Mining Facebook Data for Predictive Personality Modeling

Dejan Markovikj
Faculty of Computer Science
and Engineering
Ss. Cyril and Methodius University
Rugjer Boshkovikj 16
Skopje, Macedonia
dejan@markovikj.com

Sonja Gievska
Faculty of Computer Science
and Engineering
Ss. Cyril and Methodius University
Rugjer Boshkovikj 16
Skopje, Macedonia
sonja.gievska@finki.ukim.mk

Michal Kosinski
The Psychometrics Centre
University of Cambridge
Psychology, Free School Lane
Cambridge CB2 3RQ
United Kingdom
mk583@cam.ac.uk

David Stillwell
The Psychometrics Centre
University of Cambridge
Psychology, Free School Lane
Cambridge CB2 3RQ
United Kingdom
ds617@cam.ac.uk

Abstract

Beyond being facilitators of human interactions, social networks have become an interesting target of research, providing rich information for studying and modeling user's behavior. Identification of personality-related indicators encrypted in Facebook profiles and activities are of special concern in our current research efforts. This paper explores the feasibility of modeling user personality based on a proposed set of features extracted from the Facebook data. The encouraging results of our study, exploring the suitability and performance of several classification techniques, will also be presented.

Introduction

Social networks have become widely-used and popular mediums for information dissemination as well as facilitators of social interactions. Users' contributions and activities provide a valuable insight into individual behavior, experiences, opinions and interests. Considering that personality, which uniquely identifies each one of us, affects a lot of aspects of human behavior, mental processes and affective reactions, there is an enormous opportunity for adding new personality-based qualities to user interfaces. Personalized systems used in domains such as, e-learning, information filtering, collaboration and e-commerce could greatly benefit from a user interface that adapts the interaction (e.g., motivational strategies, presentation styles, interaction modalities and recommendations) according to user's personality. Having captured past user interactions is only a starting point in explaining the user behavior from a personality point of view.

Several well studied personality models have been proposed, the Big Five model established as the most popular one (Goldberg 1992). Regularity in someone's behavior over time and situations uniquely identifies her personality type along Big Five dimensions: Openness to experience, Neuroticism, Extraversion, Agreeableness and Conscientiousness.

This research builds upon previous interdisciplinary research works regarding personality as it pertains to the design of intelligent interactive systems. The new communication technologies have brought more information to consider, though the process of their utilization is far from straightforward. Intelligent technologies are expected to play a prominent role in bringing these data to a new level of usability.

A variety of Facebook variables were expected to play a prominent role in establishing appropriate context for our particular investigations. Facebook profiles and activities provide valuable indicators of user's personality, revealing the actual, rather than idealized or projected personality (Back et al. 2010). Our research has two interconnected objectives: (1) to identify the relevant personality-related indicators that are explicitly or implicitly present in Facebook user data; and (2) to explore the feasibility of predictive personality modeling to support future intelligent systems.

We hypothesized that increasing the relevance of what is included in the model, and considering features drawn from a variety of sources may lead to better performance of the classifiers under investigation. The choice to include a feature was based on whether the previous research had underlined the importance of such a choice and its relevance to the objectives of this research. Our research is currently focused on investigating the suitability and performance of various classification techniques for personality modeling.

Related Work

Data mining techniques play a fundamental role in extracting correlation patterns between personality and variety of user's data captured from multiple sources. Generally, two approaches were adopted for studying personality traits of social network users. The first approach uses a variety of machine learning algorithms to build models based on social network activities only. The second one extends the personality-related features with linguistic cues (Mairesse et al. 2007; Oberlander and Nowson 2006).

Several classification and regression techniques were used to build predictive personality models along the five personality dimensions using the linguistic features of a dataset comprised of few thousand essays solicited from introductory psychology students (Mairesse et al. 2007). The Linguistic Inquiry and Word Count – LIWC (http://www.liwc.net) was used as a tool for linguistic analysis. The reported precisions of the classifiers were in the range of 54-62% for all traits. In (Oberlander and Nowson 2006), SMO and Naïve Bayes were used for modeling four out of five personality dimensions by extracting n-gram features from a corpus of personal web-blogs. Their results point out to the importance of the process of feature selection in increasing the classifiers precision yielding 83%-93% for automatic feature selection. We would like to point out the differences in the datasets used in these studies compared to ours, namely different solicitation methods and the sources from which they were collected.

