have tested how prediction performances are changed while the threshold values for the number of common friends or likes are increasing. According to the results, people that are similar to their friends such as people high in openness are usually friends with open people, and people high in conscientiousness are usually friends with conscientious people and so on. Another result shows that prediction performances increase in direct proportion to the threshold values for the number of common friends. This implication satisfies our second hypothesis "*If two friends have too many common friends, their personality should be same or close to each other. Incorporation of such friends' personality in a supervised model improves the predictive performance of the model*". However, when the threshold value increases for the number of common likes, personality prediction does not increase in all traits. Therefore, we could not have any supporting evidence for our last hypothesis "*If two friends have too many common likes, their personality should be same or close to each other. Incorporation of such friends' personality in a supervised model improves the predictive performance of the model*".

CHAPTER 5

CONCLUSION AND FUTURE WORK

In this last chapter of the thesis, the results of our experiments and contributions are concluded. We also discussed the limitation of our study and presented the future study.

5.1 Discussion and Conclusion

Academics from different disciplines study on personality inference using the residuals of subjects. These residuals are sometimes a written text, recorded speech, or activities of subject in real and social life. Mairesse et al. predict subject's personality using their essay corpuses and transcribed recorded speech corpuses in their research [14]. They have made significant contribution to the literature by showing the relation between linguistic arguments and personality. There are also studies in literature which investigate subjects' social network structures in real life. According to individuals' location and activities they try to predict their personality. These researches also made important contributions to the literature. They claim that individuals' personality has an effect on extending of friendship network. Based on these studies, Golbeck et al. comes with a research that focuses on prediction of subjects' personality traits using social media / network instead of real life attributes [16]. They figure out tight relations between the content of users' status updates and their personality. Moreover, Gosling et al. study on how users' personality affects usage of social media [28]. They also come up with important inferences such as extrovert people tend to share their photos, make friends and so on. Besides, there are some other researches that study on different social media platforms such as Facebook and Twitter to predict user personality traits evaluating disclosed information shared on social media.

These researches provided fundamental arguments to motivate us to study in such topic. The main purpose of our study is to propose a better inference model to determine users' personality traits in social media. Therefore, in addition to existing methodologies, we

research on how prediction performances are being affected by the increasing number of attributes or evidences. Furthermore, we have observed tthe relations between friends' personality traits. We have studied how we may incorporate close friends' or couple's personality trait scores to predict the targeted user's personality better and hence improve the prediction performance.

To perform this study and support our claims, we used data sets provided by myPersonality Project. We have run six experiments using these data sets. According to the results of our experiments, there are tight correlations between users' personality trait scores and the use percentage of LIWC tags calculated from their status updates. However, the usage percentages of LIWC tags are not sufficient to predict users' personality trait scores accurately. We have observed improvement in prediction performance by eliminating users having limited number of status updates. With regard to our observation, prediction performances have increased while the number of evidence is increased. So we can infer that individuals using social media seldom cause noise in prediction results. Demographic information and usage statistic of Facebook features are another clues to predict users' personality traits. According to our results, people high in extraversion tend to be more social; therefore they are mostly tagged in shared photos taken during activities in their real life. Besides, it is also verified that users' personality is correlated with their usage habits. For instance, neurotic people usually share their thoughts on social media by status updates. Social media serves them a convenient platform to express themselves and spread their thoughts without getting into a direct interaction with other people in real life.

During the experiments we have also evaluated different machine learning algorithms' performances in inference of personality traits. We think that deviation in correlated attributes may make noise during the prediction if we use SMO machine learning algorithm in our models. Therefore, we have compared SMO with J48 and Random Forest. According to our results, SMO algorithm performs better than J48 and Random Forest algorithms. That's why within this study we have preferred applying SMO algorithm for personality inference.

In this study, we have also assessed which personality traits are effective in getting on well with each other, and easily starting an emotional relationship. As a result, we conclude that an individual prefers another person as a couple who has similar personality traits with him/her. For instance, people high in openness mostly start a relationship with people who are also high in openness. Conscientious people also prefer conscientious people as a couple. As an interesting but also understandable outcome of this study is, conscientious people tend

to have a relationship with people who are high in neuroticism. Since neurotic people are so anxious, assured and sensitive, conscientious people, who are extremely reliable and able to manage impulse control, can get on well with these people. Agreeable individuals also prefer agreeable individuals. And lastly, people high in neurotic mostly prefer conscientious people. For a second option, they prefer neurotic people like themselves.

Since couples' personality are correlated with each other, while constructing our prediction model we have added couples' personality traits. According to results of this model, prediction performance has increased in conscientiousness, neuroticism and openness personality traits compared to our former experiment results which make use of demographic, linguistic, usage statistics and social network attributes. However, the changes in performance are not significant.

Lastly we have examined similarities between users' personality traits in a friendship network. Just like our relationship case, people tend to have friends with similar personality traits. Therefore, in this case, we have used friends' personality trait scores while predicting users' personality trait scores. Firstly, we have observed how prediction performances change while eliminating friendships where users do not have sufficient number of common friends. According to the results; as threshold value for the number of common friends increases, the prediction performance increases and gives better results. Secondly, we have tried to observe how prediction performance changes if we eliminate the friendships having a certain number of common likes on the same contents. During the experiments 10, 25 and 50 values are selected as threshold values denoting the number of common likes between two users in a friendship. While the threshold value increased in each one of iterations, we expected the prediction performance to increase. But according to the results, there was no significant change to verify this assertion. For some personality traits, the prediction performance increased while for some personality traits' prediction performance decreased.

5.2 Limitations and Further Research

This study has limitations. First and the most important limitation is about status updates data table. myPersonality Project did not share users' status update table due to some privacy issues. They only shared data table that is populated by processing users' status updates with LIWC tool. The table contains the usage percentage of each word categories defined in LIWC Dictionary in their status updates. Therefore, we could not use different dictionaries such as General Inquirer Dictionary, AFINN 111 Dictionary and etc. Since we do not have any status updates, we could not categorize sentences by their types such as "Assertive",

"Commissive", "Declarative", "Directive" and "Expressive". Instead of looking word by word, meaning and type of sentences might be more helpful to determine users' personality traits. Also punctuation is another factor that can change the meaning of a sentence. In this study, punctuations used in sentences are not considered. Moreover, when deciding threshold values to improve prediction performances, we only considered the number of status updates since we did not have the number of words used in status updates. In addition, we did not have temporal information about status updates. The frequency of sharing status updates or sharing time in a day might be an indicator for users' personality.

As denoted in Figure 4, users are generally from the United States and they have shared status updates in English. Therefore, the distribution of users is biased. Since the status updates are analyzed by LIWC Tool using English Dictionary, the correlation results cannot be generalized for other languages.

Demographic information may be deceptive due to privacy issues. Facebook enables users to hide this information on their profile to a specific person, audience or application. Therefore, while gathering information, myPersonality application might not have gathered actual friends, joined groups, likes, tagged or shared photos, attending events and etc. During the study, we could not identify the users who have prevented the application gathering all the disclosed information on their profile.

In addition to lack of information about profiles, there are environmental factors which affect sharing of social media profiles. Some social events such as elections, public oppositions and natural disasters may affect content of status updates. For instance; when a natural disaster is occurred, users generally share status updates in order to help people. When we consider status updates collected in such terms, individuals may be generally predicted as agreeable person. And, during the election term, people generally share status updates criticizing government's politics. When we consider status updates collected in such terms, individuals may be generally predicted as neurotic person. Therefore, in personality inference, disclosed information on social media profiles may be deceptive.

As a future work, experiments can be conducted by overcoming these defined limitations. Different dictionaries can be used to find a correlation between users' personality traits and linguistic features of shared textual artifacts. Moreover, as mentioned, temporal information can be used in prediction models. Observing what time users share status updates, how often a user shares status updates, how much time a user spends in social media in a day,

frequency of sharing status updates, liking sharing, being tagged in photos, joining groups and attending events may be significant indicators of users' personality.

When we inspect correctly classified instances, we realize that users, whose personality traits score is around mean value of corresponding personality trait score, cannot be predicted successfully. Since these users do not intensively show indication of that corresponding trait, they cannot be easily predicted. Discarding such users, whose personality traits score is between plus and minus standard deviation of mean value of corresponding personality trait, may increase the performance of the models.

On the other hand, most of the social media platforms provides a unique profile where users are able to use different applications for different purposes. For instance; when a user check-ins a place in Foursquare, there is an option in order to share this information on Facebook. A twitter user may pass a favorite twit that he/she just reads in Facebook. YouTube videos that users have shared can be an important feature to determine the personality of a user. Considering these arguments, personality prediction can be done with better performance results by using attributes from different social networking sites belonging to the same user profile.

REFERENCES

[1] R. E. Wilson, S. D. Gosling and L. T. Graham, "A Review of Facebook Research in the Social Sciences," *Perspectives on Psychological Science,* vol. 7, no. 3, p. 203 –220, 2012.

[2] D. Quercia, R. Lambiotte, D. Stillwell, M. Kosinski and J. Crowcroft, "The personality of popular facebook users," in *Proceedings of the ACM 2012 conference on computer supported cooperative work*, ACM, 2012, pp. 955-964.

[3] Y. Bachrach, M. Kosinski, T. Graepel, P. Kohli and D. Stillwell, "Personality and Patterns of Facebook Usage," in *Proceedings of the 3rd Annual ACM Web Science Conference*, ACM, 2012, pp. 24-32.

[4] M. Kosinski, D. Stillwell and T. Graepel, "Private traits and attributes are predictable from digital records of human behavior," *Proceedings of the National Academy of Sciences*, vol. 110, no. 15, pp. 5802-5805, 2013.

[5] F. Celli, F. Pianesi, D. Stillwell and M. Kosinski, "Workshop on Computational Personality Recognition: Shared Task," in *Seventh International AAAI Conference on Weblogs and Social Media*, 2013.

[6] G. Farnadi, S. Zoghbi, M.-F. Moens and M. De Cock, "Recognising Personality Traits Using Facebook Status Updates," in *Proceedings of the workshop on computational personality recognition (WCPR13) at the 7th international AAAI conference on weblogs and social media (ICWSM13)*, 2013.

[7] M. T. Tomlinson, D. Hinote and D. B. Bracewell, "Predicting Conscientiousness through Semantic Analysis of Facebook Posts," *Proceedings of WCPR*, 2013.

[8] D. Markovikj, S. Gievska, M. Kosinski and D. Stillwell, "Mining Facebook Data for

Predictive Personality Modeling," in *Proceedings of the 7th international AAAI conference on Weblogs and Social Media (ICWSM 2013)*, 2013.

[9] F. Alam, E. A. Stepanov and G. Riccardi, "Personality Traits Recognition on Social Network - Facebook," in *Proc of Workshop on Computational Personality Recognition, AAAI Press, Melon Park, CA*, 2013, pp. 6-9.

[10] D. S. Appling, E. J. Briscoe, H. Hayes and R. L. Mappus, "Towards Automated Personality Identification Using Speech Acts," in *Seventh International AAAI Conference on Weblogs and Social Media*, 2013.

[11] F. Iacobelli and A. Culotta, "Too Neurotic, Not Too Friendly: Structured Personality Classification on Textual Data," in *Proc of Workshop on Computational Personality Recognition, AAAI Press, Melon Park, CA*, 2013, pp. 19-22.

[12] B. Verhoeven, W. Daelemans and T. De Smedt, "Ensemble Methods for Personality Recognition," in *Proc of Workshop on Computational Personality Recognition, AAAI Press, Melon Park, CA*, 2013, pp. 35-38.

[13] I. Cantador, I. Fernández-Tobías, A. Bellogín, M. Kosinski and D. Stillwell, "Relating Personality Types with User Preferences in Multiple Entertainment Domains," in *UMAP Workshops*, 2013.

