Inferring Neuroticism of Twitter Users by Utilizing their Following Interests

Joran Cornelisse
University of Amsterdam
joran.cornelisse@gmail.com

Raoul Grasman
University of Amsterdam
R.P.P.P.Grasman@uva.nl

Abstract

Twitter is a medium where, when used adequately, users' interests can be derived from what he follows. This characteristic can make it attractive for a source of personality derivation. We set out to test the hypothesis that, analogous to the Lexical hypothesis, which posits that word use should reveal personality, following behavior on social media should reveal personality aspects. We used a two-step approach, wherein the first stage, we selected accounts for whom it was possible to infer personality profiles to some extent using available literature on personality and interests. On these accounts, we trained a regression model and segmented the derived features using hierarchical cluster analysis. In the second stage, we obtained a small sample of users' personalities via a questionnaire and tested whether the model from stage 1 correlated with the users from step 2. The the explained variance for the neurotic and neutral neuroticism groups indicated significant results ($R^2 = .131$, $p = .0205$; $R^2 = .22$, $p = .0044$). Confirming the hypothesis that following behavior should be correlated with one's interests and that interests are correlated with the neuroticism personality dimension.

1 Introduction

With the advent of massive data streams of personal online behavior, there has been a surge in interest in relating this behavior to classical constructs in psychology. In particular, the question of whether online behavior can be related to personality profiles obtained from psychometric personality instruments, often the Big Five (Digman, 1990), has led to a field of research that has been coined *Personality computing* (Vinciarelli and Mohammadi, 2014). Studies in this field have found significant correlations between personality scores and lexical elements in written text, such as essays (Mairesse et al., 2006), blogs (Oberlander and Nowson, 2006; Yarkoni, 2010; Minamikawa and Yokoyama, 2011; Iacobelli et al., 2011), self-presentations (Batrinca et al., 2011), emails (Estival et al., 2007), texting (Holtgraves, 2011), and even the choice of email addresses (Back et al., 2008).

The idea that an author's personality is reflected in written or spoken text at all, is referred to as the *Lexical hypothesis* (Pennebaker and King, 1999; Mairesse et al., 2006), and has been confirmed in many studies. As personality involves and influences the interaction with one's environment, it should be expected to influence other facets of linguistic and non-linguistic expression. The rise of social media over the last decade resulted in this interaction being partly transposed to digital platforms. In fact, the current omnipresence of social media have compounded to the availability of records of individual expression and extends well beyond just the textual, and includes Facebook "likes" (Kosinski et al., 2013), sharing behavior (Gou et al., 2014), website visiting behavior, YouTube videos (Biel and Gatica-Perez, 2012; Farnadi et al., 2013), LinkedIn profiles (Faliagka et al., 2012), Instagram pictures (Ferwerda and Tkalcic, 2018), and tweets on Twitter (Golbeck et al., 2011; Chen et al., 2014; Li et al., 2014).

At face value, Facebook, Twitter and other social media platforms seem quite similar. Indeed lexical hypothesis-driven approaches have treated Facebook status updates as more or less equivalent to Twitter tweets. However, there is an essential distinction between the purposes of the two media. Facebook is

Proceedings of the Third Workshop on Computational Modeling of PEople's Opinions, PersonaLity, and Emotions in Social media, pages 1–10
Barcelona, Spain (Online), December 13, 2020.

generally used as a place to keep in touch with friends and acquaintances. Twitter, on the other hand, is a medium that is used for sharing and gaining information related to someone's interests. As a result, to use Twitter adequately and benefit from its full set of features, an active Twitter user is forced to follow his different interests on Twitter. This characteristic of Twitter can make it especially attractive as a source for personality derivation, as preferences of individuals can, to a significant extent, be explained by underlying personality traits (Ozer and Benet-Martinez, 2006).

It has often been contended that the social network(s) that people find themselves embedded in is reflective of many personal characteristics (Kosinski et al., 2013). Social media networks —reflected in the connections users can make to other users, e.g., by 'friending' on Facebook or 'following' on Twitter[1] —have been related to various personal characteristics such as gender, age, ethnicity, sexual orientation, religious and political views (Kosinski et al., 2013). Although it has been observed that various generic network metrics that do not take into account the node characteristics of the connected nodes, including connectedness, node centrality measures, as well as the total number of friends / followers and the number of users followed, are correlated with personality (Quercia et al., 2011; Li et al., 2014), to our knowledge, no research on computational personality has investigated whether the *specific set of accounts* followed by a Twitter user is associated with personality.