The correlation between users' social network activity and personality has been the focus of several studies in the last decade (Bai, Zhu, and Cheng 2012; Golbeck, Robles, and Turner 2011; Bachrach et al. 2012). Personality traits of the Chinese most popular social network RenRen users were analyzed in (Bai, Zhu, and Cheng 2012). C4.5 Decision Trees have shown the best results, yielding 69-72 percent accuracy, for a combination of features related to users' network activity along with affective linguistic features extracted from statuses and blog posts.

The work most closely related to our own is (Golbeck, Robles, and Turner 2011). Two regression techniques, namely m5sup/Rules and Gaussian Processes, were applied to build predictive personality models. The authors consider users' Facebook data through parameters such as structural characteristics, personal info, activities and preference, in addition to the linguistic attributes extracted with LIWC from the users' statuses. The lack of demographic diversity in participant sampling was one of the major drawbacks for generalizing the results of the last two studies, Chinese population and authors' Facebook friends respectively.

Few studies using considerably larger number of instances from the same dataset under our investigation have a rather different objective from ours, namely to examine the correlations between the personality traits and Facebook activity data (Bachrach et al. 2012) and the associations between personal attributes and Facebook Likes (Kosinski, Stillwell, and Graepel 2013). These studies were not meant to look at the rich linguistic patterns that occur in the language use on social networks, which is in the focus of this research.

Experiments

Dataset

A sample of 250 user instances from Facebook (activity and demographic data) with approximately 10,000 status updates used in our study was provided by the MyPersonality project (http://mypersonality.org/wiki; Celli et al. 2013).

Features

A set of 725 features included in the model could be broken down into five groups. The features have been derived from both, theoretical and empirical works in previous research studies. Some of these categories represent the features obtained from the Facebook dataset, namely demographics, activity-related parameters, status updates and egocentric network data. In other words, the social contexts in which people are embedded allow us to search for patterns indicative of the five personality traits. We wanted to identify those who are instrumental in answering the question posed in this research.

Another set of features have been derived by applying a set of natural language processing techniques and specific word classification schemes reported to be relevant for the problem under our investigation (Pennebaker and King 1999).

MyPersonality Data

The study of personality reflected in user's Facebook activities includes a wide range of features. Some of the most intuitively predicted indices are the statistical data for user's activities (e.g., number of likes, statuses, groups, tags, events). Demographic characteristics such as: age and gender, were accounted for since their effect is known to manifest in the context under investigation (Golbeck, Robles, and Turner 2011). Egocentric network parameters representing the number of friends, and measures such as density, brokerage, and betweenness, provide additional insight into user's social behavior instrumental in assessing personality along several dimensions (Celli et al. 2013).

Basic Linguistic Features

Several studies have pointed out to the significant correlation between personality and the spoken or written linguistic cues (Pennebaker and King 1999; Mairesse et al. 2007). A selected set of categories, which correspond to the LIWC linguistic process variables (e.g., word count, words per sentence/status, sentences per status, punctuation count and lexical diversity) have been accounted for. In addition, a list of words related to online jargon, chat acronyms, emoticons, profanities, and slang words were included to account for the specific language use exhibited by social network users.

POS Tag Parameters

Additional linguistic analysis utilizing the Brown corpus included in the NLTK (http://nltk.org) toolkit was performed resulting in extended set of linguistic features, which includes the numbers and average numbers of words in specific grammar categories (e.g., adverbs, adjectives, verbs, pronouns). Much work has been done to understand how these linguistic cues are related to the five personality traits (Pennebaker and King 1999; Mairesse et al. 2007).