[14] F. Mairesse, M. A. Walker, M. R. Mehl and R. K. Moore, "Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text," *J. Artif. Intell. Res.(JAIR)*, vol. 30, pp. 457-500, 2007.

[15] S. M. Mohammad and S. Kiritchenko, "Using Nuances of Emotion to Identify Personality," *arXiv preprint arXiv:1309.6352*, 2013.

[16] J. Golbeck, C. Robles and K. Turner, "Predicting personality with social media," in *CHI'11 Extended Abstracts on Human Factors in Computing Systems*, ACM, 2011, pp. 253-262.

[17] J. Schrammel, C. Koffel and M. Tscheligi, "Personality Traits, Usage Patterns and Information Disclosure in Online Communities," in *Proceedings of the 23rd British HCI Group Annual Conference on People and Computers: Celebrating People and*

Technology, British Computer Society, 2009, pp. 169-174.

[18] S. Adali and J. Golbeck, "Predicting Personality with Social Behavior," in *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)*, IEEE Computer Society, 2012, pp. 302-309.

[19] S. Adali, F. Sisenda and M. Magdon-Ismail, "Actions Speak as Loud as Words:Predicting Relationships from Social Behavior Data," in *Proceedings of the 21st international conference on World Wide Web,* ACM, 2012, pp. 689-698.

[20] O. P. John and S. Srivastava, "The Big-Five Trait Taxonomy: History, Measurement, and Theoretical Perspectives," *Handbook of personality: Theory and research,* vol. 2, pp. 102-138, 1999.

[21] V. Benet-Martinez and O. P. John, "Los Cinco Grandes across cultures and ethnic groups: Multitrait-multimethod analyses of the Big Five in Spanish and English," *Journal of personality and social psychology,* vol. 75, no. 3, p. 729, 1998.

[22] J. Staiano, B. Lepri, N. Aharony, F. Pianesi, N. Sebe and A. Pentland, "Friends don't Lie - Inferring Personality Traits from Social Network Structure," in *Proceedings of the 2012 ACM Conference on Ubiquitous Computing,* ACM, 2012, pp. 321-330.

[23] M. Selfhout, W. Burk, S. Branje, J. Denissen, M. Van Aken and W. Meeus, "Emerging Late Adolescent Friendship Networks and Big Five Personality Traits: A Social Network Approach," *Journal of Personality,* vol. 78, no. 2, pp. 509-538, 2010.

[24] C. A. Lampe, N. Ellison and C. Steinfield, "A familiar face (book): profile elements as signals in an online social network," in *Proceedings of the SIGCHI conference on Human factors in computing systems,* ACM, 2007, pp. 435-444.

[25] P. A. Rosen and D. H. Kluemper, "The Impact of the Big Five Personality Traits on the Acceptance of Social Networking Website," *AMCIS 2008 Proceedings,* p. 274, 2008.

[26] A. Mislove, B. Viswanath, K. P. Gummadi and P. Druschel, "You are who you know: inferring user profiles in online social networks," in *Proceedings of the third ACM international conference on Web search and data mining,* ACM, 2010, pp. 251-260.

[27] C. Sumner, A. Byers and M. Shearing, "Determining personality traits & privacy concerns from facebook activity," *Black Hat Briefings,* vol. 11, 2011.

[28] S. D. Gosling, A. A. Augustine, S. Vazire, N. Holtzman and S. Gaddis, "Manifestations of Personality in Online Social Networks: Self-Reported Facebook-Related Behaviors and Observable Profile Information," *Cyberpsychology, Behavior, and Social Networking,* vol. 14, no. 9, pp. 483-488, 2011.

[29] B. Chen and J. Marcus, "Students' self-presentation on Facebook: An examination of personality and self-construal factors," *Computers in Human Behavior,* vol. 28, pp. 2091-2099, 2012.

[30] S. Bai, T. Zhu and L. Cheng, "Big-Five Personality Prediction Based on User Behaviors at Social Network Sites," *arXiv preprint arXiv:1204.4809,* 2012.

[31] S. S. Wang and S. Michael, "Showing Off? Human Mobility and the Interplay of Traits, Self Disclosure, and Facebook Check-Ins," *Social Science Computer Review,* pp. 1-21, 2013.

[32] M. D. Back, J. M. Stopfer, S. Vazire, S. Gaddis, S. C. Schmukle, B. Egloff and S. D. Gosling, "Facebook Profiles Reflect Actual Personality, Not Self-Idealization," *Psychological Science,* vol. 21, no. 3, pp. 372-374, 2012.

[33] J. W. Pennebaker, C. K. Chung, M. Ireland, A. Gonzales and R. J. Booth, "The development and psychometric properties of LIWC2007," *Austin, TX, LIWC. Net,* 2007.

[34] R. R. a. F. E. a. H. M. a. K. R. Bouckaert, P. Reutemann, A. Seewald and D. Scuse, "WEKA Manual for Version 3-7-10," 2013.

[35] C. J. Willmott and K. Matsuura, "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance," *Climate Research,* vol. 30, no. 1, p. 79, 2005.

[36] D. J. Stillwell and M. Kosinski, "myPersonality project: Example of successful utilization of online social networks for large-scale social research," *American Psychologist,* vol. 59, no. 2, pp. 93-104, 2004.

[37] K. Audhkhasi, A. Metallinou, M. Li and S. Narayanan, "Speaker Personality Classification Using Systems Based on Acoustic-Lexical Cues and an Optimal Tree-Structured Bayesian Network," in *INTERSPEECH*, 2012.

[38] G. Chittaranjan, J. Blom and D. Gatica-Perez, "Who's who with Big-Five: Analyzing and Classifying personality traits with smartphones," in *Wearable Computers (ISWC), 2011 15th Annual International Symposium on*, IEEE, 2011, pp. 29-36.

[39] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I. H. Witten, "The WEKA data mining software: an update," *ACM SIGKDD explorations newsletter,* vol. 11, no. 1, pp. 10-18, 2009.

[40] "Weka 3: Data Mining Software in Java," The University of Waikato, [Online]. Available: http://www.cs.waikato.ac.nz/ml/weka/index.html.

[41] "Linguistic Inquiry and Word Count: How it works," [Online]. Available: http://www.liwc.net/howliwcworks.php.

[42] P. J. Stone, D. C. Dunphy and M. S. Smith, "The General Inquirer: A Computer Approach to Content Analysis," *MIT press*, 1966.

[43] F. A. Nielsen, "A new ANEW: Evaluation of a word list for sentiment analysis in microblogs," *arXiv preprint arXiv:1103.2903.*

[44] R. Kohavi and others, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *IJCAI*, 1995, pp. 1137-1145.

[45] "An Introduction To Cross-Validation," [Online]. Available: http://www.salford-systems.com/videos/tutorials/how-to/an-introduction-to-cross-validation.

[46] M. Selfhout, W. Burk, S. Branje, J. Denissen, M. Van Aken and W. Meeus, "Emerging Late Adolescent Friendship Networks and Big Five Personality Traits: A Social Network Approach," *Journal of personality*, vol. 78, no. 2, pp. 509-538, 2010.

Appendix A: Pearson Correlation Table for Experiment 1 (r: pearson correlation coefficient, t: t-distribution value and c: Cohen's conventions to interpret effect size)

		Agreeableness			Conscientiousness			Extraversion			Neuroticism			Openness		
Attribute	**Mean±StdDev**	**r**	**t**	**c**	**r**	**t**	**c**	**r**	**t**	**c**	**r**	**t**	**c**	**r**	**t**	**c**
achieve	1.1±0.77	**0.072**	24.741	-	**0.093**	31.936	-	**0.016**	5.566	-	**-0.048**	16.305	-	-0.008	2.793	-
adverb	3.2±1.51	**0.048**	16.203	-	**0.026**	8.867	-	0.004	1.473	-	**0.031**	10.551	-	**0.042**	14.213	-
affect	5.74±2.19	**0.037**	12.715	-	**0.064**	21.939	-	**0.059**	20.27	-	0.007	2.381	-	**-0.03**	10.379	-
anger	0.72±0.68	**0.057**	19.476	-	**-0.107**	36.587	S	**-0.015**	5.075	-	**0.051**	17.374	-	**0.031**	10.642	-
anx	0.2±0.27	**0.038**	12.86	-	-0.008	2.849	-	-0.008	2.621	-	**0.04**	13.666	-	**0.041**	13.855	-
article	3.74±1.59	**0.023**	7.862	-	**0.086**	29.219	-	-0.009	2.903	-	**-0.042**	14.291	-	**0.093**	31.768	-

assent	0.63±0.69	**0.05**	17.185	-	**-0.036**	12.247	-	**0.04**	13.494	-	-0.009	3.201	-	**-0.019**	6.579	-
auxverb	6.03±2.37	**0.045**	15.462	-	**0.053**	17.927	-	0.003	1.037	-	**0.022**	7.362	-	**0.057**	19.522	-
bio	2.12±1.19	**-0.028**	9.437	-	**-0.02**	6.673	-	**0.05**	16.966	-	**0.04**	13.628	-	**0.027**	9.192	-
body	0.69±0.56	**0.056**	19.078	-	**-0.058**	19.853	-	**0.017**	5.881	-	**0.04**	13.644	-	**0.039**	13.134	-
cause	0.84±0.59	**0.05**	17.173	-	**0.019**	6.6	-	**-0.019**	6.598	-	0.01	3.431	-	**0.055**	18.811	-
certain	1.07±0.78	**0.05**	17.142	-	**0.063**	21.637	-	**0.023**	7.714	-	-0.011	3.646	-	**0.02**	6.862	-
cogmech	9.74±3.74	**0.033**	11.282	-	**0.072**	24.582	-	0.008	2.655	-	**0.015**	5.055	-	**0.071**	24.329	-
conj	3.4±1.62	**0.045**	15.277	-	**0.069**	23.633	-	**0.023**	7.759	-	**0.014**	4.791	-	**0.056**	18.923	-
death	0.15±0.27	**0.073**	25.063	-	**-0.051**	17.278	-	**-0.044**	14.874	-	**0.025**	8.576	-	**0.063**	21.568	-
discrep	1.19±0.78	**0.062**	21.181	-	**0.026**	8.869	-	-0.009	2.895	-	**0.034**	11.608	-	**0.014**	4.673	-
excl	1.67±0.95	**-0.017**	5.712	-	**0.019**	6.436	-	**-0.014**	4.89	-	**0.025**	8.492	-	**0.044**	14.897	-
family	0.37±0.5	**0.057**	19.378	-	**0.073**	24.945	-	**0.024**	8.313	-	0.008	2.587	-	**-0.075**	25.52	-
feel	0.47±0.43	**0.024**	8.191	-	-0.01	3.356	-	0.003	1.129	-	**0.023**	7.952	-	**0.029**	10.007	-
friend	0.22±0.39	**-0.112**	38.394	S	**0.047**	16.184	-	**0.023**	7.786	-	-0.006	1.935	-	**-0.023**	7.87	-
funct	34.19±11.21	**0.066**	22.394	-	**0.092**	31.303	-	**0.015**	4.974	-	0.002	0.737	-	**0.067**	22.729	-
future	0.69±0.51	**0.047**	16.121	-	**0.034**	11.597	-	-0.005	1.778	-	0.01	3.535	-	**0.038**	12.968	-