In this study we aim to investigate if, and how profiles of accounts followed by a user are related to personality. We do this by deriving a predictive model from Twitter following graph data related to a set of prior chosen interests, and validating this predictive model on a sample of Twitter users from whom we obtained personality profiles using a standardized test. We focused particularly on neuroticism, as it has shown to have a reliable correlation to social media extracted features (Blackwell et al., 2017; Abbasi and Drouin, 2019). The specific interest were chosen on the basis of available literature that relates the interest to personality, and on the requirement that this interest can be relatively easily inferred from Twitter accounts using heuristic methods. We hypothesize that the following of (clusters of) nodes in the Twitter graph (that we coin 'influencer accounts' below) are significantly correlated with personality scores on a standardized personality test.

2 Methods

To build a predictive model that uses the links between nodes, a large sample of users is required as predictor variables within all the possible connections in the following graph data. While it is easy to obtain a large sample of Twitter users and the information regarding their following behavior, it is not easy to obtain their personality profiles by having these users fill out a psychometric test. We therefore developed a two stage approach: In the first stage we select a large sample of Twitter users for whom it was possible to infer personality profiles to some extend by heuristic means from their expressed interests and professions (we detail this below). On this data set, we trained a regression model. In the second stage, we obtained a small sample of Twitter users who were willing to fill out a personality questionnaire (the NEO-FFI) to provide us with a validation sample: We tested whether the predictions by model from stage 1 for the people in our stage 2 sample significantly correlated with personality profiles obtained from the questionnaire they filled out.

2.1 Using specific interests as a personality gauge

In the first stage we used heuristic methods to build a regression model. In particular, we searched for Twitter users who expressed specific interests on their Twitter accounts that have been linked to personality characteristics in previous research. For instance, we searched for Twitter users who expressed interest in *yoga* as people participating in yoga tend to score *lower* on neuroticism than the general population (Venkatesh et al., 1994). Similarly, social interests have been correlated with agreeableness (Costa and McCrae, 1985); interest in self-enhancing or affiliating humor have been found to negatively correlate with neuroticism (Greengross et al., 2012); and entrepreneurial inclinations correlate negatively with neuroticism (Zhao and Seibert, 2006). Twitter users express these specific interests not only in their

[1] Note that the difference between 'friending' and 'following' is another difference between Facebook and Twitter, in that the former is a symmetric relation, while the latter is unidirectional.

tweets, but also in the Twitter accounts that they follow. We will coin Twitter accounts followed by a user *influencer accounts*, or *influencers* for short, in line with the use of these terms in the field of marketing. Furthermore, such interests can be expressed in their profile in their occupational denomination, which is either announced in their profile or can possibly be inferred from the influencer accounts that they follow. Research showed that sort-alike interests, such as food consumption, brand-preference, and political preference could all be deducted from following behavior (Abbar et al., 2015; Chu et al., 2016; Golbeck and Hansen, 2014). Correspondingly, neuroticism is negatively associated with the *Enterprising type* of Holland's RIASEC model of occupational interest (Armstrong and Anthoney, 2009; Holland, 1997; Costa Jr and McCrae, 1992; Zhao and Seibert, 2006). This way, an Enterprising-occupational preference can be used as a parameter for the emotional stable group.

