Predicting Personality with Social Media

Jennifer Golbeck
Human-Computer Interaction Lab
University of Maryland
College Park, MD 20742 USA
jgolbeck@umd.edu

Cristina Robles
University of Maryland
College Park, MD 20742 USA
crobles@umd.edu

Karen Turner
University of Maryland
College Park, MD 20742 USA
kturner@umd.edu

CHI 2011, May 7–12, 2011, Vancouver, BC, Canada.
ACM 978-1-4503-0268-5/11/05.

Abstract
Social media is a place where users present themselves to the world, revealing personal details and insights into their lives. We are beginning to understand how some of this information can be utilized to improve the users' experiences with interfaces and with one another. In this paper, we are interested in the personality of users. Personality has been shown to be relevant to many types of interactions; it has been shown to be useful in predicting job satisfaction, professional and romantic relationship success, and even preference for different interfaces. Until now, to accurately gauge users' personalities, they needed to take a personality test. This made it impractical to use personality analysis in many social media domains. In this paper, we present a method by which a user's personality can be accurately predicted through the publicly available information on their Facebook profile. We will describe the type of data collected, our methods of analysis, and the results of predicting personality traits through machine learning. We then discuss the implications this has for social media design, interface design, and broader domains.

Keywords
personality, social networks, social media

ACM Classification Keywords
H5.3. Group and Organization Interfaces; Asynchronous interaction; Web-based interaction.

General Terms
Human Factors

Introduction

Social networking on the web has grown dramatically over the last decade. In January 2005, a survey of social networking websites estimated that among all sites on the web there were roughly 115 million members [7]. Just over five years later, Facebook alone has exceeded 500 million members. In the process of creating social networking profiles, users reveal a lot about themselves both in what they share and how they say it. Through self-description, status updates, photos, and interests, much of a user's personality comes out through their profile.

This paper attempts to bridge the gap between social media and personality research by using the information people reveal in their online profiles. Our core research question asks whether social media profiles can predict personality traits. If so, then there is an opportunity to integrate the many results on the implications of personality factors and behavior into the users' online experiences and to use social media profiles as a source of information to better understand individuals. For example, the friend suggestion system could be tailored to a user based on whether they are more introverted or extraverted.

Previous work has shown that the information in users' Facebook profiles is reflective of their actual personalities, not an "idealized" version of themselves [2]. This plus a broad user base makes Facebook an ideal platform for studying this connection.

We administered the Big Five Personality Inventory to 279 subjects through a Facebook application. In the process, we gathered all the public data from their Facebook profiles. This was aggregated, quantified, and passed through a text analysis tool to obtain a feature set. Using these statistics describing the Facebook profile of each user, we were able to develop a model that can predict personality on each of the five personality factors to within 11% of the actual values.

The ability to predict personality has implications in many areas. Existing research has shown connections between personality traits and success in both professional and personal relationships. Social media tools that seek to support these relationships could benefit from personality insights. Additionally, previous work on personality and interfaces showed that users are more receptive to and have greater trust in interfaces and information that is presented from the perspective of their own personality features (i.e. introverts prefer messages presented from an introvert's perspective). If a user's personality can be predicted from their social media profile, online marketing and applications can use this to personalize their message and its presentation.

We begin by presenting background on the Big Five Personality index and related work on personality and social media. We then present our experimental setup and methods for analyzing and quantifying Facebook profile information. To understand the relationship between personality and social media profiles, we present results on correlations between each profile feature and personality factor. Based on this, we describe the machine learning techniques used for classification and show how we achieve large and significant improvements over baseline classification on each personality factor. We conclude with a discussion of the implications that this work has for social media websites and for organizations that may utilize social media to better understand the people with whom they interact.

