Studying the Dark Triad of Personality through Twitter Behavior

Daniel Preoţiuc-Pietro, Jordan Carpenter, Salvatore Giorgi, Lyle Ungar
Positive Psychology Center
Computer & Information Science
University of Pennsylvania
danielpr@sas.upenn.edu

ABSTRACT

Research into the darker traits of human nature is growing in interest especially in the context of increased social media usage. This allows users to express themselves to a wider online audience. We study the extent to which the standard model of dark personality – the dark triad – consisting of narcissism, psychopathy and Machiavellianism, is related to observable Twitter behavior such as platform usage, posted text and profile image choice. Our results show that we can map various behaviors to psychological theory and study new aspects related to social media usage. Finally, we build a machine learning algorithm that predicts the dark triad of personality in out-of-sample users with reliable accuracy.

1. INTRODUCTION

Online spaces have increasingly become a medium for self expression and social communication. Social media websites allow users to build an online identity, post content (text updates, links or images) and interact with others. While social media platforms become ubiquitous, there is a rising concern that these offer a medium for expressing darker traits of human personality such as self-promotion, vanity, anti-social behavior, alteration of the truth or self-interest. These malevolent human traits have been operationalized in psychology research in the form of the dark triad of personality [49], which consists of three traits:

a) **Narcissism** – grandiose and inflated self-views, sense of entitlement and a craving for admiration;

b) **Machiavellianism** – self-interest, cynicism and a tendency to manipulate and exploit others;

c) **Psychopathy** – enduring antisocial behavior, impulsivity.

To date, the study of the dark triad of personality is lacking a systematic, data-driven exploration. The overwhelming majority of existing studies in the online expression of the dark triad traits were conducted through surveys about social media behavior. We use an empirical approach to directly quantify social media behaviors and relate them to users who also took a dark triad personality questionnaire. Moreover, one of the peculiarities of the dark triad traits is represented by alteration of the truth, which would bias self-reported behaviors. We thus consider a data-driven analysis of content a viable alternative of studying the dark triad.

Many characteristics of the dark triad lend themselves naturally to expression via social media. For example, narcissism is related to enhanced self-views, self-promotion and craving for admiration, which can be reflected by selfie posting, constantly seeking attention and projecting a positive view of the self. Anti-social and impulsive behavior, characteristic of psychopathy, can be expressed by an increased use of swear words. Machiavellianism could be related to attempts at ingratiating or manipulative language.

We adopt a multi-modal approach to quantifying user behavior on Twitter that captures general platform usage, such as posting frequency or social connections, publicly posted text and high-level profile image features. Recently, machine learning models were successfully deployed to study and predict a number of demographic [61] or psychological traits [56]. The main motivation for these studies was to predict different traits when they are lacking from a user's profile, which is useful for applications such as targeted advertising. Others have focused on interpreting the model features with the goal of gaining insight into group differences [65].

The scope of this study is two-fold. We first aim to directly explore the relationships between online behaviors and the three components of the dark triad, specifically aiming to highlight similarities and differences across the three components. To this end, we use a large number of features and explore correlations using univariate feature analysis, while controlling for possible effects of basic demographics, such as age and gender. We employ models from natural language processing and image analysis in order to extract features which are both predictive and interpretable, such as word topics or facial features. Further, we use existing state-of-the-art models that predict the expression of emotion and relate them to previous findings. Secondly, we aim to build a predictive model for the dark triad traits that uses only public Twitter information. Using all features, we achieve a robust predictive performance of Pearson $R \sim .25$ for all traits. Predicting the dark triad in social media could have a multitude of socially valuable uses, such as identifying potentially abusive trolls by monitoring users' psychopathy or assisting in recognizing duplicitous statements by measuring users' Machiavellianism.

CIKM'16 , October 24 - 28, 2016, Indianapolis, IN, USA

ISBN 978-1-4503-4073-1/16/10...$15.00

DOI: http://dx.doi.org/10.1145/2983323.2983822

2. BACKGROUND

The darker traits of human personality have been a popular and controversial research topic over the past decade [20]. The standard model is represented by the dark triad, a model which posits three distinct traits: narcissism, psychopathy and Machiavellianism [49]. All three traits share a core set of characteristics such as hostility or lack of empathy [31] and represent a short-term, agentic, exploitative social strategy [30]. The core of the dark triad can account for up to 50% of the variance and can be studied as a single trait [29].

A large thread of recent work in psychology has focused on the expression of the dark triad in online spaces, usually in social networking platforms. Many researchers claim that such platforms, centered around social interaction and communication, are rife for the expressions of these traits. Most studies have focused on studying narcissism as especially central to online behaviors.

Foremost, the social media platform has been shown to play a role in the types of behaviours expressed online by people high in narcissism [48]. Status update frequency on Facebook was found to correlate with narcissism [24], while Facebook users have been found to be more narcissistic in general [63]. Posting of self-promotional content was related with certain facets of narcissistic expression [39]. Anti-social behaviors were found to be mixed in relationship to narcissism [6]. When asked about the topics of their posts, narcissists mentioned their accomplishments, diet and exercise routines and self-reported a greater number of likes and comments to their posts [38]. Narcissism has been linked by lay theories to increased usage of first person pronouns [14], however an extensive study on various data showed a null relationship [5]. Finally, multiple studies have looked at images or selfies and their relationship with the dark triad, finding that users high in narcissism post more selfies [76], photos that are more attractive [14, 47], edited [19], with this behavior more prevalent in males [67].

While these approaches are suitable for studying specific hypotheses regarding social media usage, they are limited by the number of behaviors they can study. Additionally, these are measured using a survey about behavioral tendencies (e.g., asking participants to report their recalled self-promotion activities), rather than observing real behaviors. Participants may deceive the interviewer by answering untruthfully to the questions, a characteristic of the dark triad [46]. Notably, [21] studied all three dark triad traits on Facebook. However, the text analysis was limited being reduced to a single feature and lacked interpretability, however finding significant correlations with psychopathy and narcissism. Similarly, [70] used LIWC and profile features, but showed limited predictive accuracy and [23] presented a Twitter study of narcissism, using LIWC categories to uncover relationships for narcissists including high negative emotions, anger and use of swear words and fewer mentions of social processes and positive emotions. The later study is limited in several ways, including focusing only on narcissism, measuring it using a single item, no demographic controls and splitting the data set into classes rather than performing regression. Other computational data-driven studies looked at specific types of anti-social behavior [8], but did not link these to the psychological state of their authors.

Concurrently, there is a surge in interest in using data-driven methods to predict user traits or demographics. This work is motivated either by commercial and research applications, including targeted marketing or improving downstream NLP tasks [74, 27] or by the study of socio-linguistic or psychological hypotheses [57, 65].

Machiavellianism
I tend to manipulate others to get my way.
I have used deceit or lied to get my way.
I tend to exploit others towards my own end.
I have used flattery to get my way.

Narcissism
I tend to want others to admire me.
I tend to want others to pay attention to me.
I tend to seek prestige or status.
I tend to expect special favors from others.