Using a set of specific interests and occupations, we collected three large groups of together 6107 Twitter users that we could classify as more likely *neurotic* (NEU+), more likely *emotionally stable* (NEU-), and more likely *neutral* (NEU±) based on the influencers that these accounts follow. In particular, users in the *neurotic group* (NEU+) were found by searching for users interested in *self help information for stress coping* (Saper and Forest, 1987; Kessler et al., 1997), or *artistic profession* (Marchant-Haycox and Wilson, 1992; Gelade, 1997; Nowakowska et al., 2005; Srivastava and Ketter, 2010); users in the *emotionally stable* group (NEU-) where found by searching users who were interested in *entrepreneurial topics* (Zhao and Seibert, 2006; Brandstätter, 2011) or in *self-enhancing humor* (Greengross et al., 2012; Kuiper and Leite, 2010; Mendiburo-Seguel et al., 2015), while users in the *neutral* group were found by searching for users who were interested in topics from both these groups (e.g., both interested in yoga and in entrepreneurial topics). For finding the included Twitter accounts we extracted the followers of accounts resembling these particular interests via the Twitter REST API. The validity of the collection was assessed by verifying a random sample of a small number accounts from each group at face value. Example searches are presented in Table 1. The logic behind the neutral group entails that a user who is interested in both interests for the neurotic and emotional stable group, would be classified as neutral. Similar according to the NEO-FFI. However, this does result for the groups not being mutual exclusive. As a result, users who were in more than one group, were excluded from the data set. In Table 2, some examples of the used influencer accounts are given.

The collection of all these *influencer* accounts constitutes our set of predictors: Each *influencer* account defines a dummy variable that is equal to 1 for users who follow this *influencer*, and 0 for users who do not follow this *influencer*. Because this results in a very large set of dummy variables (over 20,000), we clustered them into a limited set of 100 *influencer clusters* by means of (complete linkage) hierarchical clustering (McQuitty, 1955), and scored each user on these groups by counting the number of accounts in each *influencer clusters* a user followed. Hence, our data set for stage 1 consisted of 18,000 Twitter users, divided over three inferred neuroticism groups, for each of which we had as predictive features their count scores for the 100 different *influencer clusters*.

2.2 Feature selection

Using this first stage data set, we built a regression model in order to determine which of the 100 *influencer* groups count features were important for the prediction of neuroticism. We used multinomial least absolute shrinkage and selection operator (*Lasso*) regression (Tibshirani, 1996) to predict neuroticism group membership from these features. Lasso regression penalizes regression coefficients by adding their absolute values to the least squares criterion, weighted by a regularization strength parameter. This penalty promotes coefficients to be exactly equal to zero that would have been close to zero in normal regression. The non-zero coefficients of the resulting model defines the set of relevant features. This, in effect, automatically selects the features that are relevant for distinguishing between the neuroticism groups. How many coefficients end up at zero depends on the regularization strength parameter. We used 10-fold cross-validation to determine the optimal regularization strength (Friedman et al., 2010).

2.3 Model validation

In order to validate the *influencer* group count features derived from the first stage data set, and in order to test the hypothesis that personality is correlated with Twitter following behavior, we conducted ordinary

Group	Query
Neurotic group	SELECT * FROM SAMPLE WHERE user follows >= 3 Stress coping influencers OR >= 3 Artistic profession influencers
Emotional stable group	SELECT * FROM SAMPLE WHERE user follows >= 5 Enterprising influencers OR >= 3 self-enhancing humor influencers OR >= 5 Yoga influencers
Neutral group	SELECT * FROM SAMPLE WHERE user follows >= 3 Enterprising influencers AND 3 Stress coping influencers OR >= 3 self-enhancing humor influencers AND >= 3 Artistic profession influencers

Table 1: Overview of example queries used in order to create groups reflecting different dimensions of neuroticism.

Group	Influencers
Neurotic influencers	@WakeupPeopIe, @suicidalwrxck, @sosadtoday, @depressingmsgs,@AgainstSuicide, @depression
Emotional stable influencers	@ondernemer24, @nieuws_4_zzp, @de_ondernemer, @GreenBiz, @MKBNL, @FinancialTimes

Table 2: In this table, we can find some examples of influencers used to generate the three different groups. The neurotic influencers depict anxiety, self-help, and stress-coping accounts, which aim to help people with negativity in their lives.The emotional stable influencers found were mostly connected to entrepreneurship. The word *ondernemer* is dutch for entrepreneur. ZZP an organization connected to freelancing and MKB, stands for middle and small companies, supports entrepreneurial companies. Note that these are just some examples to give a general idea of what kind of accounts were categorized as influencers.

multiple linear regression of neuroticism scores onto the derived features in a new sample of Twitter user participants who we asked to fill out the NEO-FFI.

These participants were recruited through ads on Twitter using targeted campaigns to Dutch nationals with an expressed affinity for scientific research, universities (students or professors) and/or psychology. The reason to target these individuals was to maximize the response rate. The ads were running from November 2018 through January 2019, until $n = 130$ responses were collected. The ad invited Twitter users to participate in research on personality. Users who clicked on the ad were directed to an online survey form.