Background and Related Work

The Big Five Personality Inventory

The "Big Five" model of personality dimensions has emerged as one of the most well-researched and well-regarded measures of personality structure in recent years. The Big Five traits are characterized by the following:

- Openness to Experience: curious, intelligent, imaginative. High scorers tend to be artistic and sophisticated in taste and appreciate diverse views, ideas, and experiences.
- Conscientiousness: responsible, organized, persevering. Conscientious individuals are extremely reliable and tend to be high achievers, hard workers, and planners.
- Extroversion: outgoing, amicable, assertive. Friendly and energetic, extroverts draw inspiration from social situations.
- Agreeableness: cooperative, helpful, nurturing. People who score high in agreeableness are peace-keepers who are generally optimistic and trusting of others.
- Neuroticism: anxious, insecure, sensitive. Neurotics are moody, tense, and easily tipped into experiencing negative emotions.

Personality Research and Social Media

To the best of our knowledge, this is the first study looking at the relationship between profile information provided in social networks and personality traits. However, there have been a few previous studies on how personality relates to social networking more generally.

It has been shown in [21] that extroversion extroversion and conscientiousness positively correlate with the perceived ease of use of social media websites. Extroversion was also shown to have a positive correlation with perceived usefulness of such sites. Not surprisingly, extroversion was also shown to correlate with the size of a user's social network in several studies [1, 22, 23]. There have also been mixed results for other personality traits. Work in [23] showed that individuals with high agreeableness scores were selected more often as friends and that people tended to choose friends with similar agreeableness, extroversion, and openness scores. This was not repeated in [22], but a correlation between openness and number of friends.

Data Collection

We created a Facebook application with two functions. First, it administered a 45-question version of the Big Five Personality Inventory [11] to users. At the same time, it also collected all profile information about the user that was available to Facebook applications. We gathered a total of 161 statistics which are detailed below.

Structural Features

Through a Facebook application, we are able to collect information about the user's egocentric network. We first obtained a list of friends. We were interested in density, and Facebook provides some information about links between a user's friends. A separate query must be made for each pair of users to determine if they are or are not friends. It was not possible to submit a query for each pair of friends because the Facebook application would timeout; Facebook limits the time an application can run, and since each query is sent over the network, performance becomes an issue. Thus, we sampled 2,000 unique pairs of friends from a user's egocentric network and used that to determine the density of the network, i.e. what percentage of possible edges between friends exist.

Personal Information

Users provide a wealth of personal information. We collected everything available, even though some features would turn out to have no use in our analysis. The raw data included features like the user's name, birthday, relationship status, religion, education history, gender, and hometown. Most of this information was not required, so some users did not include all information. Where possible, we created additional features that indicated whether or not the user had included the information (e.g. was a religion or hometown provided or not), or how many items were listed (e.g. how many educational experiences were listed). These added features turned out to be much more useful and predictive than the original raw data. For example, from 279 users, 111 listed a religion. Within those 111 people were 82 different entries. This creates a space too sparse to do any statistical analysis, but just knowing if a person listed a religion or not reveals insights into what they are willing to share.

Activities and Preferences

Providing lists of personal activities or favorite things has always been a part of Facebook. Users list favorite TV shows, movies, music, book, quotes, as well as political and organizational affiliations and favorite activities. As was the case with religion described above, the space is far to sparse to do any analysis over the actual entries in these fields, so we created more companion measures. For lists of favorite things and activities, we counted the number of characters in the entry, roughly measuring how much information the user provided in each field. This included values of 0 for users who did not supply any information. For organizational affiliations, we counted the number listed and for political affiliations, we simply measured whether it was shared or not.

Language Features

Similar to the activities and likes described above, users also have opportunities to share more personal written information through the ”About Me” and ”blurb” text in their profiles, and through status updates. We collected these entries and also added features to measure the character length of each entry.

Previous research has shown that linguistic features can be used to predict personality traits [13, 17]. Since there is text available on users' Facebook profiles, there is potential to apply these linguistic analysis methods to help predict personality. However, the text samples used in earlier studies are much larger than are available to us through Facebook. Data collected in [17] was used in both studies mentioned above. They had three separate sources of text, ranging from an average of 1,770 words to over 5,000 words per person.

We are in a much different position pulling text from Facebook profiles. We combined status updates, the ”About Me” text, and the ”blurb” text from the profile into a single string so there would be enough text to analyze.The average number of words in these three strings combined was 26.6. Many users did not have enough words for linguistic analysis. Fifty-four subjects had no text at all in these three fields, and 112 had fewer than 10 words. We eliminated these users to the text analysis statistics would not be too noisy. This left us with 167 subjects with an average of 42.6 words per person. Note that the elimination of subjects with too little text should have a limited impact on our results. For each personality factor, a two-tailed Student's t-test showed no significant difference in the personality score between users with 10 words or more and users with fewer than 10.