Psychopathy
I tend to lack remorse.
I tend to be callous or insensitive.
I tend to be unconcerned with the morality of my actions.
I tend to be cynical.

Table 1: Dirty Dozen questionnaire. Each item can be answered using a five point scale (from 'Strongly Disagree' to 'Strongly Agree').

Arguably the largest body of work explored linguistic differences in posts in order to predict user attributes, drawing upon socio-lingistic theory. Studies focused on a broad set of demographic and psychological features: age [61, 45], gender [4, 64], political orientation [51], location [9, 15], popularity [35], income [18, 59], social status [34, 60], occupation [57], mental illnesses [10, 12, 58] or personality [65]. Models are usually built on large social media data sets, most commonly using self-reports. Other recent studies have analysed perceived user traits [71, 73], which may lead to certain biases [17].

Others have looked at predicting user traits from other behaviors such as likes on Facebook [33], images or social networks. Social network user profiling is based on homophily, wherein similar users are more likely to connect. This was exploited for traits including location [62], political orientation [72] or age [55]. Using images, several studies have looked at Big Five personality prediction [1, 7, 25, 36]. A few studies have also combined different modalities to boost performance or supplement one source of information when the other is not existing (e.g., not enough posts) [72].

3. DATA SET

We build a data set of Twitter users who took a twelve item questionnaire of the dark triad named the 'Dirty Dozen' [30]. This uses four questions to assess each of the three traits. The trait score is the arithmetic mean of its four questions scored with values on a 1–5 scale. The questions are presented in Table 1. Additionally, we use a combined dark triad score as the arithmetic mean of all twelve questions, similarly to previous research [28].

The users were part of a larger crowdsourcing experiment – not presented here – conducted through Amazon Mechanical Turk. This required participants to fill in the Dirty Dozen scale amongst others as a qualification for performing additional tasks. Additionally, users were asked to *voluntarily* provide their Twitter handle for research purposes. Out of a total of 2093 participants, 863 entered a valid and public

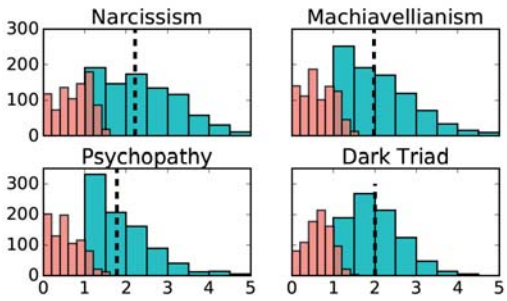


Figure 1: Distribution of the dark triad trait scores in our data. The green bars show raw values, the orange bars show log values.

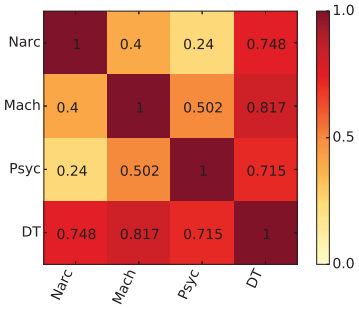


Figure 2: Intercorrelations between the log transformed traits.

Twitter handle. We found no significant differences in the distribution of the dark triad traits between participants who chose to enter the handle and those who did not. Participation in the experiment was restricted to users from the United States.

The distribution of the three traits in our data set is presented in Figure 1. We notice that the distribution is not normal and skewed towards lower values. This is expected, as the questions reference fairly extreme and socially undesirably behavior, which results in a negative skew. Based on this, we will use as the outcome of interest the natural logarithm of the dark triad traits scores for the rest of the study. The inter-correlations between the three traits and the overall score are presented in Figure 2.

As part of the required qualifications study, we also collected standard demographics of the participants, including age and gender. Both age and gender are significantly correlated with all three traits (Narcissism – $R = .07$ with Male, $R = -.162$ with age; Machiavellianism – $R = .125$ with Male, $R = -.138$ with age; Psychopathy – $R = .21$ with Male, $R = -.093$ with age; dark triad – $R = .17$ with Male, $R = -.171$ with age). Hence, for the rest of the experiments we will control for their effects using partial correlation.

4. FEATURES

We hypothesise that the dark triad is manifested online through a wide range of behaviors. We automatically extract features that capture multiple aspects of online behavior: text use, profile picture posting and general profile information.

4.1 Text Analysis

We collected all public posts of the Twitter users up to the most recent 3,200 posts in their history using the Twitter Search API, which results in a data set of 538,712 posts from 710 distinct users. In order to have certain confidence in our text analysis, we limit ourselves to users who only posted at least 500 tokens in their history. This restricts our data for text analysis to 491 users who produced a total of 536,579 tweets, which we tokenize with a Twitter-specific tokenizer. For each of these users, we extract the following text-derived features:

Unigrams.

We use the bag-of-words representation to reduce each user's posting history to a normalised frequency distribution over a vocabulary. The vocabulary consists of all words used by at least 10% of users (6,491 words).

LIWC.

Traditional psychological studies use a dictionary-based approach to representing text. The most popular method is based on Linguistic Inquiry and Word Count (LIWC) [52], and automatically counts word frequencies for 64 different categories manually constructed based on psychological theory. These include different parts-of-speech, topical categories and emotions. Each user is thereby represented as a frequency distribution over these categories.

Word Clusters.

Using all unigrams as features is likely to cause overfitting, especially on our data set, where there is an order of magnitude less observations than the vocabulary size. In order to reduce the feature space and also provide feature interpretability, we use word clusters. These clusters of words can be thought as *topics*, i.e., groups of words that are semantically and/or syntactically similar.

To create these groups of words, we use an automatic method that leverages word co-occurrence patterns in large corpora by making use of the distributional hypothesis: similar words tend to co-occur in similar contexts [26]. Based on co-occurrence statistics, each word is represented as a low dimensional vector of numbers with words closer in this space being more similar [13].

We use a *separate* reference corpus of $\sim$ 400 million tweets to compute a word to word similarity matrix using Word2Vec. This is a neural network approach to distributed word representations, where the words are projected into a lower dimensional dense vector space via a hidden layer [41]. These models can provide a better representation of words compared to traditional language models [42]. In this low dimensional vector space, words with a small Euclidean distance are considered semantically similar. We use the skip-gram model with negative sampling to learn word embeddings on the Twitter reference corpus. We use a layer size of 50 and the Gensim implementation.

We then apply spectral clustering [66, 75] to obtain hard clusters of words from the word $\times$ word similarity matrix. This first performs a dimensionality reduction using SVD on the graph Laplacian of the similarity matrix, obtaining a low-rank embedding of the words. It then performs k-mean clustering to obtain the word clusters. Each user is thus represented by a distribution over topics, where each topic score

is simply the fraction of the words belonging to that cluster. We experiment with 100, 200 and 500 clusters and only show results with the best performing number (200).

We have tried other alternatives to building clusters: using other word similarities to generate clusters (such as NPMI [35] or GloVe [53]) as proposed in [57] or using standard topic modelling approached to create soft clusters of word (e.g., Latent Dirichlet Allocation [2]), but empirical results showed our method provided best results. For brevity, we refrain from presenting all results.

Sentiment & Emotions.