health	0.6±0.55	**0.039**	13.401	-	**0.027**	9.126	-	0.005	1.855	-	**0.041**	13.801	-	-0.004	1.279	-
hear	0.4±0.41	**0.068**	23.229	-	**-0.026**	8.787	-	-0.006	1.924	-	0.01	3.255	-	**0.056**	19.012	-
home	0.45±0.48	**0.13**	44.604	S	**0.07**	23.931	-	**0.021**	7.272	-	-0.001	0.18	-	**-0.032**	10.899	-
humans	0.66±0.58	**-0.089**	30.405	-	**0.025**	8.616	-	**0.05**	17.078	-	-0.003	0.903	-	-0.001	0.466	-
i	3.92±2.1	**-0.123**	42.317	S	**-0.033**	11.245	-	**0.031**	10.555	-	**0.053**	17.925	-	**0.048**	16.399	-
incl	2.52±1.35	**0.063**	21.657	-	**0.103**	35.083	S	**0.055**	18.89	-	**-0.014**	4.626	-	**0.046**	15.841	-
ingest	0.37±0.45	**0.027**	9.066	-	0.008	2.732	-	**0.014**	4.699	-	-0.012	4.012	-	**0.036**	12.134	-
inhib	0.37±0.36	**0.015**	5.145	-	**0.044**	15.12	-	0.01	3.514	-	0.003	0.996	-	-0.002	0.517	-
insight	1.27±0.79	**0.023**	7.8	-	**0.023**	7.937	-	**-0.028**	9.435	-	**0.022**	7.624	-	**0.089**	30.563	-
ipron	3.11±1.56	**0.028**	9.432	-	**0.055**	18.744	-	-0.005	1.586	-	-0.002	0.526	-	**0.058**	19.671	-
leisure	1.23±0.89	**0.042**	14.383	-	**0.044**	15.017	-	**0.044**	14.966	-	**-0.055**	18.616	-	0.009	2.937	-
money	0.36±0.43	**0.023**	7.931	-	**0.045**	15.361	-	0.001	0.194	-	**-0.017**	5.81	-	**0.022**	7.517	-
motion	1.46±0.84	**0.08**	27.349	-	**0.086**	29.263	-	**0.036**	12.266	-	**-0.031**	10.413	-	-0.01	3.442	-
negate	1.3±0.77	**0.027**	9.237	-	-0.002	0.766	-	**-0.013**	4.324	-	**0.042**	14.262	-	**0.029**	10.031	-
negemo	1.81±1.11	**0.029**	9.894	-	**-0.1**	34.132	-	**-0.02**	6.809	-	**0.082**	28.008	-	**0.034**	11.41	-
nonfl	0.16±0.28	**0.024**	8.205	-	-0.007	2.431	-	0	0.063	-	0.01	3.475	-	-0.005	1.829	-

number	0.48±0.46	**0.014**	4.826	-	**0.033**	11.079	-	**-0.024**	8.128	-	**-0.013**	4.394	-	**0.037**	12.677	-
past	1.76±1.08	**-0.036**	12.236	-	**0.024**	8.161	-	-0.007	2.327	-	-0.001	0.496	-	0.008	2.815	-
percept	1.72±0.96	**0.014**	4.775	-	-0.011	3.621	-	0	0.007	-	**0.017**	5.64	-	**0.066**	22.431	-
posemo	3.9±1.9	**-0.016**	5.579	-	**0.133**	45.585	S	**0.081**	27.743	-	**-0.041**	13.942	-	**-0.056**	19.132	-
ppron	6.59±2.89	**0.088**	30.191	-	**0.019**	6.4	-	**0.041**	14.023	-	**0.041**	13.917	-	**0.046**	15.724	-
preps	7.59±2.89	**0.069**	23.421	-	**0.119**	40.741	S	**0.012**	4.089	-	**-0.03**	10.302	-	**0.046**	15.842	-
present	7.14±2.68	**0.044**	15.127	-	**0.053**	17.992	-	**0.028**	9.632	-	**0.027**	9.189	-	**0.026**	8.743	-
pronoun	9.7±3.92	**0.088**	30.029	-	**0.036**	12.156	-	**0.029**	9.706	-	**0.03**	10.047	-	**0.057**	19.411	-
quant	1.7±0.97	**0.034**	11.443	-	**0.084**	28.568	-	0.008	2.729	-	**-0.014**	4.88	-	**0.012**	4.245	-
relativ	10.16±3.74	**0.048**	16.289	-	**0.123**	42.309	S	**0.034**	11.659	-	**-0.038**	12.853	-	-0.001	0.411	-
relig	0.46±0.65	**0.042**	14.313	-	**0.067**	22.94	-	0.009	3.05	-	**-0.039**	13.389	-	-0.007	2.474	-
sad	0.4±0.42	**0.048**	16.429	-	**-0.018**	5.983	-	-0.007	2.305	-	**0.056**	19.209	-	0.011	3.731	-
see	0.71±0.6	**0.063**	21.418	-	0.005	1.672	-	0.001	0.384	-	0.005	1.794	-	**0.029**	9.955	-
sexual	0.53±0.6	**-0.051**	17.47	-	**-0.026**	8.75	-	**0.076**	26.024	-	**0.013**	4.479	-	-0.001	0.346	-
shehe	0.5±0.61	**0.027**	9.036	-	**0.034**	11.706	-	0.011	3.589	-	**0.02**	6.676	-	0.008	2.728	-
social	6.22±2.98	-0.011	3.825	-	**0.085**	28.985	-	**0.051**	17.535	-	-0.002	0.714	-	-0.006	2.132	-

space	3.68±1.61	0.012	4.032	-	**0.073**	24.985	-	**0.026**	9.003	-	**-0.033**	11.237	-	**0.056**	19.235	-
swear	0.34±0.54	0.008	2.615	-	**-0.095**	32.375	-	**0.017**	5.725	-	**0.027**	9.062	-	**0.014**	4.722	-
tentat	1.57±0.93	0.007	2.44	-	**0.018**	6.05	-	**-0.023**	7.864	-	**0.019**	6.347	-	**0.053**	17.997	-
they	0.29±0.35	-0.005	1.671	-	**0.051**	17.261	-	**-0.017**	5.886	-	0	0.121	-	**0.022**	7.562	-
time	5.07±2.23	0.005	1.535	-	**0.114**	38.984	S	**0.028**	9.43	-	**-0.026**	8.928	-	**-0.036**	12.271	-
verb	35.09±7.8	-0.003	1.134	-	**-0.09**	30.636	-	**-0.028**	9.491	-	0.012	3.938	-	**-0.084**	28.846	-
we	0.43±0.53	0.002	0.832	-	**0.073**	24.761	-	**0.034**	11.613	-	**-0.033**	11.136	-	-0.002	0.64	-
work	1.13±0.9	0	0.089	-	**0.075**	25.655	-	**-0.033**	11.203	-	**-0.031**	10.568	-	-0.003	1.189	-
you	1.45±1.22	0	0.088	-	**0.038**	13.103	-	**0.029**	9.913	-	0.011	3.619	-	**0.017**	5.716	-

Appendix B: Pearson Correlation Table for Experiment 3 (r: pearson correlation coefficient, t: t-distribution value and c: Cohen's conventions to interpret effect size)

		Agreeableness			Conscientiousness			Extraversion			Neuroticism			Openness		
Attribute	**Mean±StdDev**	**r**	**t**	**c**	**r**	**t**	**c**	**r**	**t**	**c**	**r**	**t**	**c**	**r**	**t**	**c**
D_age	25.54±9.98	**0.03**	4.406	-	**0.14**	20.496	S	0	0.018	-	**-0.047**	6.86	-	0.001	0.196	-
D_gender	0.62±0.48	**0.06**	8.678	-	**0.061**	8.821	-	0.016	2.303	-	**0.173**	25.435	S	**-0.043**	6.182	-
D_interested_in	0.56±0.84	-0.006	0.94	-	**-0.055**	8.026	-	**0.036**	5.27	-	**-0.055**	7.921	-	0.022	3.114	-
D_mf_dating	0.05±0.21	0.001	0.159	-	-0.014	2.092	-	0.021	3.016	-	-0.018	2.573	-	0.003	0.491	-
D_mf_friendship	0.27±0.45	0.026	3.791	-	0.015	2.222	-	**0.042**	6.07	-	-0.022	3.196	-	-0.002	0.354	-
D_mf_networking	0.08±0.28	-0.011	1.567	-	**0.035**	5.118	-	0.016	2.379	-	**-0.035**	5.1	-	**0.033**	4.846	-
D_mf_random	0.0±0.05	-0.002	0.318	-	-0.003	0.501	-	0.002	0.221	-	-0.012	1.748	-	0.008	1.197	-
D_mf_relationship	0.05±0.22	0.003	0.478	-	-0.017	2.528	-	0.019	2.706	-	-0.012	1.806	-	-0.008	1.207	-
D_mf_whatever	0.0±0.06	-0.002	0.299	-	-0.015	2.177	-	0	0.066	-	-0.005	0.672	-	0.015	2.242	-
D_network_size	314.38±261.3	**0.049**	7.037	-	**0.046**	6.655	-	**0.222**	32.992	S	**-0.088**	12.717	-	-0.007	0.978	-
D_rel_status	1.7±1.37	-0.021	3.018	-	0.021	3.04	-	**0.033**	4.807	-	0.027	3.873	-	-0.002	0.311	-
D_timezone	-1.9±5.36	**-0.045**	6.559	-	**-0.072**	10.465	-	**-0.029**	4.209	-	**0.038**	5.522	-	**-0.072**	10.375	-
F_n_concentration	0.56±0.98	0.024	3.438	-	**0.128**	18.73	S	-0.002	0.293	-	**-0.039**	5.717	-	**0.077**	11.181	-
F_n_diads	336.9±283.19	**0.033**	4.777	-	**0.042**	6.038	-	**0.206**	30.457	S	**-0.075**	10.879	-	0.015	2.139	-
F_n_education	2.0±1.3	0.02	2.936	-	**0.105**	15.229	S	0.014	1.964	-	**-0.035**	5.125	-	0.026	3.823	-

F_n_event	8.32±46.15	0.002	0.305	-	0.008	1.2	-	**0.05**	7.174	-	-0.013	1.897	-	0.025	3.618	-
F_n_group	36.88±44.8	-0.015	2.134	-	**-0.071**	10.346	-	**0.045**	6.558	-	**0.038**	5.51	-	**0.05**	7.269	-
F_n_like	181.85±357.61	**-0.037**	5.338	-	**-0.095**	13.733	-	-0.018	2.533	-	**0.086**	12.494	-	0.016	2.31	-
F_n_status	242.65±221.4	-0.016	2.363	-	-0.028	4.044	-	**0.034**	4.955	-	**0.072**	10.456	-	0.024	3.417	-
F_n_tags	159.62±232.18	**0.039**	5.637	-	-0.01	1.386	-	**0.122**	17.737	S	-0.017	2.447	-	-0.003	0.39	-
F_n_work	0.92±1.26	**0.031**	4.492	-	**0.126**	18.411	S	**0.036**	5.18	-	**-0.047**	6.854	-	0.027	3.925	-
L_achieve	1.14±0.47	**0.063**	9.096	-	**0.167**	24.452	S	0.019	2.778	-	**-0.098**	14.227	-	-0.012	1.752	-
L_adverb	3.38±1.12	**0.06**	8.753	-	**0.037**	5.303	-	-0.012	1.689	-	**0.044**	6.316	-	**0.032**	4.689	-
L_affect	5.69±1.51	**0.067**	9.711	-	**0.061**	8.84	-	**0.069**	10.068	-	0.027	3.943	-	-0.024	3.474	-
L_anger	0.74±0.49	**-0.181**	26.557	S	**-0.16**	23.472	S	**-0.037**	5.335	-	**0.079**	11.395	-	**0.033**	4.808	-
L_anx	0.21±0.13	-0.002	0.218	-	-0.014	2.056	-	**-0.043**	6.172	-	**0.082**	11.858	-	**0.074**	10.694	-
L_article	3.88±1.19	**0.05**	7.283	-	**0.107**	15.543	S	-0.017	2.406	-	**-0.066**	9.635	-	**0.128**	18.718	S
L_assent	0.66±0.47	**0.045**	6.571	-	**-0.036**	5.169	-	**0.05**	7.19	-	-0.003	0.496	-	**-0.042**	6.078	-
L_auxverb	6.32±1.86	**0.068**	9.807	-	**0.062**	8.973	-	-0.018	2.585	-	**0.028**	4.107	-	**0.064**	9.261	-
L_bio	2.23±0.82	-0.028	3.992	-	**-0.047**	6.832	-	**0.057**	8.203	-	**0.061**	8.883	-	**0.033**	4.801	-
L_body	0.75±0.37	**-0.053**	7.731	-	**-0.096**	13.92	-	0.017	2.451	-	**0.06**	8.739	-	**0.051**	7.372	-
L_cause	0.87±0.35	0.015	2.191	-	0.02	2.92	-	**-0.052**	7.587	-	0.008	1.152	-	**0.093**	13.525	-
L_certain	1.03±0.42	**0.043**	6.221	-	**0.078**	11.369	-	**0.032**	4.58	-	-0.008	1.153	-	**0.086**	12.478	-
L_cogmech	10.04±2.96	**0.066**	9.534	-	**0.075**	10.815	-	-0.012	1.774	-	0.019	2.69	-	**0.094**	13.651	-

