Predicting Dark Triad Personality Traits from Twitter Usage and a Linguistic Analysis of Tweets

Article · December 2012

DOI: 10.1109/ICMLA.2012.218

7 authors, including:

Chris Sumner
The Online Privacy Foundation
8 PUBLICATIONS **535** CITATIONS

Predicting Dark Triad Personality Traits from Twitter usage and a linguistic analysis of Tweets

Chris Sumner[1], Alison Byers[2], Rachel Boochever[3] and Gregory. J. Park[4]
[1]Online Privacy Foundation; chris@onlineprivacyfoundation.org
[2] Online Privacy Foundation; alison@onlineprivacyfoundation.org
[3]Cornell University; rcb264@cornell.edu
[4]University of Pennsylvania; gregoryjpark@gmail.com

***Abstract*—Social media sites are now the most popular destination for Internet users, providing social scientists with a great opportunity to understand online behaviour. There are a growing number of research papers related to social media, a small number of which focus on personality prediction. To date, studies have typically focused on the Big Five traits of personality, but one area which is relatively unexplored is that of the anti-social traits of narcissism, Machiavellianism and psychopathy, commonly referred to as the Dark Triad. This study explored the extent to which it is possible to determine anti-social personality traits based on Twitter use. This was performed by comparing the Dark Triad and Big Five personality traits of 2,927 Twitter users with their profile attributes and use of language. Analysis shows that there are some statistically significant relationships between these variables. Through the use of crowd sourced machine learning algorithms, we show that machine learning provides useful prediction rates, but is imperfect in predicting an individual's Dark Triad traits from Twitter activity. While predictive models may be unsuitable for predicting an individual's personality, they may still be of practical importance when models are applied to large groups of people, such as gaining the ability to see whether anti-social traits are increasing or decreasing over a population. Our results raise important questions related to the unregulated use of social media analysis for screening purposes. It is important that the practical and ethical implications of drawing conclusions about personal information embedded in social media sites are better understood.**

***Index Terms*—Personality, Social Networks, Twitter, Dark Triad.**

I. Introduction

The growth of online social networking has increased dramatically during the past decade. A prime example of this growth is the social networking tool and micro-blogging service, Twitter. In the six months from January to July 2012, the number of Twitter accounts increased 35%, from approximately 383 million to 517 million [7]. This progressive use of social networking technology provides an interesting area of study, although research is still in its infancy. Some services, such as Facebook, have been studied fairly extensively (~412 published articles) [36]; however, there is much less research on Twitter (~150 articles) [11].

Online behavioural analysis may offer important insight into how large numbers of people interact and whether those interactions are changing over time. One area of social media that has had very little research is that of anti-social personality constructs and their relation to online behaviour (Section II), and thus to address this, the present study examines the self-reported 'Dark Triad' personality traits of 2,927 Twitter users (Section III).

We present a summary of the current state of social media personality research, then provide a background on the Dark Triad constructs of narcissism, Machiavellianism and psychopathy. We examine the relationship between these traits and Twitter activities and then apply machine learning techniques to determine the predictability of Dark Triad constructs based solely on Twitter usage. To avoid incorrectly labelling an individual, we pay particular attention to the evaluation metrics of predictive models, because criteria such as Mean Average Error (MAE) and Root Mean Square Error (RMSE) can present good results over an entire model, but can mask inaccuracies when trying to predict the top and bottom percentiles. Since social media personality prediction could be used to label an individual, it is important to ensure the correct evaluation metrics are selected in research studies.

We demonstrate that there are links between Dark Triad constructs and Twitter usage and employ a variety of machine learning techniques to attempt to predict these constructs in users. While the margin of error is too great for examining individuals, there may still be a practical use in examining large groups of people, although important ethical issues must first be addressed.

II. Background and Related Work

A. Social Media and Personality

A large proportion of articles in this field focus on 'Identity Presentation', which can be defined as the examination of personality in relation to social network activity [11] [36]. These articles can be broadly divided into two categories: observer rated studies and automated feature extraction studies. Observer rated studies ask participants to self-assess their personalities, after which independent observers are asked to rate the participants' personalities based on their social media

To appear in conference proceedings at the IEEE 11th International Conference on Machine Learning and Applications ICMLA 2012
Please cite as C. Sumner, A. Byers, R. Boochever and G. J. Park. "Predicting Dark Triad Personality Traits from Twitter usage and a linguistic analysis of Tweets", unpublished.

profiles. Studies have found that observers are able to form a reasonably accurate picture of someone's Big Five personality traits with the greatest success in rating extraversion and openness [17] [26].