Following the methods used in [13, 17] as well as other studies of Facebook behavior, such as [6], we utilized the Linguistic Inquiry and Word Count (LIWC) tool [16] to analyze the text. LIWC produces statistics on 81 different features of text in five categories. These include Standard Counts (word count, words longer than six letters, number of prepositions, etc.), Psychological Processes (emotional, cognitive, sensory, and social processes), Relativity (words about time, the past, the future), Personal Concerns (such as occupation, financial issues, health), and Other dimensions (counts of various types of punctuation, swear words).

Internal Facebook Statistics

A number of features were available that described a user's experience, settings, and history with facebook. This included the user ID, an integer value that corresponds to when the user joined the network (lower values indicate an earlier join time), the unix timestamp of their last profile update, the number of notes (short messages) posted, and other features that proved less useful such as the URL of the profile picture, whether or not their profile was blocked (no one's was), whether the person was an app user (everyone was), and if they had provided a status update (everyone had).

Personality and Profile Correlations

We had 279 subjects who completed the personality inventory, but we only used data from the 167 subjects who had at least 10 words among all of their text fields so we could perform a linguistic analysis. Demographic information was pulled from their Facebook profiles. Among these subjects, the average age was 31.2 years (std dev 8.7). Of those reporting gender, 68 were female and 61 were male (38 did not report). In terms of location, 82.6% (138) were from the United States with the remaining subjects coming from India (8), Australia (7), Italy (7), and others (7).

We found many weak correlations between users' profile features and personality scores. This echos previous results of linguistic analysis and personality found in [13]. These are reported in table and statistically significant correlations ($p < 0.05$) are bolded. Below, we discuss some of the more interesting relationships.

Since linguistic features made up half of all features considered, this is where we found the highest number of correlations. They also largely make intuitive sense. Conscientiousness was the personality factor that had the most correlations with linguistic measures. The frequency of swear words is negatively correlated with conscientiousness ($\rho = 0.171$). It is also negatively correlate with words that describe perceptual processes (seeing, hearing, feeling) and the subset of words specifically about seeing (ρ = -0.195 and -0.227 respectively). This suggests that more conscientious people do not write about the things they saw or heard. On the other hand, conscientiousness is positively correlated with words surrounding social processes ($\rho = 0.264$), as well as the subset of words that describe people ($\rho = 0.203$). In other words, more conscientious people are likely to discuss other people.

Affective processes - words describing feelings - also had interesting correlations. The use of affective process words in general, and positive emotion words in particular, correlates positively with agreeableness (ρ= 0.203 and 0.167 respectively). However, the frequency of words that express anxiety, not surprisingly, correlates positively with neuroticism ($\rho = 0.192$).

Although we only measured two network structure features - number of friends and density - both showed correlations with personality features. Extroverts tended to have more friends ($\rho = 0.186$), but their networks tended to be more sparse ($\rho = -0.224$). Since we expect extroverts to reach out and make friends with many different groups of people, it would follow that their networks are less dense since those friends are less likely to know one another. Density was also negatively correlated with openness ($\rho = -0.152$), suggesting people who are more open also tend to have friends who are more dispersed socially.

Extraversion and openness were also factors with correlations to reported activities and interests. The length of reported activities correlated positively with extraversion ($\rho = 0.186$). It could be that extroverts participate in more activities or they are simply more likely to describe them. The length (in characters) of subjects' favorite books lists had a positive correlation with openness ($\rho = 0.158$).

One of the most unusual correlations we found was between neuroticism and the character length of a subject's last name. There was a significant positive correlation ($\rho = 0.184$). We offer that a lifetime of having one's long last name misspelled may lead to a person expressing more anxiety and quickness to anger. However, there may be some unseen factor at work or this could be an unusual positive, significant but false correlation.