We hypothesise that different traits express different emotions through their posts and aim to automatically quantify this in our Twitter data set. A well studied model of discrete emotions is the Ekman model [16] which posits the existence of six basic emotions: anger, disgust, fear, joy, sadness and surprise. This is also most studied in Natural Language Processing research [68, 69]. We use a publicly available crowdsourcing derived lexicon of words associated with any of the six emotions, as well as trust and anticipation and general positive and negative sentiment [43, 44]. Using these lexicons, we assign a predicted emotion to each message and then average across all users' posts to obtain user level emotion expression scores.

4.2 Image Analysis

We hypothesise that personality traits, including those from the dark triad, are expressed through posting different types of images. Previous work has shown this link for Big Five personality [1, 7, 25, 36]. Although other images may exist in user posts, we restricted ourselves to the profile image, as it is representative of the online persona of the poster and thus might contain important psychological cues [54]. State-of-the-art image recognition systems are models trained on thousands or more features. Since our main goal is interpretability, we compute high-level image cues. We expect that profile pictures will typically contain faces, hence we divide the features into general image features and facial features computed only over images that contain faces.

Image Features.

We compute a series of image aesthetic and attractiveness features, as proxies for the quality and aesthetics of the images [11]. First, we identify if an image is in grayscale or not – grayscale images are considered more artistic (**I-Grayscale**). We compute brightness (**I-Brightness**) and contrast (**I-Contrast**) as the relative variations of luminance. Saturation indicates vividness and chromatic purity, which are more appealing to the human eye [11]. We compute it by transforming images in the HSV (Hue-Saturation-Value) color space and extracting the mean saturation of the pixels (**I-Saturation**). Sharpness measures the coarseness of the degree of detail contained in an image and is a proxy for the quality of the photographer and his gear [32]. This is computed of the mean of the image Laplacian normalized by local average luminance (**I-Sharpness**). Image blur is estimated using the method from [32] (**I-Blur**). We do not explicitly detect the subject, but we use the saliency map [37] to compute a probability of each pixel to be on the subject and re-weight the image features by this probability, similarly to [22]. All the above features are computed on the re-weighted image. Finally, we use a binary indicator variable if the profile picture is the Twitter default (**I-IsDefault**) and exclude these from image analysis.

Facial Features.

We extract general facial features using the Face++ API,[1] which uses deep learning methods to identify faces and landmarks in images. First, we detect if the image contains faces and their number and encode this information in three binary features: one if at least one face is present (**I-HasFace**), one if only a single face is present (**I-OneFace**) and another feature that encodes images with more than one face (**I-MoreFaces**).

Further, we aim to capture the self-presentation characteristics of the user. Features include the face ratio (the size of the face divided by the size of the profile picture – **I-FaceRatio**), whether the face wears any type of glasses (reading – **I-Reading**; or sunglasses – **I-Sunglasses**) or not (**I-NoGlass**), the 3D face posture, which includes the pitch (**I-FacePitch**), roll (**I-FaceRoll**) and yaw angle (**I-FaceYaw**) of the face and the degree of smiling (**I-Smiling**). Where multiple faces exist in an image, we extract features only from the largest detected face.

4.3 Platform Usage

Finally, we extract a series of features which capture platform specific usage. We divide these in user profile derived features and shallow features of tweets.

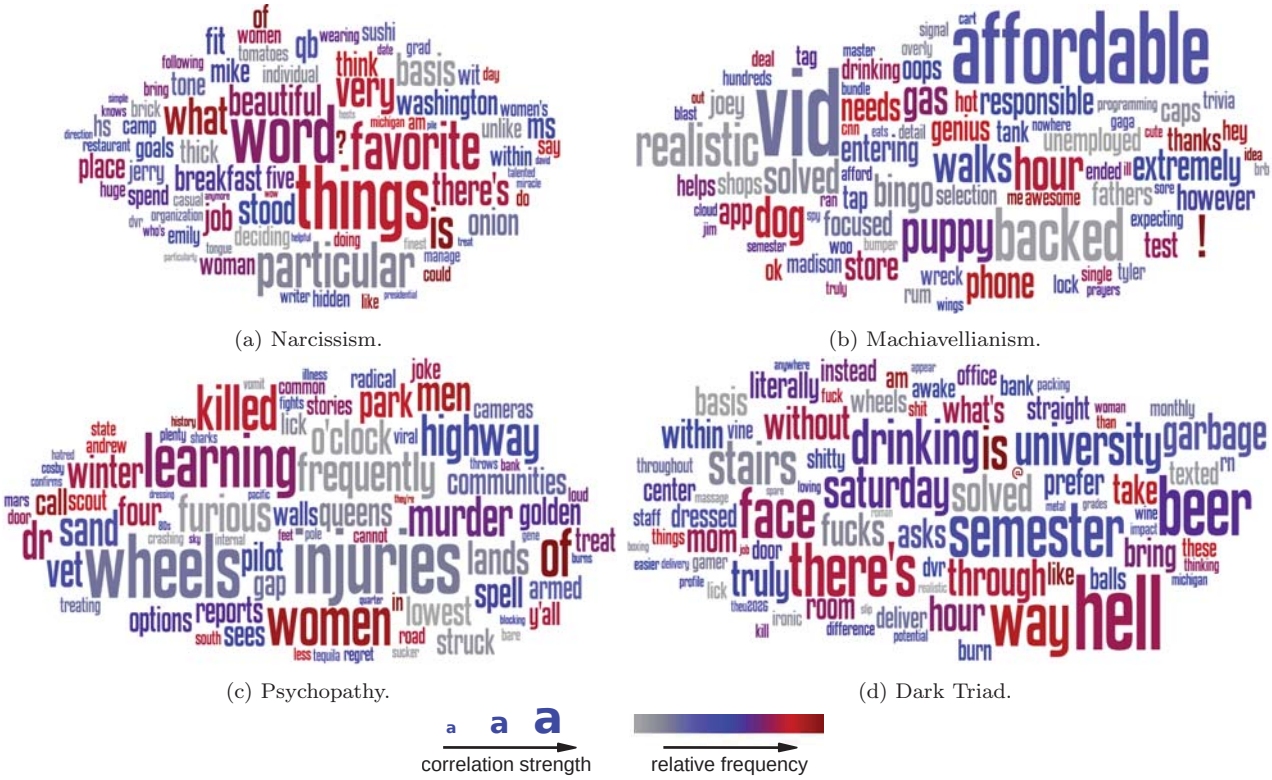
Profile Features.

We extract the total number of tweets posted by a user (**U-No**) and the average number of tweets posted per day by dividing the number of tweets by the number of days since the account was created (**U-NoAvg**). We also quantify basic social attributes of the users: the number of friends (**U-Friends**), the number of followers (**U-Followers**), the follower/friend ratio (**U-FollFriend**) and the number of times the user is listed by others (**U-Listed**). Further, for each message authored by a user we were able to retrieve (i.e., not authored by others and retweeted by the user), we compute the number of times it was retweeted or favorited by others. This information is encoded through four features: the proportion of tweets that were retweeted (**U-RTed**), the average retweet count (**AvgRTed**), the proportion of tweets that were favorited (**U-FAVed**) and the average favorite count (**U-AvgFAVed**). Finally, we code if a user profile uses the default background (**U-DefBack**) and if the account is geo-enabled (**U-Geo**).