After a brief explanation of the purpose of the study—relating Twitter activity to personality—they were asked to fill out the Dutch version of the NEO-FFI (Hoekstra et al., 2007). We asked to fill in the personality test truthfully and emphasized that their results would be kept private. The NEO-FFI measures personality based on the Big Five personality dimensions (Costa Jr and McCrae, 1992). The Dutch NEO-FFI has been evaluated and normed(Hoekstra et al., 2007). The manual reports a reliability of 0.89 (Cronbach's alpha) for Neuroticism. In the survey, we also asked participants to share their Twitter account handle. The Twitter account handles of the participants were used to extract their following behavior (*influencer accounts*).

3 Results

3.1 Feature extraction: Clustering and Lasso regression

The database queries yielded a total of 6107 twitter users divided into 2301 NEU+, 1156 NEU±, and 2650 NEU- accounts. All subsequent data processing was conducted in R (R Core Team, 2019). These

6107 Twitter users followed a total number of 669104 unique *influencer* accounts. To reduce this number of potential predictive features, we first removed the *influencer* accounts from that had near zero variance (Kuhn and Johnson, 2013, those that were either followed by less than 1% or not followed by less than 1%). This reduced the set of 669104 *influencer* accounts to 4367 potential predictive features. We then reduced the number of predictive features further by means of agglomerative hierarchical clustering with complete linkage as cluster dissimilarity measure (Venables and Ripley, 2013) on the basis of similar follower patterns across the 6107 Twitter users. The output yielded 100 *influencer* group count clusters of varying sizes, ranging from just a few Twitter accounts to nearly 200. Subsequently, we used multinomial lasso regression (Tibshirani, 1996) to find the *influencer* group count clusters that are most predictive for group membership. Before fitting the multinomial lasso regression model, the *influencer* group count features were standardized to have zero means and unit standard deviations. The lasso-penalty strength was fine tuned with 10-fold cross-validation to optimize the predictive classification accuracy. The cross-validated prediction accuracy of this model was more than 90%, indicating that the NEU+, NEU$\pm$, and NEU- groups could be well seperated.[2] Of the 100 *influencer* group count features that entered the multinomial lasso regression, 76 had non-zero coefficients on at least one of the multinomial predictive functions. Only 46 of these we considered of predictive importance: *influencer* features with a regression coefficient of at least 10% of the largest coefficient in absolute value.

3.2 Validation: Multiple linear regression

130 Twitter users filled out the personality questionnaire. Participants were excluded if they did not complete the survey, or had protected Twitter accounts.[3] In total, 98 participants were eligible.

We admitted the *influencer* group count features selected by the multinomial lasso regression to a multiple linear regression analysis in which the NEO derived neuroticism score was used as the dependent variable. We did this separately for each of the three prediction functions for the NEU+, NEU$\pm$, and NEU- groups in the previous step. While for the NEU$\pm$ and NEU- group equations the explained variance was significant at the .05 level, $R^2 = .248$, adjusted $R^2 = .131$, $F(18, 84) = 2.125$, $p = .0205$, and $R^2 = .38$, adj $R^2 = .22$, $F(20, 77) = 2.328$, $p = .0044$, respectively, the explained variance for the NEU+ equation was only marginally significant, $R^2 = .261$, adjusted $R^2 = .104$, $F(17, 80) = 1.66$, $p = .068$. Note that these tests evaluate different predictive equations because they include mostly non-overlapping sets of *influencer* group count features. The influencer Twitter handles (i.e., their '@' names) that are associated with significant coefficients are displayed in Table 3. The significant coefficients for the NEU$\pm$ and NEU- regressions were mostly negative, indicating that with increasing counts on the corresponding features (i.e., with increasing number of influencer accounts followed within the feature cluster) the neuroticism score decreased. Hence, these regression models are mostly indicative of emotional stability. The only regression coefficient that was significant in the NEU+ equation was positive.