L_conj	3.54±1.25	**0.074**	10.671	-	**0.087**	12.623	-	0.006	0.905	-	0.015	2.111	-	**0.065**	9.431	-
L_death	0.15±0.16	**-0.074**	10.802	-	**-0.081**	11.827	-	**-0.067**	9.734	-	**0.044**	6.431	-	**0.105**	15.223	S
L_discrep	1.22±0.48	0.025	3.6	-	0.017	2.402	-	-0.026	3.749	-	**0.058**	8.446	-	**0.032**	4.658	-
L_excl	1.71±0.65	**0.032**	4.679	-	0.015	2.241	-	**-0.046**	6.646	-	**0.037**	5.349	-	**0.065**	9.395	-
L_family	0.33±0.29	**0.063**	9.188	-	**0.088**	12.713	-	0.024	3.436	-	0.024	3.415	-	**-0.097**	14.159	-
L_feel	0.5±0.23	0.022	3.236	-	-0.019	2.732	-	-0.016	2.286	-	**0.049**	7.112	-	**0.043**	6.174	-
L_friend	0.19±0.16	**0.059**	8.483	-	**0.089**	12.908	-	**0.059**	8.559	-	0.004	0.608	-	-0.025	3.658	-
L_funct	35.32±9.34	**0.076**	10.956	-	**0.095**	13.857	-	-0.001	0.176	-	0	0.054	-	**0.078**	11.269	-
L_future	0.72±0.32	**0.04**	5.827	-	**0.033**	4.714	-	**-0.028**	4.091	-	0.014	2.066	-	**0.06**	8.746	-
L_health	0.6±0.3	0.005	0.727	-	**0.031**	4.421	-	-0.008	1.15	-	**0.071**	10.345	-	-0.004	0.57	-
L_hear	0.42±0.25	0.01	1.4	-	**-0.046**	6.663	-	**-0.031**	4.53	-	0.014	1.996	-	**0.099**	14.401	-
L_home	0.5±0.33	**0.073**	10.63	-	**0.106**	15.483	S	0.016	2.365	-	0.005	0.693	-	**-0.074**	10.669	-
L_humans	0.65±0.34	-0.013	1.935	-	0.022	3.211	-	**0.066**	9.562	-	-0.01	1.377	-	0.018	2.539	-
L_i	4.02±1.7	0.012	1.744	-	**-0.041**	5.988	-	0.025	3.6	-	**0.066**	9.571	-	**0.05**	7.212	-
L_incl	2.66±1.02	**0.1**	14.526	-	**0.13**	18.912	S	**0.061**	8.776	-	-0.018	2.535	-	**0.058**	8.34	-
L_ingest	0.42±0.3	0.01	1.501	-	**0.029**	4.261	-	0.013	1.855	-	-0.027	3.884	-	0.027	3.978	-
L_inhib	0.38±0.19	0.026	3.825	-	**0.073**	10.56	-	-0.009	1.243	-	0.005	0.718	-	-0.004	0.638	-
L_insight	1.31±0.52	0.026	3.819	-	0.023	3.346	-	**-0.063**	9.163	-	**0.03**	4.323	-	**0.128**	18.668	S
L_ipron	3.15±1.13	**0.06**	8.712	-	**0.057**	8.197	-	-0.026	3.744	-	-0.004	0.629	-	**0.082**	11.863	-

L_leisure	1.33±0.6	**0.068**	9.789	-	**0.08**	11.652	-	**0.071**	10.274	-	**-0.091**	13.153	-	-0.012	1.706	-
L_money	0.38±0.24	-0.016	2.328	-	**0.077**	11.172	-	0.001	0.183	-	**-0.034**	4.968	-	0.028	4.037	-
L_motion	1.55±0.55	**0.1**	14.51	-	**0.133**	19.347	S	**0.032**	4.659	-	**-0.049**	7.106	-	**-0.031**	4.517	-
L_negate	1.34±0.49	**-0.034**	4.863	-	-0.021	3.041	-	**-0.043**	6.299	-	**0.076**	11.037	-	**0.045**	6.459	-
L_negemo	1.85±0.75	**-0.144**	21.038	S	**-0.156**	22.786	S	**-0.062**	8.988	-	**0.131**	19.088	S	**0.037**	5.326	-
L_nonfl	0.16±0.17	0.014	1.997	-	-0.005	0.77	-	-0.011	1.594	-	0.016	2.283	-	**-0.029**	4.128	-
L_number	0.5±0.26	**0.044**	6.44	-	**0.06**	8.745	-	**-0.044**	6.405	-	**-0.03**	4.35	-	**0.063**	9.114	-
L_past	1.82±0.75	**0.07**	10.162	-	**0.038**	5.462	-	-0.018	2.624	-	0.001	0.207	-	0.004	0.576	-
L_percept	1.81±0.6	**0.05**	7.283	-	-0.018	2.632	-	-0.024	3.441	-	**0.028**	4.113	-	**0.105**	15.236	S
L_posemo	3.8±1.26	**0.165**	24.255	S	**0.165**	24.251	S	**0.122**	17.766	S	**-0.047**	6.805	-	**-0.052**	7.54	-
L_ppron	6.69±2.35	**0.039**	5.644	-	0	0.045	-	**0.034**	4.951	-	**0.05**	7.28	-	**0.068**	9.862	-
L_preps	7.95±2.36	**0.079**	11.431	-	**0.138**	20.146	S	0.003	0.488	-	**-0.044**	6.3	-	**0.044**	6.424	-
L_present	7.41±2.1	**0.055**	8.007	-	**0.053**	7.642	-	0.017	2.433	-	**0.041**	5.953	-	**0.029**	4.217	-
L_pronoun	9.84±3.22	**0.05**	7.176	-	0.02	2.829	-	0.016	2.308	-	**0.035**	5.1	-	**0.078**	11.363	-
L_quant	1.69±0.58	**0.063**	9.101	-	**0.116**	16.895	S	0.001	0.199	-	-0.024	3.429	-	**0.041**	5.997	-
L_relativ	10.73±3.05	**0.101**	14.64	S	**0.155**	22.755	S	**0.036**	5.262	-	**-0.055**	8.041	-	**-0.032**	4.574	-
L_relig	0.43±0.44	**0.081**	11.769	-	**0.075**	10.836	-	-0.001	0.116	-	**-0.055**	8.003	-	0.011	1.613	-
L_sad	0.4±0.22	-0.012	1.727	-	**-0.054**	7.865	-	**-0.035**	5.059	-	**0.113**	16.408	S	0.014	2.02	-
L_see	0.74±0.31	**0.052**	7.541	-	0.01	1.411	-	-0.002	0.353	-	0.01	1.443	-	**0.061**	8.845	-

L_sexual	0.53±0.37	**-0.04**	5.749	-	**-0.075**	10.941	-	**0.116**	16.855	S	**0.033**	4.846	-	0.008	1.197	-
L_shehe	0.55±0.46	**0.03**	4.282	-	**0.034**	4.974	-	0.007	0.962	-	0.027	3.979	-	0.021	3.05	-
L_social	6.06±2.18	**0.067**	9.679	-	**0.067**	9.734	-	**0.055**	8.007	-	-0.001	0.163	-	**0.033**	4.731	-
L_space	3.89±1.18	**0.056**	8.148	-	**0.099**	14.342	-	0.026	3.822	-	**-0.056**	8.115	-	**0.06**	8.698	-
L_swear	0.35±0.4	**-0.153**	22.454	S	**-0.139**	20.241	S	0.017	2.524	-	**0.04**	5.807	-	0.009	1.372	-
L_tentat	1.61±0.62	**0.036**	5.182	-	0.02	2.934	-	**-0.055**	8.033	-	0.021	3.089	-	**0.076**	11.011	-
L_they	0.28±0.21	0	0.038	-	**0.065**	9.353	-	**-0.046**	6.64	-	-0.007	1.027	-	**0.058**	8.461	-
L_time	5.35±1.72	**0.104**	15.123	S	**0.152**	22.304	S	**0.036**	5.282	-	**-0.039**	5.685	-	**-0.084**	12.179	-
L_verb	34.99±6.32	**-0.029**	4.163	-	**-0.098**	14.3	-	-0.01	1.502	-	0.02	2.921	-	**-0.105**	15.226	S
L_we	0.42±0.32	**0.075**	10.915	-	**0.094**	13.639	-	**0.046**	6.633	-	**-0.047**	6.873	-	0.016	2.382	-
L_work	1.22±0.7	**0.049**	7.067	-	**0.141**	20.615	S	**-0.05**	7.217	-	**-0.052**	7.525	-	**-0.038**	5.551	-
L_you	1.42±0.88	**0.038**	5.521	-	0.012	1.715	-	**0.034**	4.93	-	0.011	1.631	-	**0.055**	7.934	-
S_betweenness	93305±172454	0.026	3.77	-	**0.039**	5.66	-	**0.167**	24.55	S	**-0.068**	9.857	-	0.016	2.339	-
S_brokerage	95241±174711	0.026	3.818	-	**0.039**	5.626	-	**0.168**	24.606	S	**-0.068**	9.88	-	0.016	2.28	-
S_density	0.05±0.09	0	0.008	-	**-0.04**	5.83	-	**-0.037**	5.356	-	0.012	1.72	-	-0.015	2.206	-
S_n_betweenness	92.41±15.02	-0.001	0.2	-	**0.045**	6.517	-	**0.031**	4.446	-	-0.008	1.192	-	0.015	2.106	-
S_nbrokerage	0.48±0.04	-0.003	0.422	-	**0.038**	5.57	-	0.017	2.53	-	-0.007	0.966	-	0.015	2.129	-
S_network_size	336±283	**0.033**	4.752	-	**0.042**	6.034	-	**0.204**	30.222	S	**-0.075**	10.83	-	0.014	2.092	-
S_transitivity	0.14±0.15	-0.012	1.744	-	**-0.058**	8.407	-	**-0.1**	14.504	-	**0.036**	5.148	-	-0.021	3.04	-

Appendix C: Pearson Correlation Table for Experiment 5 (r: pearson correlation coefficient, t: t-distribution value and c: Cohen's conventions to interpret effect size)