Papers focusing on automated feature extraction, however, typically consist of analysing profile attribute and linguistic usage against self-reported personality tests. Profile attributes include features such as number of friends, number of groups and number of status updates. Linguistic analysis examines the frequency of words in pre-defined categories in order to identify patterns that can reveal personality traits. These studies have consistently demonstrated statistically significant relationships between profile attributes, linguistic analysis and personality, although the practical application of these results is limited. The majority of automated feature extraction studies in this area have focused on examining the relationship between social media usage and personality. We are, however, seeing the emergence of data mining and machine learning techniques in exploring the prediction of personalities through social media, rather than purely identifying relationships. Data mining may help expose the low-validity cues proposed within the Realistic Accuracy Model [14] as an explanation of the ability of observers to intuitively detect personality cues in others.

Many of the papers cited in this study have focused on using social media services to predict personality or create recommender systems. Yet recent world events, such as the 'Arab Spring' and the London riots in 2011, highlight the possible use of social media for crime prediction [35].

Since psychopathy is an important factor in understanding violence [24], papers in relation to crime prediction are also beginning to emerge examining psychopathy through language and social network activity. A 2011 paper identified statistically significant differences between the language of psychopathic and non-psychopathic murderers, indicating that language may provide an insight into the subconscious mind of the psychopath [22]. A second paper has expanded on this research, examining the relationship between undergraduate students' scores on a psychopathy inventory within the text in their emails, text messages, and Facebook messages [10]. Similarly, Boochever found that students higher in psychopathy ratings use language differently than those with less psychopathic tendencies. Given the public fascination for psychopathy and the emergence of predictive modelling, it is, perhaps, not surprising to see articles emerge asking, "Can Twitter Help Expose Psychopath Killers' Traits?" [2].

B. The Dark Triad of Personality

The personality construct of psychopathy has begun to be studied in combination with Machiavellianism and narcissism. The three constructs are "overlapping, but distinct", and have been named the Dark Triad of personality because they all focus, to varying degrees on social malevolence, self-promotion, emotional coldness, duplicity and aggressiveness [28].

Narcissism is arguably the oldest of the three constructs and originated from the Greek myth of Narcissus, who, as legend has it, was doomed to fall in love with his own reflection in a pool of water. The Diagnostic and Statistical Manual of Mental Disorders IV (DSM-IV) [1] defines a narcissistic personality as a pervasive pattern of grandiosity, need for admiration, and lack of empathy, all of which begin by early adulthood and are present in a variety of contexts. Narcissism is the only member of the Dark Triad listed in the DSM-IV. The narcissist tends to view him- or herself as intelligent, powerful, physically attractive, unique and entitled [12].

Machiavellianism takes its name from Niccolò Machiavelli (1469-1527), a Florentine diplomat, and is characterised by a tendency to deceive and manipulate other people, usually for personal gain [9]. Although previously believed to have been a sub-clinical form of psychopathy, Paulhus and Williams [28] identified that whilst there is an overlap; Machiavellianism and psychopathy are indeed "distinct constructs".

Psychopathy holds perhaps the greatest public fascination as demonstrated by the popularity of films such as The Silence of the Lambs, and books such as The Psychopath Test by Jon Ronson. Although there is still disagreement in what characterises psychopathy, the Psychopathy Check List Revised (PCL-R) [23] is arguably the most used measurement instrument. The PCL-R consists of twenty items including a lack of empathy and guilt, glib speech, pathological lying, a grandiose sense of self-worth, anti-social and promiscuous behaviour and a parasitic lifestyle. Individuals are given a psychopathy score between 0 and 40, with a generally accepted cut off between 25 and 30. Individuals receiving a score above the cut off will typically be labelled a Psychopath, with approximately 1% of the population falling into this category [23].

Research shows that neurological abnormalities are responsible for a predisposition to psychopathy [27], but also that early childhood experience is an important factor in whether an individual will turn to crime and/or violent crime [20].

C. Social Media and the Dark Triad

We were able to identify several papers on narcissism and social networking, but to our knowledge, there is only one publicly available paper, (Boochever [10]) examining psychopathy and social media usage. Boochever found that users who scored higher (more psychopathic) on the Self-Report Psychopathy Test III (SRP-III), used more swear words and words related to anger, and psychologically distanced themselves from their messages, all reflecting emotional deficits and disagreeableness fundamental to the psychopathic personality. We were unable to find any papers covering all three Dark Triad traits and social media.

D. Machine Learning and Prediction

To our knowledge, the present study is the first to examine machine prediction of all three Dark Triad personality traits using social media. We examine the predictive ability of machine learning algorithms over the importance of individual features, and by doing so we aim to present a more realistic

To appear in conference proceedings at the IEEE 11th International Conference on Machine Learning and Applications ICMLA 2012
Please cite as C. Sumner, A. Byers, R. Boochever and G. J. Park. "Predicting Dark Triad Personality Traits from Twitter usage and a linguistic analysis of Tweets", unpublished.

assessment of the predictability of these psychological traits from available social media services.