We did various diagnostic checks on these regressions (Fox, 2015; Fox and Weisberg, 2018). Because influence measures indicated some of the cases to be influential, we also ran the regression analyses with these influential cases removed. For all three equations, this only had the effect of making the observed effects larger, and the p-values smaller—indeed rendering the previously marginal significance of the NEU+ regression model highly significant ($R^2 = 0.346$, adjusted $R^2 = 0.207$, $F(16, 75) = 2.48, p = .004$), while the NEU$\pm$ and NEU- models maintained high levels of significance ($R^2 = 0.3$, adjusted $R^2 = 0.184$, $F(13, 79) = 2.6, p = 0.0047$, and $R^2 = 0.457$, adjusted $R^2 = 0.306$, $F(20, 72) = 3.025$, $p = 0.0003$, respectively).

Because it has been previously reported that the number of Facebook 'friends' is associated with personality, we verified that adding a total number of following accounts feature did not change any of the models, $\Delta R^2 = 0.01$, $F(1, 79) = 1.134, p = .29$ for the NEU+ model, $\Delta R^2 = 0.017$, $F(1, 83) =$

[2] Note that this high prediction accuracy indicates that the lasso model has learned the differences between the queries that were used to create our NEU+, NEU$\pm$ and NEU- groups. By extension, it presumably learns to distinguish between Twitter users that score high, medium, or low on the neuroticism personality axis.

[3] Twitter allows users to make their account private. As a result no information can be extracted for that account by third parties.

$1.957, p = .165$ for the NEU± model, $\Delta R^2 = 0.007$, $F(1, 76) = 0.812, p = .371$ for the NEU- model. Also, a separate regression of the neuroticism score on this feature did not yield a significant explained variance, $F(1, 96) = 0.03061$, $p = 0.862$. In addition, because the *influencer* accounts that are associated with significant coefficients were mostly related to entrepreneurial interest and self-employment, we checked whether the significant predictive power of the models simply resulted from entrepreneurial interest by adding the count totals of the number of accounts participants followed that were also followed by Twitter users in the NEU- group. This also did not lead to a significant change of models, $\Delta R^2 = .002, F(1, 79) = 0.254, p = .642$ for the NEU+ model, $\Delta R^2 = .01, F(1, 76) = 1.065, p = .305$ for the NEU± model, and $\Delta R^2 = .002, F(1, 76) = 0.217, p = .642$ for the NEU- model; nor did a separate simple regression of the neuroticism scores on this feature yield a significant effect ($F(1, 96) = 0.031, p = .861$). Hence, taken together these results indicate that it is not merely the number of account followed, nor merely an entrepreneurial interest that is able to explain the variance in the neuroticism scores, but the specific pattern of *influencer* groups that are followed. A series of similar regressions in which the other personality scores (openness, conscientiousness, extraversion, and agreeableness) were used as dependent variables did not result in significant omnibus tests, which shows that the results are particular to neuroticism for which we constructed our *influencer* group count features.

Group	**Relation**	**(example) Accounts**	**Cluster Size**
Neu+	Positive	@9GAG, @9GAGTweets, @Rosssen, @OhDailyJustin	4
Neu±	Negative	@JOR_ID, @BoogerdLive, @TeamSunweb, @RobScheepers, @lars_boom	14
Neu±	Negative	@MINOCW, @Leraar24, @SanderDekker, @LerarenMetLef, @JelleJolles	25
Neu±	Positive	@bibliotheek, @NPO2extra, @NOGvacatures, @taalmissers, @vangoghmuseum	21
Neu-	Negative	@StephenRCovey, @MarcStijfs, @woutsmelt, @TonyRobbins, @Upgres	17
Neu-	Negative	@Politie_Zeeland, @zeelandzakelijk, @JoAnnesdeBat, @hvzeeland, @PetradeBoevere	22
Neu-	Negative	@OP_Nederland, @KVK_NL, @AccWeek, @Taxence, @BDONederland	32
Neu-	Positive	@Sebastiaan_IMG, @Brandpunt_plus, @nieuwelente, @StudioLizix, @TL_070	19

Table 3: Influencer account handles ('@' names) associated with the *influencer* group count features with significant multiple regression coefficients in the validation data set. *Top row*: Features that predict NEU+ membership. *Middle rows*: Features that predict NEU± membership. *Bottom rows*: Features that predict NEU- membership. Only features with significant multiple regression coefficients are displayed. For each cluster, we indicate whether it had a negative (i.e., indicate lower neuroticism score) or positive (i.e., indicate higher neuroticism score) coefficient. We show five arbitrary chosen accounts per cluster.