In addition to computing a series of correlation coefficients, we also compared values between groups and found significant differences. Table 3 shows the two significantly different populations we found. Women were found to be significantly more conscientious, agreeable, neutroic than men. This matches and differs from some previous results; [4] found women had higher rates of agreeableness and neuroticism. The previous work did not find the conscientiousness difference we found here. This could be because the population of Facebook users differs from the general public or there is a self selection effect in our subjects.

We also created many profile features that indicated the absence of presence of a particular type of information. For example, users can choose to share a URL to an external personal website. When we compared personality test results for users who provided a website URL to those who do not. Users providing a website were significantly more open than those who did not.

Table 1: Average scores on each personality factor on a normalized 0-1 scale

	Open.	Consc.	Extra.	Agree.	Neuro.
Avg	0.70	0.62	0.58	0.68	0.47
StDev	0.15	0.16	0.21	0.14	0.18

Table 3: T-tests for several features in our datasets. Bolded values are significant for $p < 0.05$.

Value	Open.	Consc.	Extra.	Agree.	Neur.
Male	3.841	**3.313**	3.145	**3.638**	**2.680**
Female	3.671	**3.582**	3.476	**3.806**	**2.996**
p	*0.101*	*0.018*	*0.018*	*0.095*	*0.018*
No Website	**3.710**	3.495	3.264	3.697	2.900
Website	**4.010**	3.498	3.508	3.773	2.770
p	*0.003*	*0.978*	*0.071*	*0.382*	*0.275*

Predicting Personality

Our feature set for each user included all meaningful features. We excluded those which could not be quantified (e.g. picture URL), for which the value was the same for all users (e.g. if their profile was blocked), or where the data was so sparse that it would not be predictive (e.g. personal website URL). Where possible, we included our companion statistics on these features (e.g. while the actual website URL was not used, a feature indicating presence or absence of the URL was included). Linguistic features were included as described above.

We also added five additional features. We ran a multiple linear regression analysis for each personality factor, producing a vector of weights for each feature. The dot product of the weight vector and the feature vector was computed for each user and for each personality feature to create five composite features.

In total, we had 74 features per user. To predict the score of a given personality feature, we performed a regression analysis in Weka [8] with a 10-fold cross-validation with 10 iterations using two algorithms: M5$'$Rules [10], a rule-based variation of the M5$'$ algorithm [18], and Gaussian Processes.

On a normalized 0-1 scale, the Mean Absolute Error for each personality factor was roughly 11%. Results are shown in

Table 2: Pearson correlation values between feature scores and personality scores. Significant correlations are shown in bold for $p < 0.05$. Only features that correlate significantly with at least one personality trait are shown.

	Open.	Consc.	Extra.	Agree.	Neuro.
Linguistic Features					
Swear Words	0.006	**-0.171**	0.032	-0.084	-0.120
Social Processes (e.g. Mate, talk, they, child)	0.010	**0.264**	0.091	-0.022	-0.142
Human Words (e.g. baby, man)	0.078	**0.203**	0.070	-0.050	-0.062
Affective Processes (e.g. Happy, cried, abandon)	0.105	-0.009	0.136	**0.203**	0.038
Positive Emotions (e.g. Love, nice, sweet)	0.052	0.045	0.117	**0.167**	-0.013
Anxiety Words (e.g. Worried, fearful, nervous)	0.044	-0.150	0.008	0.101	**0.192**
Perceptual Processes (e.g. Observing, heard, feeling)	-0.040	**-0.195**	**-0.163**	-0.027	0.096
Seeing Words (e.g. View, saw, seen)	0.060	**-0.227**	-0.112	0.013	0.067
Biological Processes (e.g. Eat, blood, pain)	-0.014	0.042	0.038	**0.154**	0.067
Ingestion Words (e.g. Dish, eat, pizza)	-0.098	-0.050	0.029	0.031	**0.207**
Work Words (e.g. Job, majors, xerox)	0.134	0.096	**0.154**	0.048	-0.044
Money Words(e.g. Audit, cash, owe)	**-0.161**	0.024	0.012	-0.006	0.029
Structural Features					
Number of Friends	-0.094	-0.078	**0.186**	0.013	-0.069
Egocentric Network Density	**-0.152**	0.050	**-0.224**	0.059	0.032
Activities and Preferences					
Activities (char length)	0.115	0.095	**0.188**	0.066	-0.145
Favorite Books (char length)	**0.158**	-0.093	0.019	0.082	0.028
Personal Information					
Relationship Status (none listed,single, not single)	0.093	0.071	**0.194**	0.040	-0.036
Last Name length in characters	0.012	-0.111	0.000	-0.044	**0.184**

table 2. This means we can predict a user's score for a personality trait to within just more than one tenth of its actual value.