Afinn Parameters

The inclusion of Afinn attributes substantially greater breadth to the personality models, namely, the consideration of personality indices related to psychological processes. Work relating the affective processes with personality originates with (Pennebaker and King 1999). From the complete list of 2,500 Afinn words (http://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=6010), each annotated with emotional valence ranking in the range of -5 to 5, only a selected number of words were deemed relevant to our study. In particular, the average and summative valence of affective and non-affective words in user's statuses, affective word count, and the number of words with specific valence.

H4Lvd Parameters

General Inquirer is a tool for content analyses of textual data, which include 182 word tag categories merging four sources (www.wjh.harvard.edu/~inquirer/Home.html), including HIV-4 and Lasswell value dictionary utilized in our set of features. A word is classified using an intensity level scale, which is a combination of different valence categories such as, positive vs. negative, strong vs. weak and active vs. passive. The H4Lvd categories span from affective words and motivation to socially-related and communication-specific ones. The rationale for including the H4Lvd as opposed to the more frequently used LIWC tool was the larger number of words per category and finer sophistication of subcategories.

Results

A number of studies have shown that restricting features used for classification to those that come through as strongly correlated, improves performance in predictive modeling. For that reason, Pearson correlation analysis was conducted for the entire set of features. A discussion of statistically significant correlations of interest follows (*, ** and *** are used to indicate p-values at 0.05, 0.01 and 0.001 respectively). As hypothesized, the most significant correlation effect was found for the number of friends and extroversion (r=0.4***), which is in line with the results of other researchers. A very accurate and realistic indicator of personality traits was the examination of the egocentric network parameters. Significant correlations were found between transitivity and extraversion (r=0.29***) as well as transitivity and agreeableness (r=0.21***). The results have shown that extroversion is also correlated with density, brokerage and betweenness with a coefficient of 0.31***, 0.28*** 0.27***, respectively. This pattern of consistency found for extroversion did not apply to the rest of the personality traits. The nature of the modeled problem has revealed itself consistently in other Facebook activity parameters, namely the correlation between extraversion and the number of tags (r=0.27***), as well as the ones between the number of events and conscientiousness (r=0.28*). It was intuitively assumed that gender would be a factor of significance, though a single significant correlation was found for neuroticism (r=0.22***).

Not surprisingly, the linguistic cues represented as Afinn affective words were correlated to the personality traits with the exception of openness to experiences; the remaining four personality traits exhibited correlation with the average valence of words (r between 0.14** – 0.21**). A significant trend was observed between conscientiousness and average number of words with valence +2 (r=0.24**).

The strongest correlation for the H4Lvd word categories could be summarized as follows: qualities that can be detected by human eye and agreeableness (r=0.17**); affective words and neuroticism (r=0.16**); and adjectives describing people apart from their relationships and extraversion (r=0.14*).

Evaluation of Classification Models

To validate our models, we run a set of experiments to investigate how accurate they are at predicting personality traits. We tested a number of popular classification algorithms, but Support Vector Machines (SVM) and their more efficient and optimized versions, Simple Minimal Optimization (SMO) and Boost algorithms (MultiBoostAB and AdaBoostM1), have shown significant precision advantage, therefore the following discussion is restricted to them. Not surprisingly, most of the algorithms performed bellow satisfactory precision levels when using all features (authors' self-imposed precision was set to 75%).

We then proceeded with refining the classification mechanisms by considering better sampling of features based on the Pearson correlation coefficient. The choice to include features maximizing the correlation with a trait and minimizing the correlation with other features, has limited the number of features to 5-16, and yielded precision improvement of up to 78%. The improvement was spread across other algorithms, such as decision trees and rule-based algorithms; an expected result since these algorithms are proven to perform better for selection of features with higher information gains.

There was a great variance in the types of features associated with each personality trait, confirming the complexity and multidimensional aspects of the problem under research. What was perhaps more informative for our models was the selection of features within particular group varied as well. The differences in Afinn and H4Lvd groups could be explained by the existing semantic separation between word categories.