We were able, however, to identify a number of papers that have used machine prediction in the context of the Big Five personality traits and social network sites [15][16][31].

Two of these papers [15][16] assert that it is possible to predict personality to within approximately 10% of an individual's self-reported personality. The evaluation metric used is the Mean Average Error (MAE). Evaluation methods such as MAE can produce impressive results if predictions focus around mean or majority values in unimodal distributions. The practical performance of models, however, may highlight poor results when examining the True Positive Rate (TPR) and the True Negative Rate (TNR).

To ensure critical analysis of performance it is therefore important to use a number of evaluation criteria. Golbeck et al also supply the correlation coefficient [16], which indicates reasonable overall predictive performance. This is certainly performance worth investigating in future studies; however, it is still not possible to determine how well the models work in terms of identifying the top and bottom extremes of the study's population distribution.

To date, there is inconsistency in the reporting of machine prediction performance in the context of social media usage and personality. Given the potential real-world uses of this information, e.g. pre-employment screening [3], and despite discrimination and invasion of privacy issues, this inconsistency should be addressed.

III. DATA COLLECTION

2,927 Twitter users from 89 countries participated in the present study. Twitter profile and time zone information indicated that the majority of participants resided in Great Britain (N=876) and the United States (N=609), with 1,442 participants residing in 87 other countries. It was not possible to determine the age or sex of participants as Twitter does not collect this information when users register for the service.

Participants were initially recruited through Tweets sent out by the Online Privacy Foundation (@The_OPF), and through the distribution of leaflets in Basingstoke and London, United Kingdom.

The majority of participants volunteered for this study following Tweets from celebrity Tweeters including British personality Stephen Fry (@StephenFry) and US Skateboarder Tony Hawk (@TonyHawk). This may have introduced a selection bias and is discussed further in Section VI.

With the exception of the first 200 participants, who were eligible to win an iPad, participants were not incented or compensated for their participation.

A purpose-built Twitter application was developed to collect self-reported ratings on the Short Dark Triad (SD3) questionnaire [29] providing measures of narcissism, Machiavellianism and psychopathy; and the Ten Item Personality Inventory (TIPI), providing measures of openness, conscientiousness, extraversion, agreeableness and emotional stability (reversed as neuroticism in this paper) [18].

A number of additional applications were created to download and process information from each participant, including their Tweet history, up to the maximum imposed by the Twitter API of approximately 3,200 Tweets. Each participant's historic Twitter post content was analysed using the standard categories provided in the Linguistic Inquiry and Word Count (LIWC) 2007 software [30]. This was further divided into original Tweets, replies and Retweets.

Processing historic Twitter post data in this manner resulted in 586 features, such as friends, followers, number of Tweets and the frequency of pre-defined words for each individual. After removing personally identifiable information, a subset of 337 features were selected for use in machine prediction.

A. Machine Learning Methodology

Prediction of personality traits can be performed in two ways, both of which are examined in this paper:

- A classification task, where the goal is to identify individuals with particularly high or low values of a trait according to some predetermined cut-off.
- A regression task, where the goal is to predict an individual's score for each of the eight personality traits based on their Twitter usage.

We considered two cut-off values for classification: the median and the 90th percentile. The median value of a trait is the value which splits the sample in half, and the 90th percentile represents the value which 90% of the sample falls below.

To identify individuals with above-median levels of each trait, we first labelled all individuals in the full data set (training and test set combined) as above (high) or below (low) the median value of each of the eight traits. Next, we used four "off-the-shelf" classification methods from the WEKA machine learning toolkit [21] to classify the 1,172 individuals in the test set as either above- or below-median on each trait after training each algorithm on the training set of 1,755 individuals. Default settings were used for the following algorithms from WEKA:

- Support Vector Machine (SVM) using sequential minimal optimization (SMO) and a polynomial kernel.
- Random Forest, an ensemble method that combines multiple decision trees.
- J48, an implementation of the C4.5 decision tree algorithm.
- Naïve Bayes (NB) classifier.

The above four algorithms were complemented by several models generated through Kaggle [4], a web-based service for hosting data science competitions. We hosted two separate competitions: one to predict psychopathic traits specifically [6] and another to predict the remaining seven personality traits [5]. This resulted in two sets of winning models, from 1,715 submissions.

To compare the Kaggle models with the binary classifiers from WEKA, we used the continuous predictions as a classifier, varying the cut-off used to classify individuals across the entire range of the continuous predictions and assessing performance at each possible cut-off value.