Shallow Text Features.

These features capture general tweet-related behaviour. We first compute the average number of characters/tweets (**U-Char**) and tokens/tweets (**U-Tokens**) as simple proxies for message complexity. We then extract the proportion of the user messages which were retweets (authored by others and reposted on the users' page) either using Twitter's automated retweet button or manual retweets, which we extract using regular expressions e.g., starting with 'RT USER' or ending in 'via USER' (**U-RTs**). Separately, we consider duplicate messages those which contain the same first five tokens except @-mentions, as these are usually either spam or the output of automated apps that post to Twitter and

[1] http://www.faceplusplus/



(a) Narcissism.

(b) Machiavellianism.

(c) Psychopathy.

(d) Dark Triad.

Figure 3: The word clouds show the unigrams with the highest Pearson correlation to each of the dark triad traits. The size of the unigram is scaled by its correlation with the outcome. The color indexes relative frequency, from gray (rarely used) through blue (moderately used) to red (frequently used). Correlations are controlled for age and gender in all images, and in (a-c) in addition for the other two traits.

measure their ratio in a users' posts (**U-NonDup**). For measuring interactions with others, we count the proportion of tweets with hashtags (**U-Hash**), the proportion of @-replies (**U-@Reply**), the proportion of tweets containing @-mentions (**U-@Mention**), the number of different users @-mentioned (**U-@Users**) and the proportion of tweets with URLs (**U-URLs**). Finally, we count the proportion of the tweets in which a user actively asks for followers (**U-Asking)** using a list of tokens (e.g., follow, #mustfollow, teamfollowback, #ff).

5. ANALYSIS

In this section, we explore the relationships between Twitter behaviors and the three dark triad traits plus the overall dark triad score. We use univariate Pearson correlations between each feature and the logarithm of the dark triad scores. In order to factor out the influence of basic demographics (i.e., age and gender), we use partial linear correlation.

The results of the analysis of textual features are presented for unigram features in Figure 3 and for LIWC features, word clusters and emotions in Tables 2–5. In order to uncover the peculiarities of each trait given the strong inter-correlations between traits presented in Figure 2, we control in our analysis for the other two traits in addition to age and gender. The average dark triad score captures the core similarities of the dark triad traits. Image feature results are presented in Table 6 and profile feature results in Table 7. We analyse the results by psychological trait below.

5.1 Dark Triad

Aggregate dark triad scores were associated with a range of features and behaviors. When analyzing single words, users high in dark triad post about drinking ('drinking', 'beer'), negative words and aspects of their lives ('hell', 'garbage', 'shitty', 'fuck', 'burn'). Although we are controlling for age, some of the most strongly correlated words are related to college and activities specific to teenagers ('university', 'semester', 'mom', 'texted'). The emotions most associated with the overall dark triad score are negative sentiment, followed by disgust – albeit in part driven by similar words – and trust. By analysing the most frequent trust words, we notice these capture a specific type of positive words. LIWC categories associated with the overall dark triad are, perhaps unsurprisingly, swear words, followed by anger – although most frequent words overlap with the swear words category – words that refer to spatial locations, function words related to present tense and filler words. Present tense and spatial words elude to a concern in the present actions of oneself, specific of both self-promoting and impulsive behaviors specific of the dark triad. The most striking topic is related to sexual activities and pornography, followed by work related words and finally by mentions of driving and addresses, which usually refer to the whereabouts of the author.

The only significant relationships for profile pictures are that users high in the dark triad are less likely to post grayscale images and photos that are less sharp, one of the proxies for photo quality. While there are no significant cor-

Label	**Topic** (most frequent words)	r
LIWC		
Swear	fuck, hell, ass, sucks, suck, butt, dick, crap, heck, dang	.127
Anger	hate, fuck, hell, sucks, war, cut, mad, suck, wars, crap	.123
Space	in, on, at, out, up, off, right, over, where, down	.119
Present	is, have, are, get, i'm, do, can, love, it's, go	.119
Filler	like, blah	.106
Topics		
Porn	hot, sex, naked, teen, adult, porn, petite, foreplay, tits, anal	.152
Work	work, working, doing, stuff, hard, making, done, full, taking, thinking	.126
Driving	car, seat, drive, window, door, horn, air, ride, gas, driving	.126
Addresses	park, tower, lake, gate, hills, street, grand, studios, river, garden	.117
Emotions		
Negative	lost, wait, bad, shit, vote, hate, black, damn, feeling, hell	.108
Disgust	bad, shit, hate, finally, damn, feeling, hell, lose, rob, sick	.122
Trust	good, happy, ground, save, money, show, hope, school, favorite, pretty	.093

Table 2: Pearson correlations between the aggregate dark triad score and textual features extracted from tweets, when controlling for age and gender. Topic labels are manually created. All features are significant at $p < .01$, two tailed t-test.

Label	**Topic** (most frequent words)	r
Topics		
TV Shows	favorite, vote, part, who's, model, voted, fave, fav, stars, entertainment	.119
Daily activities	day, ready, weekends, rest, fresh, #blessed, spend, spent, headed, lazy	.111
Positivity	beautiful, such, loving, rare, fantastic, beyond, simply, unbiased, inspired, spectacular	.110
Future activities	today, week, after, let's, next, season, until, another, coming, weekend	.104
Emotions		
Trust	good, happy, ground, save, money, show, hope, school, favorite, pretty	.130
Positive	love, good, happy, working, save, money, hope, gift, join, favorite	.104

Table 3: Pearson correlations between narcissism and textual features extracted from tweets, when controlling for age, gender, psychopathy and Machiavellianism. Topic labels are manually created. All features are significant at $p < .01$, two tailed t-test.

relations in terms of social connections and other profile features, the shallow textual features reveal a number of patterns. Users high in the dark triad post shorter tweets, albeit not having significantly fewer tokens, showing a preference for shorter tokens. These users endorse far fewer messages from others in the form of retweets, showing thus both fewer appreciation of others' and a tendency to self-promote their own content. Fewer hashtags and URLs in tweets also hint towards a lack of interest in referring to other content. However, examining the behaviors of each trait separately can provide more specific insight.