		Agreeableness			Conscientiousness			Extraversion			Neuroticism			Openness		
Attribute	**Mean±StdDev**	**r**	**t**	**c**	**r**	**t**	**c**	**r**	**t**	**c**	**r**	**t**	**c**	**r**	**t**	**c**
D_age	25.84±8.99	-0.008	0.282	-	0.071	2.387	-	0.014	0.466	-	**-0.093**	3.135	-	0.014	0.483	-
D_gender	0.49±0.5	0.055	1.837	-	-0.011	0.36	-	0.008	0.274	-	**0.275**	9.595	S	-0.032	1.072	-
D_interested_in	0.48±0.82	-0.01	0.32	-	-0.029	0.98	-	0.003	0.115	-	-0.043	1.451	-	-0.002	0.069	-
D_relat_status	2.57±0.98	-0.038	1.275	-	0.06	2.027	-	0.014	0.459	-	0.038	1.277	-	0.003	0.107	-
D_timezone	-2.52±4.94	-0.031	1.048	-	**-0.093**	3.131	-	-0.001	0.029	-	0.007	0.249	-	-0.082	2.747	-
F_n_concentration	0.7±1.11	0.025	0.85	-	**0.124**	4.177	S	0.028	0.948	-	-0.041	1.392	-	**0.093**	3.118	-
F_n_diads	284.66±250.47	**0.085**	2.873	-	0.041	1.375	-	**0.182**	6.218	S	-0.071	2.384	-	-0.033	1.103	-
F_n_education	2.04±1.36	0.068	2.285	-	**0.093**	3.141	-	0.043	1.443	-	-0.069	2.324	-	0.052	1.752	-
F_n_event	5.76±31.59	-0.003	0.094	-	-0.019	0.638	-	**0.075**	2.529	-	0.036	1.2	-	0.037	1.24	-
F_n_group	33.19±43.81	0.016	0.544	-	**-0.119**	4.008	S	0.05	1.681	-	0.064	2.162	-	0.063	2.109	-
F_n_like	151.82±347.27	-0.034	1.149	-	**-0.105**	3.536	S	-0.005	0.169	-	**0.085**	2.852	-	0.044	1.484	-
F_n_status	198.52±221.3	-0.035	1.171	-	**-0.087**	2.937	-	**0.069**	2.309	-	**0.15**	5.104	S	0.057	1.923	-
F_n_tags	168.91±241.6	0.055	1.849	-	-0.035	1.188	-	**0.105**	3.542	S	0.011	0.359	-	-0.068	2.299	-
F_n_work	0.99±1.21	0.016	0.53	-	**0.093**	3.131	-	0.011	0.381	-	-0.058	1.954	-	0.021	0.7	-

L_achieve	1.16±0.64	0.037	1.246	-	**0.09**	3.021	-	0.024	0.798	-	**-0.092**	3.099	-	-0.041	1.374	-
L_adverb	3.32±1.31	0.035	1.183	-	0.009	0.287	-	-0.009	0.318	-	0.067	2.241	-	0.003	0.097	-
L_affect	5.53±1.74	0.057	1.911	-	0.007	0.243	-	0.045	1.516	-	-0.018	0.613	-	-0.054	1.823	-
L_anger	0.74±0.58	**-0.135**	4.569	S	**-0.149**	5.066	S	-0.033	1.12	-	0.08	2.678	-	0.033	1.091	-
L_anx	0.21±0.24	0.012	0.405	-	-0.032	1.083	-	0.008	0.267	-	0.043	1.454	-	**0.09**	3.031	-
L_article	4.08±1.42	0.004	0.128	-	0.071	2.385	-	**-0.093**	3.137	-	-0.02	0.661	-	**0.106**	3.57	S
L_assent	0.61±0.48	0.077	2.584	-	-0.009	0.311	-	**0.09**	3.014	-	-0.026	0.874	-	-0.033	1.093	-
L_auxverb	6.33±2.09	0.038	1.265	-	0	0.005	-	-0.024	0.797	-	0.044	1.462	-	0.054	1.813	-
L_bio	2.21±1.24	-0.06	2.024	-	**-0.09**	3.044	-	-0.006	0.213	-	0.071	2.387	-	0.039	1.321	-
L_body	0.75±0.8	-0.075	2.53	-	**-0.088**	2.966	-	-0.009	0.303	-	0.016	0.526	-	0.069	2.315	-
L_cause	0.84±0.46	0.007	0.233	-	0.013	0.426	-	-0.034	1.126	-	0.055	1.861	-	0.054	1.828	-
L_certain	1.01±0.71	-0.026	0.887	-	-0.02	0.686	-	-0.022	0.743	-	-0.022	0.746	-	-0.015	0.486	-
L_cogmech	10.02±3.26	0.018	0.606	-	0.026	0.87	-	-0.043	1.459	-	0.052	1.738	-	0.078	2.628	-
L_conj	3.53±1.44	0.02	0.661	-	0.058	1.937	-	-0.01	0.332	-	0.067	2.24	-	**0.088**	2.979	-
L_death	0.18±0.28	-0.013	0.421	-	-0.039	1.31	-	-0.033	1.1	-	-0.03	1.02	-	**0.1**	3.37	S
L_discrep	1.2±0.7	-0.011	0.368	-	-0.011	0.359	-	-0.044	1.468	-	**0.083**	2.787	-	-0.03	1.006	-
L_excl	1.69±0.93	-0.029	0.987	-	-0.015	0.5	-	**-0.062**	2.095	-	**0.087**	2.933	-	0.063	2.131	-
L_family	0.32±0.44	-0.001	0.032	-	**0.118**	3.986	S	0.029	0.965	-	-0.034	1.146	-	-0.065	2.193	-
L_feel	0.48±0.3	0.028	0.938	-	-0.009	0.303	-	0.017	0.585	-	0.011	0.375	-	**0.093**	3.139	-

L_friend	0.19±0.26	0.014	0.459	-	0.017	0.556	-	0.017	0.572	-	-0.053	1.766	-	0.005	0.183	-
L_funct	35.6±9.74	0.04	1.35	-	0.055	1.862	-	-0.044	1.492	-	0.027	0.892	-	0.069	2.312	-
L_future	0.7±0.44	-0.023	0.788	-	0.007	0.243	-	0.003	0.095	-	0.04	1.348	-	0.056	1.882	-
L_health	0.58±0.5	-0.017	0.581	-	0.002	0.079	-	-0.009	0.318	-	**0.085**	2.858	-	0.033	1.091	-
L_hear	0.39±0.29	0.019	0.638	-	-0.064	2.156	-	-0.052	1.753	-	0.018	0.61	-	**0.112**	3.784	S
L_home	0.5±0.43	0.067	2.25	-	0.076	2.556	-	0.012	0.404	-	-0.037	1.24	-	-0.008	0.277	-
L_humans	0.64±0.47	-0.016	0.53	-	0.024	0.808	-	0.028	0.937	-	0.021	0.7	-	-0.029	0.984	-
L_i	3.81±1.87	0.012	0.399	-	-0.05	1.664	-	0.017	0.557	-	**0.083**	2.79	-	0.042	1.393	-
L_incl	2.78±1.28	0.056	1.895	-	0.079	2.644	-	0.017	0.553	-	-0.013	0.443	-	**0.088**	2.948	-
L_ingest	0.45±0.45	0.037	1.246	-	-0.024	0.794	-	-0.032	1.072	-	0.066	2.212	-	-0.01	0.323	-
L_inhib	0.38±0.33	0.059	1.973	-	0.044	1.461	-	0.019	0.624	-	-0.044	1.469	-	-0.017	0.586	-
L_insight	1.27±0.65	0.012	0.393	-	0.014	0.457	-	-0.043	1.443	-	0.027	0.89	-	0.081	2.732	-
L_ipron	3.06±1.33	0.027	0.921	-	0.002	0.078	-	-0.033	1.102	-	0.001	0.049	-	0.069	2.323	-
L_leisure	1.34±0.76	0.075	2.508	-	0.03	0.998	-	0.034	1.154	-	-0.06	2.02	-	-0.017	0.568	-
L_money	0.4±0.3	-0.028	0.955	-	0.072	2.432	-	0.011	0.363	-	-0.071	2.373	-	0.058	1.933	-
L_motion	1.55±0.73	0.078	2.611	-	0.074	2.483	-	0.039	1.31	-	-0.055	1.843	-	0.004	0.118	-
L_negate	1.3±0.8	-0.051	1.708	-	**-0.103**	3.487	S	-0.015	0.514	-	**0.128**	4.316	S	0.046	1.534	-
L_negemo	1.79±0.89	**-0.088**	2.968	-	**-0.175**	5.942	S	-0.025	0.842	-	**0.127**	4.293	S	0.063	2.116	-
L_nonfl	0.15±0.18	0.011	0.364	-	-0.029	0.988	-	0.021	0.718	-	0.016	0.536	-	-0.013	0.424	-

L_number	0.56±0.5	-0.002	0.068	-	0.056	1.865	-	**-0.1**	3.383	S	-0.028	0.936	-	0.001	0.032	-
L_past	1.88±0.94	0.074	2.487	-	0.007	0.23	-	**-0.064**	2.167	-	0.052	1.753	-	-0.01	0.327	-
L_percept	1.76±0.78	0.076	2.564	-	-0.026	0.875	-	-0.045	1.497	-	0.028	0.94	-	**0.117**	3.94	S
L_posemo	3.71±1.5	**0.119**	4.006	S	**0.115**	3.866	S	**0.069**	2.318	-	**-0.1**	3.386	S	**-0.102**	3.443	S
L_ppron	6.47±2.64	0.013	0.448	-	-0.003	0.102	-	0.008	0.28	-	**0.084**	2.813	-	0.047	1.574	-
L_preps	8.28±2.6	0.064	2.148	-	**0.117**	3.965	S	-0.044	1.467	-	-0.053	1.791	-	0.033	1.123	-
L_present	7.31±2.31	0.043	1.442	-	-0.001	0.039	-	0.009	0.303	-	0.034	1.132	-	-0.012	0.387	-
L_pronoun	9.53±3.51	0.021	0.687	-	-0.001	0.049	-	-0.006	0.208	-	0.063	2.13	-	0.062	2.068	-
L_quant	1.7±0.85	0.007	0.251	-	0.025	0.826	-	-0.055	1.859	-	-0.023	0.767	-	-0.007	0.242	-
L_relativ	11.01±3.52	0.077	2.583	-	**0.091**	3.051	-	0	0.007	-	-0.039	1.323	-	-0.042	1.394	-
L_relig	0.4±0.5	0.058	1.958	-	0.064	2.135	-	0.04	1.355	-	-0.035	1.17	-	0.009	0.29	-
L_sad	0.38±0.32	-0.024	0.794	-	**-0.108**	3.646	S	-0.03	1.014	-	0.064	2.152	-	0.037	1.242	-
L_see	0.75±0.49	0.072	2.434	-	0.01	0.351	-	-0.039	1.325	-	0.014	0.483	-	0.033	1.09	-
L_sexual	0.53±0.67	**-0.114**	3.853	S	-0.079	2.645	-	0.04	1.344	-	0.002	0.065	-	-0.004	0.136	-
L_shehe	0.6±0.69	-0.061	2.062	-	0.051	1.724	-	-0.024	0.821	-	0.082	2.771	-	-0.015	0.511	-
L_social	6.02±2.55	0.018	0.588	-	0.058	1.934	-	0.014	0.46	-	0.003	0.114	-	0.018	0.613	-
L_space	4.06±1.46	0.033	1.091	-	0.058	1.938	-	-0.015	0.502	-	-0.056	1.885	-	0.03	0.992	-
L_swear	0.36±0.45	**-0.146**	4.948	S	**-0.136**	4.59	S	0.005	0.156	-	0.046	1.559	-	0.031	1.055	-
L_tentat	1.58±0.76	-0.003	0.085	-	0	0.001	-	-0.051	1.719	-	0.063	2.108	-	**0.084**	2.834	-