To appear in conference proceedings at the IEEE 11th International Conference on Machine Learning and Applications ICMLA 2012
Please cite as C. Sumner, A. Byers, R. Boochever and G. J. Park. "Predicting Dark Triad Personality Traits from Twitter usage and a linguistic analysis of Tweets", unpublished.

We then compare the performance of all six methods (the four WEKA algorithms, a benchmark Kaggle model, and the respective winning Kaggle models) using several performance metrics and create visual comparisons using receiver operating characteristic (ROC) plots.

This entire procedure was repeated for the classification of individuals using the 90th percentile cut-off for each trait.

IV. STATISTICAL ANALYSIS

A simple, zero-order Spearman's correlation was conducted on the self-reported Dark Triad and Big Five personality scores and values obtained through analysing Twitter profile and language data. Significant correlations from the linguistic analysis are presented in Table I, with Twitter profile information presented in Table II.

The most statistically significant results were found in the relationship between Dark Triad traits and language.

Narcissistic traits were significantly positively correlated with 'other punctuation' (OtherP), which includes the @ and # characters ($r(2,614) = 0.073$, $p = < 0.001$). The @ and # characters have special significance when used in Twitter. The @ character is used before other characters to signify a Twitter username and is typically used in replies and Tweets which mention other users, while the # character indicates a "hashtag", something that facilitates search. Narcissistic traits were also significantly positively correlated with words associated with sex ($r(2,614) = 0.061$, $p = 0.002$). This could be explained by an increased urge to fulfil basic needs, possibly as a reaction to not having had these basic needs satisfactorily fulfilled earlier in life [33], or may be a result of the narcissistic need for triumph and domination.

Machiavellian traits were significantly positively correlated with swear words ($r(2,614) = 0.129$, $p = < 0.001$), anger ($r(2,614) = 0.116$, $p = < 0.001$) and negative emotions (negemo) ($r(2,614) = 0.073$, $p = < 0.001$), suggesting that as levels of Machiavellianism increase, so does the use of negative and hostile language. Machiavellian traits were significantly negatively correlated with positive emotion ($r(2,614) = -0.118$, $p = < 0.001$) and the use of the word "we" ($r(2,614) = -0.070$, $p = < 0.001$), showing that as levels of Machiavellianism increase, references to other people, i.e. "we", decreases. This supports the assertion that Machiavellianism is related to an increased self-focus [28].

Psychopathic traits were significantly positively correlated with swear words ($r(2,614) = 0.187$, $p = < 0.001$), anger ($r(2,614) = 0.151$, $p = < .001$), death ($r(2,614) = 0.094$, $p = < 0.001$) and negative emotion (negemo) ($r(2,614) = 0.084$, $p = < 0.001$). We also saw significantly positive correlations between psychopathic traits and filler words ($r(2,614) = 0.073$, $p = < 0.001$).

In no cases did all three Dark Triad traits share statistically significant results. For example, while we saw increased levels of swearing and anger in relation to Machiavellianism and psychopathy, these relationships did not appear with narcissism.

V. MACHINE LEARNING RESULTS

In order to examine the predictability of Dark Triad traits, we used a number of machine learning techniques, as described in Section III-A. The best performing models came from the winners of a data science competition hosted on Kaggle.com and are:

- For the prediction of narcissism and Machiavellianism: An ensemble combining random forests, gradient boosted trees, support vector machines, and multivariate adaptive regression splines.
- For the prediction of psychopathy: An ensemble of thousands of gradient boosted trees.

As stated in Section II-D, while a model may appear accurate, the practical performance may still result in an unacceptable number of errors. The Area Under the Curve (AUC) values therefore provide a reasonable indication of the practical performance. Wald et al [34], offer further evaluation criteria including the Geometric and Arithmetic Mean of the TPR and TNR. Finally, we also provide the Accuracy (Acc) for both the maximum Geometric and Arithmetic means. These are presented in Table III as AUC, G-Mean, G-TPR, G-TNR, A-Mean, A-TPR, and A-TNR and are displayed for both median split and 90^{th} percentile split classification.

Results show that classification accuracies are slightly above chance for the best performing models. In the case of the 90^{th} percentile split the high accuracy results were achieved by classifying most cases as negatives, with all models being poor at detecting true positives. In practical terms, this results in the correct identification of 2 out of 125 individuals making up the top 10% of the distribution, without incurring any false positives.

The winning Kaggle models (from the psychopathy contest [6] and the contest for the remaining seven traits [5]) accounted for the greatest amount of trait variance compared to all other models. Results indicate that there is significant variation in the between traits in their respective predictability from Twitter, indicating that some traits (such as psychopathy and extraversion) may be easier to predict than others (such as openness and conscientiousness).