Label	**Topic** (most frequent words)	r
LIWC		
Death	dead, die, war, died, alive, dying, dies, bury, buried, grief	.153
Anger	hate, fuck, hell, sucks, war, cut, mad, suck, wars, crap	.138
NegEmo	lost, bad, miss, sorry, hate, fuck, hell, sad, lose, seriously	.110
Body	face, ass, head, heart, hand, wear, hearts, hands, fat, wake	.101
Topics		
Fight	fight, kill, killed, treated, died, angry, killing, himself, attack, brave	.144
Names	st, hall, hill, franklin, warren, louis, elizabeth, pete, wiggins, carson	.142
Assaults	fire, police, force, victims, shooting, zone, military, mass, guns, streets	.123
Crime	marijuana, assault, court, guilty, jail, murder, officer, charges, rape, victim	.123
'Positive' aggression	courageous, war, freedom, human, rule, justice, religious, nation, equality, notion	.116
Research	learning, science, project, research, instruction, skills, master, lessons, theory, real-world	.113
	empire, across, waters, harvest, museum, forest, hidden, tale, liberty, ancient	.110
Sport Scores	second, lose, record, point, rules, challenge, earns, lead, major, upping	.108
Trains	between, near, road, outage, train, trains, directions, lane, crash, closed	.103
Weather	o, south, tx, north, west, fair, f, international, airport, calm	.102
Addresses	park, tower, lake, gate, hills, street, grand, studios, river, garden	.099
Emotions		
Negative	lost, wait, bad, shit, vote, hate, black, damn, feeling, hell	.189
Disgust	bad, shit, hate, finally, damn, feeling, hell, lose, rob, sick	.177
Fear	watch, bad, god, hate, feeling, hell, change, chicken, crazy, cash	.174
Anger	money, bad, shit, vote, hot, hate, damn, feeling, hell, crazy	.173
Sadness	lost, bad, music, vote, hate, black, feeling, hell, crazy, lose	.169

Table 4: Pearson correlations between psychopathy and textual features extracted from tweets, when controlling for age, gender, narcissism and psychopathy. Topic labels are manually created. All features are significant at $p < .01$, two tailed t-test.

Label	**Topic** (most frequent words)	r
Topics		
Driving	car, seat, drive, window, door, horn, air, ride, gas, driving	.102
Gratitude	good, great, thanks, thank, everyone, support, enjoy, proud, huge, sharing	.096

Table 5: Pearson correlations between Machiavellianism and high level textual features extracted from tweets, when controlling for age, gender, narcissism and psychopathy. Topic labels are manually created. All features are significant at $p < .01$, two tailed t-test.

5.2 Narcissism

Narcissism is associated with expression that is both positive ('favorite', 'beautiful') and somewhat banal ('breakfast', 'place'). The topics positively associated with narcis-

Feature	Narc	Mach	Psyc	DT
I-IsDefault	-.033	-.015	-.026	-.030
I-Grayscale	-.078	-.049	-.067	-.088
I-Brightness	.012	.018	.065	.033
I-Contrast	-.028	.073	.044	.028
I-Saturation	-.025	.-.048	-.093	-.060
I-Sharpness	-.062	-.059	-.068	-.085
I-Blur	.035	.059	.036	.056
I-HasFace	.041	-.050	-.052	-.018
I-OneFace	.082	-.058	-.029	.003
I-MoreFaces	-.076	.014	-.045	-.043
I-FaceRatio	.010	-.027	.035	.011
I-FacePitch	.006	-.031	-.032	-.028
I-FaceRoll	-.095	-.023	-.096	-.101
I-FaceYaw	-.077	-.062	-.059	-.089
I-NoGlass	-.053	.061	.071	.031
I-Reading	.066	-.006	-.019	.021
I-Sunglasses	-.011	-.102	-.101	-.094
I-Smiling	.117	.005	-.005	.046

Table 6: Pearson correlations between the dark triad personality traits and features extracted from the profile picture, when controlling for age and gender. Positive correlations are highlighted with green ($p < .05$, two-tailed t-test) and negative correlations with red ($p < .05$, two-tailed t-test).

Feature	Narc	Mach	Psyc	DT
U-No	.014	-.027	-.003	-.009
U-NoAvg	.027	-.024	-.013	-.007
U-Friends	-.016	-.028	.056	.010
U-Followers	.001	-.020	.041	.016
U-FollFriend	-.004	.005	.028	.009
U-Listings	.031	-.001	.032	.031
U-DefBack	-.071	.016	.011	-.027
U-Geo	.078	.029	.030	.055
U-RTed	.053	-.020	.006	.022
U-AvgRTed	.010	.002	.036	.019
U-FAVed	.098	-.011	-.018	.037
U-AvgFAVed	.035	.006	.009	.023
U-Char	-.067	-.067	-.064	-.088
U-Tokens	-.027	-.017	-.032	-.034
U-RTs	-.048	-.101	-.041	-.083
U-NonDup	.075	.0	.040	.048
U-Hash	-.080	-.055	-.043	-.082
U-@Reply	-.041	.039	-.003	-.004
U-@Mention	-.075	-.024	-.054	-.071
U-@Users	.059	.015	.054	.055
U-URLs	-.071	-.119	-.075	-.116
U-Asking	-.015	.010	-.101	-.034

Table 7: Pearson correlations between the dark triad personality traits and features extracted from the user's profile, when controlling for age and gender. Positive correlations are highlighted with green ($p < .05$, two-tailed t-test) and negative correlations with red ($p < .05$, two-tailed t-test).

sism also display facile, sanguine language: discussions of reality TV competitions, weekend plans and cheerfulness. Together, these patterns suggest a surface-level pleasantness characterized by narcissistic personalities, as well as a possible chronic assumption that others are interested in their mundane activities and interests. Confirming this, both trust and positive sentiment are correlated with narcissism, although the words that drive both emotions are similar. No LIWC categories were correlated with narcissism beyond age, gender, Machiavellianism and psychopathy.

Users high in narcissism have profile images that are less likely to be grayscale and more likely to feature a single face as opposed to multiple ones and to include smiling. Again, these behaviors characterize a desire to present oneself positively and try to be in the center of attention.

For profile features, narcissism was positively associated with geo-enabled tweets – suggesting Twitter use from mobile – and it was negatively associated with duplicated posts, @-mentions and hashtags. Lack of duplicated posts hints towards a careful curation of the Twitter timeline. Hashtags and @-mentions are ways of inviting others to participate in Twitter interactions, and the negative relationship implies that narcissists prefer to more tightly control their social media spaces, perhaps because of a fear of external criticism. Importantly, they also have a higher proportion of tweets that are favorited at least once; this is consistent with narcissists' preference for regular, positive feedback.

5.3 Psychopathy

Psychopathy shows a very distinct pattern from narcissism and Machiavellianism in online behaviors, with the highest number of correlated word categories. The content of posts associated with users high in psychopathy is coarse, angry, and violent ('killed', 'injuries', 'furious') and characterized by negative emotionality. The correlated LIWC features carry a high level of negativity and morbidity; they are also not driven by the same words in each category, implying that psychopathy is not simply associated with a specific kind of negativity but rather a wide range. However, although psychopathy was significantly associated with all four negative Ekman's emotions, many of these were driven by the same words, e.g., 'lost' or 'bad.' Certain topics suggest an interest in violent world events such as wars or criminal acts. Noteworthy is that one of the topics associated with psychopathy describes a more 'positive' form of aggression ('courageous', 'freedom'). These topics are a good reflection of the sensation-seeking and impulsivity displayed by people high in psychopathy. Similarly, psychopathy was also associated with talking about location and place, which may also indicate impulsivity and recklessness in publicly mentioning physical locations they inhabit.

Psychopaths' profile pictures are only associated with less saturation, which may suggest less planning or conscientiousness in taking or choosing a photo. In profile features, psychopathy was negatively associated with posting URLs and with explicitly asking for followers. This latter finding may indicate a lower level of concern with social approval.