We also find good results in the correlation coefficients, as shown in table 4. M5′Rules produce results with strong correlations ($\rho >= 0.5$) on Openness, Conscientiousness, Extroversion and Neuroticism, and a medium correlation ($0.3 <= \rho < 0.5$) on Agreeableness.

The Gaussian Processes correlations were not quite as impressive. They were smaller than the correlations for M5′Rules in all cases, with two weak correlations ($0.1 <= \rho < 0.3$) on Openness, Agreeableness and Neuroticism, and no real correlation on Extroversion and Conscientiousness.

Table 4: Mean Absolute Error and correlation coefficients of predicted personality values for each factor.

Factor	M5′Rules	Gaussian	M5′Rules	Gaussian
	MAE		Correlation Coefficient	
Open.	0.099	0.117	0.653	0.179
Consc.	0.104	0.117	0.595	0.094
Extra.	0.138	0.124	0.553	0.050
Agree.	0.109	0.117	0.482	0.150
Neuro.	0.127	0.117	0.531	0.106

Discussion

The question that arises from this research is how the results can be used. Drawing on research results that connect personality type to behavior and preferences, there is potential to integrate previous personality results into social media as a way to enhance the accuracy of certain features or the user's experience.

Research on interface preference and personality typeshowed that users preferred interfaces designed to represent personalities that most closely matched their own [14, 12]. This has significant implications for this work. With the ability to infer a user's personality, social media websites, e-commerce retailers, and even ad servers can be tailored to reflect the user's personality traits and present information such that users will be most receptive to it. For example, the presentation of Facebook ads could be adjusted based on the personality of the user. Similarly, product reviews from authors with personality traits similar to the user could be highlighted to increase trust and perceived usefulness by the user. Customized website "skins" could be created for different user personality types, as suggested in [3]. Our methods provide a straightforward way to obtain personality profiles of users without the burden of tests, and this will make it much easier to create personality-oriented interfaces.

This same idea can be extended even further to advertising. While results of integrating personality to marketing have been mixed, some work has demonstrated connections between marketing techniques and consumer personality [15]. For e-commerce marketers, both those who advertise on Facebook and elsewhere, utilizing social media profiles as a way to determine consumer personality can make it easy to implement existing techniques that benefit from this knowledge of consumer background.

Recommender systems may also benefit from integrating predicted personality values. Results showing correlations between personality and music taste are well established in the literature [20, 5, 9, 19]. Inferring personality traits from Facebook profiles may allow recommender systems to improve their accuracy by recommending music, and possibly other items, that are tailored to the user's personality profile.

Conclusions

In this paper, we have shown that a users' Big Five personality traits can be predicted from the public information they share on Facebook. Our subjects completed a personality test and through the Facebook API, we collected publicly accessible information from their profiles. After processing this data, we found many small correlations in the data. Using the profile data as a feature set, we were able to train two machine learning algorithms - m5sup′Rules and Gaussian Processes - to predict each of the five personality traits to within 11% of its actual value.

With the ability to guess a user's personality traits, many opportunities are opened for personalizing interfaces and information. We discussed some of these opportunities for marketing and interface design above. However, there is much work to be pursued in this area.

One area that deserves attention is the connection between personality and the actual social network. We considered two structural features - number of friends and network density - but we did not look at personality scores between friends. Understanding the connections between personality, tie strength [6], trust [7], and other related factors is an open space for research. By improving our knowledge of these relationships, we can begin to answer more sophisticated questions about how to present trusted, socially-relevant, and well-presented information to users.