To appear in conference proceedings at the IEEE 11^{th} International Conference on Machine Learning and Applications ICMLA 2012
Please cite as C. Sumner, A. Byers, R. Boochever and G. J. Park. "Predicting Dark Triad Personality Traits from Twitter usage and a linguistic analysis of Tweets", unpublished.

Table I : Spearman's correlations between linguistic variables and personality scores. Significant correlations at the 0.01 level (2-tailed) are shown in bold.

Category	Abbrev	Examples	Na	Ma	Ps	Op	Co	Ex	Ag	Ne
Linguistic Processes										
Words > 6 letters	Sixltr	Words > 6 letters	0.047	**-0.078**	-0.047	0.035	0.042	0.042	0.017	**0.079**
Dictionary words	Dic	Dictionary words	**-0.088**	0.001	-0.050	**-0.073**	**-0.051**	-0.031	-0.007	**-0.113**
Total function words	funct	Total Function words	**-0.093**	0.021	-0.033	**-0.076**	**-0.094**	-0.046	-0.033	**-0.123**
Total pronouns	pronoun	I, them, itself	-0.043	0.016	-0.023	-0.027	**-0.100**	-0.004	-0.003	**-0.142**
Personal pronouns	ppron	I, them, her	-0.021	0.021	-0.017	-0.016	**-0.092**	0.022	0.013	**-0.145**
1st pers singular	i	I, me, mine	-0.017	0.050	-0.001	-0.011	**-0.116**	-0.028	-0.023	**-0.171**
1st pers plural	we	We, us, our	0.036	**-0.070**	**-0.071**	-0.006	**0.052**	**0.063**	**0.068**	0.044
Impersonal pronouns	ipron	It, it's, those	**-0.077**	0.008	-0.033	**-0.051**	**-0.099**	**-0.060**	-0.035	**-0.115**
[Common verbs]	verb	Walk, went, see	**-0.084**	0.014	-0.028	**-0.083**	**-0.071**	-0.013	0.002	**-0.132**
Auxiliary verbs	auxverb	Am, will, have	**-0.078**	0.025	-0.018	**-0.074**	**-0.098**	-0.021	-0.021	**-0.131**
Past tense	past	Went, ran, had	**-0.069**	-0.003	-0.040	**-0.066**	**-0.051**	0.001	0.000	**-0.074**
Present tense	present	Is, does, hear	**-0.068**	0.017	-0.024	**-0.061**	**-0.071**	-0.002	0.002	**-0.140**
Adverbs	adverb	Very, really, quickly	**-0.088**	0.026	-0.035	**-0.077**	**-0.071**	-0.034	-0.022	**-0.119**
Prepositions	preps	To, with, above	**-0.065**	-0.030	**-0.086**	**-0.059**	**0.062**	-0.024	0.021	0.009
Conjunctions	conj	And, but, whereas	**-0.057**	0.031	-0.030	**-0.070**	**-0.085**	**-0.056**	-0.014	**-0.109**
Negations	negate	No, not, never	**-0.073**	**0.068**	0.034	**-0.080**	**-0.068**	-0.034	**-0.069**	**-0.124**
Quantifiers	quant	Few, many, much	**-0.056**	0.010	-0.012	-0.042	**-0.064**	-0.045	-0.032	**-0.085**
Numbers	number	Second, thousand	0.036	**0.064**	0.030	-0.043	**0.082**	0.038	-0.046	**0.117**
Swear words	swear	Damn, piss, fuck	0.040	**0.129**	**0.187**	-0.028	**-0.171**	0.025	**-0.136**	**-0.097**
Psychological Processes										
Social processes	social	Mate, talk, they, child	0.014	**-0.060**	-0.050	-0.007	-0.021	**0.101**	**0.096**	-0.036
Family	family	Daughter, husband,	0.008	-0.036	**-0.076**	0.012	-0.007	0.041	**0.102**	-0.020
Friends	friend	Buddy, friend	**0.073**	-0.030	0.001	0.048	-0.032	**0.121**	**0.069**	-0.049
Affective processes	affect	Happy, cried, abandon	-0.038	**-0.062**	-0.050	-0.017	-0.009	**0.069**	**0.091**	**-0.095**