Together, the findings specific to narcissism and psychopathy explain the seemingly paradoxical nature of the overall dark triad results. Narcissists talk about prosaic events with a veneer of positivity, while psychopaths talk about violence and death in an angry manner. Although these two traits are positively correlated, they reveal distinct types of behavior.

5.4 Machiavellianism

Machiavellianism shows the fewest relationships out of the three traits. This may be because it was rather strongly intercorrelated with the other two dark triad traits, leaving less variance to be explained by other features when narcissism

and psychopathy were entered as covariates. The distinctive words associated with Machiavellianism appear to feature a greater amount of spam advertisement posts, suggested by words like 'affordable', 'bingo', 'app', and 'entering'. It may be the case that being high in Machiavellianism is associated with fewer scruples about allowing advertisements to be part of their personal communications. The topics associated with Machiavellianism include the topic associated with the general dark triad focused on cars and driving, which has no theoretical association with Machiavellianism. More notable is a topic focused on expressing gratitude, which could be considered a form of gaining social capital. No LIWC categories and emotions correlated with Machiavellianism above-and-beyond age, gender, psychopathy and narcissism and there are also no relationships for profile pictures and profile features.

Users high in Machiavellianism post fewer URLs and fewer retweets, which confirms the previous findings about these users posting advertisements and more personal messages such as thank-yous. In general, little insight could be provided about the nature of Machiavallianism online above-and-beyond the other two subscales. This is consistent with the theory of the dark triad: Machiavellianism contains surface-level charm and a social orientation in common with narcissism, while it also has a cynicism and lack of concern for ethics in common with psychopathy. It may simply be that few behaviors distinguish Machiavellianism above-and-beyond its more extreme counterparts within the dark triad.

6. PREDICTION

Finally, we use all the previously derived features to create a predictive model of the three dark triad traits, as well as a composite score. We consider this a regression problem to which we can apply machine learning algorithms. We use a linear regression algorithm with an Elastic Net regularizer [77] with the ScikitLearn implementation [50].

To evaluate our results, we split our data into 10 stratified folds and performed cross-validation on one held-out fold at a time. For all our methods we tune the parameters of our models on a separate validation fold. The overall performance is assessed using Pearson correlation of the predicted value to the survey-derived score. In order to emulate a real-world scenario where there is no gender or age skew and uncover the predictive power of our features beyond basic demographics, we predict the residual of each trait after adjusting for the effect of age and gender. Results are presented in Table 8. The same patterns hold when evaluating the results with Root Mean Squared Error (RMSE).

We observe that, with the exception of psychopathy, the textual topic features obtain the best prediction results. In case of psychopathy the LIWC textual categories obtain a better performance. As seen through the feature analysis section, this is probably due to the more syntactical patterns of usage by uses high in psychopathy, which are captured through some LIWC categories. In rest, the predominantly semantic information encapsulated by the Word2Vec word clusters performs best. Other consistently good predictive features are the shallow textual features and the emotions, albeit these do not offer any significant performance when predicting Machiavellianism. Image features perform poorly over the three traits, but provide significant accuracy for predicting the combined score. Profile features also perform poorly in all traits except narcissism.

When combining the features from different modalities, our prediction model obtains similar results across all three traits and the combined score of $R \sim .25$. In order to contextualize our results, subjective psychological variables typically have a *correlational upper-bound* in the range of $R = .3 - .4$ with human behaviors [40]. For example, a model trained on 70,000 Facebook users' posts reports an average Pearson correlation of $R = .351$ across the Big Five personality traits [65]. However, our data set is much more limited being two order of magnitude lower. We expect our method to fare better if more data is available.

Features	# Feat.	Narc	Mach	Psyc	DT
Unigrams	6492	.151	.140	.161	.156
LIWC	64	.146	.091	.245	.184
Word Clusters	200	.225	.211	.205	.194
Emotions	10	.164	.020	.201	.155
Image	18	.040	.012	.044	.100
Profile	12	.087	.014	.002	.047
Shallow	10	.136	.018	.123	.107
All w/o Unigrams	314	.247	.248	.249	.243

Table 8: Prediction results using different types of online behaviors. Performance is measured using Pearson correlation on 10-fold cross-validation.

7. DISCUSSION

We have presented a data-driven, multi-modal exploration of the expression of the dark triad in social media. Using a sample of Twitter users and their dark triad personality scores assessed through questionnaires, we have studied the relationships between Twitter usage, text and profile image features and the three dark triad traits separately, as well as together. Our results represent another perspective that compliments psychological studies conducted using questionnaires about social media usage. Finally, using public behaviour on Twitter, we managed to build a predictive model of the three dark triad traits which achieves robust predictive performance on out-of-sample testing.

As other studies using social media data, we note potential limitations to our approach [59]. The population using social media, here Twitter, and in addition Amazon Mechanical Turk is not representative of the general population. However, traditional psychology research also uses non-representative samples and, in addition, we also control for basic demographics, here age and gender. Behavior on Twitter is public and non-anonymous by default, lending itself to enhanced self-presentation of users [3]. While this is of particular interest especially in the analysis of the dark triad of personality, our results need to be contextualised.

Our study represents the first comprehensive study using observed social media behaviors of the dark triad. As directions of future work, the dark triad of personality can be studied through more complex questionnaires. For example, Narcissism can be assessed using the Narcissistic Personality Inventory, which leads to the division of the trait into three sub factors, some of which have been differently linked to behaviors [6]. By analyzing these subfactors, we can study differences in behavior or mediation effects. Our data and models are freely available online.[2]

[2]http://wwbp.org/

Acknowledgments

The authors acknowledge the support from Templeton Religion Trust, grant TRT-0048.

8. REFERENCES

[1] N. Al Moubayed, Y. Vazquez-Alvarez, A. McKay, and A. Vinciarelli. Face-based automatic personality perception. In *Proceedings of the 22nd ACM International Conference on Multimedia*, MM, pages 1153–1156, 2014.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[3] L. E. Buffardi and W. K. Campbell. Narcissism and Social Networking Web Sites. *Personality and Social Psychology Bulletin*, 34(10):1303–1314, 2008.

[4] D. J. Burger, J. Henderson, G. Kim, and G. Zarrella. Discriminating Gender on Twitter. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, EMNLP, pages 1301–1309, 2011.

[5] A. L. Carey, M. S. Brucks, A. C. Küfner, N. S. Holtzman, M. D. Back, M. B. Donnellan, J. W. Pennebaker, and M. R. Mehl. Narcissism and the Use of Personal Pronouns Revisited. *Journal of Personality and Social Psychology*, 109, 2015.

[6] C. J. Carpenter. Narcissism on Facebook: Self-promotional and Anti-social Behavior. *Personality and Individual Differences*, 52(4):482–486, 2012.

[7] F. Celli, E. Bruni, and B. Lepri. Automatic Personality and Interaction Style Recognition from Facebook Profile Pictures. In *Proceedings of the 22nd ACM International Conference on Multimedia*, MM, pages 1101–1104, 2014.