L_they	0.28±0.29	-0.039	1.322	-	0.049	1.661	-	-0.038	1.26	-	-0.003	0.109	-	0.035	1.188	-
L_time	5.45±2.22	0.075	2.52	-	0.069	2.32	-	-0.007	0.231	-	-0.006	0.212	-	**-0.095**	3.213	-
L_verb	34.72±6.65	0.042	1.395	-	**-0.103**	3.467	S	-0.022	0.75	-	0.003	0.108	-	**-0.162**	5.499	S
L_we	0.42±0.51	0.043	1.456	-	0.049	1.642	-	0.017	0.563	-	-0.06	2.029	-	0.03	0.997	-
L_work	1.31±1.0	0.063	2.118	-	0.076	2.546	-	-0.023	0.758	-	-0.028	0.943	-	-0.034	1.157	-
L_you	1.35±1.05	0.042	1.408	-	0.01	0.323	-	0.009	0.317	-	0.038	1.288	-	0.03	0.994	-
RS_euc_dist	2.14±0.82	-0.058	1.959	-	-0.04	1.339	-	-0.038	1.285	-	0.042	1.399	-	-0.055	1.862	-
RS_n_diads1	284.66±250.47	**0.085**	2.873	-	0.041	1.375	-	**0.182**	6.218	S	-0.071	2.384	-	-0.033	1.103	-
RS_n_diads2	65.11±161.01	0.015	0.494	-	0.007	0.22	-	0.044	1.478	-	-0.046	1.557	-	-0.029	0.968	-
RS_n_group1	33.19±43.81	0.016	0.544	-	**-0.119**	4.008	S	0.05	1.681	-	0.064	2.162	-	0.063	2.109	-
RS_n_group2	7.13±28.05	-0.052	1.74	-	-0.081	2.713	-	-0.005	0.151	-	0.024	0.806	-	-0.01	0.323	-
RS_n_like1	151.82±347.27	-0.034	1.149	-	**-0.105**	3.536	S	-0.005	0.169	-	**0.085**	2.852	-	0.044	1.484	-
RS_n_like2	31.82±101.72	-0.052	1.76	-	-0.054	1.811	-	-0.021	0.717	-	0.021	0.69	-	0.017	0.578	-
RS_n_tags1	168.91±241.6	0.055	1.849	-	-0.035	1.188	-	**0.105**	3.542	S	0.011	0.359	-	-0.068	2.299	-
RS_n_tags2	45.54±106.07	0.021	0.698	-	0.02	0.677	-	0.027	0.896	-	**-0.093**	3.12	-	-0.03	1.002	-
RS_rel_status	2.57±0.98	-0.038	1.275	-	0.06	2.027	-	0.014	0.459	-	0.038	1.277	-	0.003	0.107	-
RS_s_friends	19.13±29.05	-0.02	0.67	-	**0.089**	3.001	-	0.021	0.714	-	-0.039	1.316	-	0.003	0.09	-
RS_s_group	0.63±3.31	-0.022	0.728	-	-0.072	2.435	-	0.026	0.869	-	0.008	0.267	-	0.014	0.485	-
RS_s_like	3.28±13.81	-0.069	2.322	-	-0.07	2.339	-	-0.007	0.246	-	0.014	0.455	-	0.073	2.446	-

RS_s_tags	22.92±44.82	0.019	0.643	-	0.038	1.269	-	0.023	0.785	-	-0.05	1.676	-	-0.05	1.676	-
RS_second_agr	3.52±0.69	0.039	1.293	-	0.066	2.232	-	0.001	0.032	-	-0.012	0.392	-	0.031	1.051	-
RS_second_con	3.49±0.71	0.062	2.093	-	**0.148**	5.009	S	**0.06**	2.023	-	**-0.125**	4.229	S	-0.032	1.08	-
RS_second_ext	3.49±0.81	0.028	0.954	-	0.041	1.384	-	**0.115**	3.874	S	**-0.085**	2.867	-	-0.029	0.958	-
RS_second_neu	2.75±0.82	-0.013	0.448	-	**-0.101**	3.405	S	-0.047	1.566	-	-0.039	1.309	-	0.051	1.703	-
RS_second_ope	3.92±0.65	-0.013	0.429	-	-0.041	1.375	-	**-0.045**	1.507	-	0.033	1.111	-	**0.119**	4.001	S
S_betweenness	68541±138217	0.08	2.692	-	0.028	0.955	-	**0.151**	5.119	S	-0.077	2.577	-	-0.028	0.935	-
S_brokerage	70444±141103	0.081	2.711	-	0.029	0.958	-	**0.15**	5.092	S	-0.076	2.563	-	-0.028	0.939	-
S_density	0.06±0.09	0.009	0.309	-	-0.03	1.02	-	-0.032	1.09	-	0.024	0.819	-	-0.001	0.05	-
S_n_betweenness	90.18±16.86	-0.009	0.309	-	-0.002	0.074	-	0.011	0.364	-	-0.007	0.245	-	-0.009	0.294	-
S_nbrokerage	0.48±0.04	-0.013	0.448	-	0.03	1.014	-	0.019	0.644	-	-0.024	0.816	-	0.004	0.124	-
S_network_size	284.53±250.1	**0.084**	2.819	-	0.041	1.389	-	**0.181**	6.156	S	-0.07	2.359	-	-0.033	1.098	-
S_transitivity	0.16±0.16	-0.022	0.749	-	-0.036	1.205	-	**-0.092**	3.097	-	0.043	1.437	-	-0.033	1.103	-

Appendix D: Pearson Correlation Table for Experiment 6 (r: pearson correlation coefficient, t: t-distribution value and c: Cohen's conventions to interpret effect size)

		Agreeableness			Conscientiousness			Extraversion			Neuroticism			Openness		
Attribute	**Mean±StdDev**	**r**	**t**	**c**	**r**	**t**	**c**	**r**	**t**	**c**	**r**	**t**	**c**	**r**	**t**	**c**
D_age	22.72±8.22	**0.024**	4.488	-	**0.067**	12.51	-	0.014	2.513	-	**-0.059**	10.974	-	**0.051**	9.371	-
D_gender	0.57±0.5	**0.063**	11.781	-	**0.066**	12.331	-	0.01	1.899	-	**0.193**	36.354	S	**-0.042**	7.864	-
D_interested_in	0.63±0.87	0.012	2.172	-	**-0.035**	6.416	-	**0.042**	7.711	-	**-0.08**	14.793	-	**0.034**	6.388	-
D_relationship_status	1.55±1.35	**-0.037**	6.832	-	0.013	2.397	-	**0.04**	7.457	-	**0.033**	6.096	-	**0.023**	4.269	-
D_timezone	-0.89±5.84	**-0.094**	17.398	-	**-0.083**	15.384	-	**-0.065**	12.123	-	**0.06**	11.197	-	**-0.111**	20.694	S
F_n_concentration	0.52±0.92	**0.027**	4.987	-	**0.119**	22.199	S	0.015	2.727	-	**-0.048**	8.889	-	**0.07**	13.021	-
F_n_diads	468.49±334.64	**0.023**	4.333	-	**0.055**	10.215	-	**0.215**	40.684	S	**-0.072**	13.423	-	0.005	1.003	-
F_n_education	2.03±1.25	**0.023**	4.296	-	**0.082**	15.23	-	0.005	0.94	-	**-0.042**	7.693	-	0.02	3.649	-
F_n_event	11.23±54.53	0.02	3.751	-	0.01	1.766	-	**0.056**	10.37	-	-0.017	3.238	-	**0.046**	8.437	-
F_n_group	45.96±51.27	-0.021	3.807	-	**-0.066**	12.303	-	**0.061**	11.344	-	**0.036**	6.593	-	**0.063**	11.748	-
F_n_like	220.84±418.39	**-0.039**	7.151	-	**-0.091**	16.942	-	-0.006	1.056	-	**0.114**	21.157	S	0.022	4.039	-
F_n_status	252.81±259.82	-0.008	1.557	-	**-0.05**	9.355	-	**0.063**	11.754	-	**0.078**	14.516	-	**0.03**	5.488	-
F_n_tags	234.79±294.28	**0.049**	9.111	-	-0.002	0.43	-	**0.133**	24.836	S	0.003	0.489	-	-0.018	3.33	-
F_n_work	0.8±1.19	**0.033**	6.123	-	**0.116**	21.597	S	**0.072**	13.316	-	**-0.047**	8.75	-	**0.049**	9.043	-

FS_c_n_of_concent	0.0±0.06	0.001	0.263	-	**0.023**	4.238	-	-0.016	2.887	-	-0.002	0.328	-	0.002	0.368	-
FS_c_n_of_event	0.01±0.3	0.009	1.591	-	-0.003	0.497	-	0.01	1.92	-	-0.008	1.433	-	0.009	1.634	-
FS_c_n_of_friend	30.31±35.24	0.008	1.393	-	0.007	1.352	-	**0.053**	9.906	-	**-0.034**	6.379	-	**-0.026**	4.903	-
FS_c_n_of_group	0.37±1.81	0	0.004	-	-0.008	1.432	-	-0.005	0.861	-	-0.006	1.091	-	-0.009	1.685	-
FS_c_n_of_like	2.23±21.06	-0.001	0.196	-	-0.012	2.274	-	-0.007	1.319	-	0.013	2.428	-	-0.002	0.347	-
FS_c_n_of_school	0.12±0.37	0.007	1.265	-	0.014	2.578	-	-0.02	3.668	-	-0.005	0.93	-	-0.011	2.027	-
FS_c_n_of_work	0.0±0.07	-0.002	0.347	-	0.016	3.023	-	-0.001	0.106	-	-0.007	1.245	-	-0.003	0.636	-
FS_eucl_dist	2.08±0.76	**-0.132**	24.718	S	**-0.048**	8.9	-	**-0.166**	31.104	S	**0.097**	18.088	-	**-0.044**	8.125	-
FS_f_n_of_concent	0.52±0.92	**0.027**	4.987	-	**0.119**	22.199	S	0.015	2.727	-	**-0.048**	8.889	-	**0.07**	13.021	-
FS_f_n_of_event	11.23±54.53	0.02	3.751	-	0.01	1.766	-	**0.056**	10.37	-	-0.017	3.238	-	**0.046**	8.437	-
FS_f_n_of_friend	468.49±334.64	**0.023**	4.333	-	**0.055**	10.215	-	**0.215**	40.684	S	**-0.072**	13.422	-	0.005	1.003	-
FS_f_n_of_group	45.96±51.27	-0.021	3.807	-	**-0.066**	12.303	-	**0.061**	11.344	-	**0.036**	6.593	-	**0.063**	11.748	-
FS_f_n_of_like	220.85±418.39	**-0.039**	7.153	-	**-0.091**	16.94	-	-0.006	1.054	-	**0.113**	21.153	S	0.022	4.042	-
FS_f_n_of_school	2.03±1.25	**0.023**	4.296	-	**0.082**	15.23	-	0.005	0.94	-	**-0.042**	7.693	-	0.02	3.649	-
FS_f_n_of_work	0.75±1.15	**0.027**	5.053	-	**0.107**	19.849	S	**0.068**	12.578	-	**-0.046**	8.443	-	**0.045**	8.27	-
FS_n_c_n_of_concent	0.0±0.08	-0.002	0.393	-	0.02	3.739	-	-0.014	2.646	-	0	0.011	-	-0.006	1.16	-
FS_n_c_n_of_event	0.0±0.04	0.003	0.536	-	0.001	0.246	-	-0.001	0.234	-	-0.004	0.772	-	0.015	2.721	-
FS_n_c_n_of_friend	0.43±0.36	0.019	3.491	-	-0.006	1.116	-	**0.028**	5.124	-	**-0.024**	4.535	-	-0.014	2.562	-
FS_n_c_n_of_group	0.02±0.09	0.002	0.285	-	0.006	1.062	-	-0.018	3.329	-	-0.007	1.292	-	**-0.023**	4.174	-