Positive emotion	posemo	Love, nice, sweet	-0.020	**-0.118**	**-0.124**	0.006	**0.077**	**0.108**	**0.183**	-0.042
Negative emotion	negemo	Hurt, ugly, nasty	-0.023	**0.073**	**0.083**	-0.049	**-0.135**	-0.027	**-0.109**	**-0.138**
Anxiety	anx	Worried, fearful	**-0.054**	-0.021	-0.042	-0.034	**-0.056**	-0.014	-0.007	**-0.179**
Anger	anger	Hate, kill, annoyed	0.004	**0.116**	**0.151**	-0.034	**-0.144**	-0.016	**-0.145**	**-0.103**
Sadness	sad	Crying, grief, sad	-0.038	-0.024	-0.036	-0.035	**-0.078**	-0.009	-0.024	**-0.111**
Cognitive processes	cogmech	cause, know, ought	**-0.074**	0.010	-0.029	-0.046	**-0.090**	**-0.061**	-0.030	**-0.099**
Insight	insight	think, know, consider	-0.048	-0.022	-0.042	-0.001	**-0.082**	**-0.060**	-0.008	**-0.094**
Discrepancy	discrep	should, would, could	**-0.059**	0.012	-0.013	**-0.053**	**-0.069**	-0.024	-0.015	**-0.094**
Tentative	tentat	maybe, perhaps, guess	**-0.102**	0.020	-0.023	-0.036	**-0.088**	**-0.081**	-0.046	**-0.078**
Inclusive	incl	And, with, include	0.004	-0.015	**-0.054**	-0.031	-0.001	0.032	0.027	-0.016
Exclusive	excl	But, without, exclude	**-0.084**	0.044	-0.002	**-0.056**	**-0.100**	**-0.062**	**-0.068**	**-0.110**
Perceptual processes	percept	Heard, feeling	**-0.052**	-0.040	**-0.092**	-0.015	-0.020	-0.031	**0.070**	**-0.089**
See	see	View, saw, seen	-0.049	-0.036	**-0.082**	-0.043	0.038	-0.004	**0.062**	-0.039
Biological processes	bio	Eat, blood, pain	0.001	0.009	0.012	-0.026	**-0.066**	0.023	0.014	**-0.136**
Body	body	Cheek, hands, spit	0.006	**0.053**	**0.066**	-0.021	**-0.077**	0.017	-0.029	**-0.112**
Health	health	Clinic, flu, pill	-0.005	-0.032	-0.019	0.007	-0.010	-0.005	0.032	**-0.106**
Sexual	sexual	Horny, love, incest	**0.068**	0.030	**0.051**	0.029	**-0.083**	**0.082**	0.008	**-0.156**
Relativity	relativ	Area, bend, exit, stop	-0.028	0.021	**-0.059**	**-0.060**	**0.088**	0.019	0.021	0.006
Motion	motion	Arrive, car, go	-0.020	-0.008	**-0.060**	-0.029	**0.075**	0.046	**0.055**	0.000
Time	time	End, until, season	-0.050	0.041	**-0.063**	**-0.070**	**0.061**	-0.005	0.026	**-0.054**
Personal Concerns										
Work	work	Job, majors, Xerox	-0.006	-0.046	**-0.077**	-0.009	**0.061**	-0.015	-0.016	0.048
Death	death	Bury, coffin, kill	0.021	0.039	**0.094**	0.050	**-0.070**	**-0.051**	**-0.085**	-0.003
Spoken categories										
Assent	assent	Agree, OK, yes	0.011	0.013	0.038	0.024	**-0.067**	**0.060**	0.016	**-0.125**
Nonfluencies	nonfl	Er, hm, umm	-0.023	-0.027	0.005	-0.008	-0.041	**0.051**	-0.015	-0.041
Fillers	filler	Blah, Imean, youknow	0.035	**0.102**	**0.073**	0.000	**-0.121**	-0.024	**-0.066**	**-0.127**
Punctuation										
Total Punctuation	AllPct		0.043	-0.039	-0.031	**0.093**	0.028	0.000	**0.055**	0.024
Exclam	Exclam	!	0.018	-0.003	**-0.054**	-0.016	0.050	**0.107**	**0.104**	-0.011
Dash	Dash	-	-0.006	**-0.052**	-0.025	0.046	0.042	**-0.057**	0.045	0.048
Quote	Quote	""	**0.057**	0.020	-0.035	0.050	**-0.055**	**-0.098**	-0.020	-0.034
Apostro	Apostro	'	**-0.088**	0.012	-0.022	-0.028	**-0.101**	**-0.085**	-0.017	**-0.129**
OtherP	OtherP	@ #	**0.070**	-0.015	0.016	**0.051**	0.046	**0.062**	0.033	**0.060**