4 Discussion

We set out to test the hypothesis that, analogous to the Lexical hypothesis which posits that word use should reveal personality, following behavior on social media should reveal aspects of personality. This hypothesis was motivated by the fact that following behavior should be highly correlated with one's interests and occupation, and that interests and occupation are correlated with personality dimensions. In particular, we aimed to test this hypothesis with respect to the degree of neuroticism of an individual. To do so, we introduced a novel technique for constructing predictive independent variables: On the basis of these well established correlations between personality on the one hand, and interests and occupation on the other, we were able to gather a sufficiently large database of Twitter users and their following behavior to extract clusters of *influencer* accounts that discriminate between Twitter users that have a lower or higher propensity to score high on the neuroticism dimension of the Big Five. Using these predictors we were then able to confirm the hypothesis by showing that these predictors were significantly related to neuroticism scores in a relatively small sample of 98 participants recruited to fill out the NEO-FFI.

On the basis of these results we can conclude that it is possible to derive estimates of neuroticism from following behavior. A significant consequence of these findings is that, in contrast to personality profiles extracted on the basis of lexical hypothesis which requires active engagement of social media platform users, our results are purely based on information that is passively conveyed by Twitter users by their following behavior. Hence, a neuroticism score can be obtained even from social media users who do not actively engage in information sharing on these platforms. One might object that the explained variance is rather low and consequently, very imprecise. This is certainly true for assessing specific individuals reliably—that is, our results do not really support the notion that it is possible, e.g., for employers to screen a specific applicant for a job. However, it is sufficient for targeting population segments; e.g., it allows for talent recruiters to target subgroups of the population that are more likely to have certain personality profiles, or for marketeers to target population segments, e.g., in attempts to sway an election.

4.1 Time and Location

A relevant note for this study is that its results are subject to time and location. This is because individual interests can be prevalent during a specific period or at a particular location. For instance, yoga or meditation are rising interests, which are now much more popular than several decades ago. These interests were used as a *neurotic features* (i.e., influencers). However, regarding their rising popularity, they might not be *significant* features in the future, as then *all kinds* of people would be interested in performing yoga or meditation. In respect to location, the interests used are subjective to Dutch nationals, and more or less the Western world. The *neurotic interests* might be different in other parts of the world. Not to mention that personality determination in itself works differently in other parts of the world, for instance, Asia (Markus and Kitayama, 1998). Therefore, the results from this study are dependent of time and location.

4.2 Caveats

It is not clear what population our sample represents, as it is a self-selected group. Also, Twitter suggests a user which accounts to follow based on the accounts the user already follows. This causes accounts to be clustered due to Twitter's recommender system. Furthermore, concerning the choice of interests and occupations for deriving our prediction variables, we limited our database to Twitter users consisting of entrepreneurial or artistic users, or expressed interest in yoga, stress coping, or certain types of humor. Needless to say, this is a minimal set of expressed interests and/or occupations. Although the validation set was not selected on the basis of these criteria and the results should generalize to a broader population of Twitter users, we anticipate that a broader range of topics of interest and/or occupations that have been shown to be correlated with personality aspects will potentially improve and extend personality profiling concerning Twitter following behavior.

4.3 Privacy

Lastly, an interesting point worth mentioning is that more and more aspects of the digital footprint of individuals present pile up to attributing variance in the process of personality computation. This way, it becomes increasingly likely that a complete combination of all an individual's social media activity could be quite an accurate determinator for one's personality. Because of this, it becomes increasingly important to think about the privacy legislation of different media. Twitter provides the opportunity to protect your account, and this way, your shared information cannot be extracted using the Twitter API. Every media should offer the option for a user to make his or her data private. Hence, if it is not in human's interest to have every individual's personality available on the web or in the hands of a few corporations, it will become imperative that legislation forces this option to be mandatory.

References

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

Irum Abbasi and Michelle Drouin. 2019. Neuroticism and facebook addiction: How social media can affect mood? *The American Journal of Family Therapy*, 47(4):199–215.