FS_n_c_n_of_like	0.02±0.07	-0.002	0.346	-	0	0.061	-	**-0.024**	4.382	-	0.004	0.75	-	-0.008	1.551	-
FS_n_c_n_of_school	0.14±0.43	0.005	0.933	-	0.005	0.852	-	-0.017	3.207	-	-0.004	0.69	-	-0.017	3.205	-
FS_n_c_n_of_work	0.01±0.1	-0.006	1.135	-	0.013	2.42	-	-0.001	0.173	-	-0.003	0.575	-	-0.002	0.334	-
FS_s_n_of_concent	0.13±0.54	0.013	2.433	-	**0.045**	8.319	-	-0.001	0.16	-	-0.015	2.757	-	**0.025**	4.667	-
FS_s_n_of_event	1.74±19.48	-0.001	0.159	-	0.007	1.336	-	0.012	2.167	-	-0.002	0.393	-	0.009	1.749	-
FS_s_n_of_friend	163.33±266.56	-0.012	2.252	-	**0.023**	4.255	-	0.003	0.525	-	-0.014	2.613	-	**-0.023**	4.238	-
FS_s_n_of_group	8.22±27.71	-0.008	1.514	-	0.003	0.471	-	-0.006	1.068	-	0	0.009	-	-0.001	0.213	-
FS_s_n_of_like	46.4±201.13	-0.009	1.589	-	-0.015	2.7	-	-0.013	2.465	-	0.013	2.452	-	0.003	0.525	-
FS_s_n_of_school	0.58±1.07	0.003	0.627	-	**0.044**	8.227	-	-0.013	2.36	-	-0.013	2.318	-	0	0.08	-
FS_s_n_of_work	0.24±0.77	0.008	1.395	-	**0.038**	6.963	-	0.005	0.953	-	-0.015	2.794	-	**0.023**	4.211	-
FS_second_agr	3.61±0.67	**0.043**	8.043	-	**0.029**	5.451	-	0.006	1.117	-	**-0.023**	4.35	-	0.004	0.812	-
FS_second_con	3.42±0.71	**0.029**	5.283	-	**0.065**	12.078	-	0.015	2.819	-	**-0.045**	8.347	-	-0.001	0.125	-
FS_second_ext	3.79±0.75	0.009	1.659	-	0.017	3.067	-	**0.07**	13.087	-	**-0.029**	5.462	-	-0.003	0.491	-
FS_second_neu	2.73±0.79	-0.015	2.696	-	**-0.036**	6.63	-	-0.017	3.225	-	**0.047**	8.716	-	0.003	0.473	-
FS_second_ope	3.87±0.65	0.002	0.403	-	**0.019**	3.548	-	0.003	0.534	-	-0.012	2.144	-	**0.102**	19.072	S
L_achieve	1.12±0.56	**0.064**	11.889	-	**0.112**	20.783	S	0.006	1.105	-	**-0.071**	13.183	-	0.006	1.104	-
L_adverb	3.37±1.15	**0.059**	11.034	-	-0.013	2.382	-	0.003	0.5	-	**0.044**	8.071	-	**0.075**	13.846	-
L_affect	5.69±1.56	**0.04**	7.346	-	0.02	3.695	-	**0.059**	10.912	-	**0.047**	8.771	-	0.001	0.108	-
L_anger	0.77±0.57	**-0.161**	30.221	S	**-0.146**	27.297	S	-0.002	0.338	-	**0.077**	14.374	-	0.002	0.317	-

L_anx	0.22±0.17	-0.014	2.502	-	-0.013	2.372	-	**-0.03**	5.504	-	**0.061**	11.389	-	**0.062**	11.486	-
L_article	3.76±1.25	**0.057**	10.518	-	**0.055**	10.275	-	0.005	0.99	-	**-0.059**	11.028	-	**0.171**	32.144	S
L_assent	0.68±0.52	**0.042**	7.798	-	**-0.027**	5.045	-	**0.046**	8.578	-	-0.002	0.317	-	**-0.044**	8.214	-
L_auxverb	6.3±1.81	**0.055**	10.174	-	0.005	0.96	-	0.008	1.468	-	**0.032**	6.001	-	**0.117**	21.902	S
L_bio	2.21±0.95	**-0.029**	5.306	-	**-0.08**	14.778	-	**0.081**	15.135	-	**0.057**	10.481	-	**0.063**	11.615	-
L_body	0.76±0.45	**-0.044**	8.188	-	**-0.102**	19.036	S	**0.042**	7.817	-	**0.053**	9.888	-	**0.058**	10.677	-
L_cause	0.87±0.43	-0.002	0.346	-	**-0.031**	5.747	-	**-0.033**	6.028	-	0.017	3.085	-	**0.116**	21.564	S
L_certain	1.03±0.53	**0.027**	4.915	-	**0.027**	4.936	-	0.014	2.591	-	0.005	0.956	-	**0.093**	17.285	-
L_cogmech	9.92±2.91	**0.05**	9.321	-	0.015	2.866	-	0.011	2.057	-	**0.026**	4.771	-	**0.159**	29.872	S
L_conj	3.44±1.26	**0.051**	9.517	-	**0.032**	5.989	-	**0.042**	7.774	-	0.02	3.659	-	**0.122**	22.767	S
L_death	0.15±0.17	**-0.077**	14.248	-	**-0.086**	16.036	-	**-0.063**	11.781	-	**0.03**	5.58	-	**0.088**	16.321	-
L_discrep	1.23±0.61	0	0.038	-	-0.01	1.869	-	-0.015	2.856	-	**0.03**	5.619	-	**0.062**	11.543	-
L_excl	1.7±0.74	0.021	3.926	-	**-0.037**	6.866	-	-0.021	3.868	-	**0.026**	4.906	-	**0.085**	15.866	-
L_family	0.28±0.28	**0.059**	11.03	-	**0.053**	9.867	-	**0.044**	8.117	-	0.007	1.33	-	**-0.038**	7.058	-
L_feel	0.5±0.29	0.012	2.161	-	**-0.034**	6.379	-	**0.024**	4.407	-	**0.03**	5.507	-	**0.066**	12.263	-
L_friend	0.18±0.2	**0.027**	5.06	-	**0.03**	5.504	-	**0.03**	5.517	-	0.02	3.711	-	0.013	2.4	-
L_funct	34.73±8.52	**0.071**	13.245	-	**0.037**	6.771	-	**0.028**	5.215	-	0.006	1.177	-	**0.158**	29.654	S
L_future	0.72±0.4	0.02	3.697	-	-0.022	4.016	-	-0.014	2.665	-	**0.027**	4.993	-	**0.066**	12.23	-
L_health	0.59±0.4	-0.014	2.666	-	0	0.083	-	0.021	3.942	-	**0.04**	7.393	-	0.02	3.72	-

L_hear	0.43±0.32	0.015	2.794	-	**-0.056**	10.373	-	-0.014	2.516	-	**0.029**	5.316	-	**0.099**	18.356	-
L_home	0.45±0.34	**0.073**	13.6	-	**0.06**	11.192	-	**0.036**	6.709	-	-0.012	2.162	-	**-0.026**	4.732	-
L_humans	0.63±0.39	-0.018	3.291	-	-0.008	1.489	-	**0.056**	10.301	-	-0.006	1.055	-	**0.051**	9.491	-
L_i	3.97±1.76	0.014	2.597	-	**-0.072**	13.409	-	**0.043**	8.043	-	**0.065**	12.138	-	**0.07**	12.932	-
L_incl	2.57±1.03	**0.087**	16.146	-	**0.088**	16.377	-	**0.105**	19.494	S	0.003	0.541	-	**0.13**	24.346	S
L_ingest	0.4±0.34	0.022	3.993	-	-0.004	0.714	-	0.013	2.421	-	-0.013	2.479	-	**0.049**	9.16	-
L_inhib	0.37±0.27	0.005	0.96	-	**0.029**	5.32	-	0.018	3.36	-	0.005	0.935	-	**0.028**	5.248	-
L_insight	1.3±0.6	0.021	3.849	-	**-0.025**	4.613	-	**-0.04**	7.388	-	0.017	3.206	-	**0.151**	28.202	S
L_ipron	3.1±1.18	**0.049**	9.05	-	0.004	0.705	-	-0.021	3.874	-	0.003	0.619	-	**0.119**	22.244	S
L_leisure	1.29±0.67	**0.078**	14.544	-	**0.044**	8.079	-	**0.072**	13.429	-	**-0.071**	13.171	-	**0.036**	6.65	-
L_money	0.37±0.31	**-0.027**	5.03	-	**0.044**	8.186	-	0.012	2.13	-	-0.016	2.916	-	**0.072**	13.356	-
L_motion	1.51±0.61	**0.098**	18.282	-	**0.088**	16.409	-	**0.071**	13.183	-	**-0.049**	9.173	-	0.02	3.692	-
L_negate	1.32±0.59	**-0.025**	4.567	-	**-0.052**	9.624	-	-0.017	3.123	-	**0.064**	11.844	-	**0.057**	10.507	-
L_negemo	1.92±0.88	**-0.133**	24.842	S	**-0.142**	26.607	S	**-0.028**	5.15	-	**0.123**	22.886	S	0.015	2.864	-
L_nonfl	0.15±0.22	0.012	2.138	-	-0.004	0.767	-	-0.002	0.297	-	-0.003	0.484	-	-0.003	0.53	-
L_number	0.5±0.34	**0.026**	4.803	-	**0.043**	8.043	-	**-0.054**	10.005	-	**-0.031**	5.763	-	**0.069**	12.884	-
L_past	1.79±0.81	**0.086**	16.061	-	0.02	3.77	-	0.014	2.505	-	-0.003	0.479	-	**0.038**	6.96	-
L_percept	1.8±0.69	**0.044**	8.164	-	**-0.051**	9.518	-	0.007	1.215	-	**0.036**	6.601	-	**0.142**	26.572	S
L_posemo	3.73±1.31	**0.138**	25.762	S	**0.12**	22.329	S	**0.091**	16.877	-	**-0.027**	5.054	-	-0.01	1.828	-