To appear in conference proceedings at the IEEE 11th International Conference on Machine Learning and Applications ICMLA 2012
Please cite as C. Sumner, A. Byers, R. Boochever and G. J. Park. "Predicting Dark Triad Personality Traits from Twitter usage and a linguistic analysis of Tweets", unpublished.

Table II : Spearman's correlations between profile attributes and personality scores. Significant correlations at the 0.01 level (2-tailed) are shown in bold.

Twitter Attribute	Mean	Std. deviation	Na	Ma	Ps	Op	Co	Ex	Ag	Ne
Description Length	67.949	53.856	**0.088**	-0.047	-0.017	**0.117**	0.044	**0.121**	0.024	-0.006
Followers[1]	1277.999	54116.514	**0.158**	-0.020	0.039	**0.117**	-0.023	**0.149**	-0.018	-0.017
Friends[2]	243.513	709.638	**0.116**	-0.035	0.015	**0.068**	-0.026	**0.115**	0.006	-0.007
Number of Followers per Friend	1.229	20.266	**0.129**	0.006	**0.044**	**0.118**	-0.012	**0.129**	**-0.038**	-0.015
Lifetime Tweets	3502.576	9137.265	**0.061**	0.006	0.044	**0.066**	**-0.054**	0.046	-0.039	0.048
Tweets in last 3 months	707.953	972.011	**0.082**	-0.006	**0.057**	**0.094**	**-0.051**	**0.073**	-0.037	0.008
Number of original Tweets	0.529	0.220	0.027	0.011	**0.050**	**0.050**	**-0.064**	0.001	**-0.051**	**0.076**
Number of Retweets	0.170	0.157	**0.047**	-0.034	0.018	0.039	**-0.049**	0.040	-0.004	**0.048**
Number of Replies	0.300	0.204	-0.004	-0.028	0.027	0.026	**-0.077**	**0.057**	-0.041	**0.070**
Number of "Follow Fridays"	0.001	0.004	**0.049**	**-0.075**	0.016	**0.070**	0.006	**0.044**	0.019	0.006
Number of Favourites	105.655	646.512	**0.074**	-0.012	0.025	**0.107**	**-0.062**	0.014	-0.020	**0.067**
Number Listed	16.054	320.005	**0.067**	**-0.050**	-0.004	**0.097**	-0.017	**0.049**	-0.005	-0.004
Klout score	26.023	13.078	**0.088**	-0.016	**0.051**	**0.055**	-0.036	**0.121**	**-0.055**	0.025

[1]Number of people who follow that user
[2]Number of people who that user follows

Table III : Evaluation metrics for median and 90th percentile splits against Dark Triad and Big Five personality traits.

	Median Split									90th Percentile Split								
Trait	***AUC***	***Acc*** *Max A-Mean*	***Acc*** *Max G-Mean*	***G-Mean***	***G-TPR***	***G-TNR***	***A-Mean***	***A-TPR***	***A-TNR***	***AUC***	***Acc*** *Max A-Mean*	***Acc*** *Max G-Mean*	***G-Mean***	***G-TPR***	***G-TNR***	***A-Mean***	***A-TPR***	***A-TNR***
Psy	0.641	0.610	0.610	0.611	0.642	0.582	0.610	0.641	0.583	0.678	0.896	0.660	0.651	0.640	0.663	0.896	0.024	1.000
Mac	0.602	0.586	0.586	0.571	0.640	0.509	0.586	0.400	0.740	0.609	0.919	0.641	0.596	0.547	0.649	0.919	0.000	1.000
Nar	0.612	0.598	0.598	0.577	0.589	0.565	0.598	0.370	0.789	0.625	0.890	0.683	0.596	0.504	0.705	0.890	0.023	0.999
Op	0.591	0.582	0.582	0.564	0.556	0.573	0.582	0.318	0.820	0.603	0.877	0.600	0.587	0.571	0.604	0.877	0.020	1.000
Co	0.605	0.593	0.593	0.588	0.671	0.516	0.593	0.683	0.506	0.611	0.858	0.541	0.594	0.681	0.518	0.858	0.000	1.000
Ex	0.644	0.623	0.623	0.611	0.619	0.602	0.623	0.475	0.753	0.668	0.899	0.667	0.629	0.585	0.676	0.899	0.000	1.000
Ag	0.600	0.587	0.587	0.566	0.545	0.588	0.587	0.368	0.775	0.647	0.876	0.696	0.618	0.531	0.720	0.876	0.000	1.000
Ne	0.612	0.598	0.598	0.582	0.576	0.589	0.598	0.412	0.765	0.624	0.837	0.625	0.595	0.554	0.639	0.837	0.021	1.000

VI. Discussion

The present paper sought to examine the relationship between Dark Triad personality traits and Twitter activity and examine whether machine learning could be used to predict these constructs based solely on Twitter usage.

Our results identify a number of statistically significant correlations between Dark Trait traits and Twitter usage. In terms of linguistic analysis, it was noted that people higher in scores of psychopathy and Machiavellianism tend to use more swear words and more words associated with anger. Both traits were also significantly negatively correlated with first person plurals and words associated with positive emotion. The major differences between psychopathy and Machiavellianism are in the frequency of words associated with sex, relativity, motion and time. Narcissism displayed far less overlap with psychopathy and Machiavellianism, most notably in the significant negative correlations between narcissism and common verbs. The relationships between language and psychopathy, specifically in relation to swearing, anger and negative emotion, support the findings of both Hancock et al [22] and Boochever [10].