[8] J. Cheng, C. Danescu-Niculescu-Mizil, and J. Leskovec. Antisocial behavior in online discussion communities. In *Proceedings of the 9th International AAAI Conference on Weblogs and Social Media*, ICWSM, 2015.

[9] Z. Cheng, J. Caverlee, and K. Lee. You are where you Tweet: A Content-Based Aproach to Geo-Locating Twitter Users. In *Proceedings of the 19th ACM Conference on Information and Knowledge Management*, CIKM, pages 759–768, 2010.

[10] G. Coppersmith, M. Dredze, and C. Harman. Quantifying Mental Health Signals in Twitter. In *Proceedings of the CLPsych Workshop*, ACL, pages 51–60, 2014.

[11] R. Datta, D. Joshi, J. Li, and J. Z. Wang. Studying Aesthetics in Photographic Images Using a Computational Approach. In *Proceedings of the 9th European Conference on Computer Vision*, ECCV, pages 288–301, 2006.

[12] M. De Choudhury, S. Counts, and E. Horvitz. Social Media as a Measurement Tool of Depression in Populations. In *Proceedings of the 5th ACM Conference on Web Science*, WebScience, pages 47–56, 2013.

[13] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990.

[14] C. N. DeWall, L. E. Buffardi, I. Bonser, and W. K. Campbell. Narcissism and Implicit Attention Seeking: Evidence from Linguistic Analyses of Social Networking and Online Presentation. *Personality and Individual Differences*, 51(1):57–62, 2011.

[15] J. Eisenstein, B. O'Connor, N. A. Smith, and E. P. Xing. A Latent Variable Model for Geographic Lexical Variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP, pages 1277–1287, 2010.

[16] P. Ekman. An argument for basic emotions. *Cognition & Emotion*, 6(3-4):169–200, 1992.

[17] L. Flekova, J. Carpenter, S. Giorgi, L. Ungar, and D. Preoţiuc-Pietro. Analyzing Biases in Human Perception of User Age and Gender from Text. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, ACL, pages 843–854, 2016.

[18] L. Flekova, L. Ungar, and D. Preoţiuc-Pietro. Exploring Stylistic Variation with Age and Income on Twitter. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, ACL, pages 313–319, 2016.

[19] J. Fox and M. C. Rooney. The Dark Triad and Trait Self-objectification as Predictors of Men's Use and Self-presentation Behaviors on Social Networking Sites. *Personality and Individual Differences*, 76:161–165, 2015.

[20] A. Furnham, S. C. Richards, and D. L. Paulhus. The Dark Triad of Personality: A 10 Year Review. *Social and Personality Psychology Compass*, 7(3):199–216, 2013.

[21] D. Garcia and S. Sikstrom. The Dark Side of Facebook: Semantic Representations of Status Updates Predict the Dark Triad of Personality. *Personality and Individual Differences*, 67:92–96, 2014.

[22] B. Geng, L. Yang, C. Xu, X.-S. Hua, and S. Li. The Role of Attractiveness in Web Image Search. In *Proceedings of the 19th ACM International Conference on Multimedia*, MM, pages 63–72, 2011.

[23] J. Golbeck. Negativity and Anti-social Attention Seeking among Narcissists on Twitter: A Linguistic Analysis. *First Monday*, 2016.

[24] F. Große Deters, M. R. Mehl, and M. Eid. Narcissistic Power Poster? On the Relationship between Narcissism and Status Updating Activity on Facebook. *Journal of Research in Personality*, 53:165–174, 2014.

[25] S. C. Guntuku, L. Qiu, S. Roy, W. Lin, and V. Jakhetiya. Do Others Perceive You As You Want Them To?: Modeling Personality Based on Selfies. In *Proceedings of the 1st International Workshop on Affect & Sentiment in Multimedia*, ASM, pages 21–26, 2015.

[26] Z. Harris. Distributional structure. *Word*, 10(23):146 – 162, 1954.

[27] D. Hovy and A. Søgaard. Tagging Performance Correlates with Author Age. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, ACL, 2015.

[28] P. K. Jonason, S. B. Kaufman, G. D. Webster, and G. Geher. What Lies Beneath the Dark Triad Dirty Dozen: Varied Relations with the Big Five. *Individual Differences Research*, 11:81–90, 2013.

[29] P. K. Jonason, N. P. Li, and D. M. Buss. The costs and benefits of the dark triad: Implications for mate poaching and mate retention tactics. *Personality and Individual Differences*, 48(4):373–378, 2010.

[30] P. K. Jonason and G. D. Webster. The Dirty Dozen: A Concise Measure of the Dark Triad. *Psychological Assessment*, 22(2):420, 2010.

[31] D. N. Jones and D. L. Paulhus. The Role of Impulsivity in the Dark Triad of Personality. *Personality and Individual Differences*, 51(5):679–682, 2011.

[32] Y. Ke, X. Tang, and F. Jing. The Design of High-level Features for Photo Quality Assessment. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, CVPR, pages 419–426, 2006.

[33] 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 of the United States of America*, 110(15):5802–f–5805, 2013.

[34] V. Lampos, N. Aletras, J. K. Geyti, B. Zou, and I. J. Cox. Inferring the Socioeconomic Status of Social Media Users based on Behaviour and Language. In *Proceedings of the 38th European Conference on Information Retrieval*, ECIR, pages 689–695, 2016.

[35] V. Lampos, N. Aletras, D. Preoţiuc-Pietro, and T. Cohn. Predicting and Characterising User Impact on Twitter. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, EACL, pages 405–413, 2014.

[36] L. Liu, D. Preoţiuc-Pietro, Z. Riahi Samani, M. E. Moghaddam, and L. Ungar. Analyzing Personality through Social Media Profile Picture Choice. In *Proceedings of the 10th International AAAI Conference on Web and Social Media*, ICWSM, 2016.

[37] Y.-F. Ma, X.-S. Hua, L. Lu, and H.-J. Zhan. A Generic Framework of User Attention Model and its Application in Video Summarization. *IEEE Transactions on Multimedia*, 7(5):907–919, 2005.

[38] T. C. Marshall, K. Lefringhausen, and N. Ferenczi. The Big Five, Self-esteem, and Narcissism as Predictors of the Topics People Write about in Facebook Status Updates. *Personality and Individual Differences*, 85:35–40, 2015.

[39] S. Mehdizadeh. Self-presentation 2.0: Narcissism and Self-esteem on Facebook. *Cyberpsychology, Behavior, and Social Networking*, 13(4):357–364, 2010.

[40] G. J. Meyer, S. E. Finn, L. D. Eyde, G. G. Kay, K. L. Moreland, R. R. Dies, E. J. Eisman, T. W. Kubiszyn, and G. M. Reed. Psychological Testing and Psychological Assessment: A Review of Evidence and Issues. *American Psychologist*, 56(2):128–165, 2001.

[41] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems*, NIPS, pages 3111–3119, 2013.

[42] T. Mikolov, W. tau Yih, and G. Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2010 annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL, pages 746–751, 2013.