Acknowledgments

Research was sponsored, in part, by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-09-2-0053. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

References

[1] A. Acar and M. Polonsky. Online Social Networks and Insights into Marketing Communications. *Journal of Internet Commerce*, 6(4):55–72, 2008.

[2] M. Back, J. Stopfer, S. Vazire, S. Gaddis, S. Schmukle, B. Egloff, and S. Gosling. Facebook Profiles Reflect Actual Personality, Not Self-Idealization. *Psychological Science*, 21(3):372, 2010.

[3] W.-P. Brinkman and N. Fine. Towards customized emotional design: an explorative study of user personality and user interface skin preferences. In *EACE '05: Proceedings of the 2005 annual conference on European association of cognitive ergonomics*, pages 107–114. University of Athens, 2005.

[4] P. Costa Jr, A. Terracciano, and R. McCrae. Gender differences in personality traits across cultures: Robust and surprising findings. *Journal of Personality and Social Psychology*, 81(2):322–331, 2001.

[5] S. Dollinger. Research Note: Personality and Music Preference: Extraversion and Excitement Seeking or Openness to Experience? *Psychology of Music*, 21(1):73, 1993.

[6] E. Gilbert and K. Karahalios. Predicting tie strength with social media. In *Proceedings of the 27th international conference on Human factors in computing systems*, pages 211–220. ACM New York, NY, USA, 2009.

[7] J. Golbeck. *Computing and Applying Trust in Web-based Social Networks*. PhD thesis, University of Maryland, College Park, MD, USA, April 2005.

[8] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. Witten. The WEKA data mining soft-

ware: An update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18, 2009.

[9] C. Hansen and R. Hansen. Constructing personality and social reality through music: Individual differences among fans of punk and heavy metal music. *Journal of broadcasting & electronic media*, 35(3):335–350, 1991.

[10] G. Holmes, M. Hall, and E. Prank. Generating rule sets from model trees. *Advanced Topics in Artificial Intelligence*, pages 1–12, 1999.

[11] O. D. John. Big five inventory, 2000.

[12] A. Karsvall. Personality preferences in graphical interface design. In *NordiCHI '02: Proceedings of the second Nordic conference on Human-computer interaction*, pages 217–218, New York, NY, USA, 2002. ACM.

[13] F. Mairesse, M. Walker, M. Mehl, and R. Moore. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of Artificial Intelligence Research*, 30(1):457–500, 2007.

[14] C. Nass and K. M. Lee. Does computer-generated speech manifest personality? an experimental test of similarity-attraction. In *CHI '00: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 329–336, New York, NY, USA, 2000. ACM.

[15] G. Odekerken-Schr
"oder, K. De Wulf, and P. Schumacher. Strengthening outcomes of retailer-consumer relationships:: The dual impact of relationship marketing tactics and consumer personality. *Journal of Business Research*, 56(3):177–190, 2003.

[16] J. Pennebaker, M. Francis, and R. Booth. Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates*, 2001.

[17] J. Pennebaker and L. King. Linguistic styles: Language use as an individual difference. *Journal of personality and social psychology*, 77(6):1296–1312, 1999.

[18] J. Quinlan. Learning with continuous classes. In *5th Australian joint conference on artificial intelligence*, pages 343–348. Citeseer, 1992.

[19] D. Rawlings and V. Ciancarelli. Music preference and the five-factor model of the NEO Personality Inventory. *Psychology of Music*, 25(2):120, 1997.

[20] P. Rentfrow and S. Gosling. The do re mi's of everyday life: The structure and personality correlates of music preferences. *Journal of Personality and Social Psychology*, 84(6):1236–1256, 2003.

[21] P. Rosen and D. Kluemper. The Impact of the Big Five Personality Traits on the Acceptance of Social Networking Website. *AMCIS 2008 Proceedings*, page 274, 2008.

[22] J. Schrammel, C. Köffel, and M. Tscheligi. Personality traits, usage patterns and information disclosure in online communities. In *BCS HCI '09: Proceedings of the 2009 British Computer Society Conference on Human-Computer Interaction*, pages 169–174, Swinton, UK, UK, 2009. British Computer Society.

[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*, 78(2):509–538, 2010.