An employment of ranking algorithms attributed significant performance gains for the SMO classifier as shown in Table 1, by improving the relevance of the selected features.

trait/ measure	TP Rate	FP Rate	Precision	Recall	ROC Area
OPE	0.948	0.1	0.948	0.948	0.924
CON	0.92	0.08	0.92	0.92	0.92
EXT	0.928	0.092	0.928	0.928	0.918
AGR	0.86	0.144	0.86	0.86	0.858
NEU	0.864	0.162	0.864	0.864	0.851

Table 1: SMO classifier measures for feature selection using ranking algorithms. OPE - 150; NEU – 70; CON – 130; EXT – 150; AGR – 140 best rank features.

Closer inspection of data revealed that the best ranking features varied between traits as presented:

Openness to experience: geo-location, number of groups, punctuation, H4Lvd (weakness, wealth, enlightenment participants, aesthetic skills)

Conscientiousness: H4Lvd (reaping affect, submission to authority, dependence of others, vulnerability to others, social relations adjectives)

Extraversion: brokerage, network size, punctuation, adjectives, verbs, H4Lvd (reaping affect, decrease as process, other relations)

Agreeableness: geo-location. Afinn number of words with valence +3/+5, transitivity, punctuation, “to”, H4Lvd (qualities, senses-detectable degrees of qualities, shame)

Neuroticism: gender, network size, number of groups, number of tags, “to”, H4Lvd (positive feelings, acceptance, appreciation, emotional support, enjoyment of a feeling, confidence, interest and commitment).

Conclusions

The results of our initial investigation in personality modeling based on Facebook data are encouraging evidence that by selecting the most indicative features the precision of the classifiers could be improved. Extracting qualitative knowledge from the large quantities of data is just the beginning of our search for meaning and plausible explanation of personality-determined social network activities.

The challenges in providing similar and even better performance results for larger datasets may require consideration of additional features, more sophisticated data processing and classification techniques. Our future research efforts are directed toward augmenting the personality models with a more qualitative features previously only quantitatively accounted for (e.g., pages, groups, events, likes). We deferred to future analysis the investigation of the relevance and sensitivity of various indicators on personality traits prediction. Our long-term goal is to demonstrate the usefulness of the predictive models in terms of exploratory scenario-based case studies for selected domains.

References

Bachrach, Y., Kosinski, M., Graepel, T., Kohli, P., and Stillwell, D. 2012. Personality and Patterns of Facebook Usage. In *Proceedings of the Third Annual ACM Web Science Conference,* 36-44. New York, NY, USA: ACM Press.

Back, M.; Stopfer, J.; Vazire, S.; Gaddis, S.; Schmukle, S.; Egloff, B.; and Gosling, S. 2010. Facebook Profiles Reflect Actual Personality, Not Self-Idealization. *Psychological Science* 21(3): 372-374.

Bai, S., Zhu, T., and Cheng, L. 2012. Big-Five Personality Prediction Based on User Behaviors at Social Network Sites. eprint arXiv:1204.4809.

Celli, F., Pianesi, F., Stillwell, D. S., and Kosinski, M. 2013. Workshop on Computational Personality Recognition (Shared Task). The Seventh International AAAI Conference on Weblogs and Social Media. Boston, MA, USA.

Golbeck, J., Robles, C., and Turner, K. 2011. Predicting Personality with Social Media. In *Proceedings of the 2011 Annual Conference on Human Factors in Computing Systems,* 253–262. New York, NY, USA: ACM Press.

Goldberg, L. R. 1992. The Development of Markers for the Big Five Factor Structure. *Psychological Assessment* 4(1): 26-42.

Kosinski, M., Stillwell D.J., and Graepel, T. 2013. (in press) Private Traits and Attributes are Predictable from Digital Records of Human Behavior. In Proceedings of the National Academy of Sciences of USA: US National Academy of Science.

Mairesse, F.; Walker, M. A.; Mehl, M. R.; and Moore, R. K. 2007. Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text. *Journal of Artificial Intelligence Research* 30(1): 457–500.

Oberlander, J., and Nowson, S. 2006. Whose Thumb Is It Anyway? Classifying Author Personality from Weblog Text. In *Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics,* 627-634. Stroudsburg, PA, USA: ACM Press.

Pennebaker, J. W.; and King, L. A. 1999. Linguistic Styles: Language Use As an Individual Difference. *Journal of Personality and Social Psychology* 77: 1296–1312.