[43] S. M. Mohammad and P. D. Turney. Emotions Evoked by Common Words and Phrases: Using Mechanical Turk to Create an Emotion Lexicon. In *Proceedings of the Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, NAACL, pages 26–34, 2010.

[44] S. M. Mohammad and P. D. Turney. Crowdsourcing a Word-Emotion Association Lexicon. *Computational Intelligence*, 29(3):436–465, 2013.

[45] D. Nguyen, N. A. Smith, and C. P. Rosé. Author Age Prediction from Text Using Linear Regression. In *Proceedings of the 5th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, ACL, pages 115–123, 2011.

[46] E. H. O'Boyle Jr, D. R. Forsyth, G. C. Banks, and M. A. McDaniel. A Meta-analysis of the Dark Triad and Work Behavior: A Social Exchange Perspective. *Journal of Applied Psychology*, 97(3):557, 2012.

[47] E. Y. Ong, R. P. Ang, J. C. Ho, J. C. Lim, D. H. Goh, C. S. Lee, and A. Y. Chua. Narcissism, Extraversion and Adolescents' Self-presentation on Facebook. *Personality and Individual Differences*, 50(2):180–185, 2011.

[48] E. T. Panek, Y. Nardis, and S. Konrath. Mirror or Megaphone?: How Relationships between Narcissism and Social Networking Site Use Differ on Facebook and Twitter. *Computers in Human Behavior*, 29(5):2004–2012, 2013.

[49] D. L. Paulhus and K. M. Williams. The Dark Triad of Personality: Narcissism, Machiavellianism, and Psychopathy. *Journal of Research in Personality*, 36(6):556–563, 2002.

[50] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[51] M. Pennacchiotti and A.-M. Popescu. A Machine Learning Approach to Twitter User Classification. In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*, ICWSM, pages 281–288, 2011.

[52] J. W. Pennebaker, M. E. Francis, and R. J. Booth. *Linguistic Inquiry and Word Count*. 2001.

[53] J. Pennington, R. Socher, and C. D. Manning. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, EMNLP, pages 1532–1543, 2014.

[54] I. S. Penton-Voak, N. Pound, A. C. Little, and D. I. Perrett. Personality Judgments from Natural and Composite Facial Images: More Evidence for a "Kernel of Truth" in Social Perception. *Social Cognition*, 24(5):607–640, 2006.

[55] B. Perozzi and S. Skiena. Exact Age Prediction in Social Networks. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15 Companion, pages 91–92, 2015.

[56] D. Preoţiuc-Pietro, J. Eichstaedt, G. Park, M. Sap, L. Smith, V. Tobolsky, H. A. Schwartz, and L. H. Ungar. The Role of Personality, Age and Gender in Tweeting about Mental Illnesses. In *Proceedings of the CLPsych Workshop*, NAACL, 2015.

[57] D. Preoţiuc-Pietro, V. Lampos, and N. Aletras. An analysis of the user occupational class through twitter content. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, ACL, pages 1754–1764, 2015.

[58] D. Preoţiuc-Pietro, M. Sap, H. A. Schwartz, and L. H. Ungar. Mental Illness Detection at the World Well-Being Project for the CLPsych 2015 Shared Task. In *Proceedings of the CLPsych Workshop*, NAACL, 2015.

[59] D. Preoţiuc-Pietro, S. Volkova, V. Lampos, Y. Bachrach, and N. Aletras. Studying User Income through Language, Behaviour and Affect in Social Media. *PLoS ONE*, 10(9), 2015.

[60] D. Preoţiuc-Pietro, W. Xu, and L. Ungar. Discovering User Attribute Stylistic Differences via Paraphrasing. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI, 2016.

[61] D. Rao, D. Yarowsky, A. Shreevats, and M. Gupta. Classifying Latent User Attributes in Twitter. In *Proceedings of the 2nd International Workshop on Search and Mining User-generated Contents*, SMUC, pages 37–44, 2010.

[62] D. Rout, D. Preoţiuc-Pietro, B. Kalina, and T. Cohn. Where's @wally: A Classification Approach to Geolocating Users based on their Social Ties. In *Proceedings of the 24th ACM Conference on Hypertext and Social Media*, HT, pages 11–20, 2013.

[63] T. Ryan and S. Xenos. Who uses Facebook? An Investigation into the Relationship between the Big Five, Shyness, Narcissism, Loneliness, and Facebook usage. *Computers in Human Behavior*, 27(5):1658–1664, 2011.

[64] M. Sap, G. Park, J. Eichstaedt, M. Kern, L. Ungar, and H. A. Schwartz. Developing Age and Gender Predictive Lexica over Social Media. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, EMNLP, pages 1146–1151, 2014.

[65] H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, L. Dziurzynski, S. M. Ramones, M. Agrawal, A. Shah, M. Kosinski, D. Stillwell, M. E. Seligman, and L. H. Ungar. Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach. *PLoS ONE*, 8(9), 09 2013.

[66] J. Shi and J. Malik. Normalized Cuts and Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.

[67] P. Sorokowski, A. Sorokowska, A. Oleszkiewicz, T. Frackowiak, A. Huk, and K. Pisanski. Selfie Posting Behaviors are Associated with Narcissism among Men. *Personality and Individual Differences*, 85:123–127, 2015.

[68] C. Strapparava and R. Mihalcea. Learning to Identify Emotions in Text. In *Proceedings of the 2008 ACM Symposium on Applied Computing*, pages 1556–1560, 2008.

[69] C. Strapparava, A. Valitutti, et al. Wordnet affect: an affective extension of wordnet. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation*, volume 4 of *LREC*, pages 1083–1086, 2004.

[70] C. Sumner, A. Byers, R. Boochever, and G. Park. Predicting Dark Triad Personality Traits from Twitter usage and a Linguistic Analysis of Tweets. In *The 11th International Conference on Machine Learning and Applications*, ICMLA, pages 386–393, 2012.

[71] S. Volkova, Y. Bachrach, M. Armstrong, and V. Sharma. Inferring Latent User Properties from Texts Published in Social Media. In *Proceedings of the Twenty-Ninth Conference on Artificial Intelligence*, AAAI, pages 4296–4297, 2015.

[72] S. Volkova, G. Coppersmith, and B. Van Durme. Inferring User Political Preferences from Streaming Communications. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, ACL, pages 186–196, 2014.

[73] S. Volkova and B. Van Durme. Online Bayesian Models for Personal Analytics in Social Media. In *Proceedings of the Twenty-Ninth Conference on Artificial Intelligence*, AAAI, pages 2325–2331, 2015.

[74] S. Volkova, T. Wilson, and D. Yarowsky. Exploring Demographic Language Variations to Improve Multilingual Sentiment Analysis in Social Media. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, EMNLP, pages 1815–1827, 2013.

[75] U. von Luxburg. A Tutorial on Spectral Clustering. *Statistics and Computing*, 17(4):395–416, 2007.

[76] E. B. Weiser. # Me: Narcissism and its Facets as Predictors of Selfie-posting Frequency. *Personality and Individual Differences*, 86:477–481, 2015.

[77] H. Zou and T. Hastie. Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society, Series B*, 67:301–320, 2005.