L_ppron	6.61±2.39	**0.03**	5.514	-	**-0.043**	7.942	-	**0.055**	10.256	-	**0.059**	10.882	-	**0.111**	20.738	S
L_preps	7.8±2.21	**0.073**	13.473	-	**0.092**	17.142	-	**0.039**	7.205	-	**-0.039**	7.152	-	**0.124**	23.055	S
L_present	7.43±2.02	**0.036**	6.631	-	0.001	0.262	-	**0.037**	6.765	-	**0.053**	9.754	-	**0.086**	15.982	-
L_pronoun	9.71±3.18	**0.041**	7.518	-	**-0.031**	5.719	-	**0.034**	6.261	-	**0.045**	8.417	-	**0.128**	23.928	S
L_quant	1.66±0.66	**0.032**	5.845	-	**0.063**	11.769	-	0.007	1.226	-	**-0.024**	4.483	-	**0.084**	15.562	-
L_relativ	10.53±2.82	**0.11**	20.5	S	**0.119**	22.142	S	**0.063**	11.678	-	**-0.053**	9.885	-	**0.042**	7.73	-
L_relig	0.42±0.47	**0.07**	13.023	-	**0.026**	4.861	-	0.009	1.669	-	**-0.069**	12.742	-	**0.033**	6.117	-
L_sad	0.42±0.32	-0.014	2.513	-	**-0.038**	7.034	-	-0.015	2.819	-	**0.104**	19.304	S	0.007	1.226	-
L_see	0.73±0.39	**0.035**	6.515	-	**-0.025**	4.637	-	0.009	1.674	-	**0.025**	4.706	-	**0.08**	14.797	-
L_sexual	0.56±0.45	**-0.033**	6.036	-	**-0.083**	15.436	-	**0.116**	21.702	S	**0.038**	6.986	-	0.019	3.478	-
L_shehe	0.53±0.51	-0.004	0.693	-	**0.034**	6.213	-	**0.026**	4.748	-	0.021	3.966	-	**0.067**	12.502	-
L_social	6.0±2.18	**0.051**	9.442	-	**0.022**	4.082	-	**0.065**	11.978	-	0.012	2.218	-	**0.101**	18.886	S
L_space	3.82±1.2	**0.057**	10.497	-	**0.048**	8.983	-	**0.073**	13.547	-	**-0.051**	9.45	-	**0.128**	23.832	S
L_swear	0.37±0.46	**-0.129**	24.049	S	**-0.148**	27.795	S	**0.037**	6.792	-	**0.038**	7.07	-	-0.019	3.536	-
L_tentat	1.6±0.71	0.018	3.36	-	-0.019	3.488	-	**-0.04**	7.383	-	0.004	0.686	-	**0.101**	18.819	S
L_they	0.26±0.24	-0.009	1.638	-	0.009	1.68	-	**-0.027**	5.032	-	0.006	1.047	-	**0.07**	12.967	-
L_time	5.28±1.7	**0.1**	18.596	-	**0.119**	22.178	S	**0.028**	5.113	-	**-0.026**	4.855	-	**-0.024**	4.361	-
L_verb	36.37±6.28	**-0.039**	7.135	-	**-0.063**	11.627	-	**-0.025**	4.614	-	**0.028**	5.256	-	**-0.142**	26.508	S
L_we	0.41±0.4	**0.05**	9.276	-	**0.034**	6.269	-	**0.031**	5.709	-	-0.012	2.304	-	**0.034**	6.388	-

L_work	1.29±0.81	**0.035**	6.469	-	**0.134**	24.997	S	**-0.048**	8.842	-	**-0.039**	7.293	-	**-0.036**	6.597	-
L_you	1.45±1.0	**0.03**	5.647	-	-0.008	1.543	-	**0.037**	6.876	-	0.018	3.379	-	**0.078**	14.557	-
S_betweenness	158341±237667	**0.025**	4.666	-	**0.047**	8.786	-	**0.183**	34.455	S	**-0.075**	13.886	-	0.006	1.197	-
S_brokerage	162649±242251	**0.026**	4.766	-	**0.047**	8.773	-	**0.183**	34.428	S	**-0.075**	13.976	-	0.005	0.913	-
S_density	0.04±0.09	0.005	0.899	-	-0.015	2.755	-	**-0.027**	4.991	-	-0.006	1.123	-	-0.019	3.443	-
S_n_betweenness	91.41±16.1	0.007	1.229	-	0.011	2.1	-	**0.038**	6.997	-	0.005	0.911	-	0.021	3.862	-
S_nbrokerage	0.48±0.04	-0.006	1.079	-	0.013	2.341	-	0.018	3.302	-	0.007	1.387	-	0.019	3.564	-
S_network_size	464.95±336.87	**0.025**	4.565	-	**0.056**	10.37	-	**0.21**	39.841	S	**-0.071**	13.11	-	0.004	0.784	-
S_network_size_inc	462.35±336.75	**0.025**	4.568	-	**0.056**	10.411	-	**0.211**	39.875	S	**-0.071**	13.123	-	0.004	0.673	-
S_transitivity	0.14±0.16	0	0.043	-	**-0.041**	7.522	-	**-0.092**	17.182	-	**0.022**	4.095	-	**-0.043**	7.99	-

Appendix E: Information Gain Ranking Result

	Agreeableness	Conscientiousness	Extraversion	Neuroticism	Openness
achieve	0.002683	0.009757	0.00286	0.00276	0.00252
adverb	0.001941	0.000555	0.00355	0.001282	0.00298
affect	0.002944	0.002767	0.0037	0.000453	0.00153
anger	0.012157	0.009102	0.00265	0.002779	0.00253
anx	0.000249	0.000157	0.00434	0.002366	0.00365
article	0.002255	0.00437	0.00316	0.001091	0.00774
assent	0.00088	0.000939	0.00323	0.000506	0.00236
auxverb	0.002549	0.001628	0.0026	0.00067	0.00459
bio	0.001612	0.00033	0.00432	0.001471	0.00164
body	0.002311	0.002886	0.003	0.002068	0.00254
cause	0.000937	0.000146	0.00451	0.000612	0.00488
certain	0.001773	0.003058	0.00307	0.000464	0.00381
cogmech	0.002526	0.002631	0.00253	0.000429	0.00652
conj	0.00276	0.002815	0.00235	0.000207	0.00316
death	0.003402	0.005131	0.00625	0.001291	0.00866
discrep	0.001292	0.000416	0.0034	0.001273	0.00264
excl	0.001355	0.000361	0.00406	0.000686	0.00375
family	0.001739	0.004193	0.00458	0.000138	0.00695
feel	0.000687	0.000229	0.00409	0.000998	0.00247
filler	0.001076	0.003071	0.00367	0.000692	0.00297
friend	0.002235	0.002994	0.00554	0.000192	0.00245
funct	0.003637	0.00549	0.0021	0	0.00693
future	0.001087	0.000719	0.00338	0.000449	0.00265
health	0.000712	0.000546	0.00409	0.001567	0.00195
hear	0.000453	0.000986	0.00388	0.000333	0.00458
home	0.002738	0.004462	0.00404	0	0.00465
humans	0.000938	0.000788	0.00486	0	0.00252
i	0.000591	0.00092	0.00249	0.001851	0.00207

incl	0.004029	0.006845	0.00323	0.000275	0.00323
ingest	0.000242	0.000212	0.00385	0.000371	0.00315
inhib	0.000753	0.002161	0.00383	0.000567	0.00213
insight	0.000908	0.000456	0.00501	0.000766	0.00839
ipron	0.002134	0.001654	0.00331	0	0.00497
leisure	0.001812	0.001775	0.00405	0.002559	0.00226
money	0.000533	0.002607	0.00387	0.00052	0.00275
motion	0.00396	0.006536	0.00344	0.001017	0.00347
negate	0.001122	0	0.00354	0.002058	0.002
negemo	0.008641	0.008614	0.0031	0.005949	0.00223
nonfl	0.000453	0	0.0034	0.000438	0.00245
number	0.001221	0.001072	0.00442	0.000262	0.00352
past	0.002464	0.000614	0.00454	0.000392	0.00249
percept	0.001183	0.001055	0.00325	0.00061	0.00552
posemo	0.011073	0.012322	0.00692	0.001504	0.00338
ppron	0.00141	0.000231	0.00208	0.001351	0.00226
preps	0.003784	0.010252	0.00238	0.000614	0.00393
present	0.002149	0.001818	0.00162	0.001321	0.00228
pronoun	0.001867	0.000804	0.00172	0.000785	0.00375
quant	0.002977	0.006204	0.00257	0.000279	0.00207
relativ	0.005125	0.011841	0.00223	0.001148	0.00392
relig	0.002822	0.002973	0.00333	0.001863	0.00378
sad	0.000405	0.000632	0.0041	0.004012	0.00189
see	0.000971	0	0.004	0.000243	0.00309
sexual	0.00085	0.00051	0.0091	0.000351	0.00178
shehe	0.000753	0.000886	0.0037	0.000524	0.00181
social	0.002292	0.004316	0.00379	0	0.00302
space	0.001831	0.003774	0.00312	0.001059	0.00456
swear	0.008793	0.0069	0.00349	0.001592	0.00266
tentat	0.001305	0.000346	0.0048	0.000372	0.00443
they	0.000457	0.001827	0.00433	0.000207	0.00296
time	0.005172	0.009625	0.00243	0.00062	0.00367
verb	0.000799	0.005347	0.00107	0.000252	0.00483

we	0.002284	0.004381	0.00477	0.001285	0.00147
work	0.001645	0.004457	0.00351	0.000826	0.00142
you	0.001012	0.000874	0.00311	0	0.00367

Appendix F: LIWC2007 Output Variable Information

Category	Abbrev	Examples	Words In Category
Linguistic Processes			
Word count	wc		
words/sentence	wps		
Dictionary words	dic		
Words>6 letters	sixltr		
Total function words	funct		464
Total pronouns	pronoun	I, them, itself	116
Personal pronouns	ppron	I, them, her	70
1st pers singular	i	I, me, mine	12
1st pers plural	we	We, us, our	12
2nd person	you	You, your, thou	20
3rd pers singular	shehe	She, her, him	17
3rd pers plural	they	They, their, they'd	10
Impersonal pronouns	ipron	It, it's, those	46
Articles	article	A, an, the	3
Common verbs	verb	Walk, went, see	383
Auxiliary verbs	auxverb	Am, will, have	144
Past tense	past	Went, ran, had	145
Present tense	present	Is, does, hear	169
Future tense	future	Will, gonna	48
Adverbs	adverb	Very, really, quickly	69
Prepositions	prep	To, with, above	60
Conjunctions	conj	And, but, whereas	28
Negations	negate	No, not, never	57
Quantifiers	quant	Few, many, much	89
Numbers	number	Second, thousand	34
Swear words	swear	Damn, piss, fuck	53
Psychological Processes			
Social processes	social	Mate, talk, they, child	455
Family	family	Daughter, husband, aunt	64
Friends	friend	Buddy, friend, neighbor	37
Humans	human	Adult, baby, boy	61
Affective processes	affect	Happy, cried, abandon	915
Positive emotion	posemo	Love, nice, sweet	406
Negative emotion	negemo	Hurt, ugly, nasty	499
Anxiety	anx	Worried, fearful, nervous	91
Anger	anger	Hate, kill, annoyed	184
Sadness	sad	Crying, grief, sad	101
Cognitive processes	cogmech	cause, know, ought	730

Insight	insight	think, know, consider	195
Causation	cause	because, effect, hence	108
Discrepancy	discrep	should, would, could	76
Tentative	tentat	maybe, perhaps, guess	155
Certainty	certain	always, never	83
Inhibition	inhib	block, constrain, stop	111
Inclusive	incl	And, with, include	18
Exclusive	excl	But, without, exclude	17
Perceptual processes	percept	Observing, heard, feeling	273
See	see	View, saw, seen	72
Hear	hear	Listen, hearing	51
Feel	feel	Feels, touch	75
Biological processes	bio	Eat, blood, pain	567
Body	body	Cheek, hands, spit	180
Health	health	Clinic, flu, pill	236
Sexual	sexual	Horny, love, incest	96
Ingestion	ingest	Dish, eat, pizza	111
Relativity	relativ	Area, bend, exit, stop	638
Motion	motion	Arrive, car, go	168
Space	space	Down, in, thin	220
Time	time	End, until, season	239
Personal Concerns			
Work	work	Job, majors, xerox	327
Achievement	achieve	Earn, hero, win	186
Leisure	leisure	Cook, chat, movie	229
Home	home	Apartment, kitchen, family	93
Money	money	Audit, cash, owe	173
Religion	relig	Altar, church, mosque	159
Death	death	Bury, coffin, kill	62
Spoken categories			
Assent	assent	Agree, OK, yes	30
Nonfluencies	nonflu	Er, hm, umm	8
Fillers	filler	Blah, Imean, youknow	9