Patrick Ian Armstrong and Sarah Fetter Anthoney. 2009. Personality facets and riasec interests: An integrated model. *Journal of Vocational Behavior*, 75(3):346–359.

Mitja D Back, Stefan C Schmukle, and Boris Egloff. 2008. How extraverted is honey. bunny77@ hotmail. de? inferring personality from e-mail addresses. *Journal of Research in Personality*, 42(4):1116–1122.

Ligia Maria Batrinca, Nadia Mana, Bruno Lepri, Fabio Pianesi, and Nicu Sebe. 2011. Please, tell me about yourself: automatic personality assessment using short self-presentations. In *Proceedings of the 13th international conference on multimodal interfaces*, pages 255–262. ACM.

Joan-Isaac Biel and Daniel Gatica-Perez. 2012. The youtube lens: Crowdsourced personality impressions and audiovisual analysis of vlogs. *IEEE Transactions on Multimedia*, 15(1):41–55.

David Blackwell, Carrie Leaman, Rose Tramposch, Ciera Osborne, and Miriam Liss. 2017. Extraversion, neuroticism, attachment style and fear of missing out as predictors of social media use and addiction. *Personality and Individual Differences*, 116:69–72.

Hermann Brandstätter. 2011. Personality aspects of entrepreneurship: A look at five meta-analyses. *Personality and individual differences*, 51(3):222–230.

Jilin Chen, Gary Hsieh, Jalal U Mahmud, and Jeffrey Nichols. 2014. Understanding individuals' personal values from social media word use. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 405–414. ACM.

Shu-Chuan Chu, Hsuan-Ting Chen, and Yongjun Sung. 2016. Following brands on twitter: An extension of theory of planned behavior. *International Journal of Advertising*, 35(3):421–437.

Paul T Costa and Robert R McCrae. 1985. The neo personality inventory.

Paul T Costa Jr and Robert R McCrae. 1992. Four ways five factors are basic. *Personality and individual differences*, 13(6):653–665.

John M Digman. 1990. Personality structure: Emergence of the five-factor model. *Annual review of psychology*, 41(1):417–440.

Dominique Estival, Tanja Gaustad, Son Bao Pham, Will Radford, and Ben Hutchinson. 2007. Author profiling for english emails. In *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics*, pages 263–272.

Evanthia Faliagka, Athanasios Tsakalidis, and Giannis Tzimas. 2012. An integrated e-recruitment system for automated personality mining and applicant ranking. *Internet research*, 22(5):551–568.

Golnoosh Farnadi, Susana Zoghbi, Marie-Francine Moens, and Martine De Cock. 2013. Recognising personality traits using facebook status updates. In *Seventh International AAAI Conference on Weblogs and Social Media*.

Bruce Ferwerda and Marko Tkalcic. 2018. Predicting users' personality from instagram pictures: Using visual and/or content features? In *Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization*, pages 157–161. ACM.

John Fox and Sanford Weisberg. 2018. *An R companion to applied regression*. Sage Publications.

John Fox. 2015. *Applied regression analysis and generalized linear models*. Sage Publications.

Jerome Friedman, Trevor Hastie, and Rob Tibshirani. 2010. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1.

Garry A Gelade. 1997. Creativity in conflict: The personality of the commercial creative. *The Journal of genetic psychology*, 158(1):67–78.

Jennifer Golbeck and Derek Hansen. 2014. A method for computing political preference among twitter followers. *Social Networks*, 36:177–184.

Jennifer Golbeck, Cristina Robles, Michon Edmondson, and Karen Turner. 2011. Predicting personality from twitter. In *2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing*, pages 149–156. IEEE.

Liang Gou, Michelle X Zhou, and Huahai Yang. 2014. Knowme and shareme: understanding automatically discovered personality traits from social media and user sharing preferences. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 955–964. ACM.

Gil Greengross, Rod A Martin, and Geoffrey Miller. 2012. Personality traits, intelligence, humor styles, and humor production ability of professional stand-up comedians compared to college students. *Psychology of Aesthetics, Creativity, and the Arts*, 6(1):74.

Harold Hoekstra, J. Ormel, and F Fruyt. 2007. Handleiding neo-pi-r en neo-ffi persoonlijkheidsvragenlijsten.

John L Holland. 1997. *Making vocational choices: A theory of vocational personalities and work environments*. Psychological Assessment Resources.

Thomas Holtgraves. 2011. Text messaging, personality, and the social context. *Journal of research in personality*, 45(1):92–99.

Francisco Iacobelli, Alastair J Gill, Scott Nowson, and Jon Oberlander. 2011. Large scale personality classification of bloggers. In *international conference on affective computing and intelligent interaction*, pages 568–577. Springer.

RC Kessler, KD Mickelson, and S Zhao. 1997. Patterns and correlates of self-help group membership. *American Psychologist*, 44(27):27–46.

Michal Kosinski, David Stillwell, and Thore Graepel. 2013. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, 110(15):5802–5805.

Max Kuhn and Kjell Johnson. 2013. *Applied predictive modeling*, volume 26. Springer.

Nicholas A Kuiper and Catherine Leite. 2010. Personality impressions associated with four distinct humor styles. *Scandinavian Journal of Psychology*, 51(2):115–122.

Lin Li, Ang Li, Bibo Hao, Zengda Guan, and Tingshao Zhu. 2014. Predicting active users' personality based on micro-blogging behaviors. *PloS one*, 9(1):e84997.

Franc Mairesse, Marilyn Walker, et al. 2006. Words mark the nerds: Computational models of personality recognition through language. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 28.

Susan E Marchant-Haycox and Glenn D Wilson. 1992. Personality and stress in performing artists. *Personality and individual differences*, 13(10):1061–1068.

Hazel Rose Markus and Shinobu Kitayama. 1998. The cultural psychology of personality. *Journal of cross-cultural psychology*, 29(1):63–87.

Louis L McQuitty. 1955. A method of pattern analysis for isolating typological and dimensional constructs. Technical report, ILLINOIS UNIV AT URBANA TRAINING RESEARCH LAB.

Andrés Mendiburo-Seguel, Darío Páez, and Francisco Martínez-Sánchez. 2015. Humor styles and personality: A meta-analysis of the relation between humor styles and the big five personality traits. *Scandinavian journal of psychology*, 56(3):335–340.

Atsunori Minamikawa and Hiroyuki Yokoyama. 2011. Personality estimation based on weblog text classification. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, pages 89–97. Springer.

Cecylia Nowakowska, Connie M Strong, Claudia M Santosa, PO W Wang, and Terence A Ketter. 2005. Temperamental commonalities and differences in euthymic mood disorder patients, creative controls, and healthy controls. *Journal of Affective Disorders*, 85(1-2):207–215.

Jon Oberlander and Scott Nowson. 2006. Whose thumb is it anyway? classifying author personality from weblog text. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 627–634.

Daniel J Ozer and Veronica Benet-Martinez. 2006. Personality and the prediction of consequential outcomes. *Annu. Rev. Psychol.*, 57:401–421.

James W Pennebaker and Laura A King. 1999. Linguistic styles: Language use as an individual difference. *Journal of personality and social psychology*, 77(6):1296.

Daniele Quercia, Michal Kosinski, David Stillwell, and Jon Crowcroft. 2011. Our twitter profiles, our selves: Predicting personality with twitter. In *2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing*, pages 180–185. IEEE.

R Core Team, 2019. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Zalman Saper and James Forest. 1987. Personality variables and interest in self-help books. *Psychological Reports*, 60(2):563–566.

Shefali Srivastava and Terence A Ketter. 2010. The link between bipolar disorders and creativity: evidence from personality and temperament studies. *Current psychiatry reports*, 12(6):522–530.

Robert Tibshirani. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

William N Venables and Brian D Ripley. 2013. *Modern applied statistics with S-PLUS*. Springer Science & Business Media.

S Venkatesh, Madan Pal, BS Negi, VK Varma, et al. 1994. A comparative study of yoga practitioners and controls on certain psychological variables. *Indian Journal of Clinical Psychology*.

Alessandro Vinciarelli and Gelareh Mohammadi. 2014. A survey of personality computing. *IEEE Transactions on Affective Computing*, 5(3):273–291.

Tal Yarkoni. 2010. Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers. *Journal of research in personality*, 44(3):363–373.

Hao Zhao and Scott E Seibert. 2006. The big five personality dimensions and entrepreneurial status: A meta-analytical review. *Journal of applied psychology*, 91(2):259.