A statistical analysis of Twitter profile attributes, such as the number of friends and followers showed the majority of statistically significant relationships related to narcissism. People with higher scores in narcissism tended to have both more followers and friends and a greater numbers of followers per friend. Narcissism was also positively correlated with Klout scores, a score associated with exerting influence over other users through online behaviour.

Overall, our results support Paulhus et al's observation that while the Dark Triad constructs are related, they are not equivalent [13]. An examination of the cross-correlations between Dark Triad and Big Five traits highlights that the only area of commonality between Dark Triad and Big Five traits was low agreeableness. This observation is also consistent with those from Paulhus et al.

In conducting this research, our observation is that the machine learning evaluation criteria provided in prior papers may not be sufficient to support the conclusions within those papers. Examples include the use of both RMSE [31] and MAE [15] [16]. This makes it difficult to compare and critically evaluate practical performance. Evaluation methods such as MAE and RMSE can mask larger errors at the extremes of a unimodal population distribution by predicting the majority of instances around the mean value. In practical terms this means that the people who are likely to be of most interest, i.e. those furthest from the mean, can easily be mislabelled, e.g. the model may predict a high scoring extrovert as a low scoring introvert without substantially affecting the overall MAE. To overcome these limitations, we provided a number of evaluation criteria.

To appear in conference proceedings at the IEEE 11th International Conference on Machine Learning and Applications ICMLA 2012
Please cite as C. Sumner, A. Byers, R. Boochever and G. J. Park. "Predicting Dark Triad Personality Traits from Twitter usage and a linguistic analysis of Tweets", unpublished.

Results from machine learning highlight that it is possible to correctly identify 2 out of the top 10% of scorers on the SD3 psychopathy scale while incurring no false positives. Increasing the true positive rate greatly increases the false positive rate, therefore, using Twitter alone for personality prediction is likely to be both insufficient and error prone. The picture gets a little more complex if we consider using Twitter usage with other observations and information. Using the Big Five, one study has observed that a multisource approach enhances predictive validity [8]. Therefore, in combining personality prediction from social media, other observable behaviour, and criminal records, the ability to identify people scoring highly in anti-social Dark Triad traits is not out of the question.

It is, however, important to stress that high-scores in anti-social traits do not always lead to anti-social behaviour such as violence and criminal activity [13].

A more practical use of the methods outlined in the present study is to better understand whether levels of anti-social traits vary between geographies and whether they are varying over time. Research already shows that anti-social traits are lower in communal countries as opposed to agentic countries and suggests that higher levels of Dark Triad traits may be beneficial for success within agentic societies [25] and conversely that they may have contributed to the Global Financial Crisis [9]. Theories related to anti-social behaviour within social groups are not new. For example, in "Unmasking the Psychopath" we find the text "As early as 1835, James C. Pritchard recognized that, among other factors, industrialization, along with its many consequences, was an important cause of psychopathy ('moral insanity')" [32]. The ability to observe sociological changes in anti-social traits represents an important area of future study and may also offer an insight into the motivations behind "trolling" and cyber bullying.

An important consideration that must be addressed is in what circumstances, if any, is social media based personality prediction ethically acceptable.

Finally, this study has a number of limitations. The sample set largely consists of followers of British celebrity Stephen Fry and US skateboarder Tony Hawk. This may have introduced a selection bias, however, through the descriptive statistics, we can see that our sample is generally consistent with those used by Paulhus [29] and Gosling [18]; therefore any selection bias may be negligible.

A further limitation is the use of the 2007 LIWC dictionary [30], as language on Twitter is likely to be different to language in correctly written speech, due to the 140 character limit. While this is a limitation, the most significant results come from words associated with swearing, anger, positive and negative emotions, which likely are the same in speech, blogs and on Twitter. However, this could be masking the fact that other correlations were not significant because the words themselves were manipulated into unrecognisable words, to fit into the Twitter character limit. The 140 character limit of Tweets may also result in more a more direct use of language. A future area of research could be further analysis of the linguistics used in social media.

Another limitation was introduced by processing the historic Twitter data using LIWC to extract pre-defined word. This resulted in a relatively small number of features for subsequent machine learning. Generating features from the raw text using techniques from Natural Language Processing may result in many more features and better predictive models.

The most notable limitation is that the research was based solely on self-assessment questionnaires, which people could easily manipulate to produce a measurement error [19]. Self-reporting, however, is a widely used method and with such a large sample size it is unlikely that the results would be significantly skewed by individuals wishing to manipulate their scores. Additionally, since there were no consequences for the volunteers who participated in the present study, there seems little benefit in manipulating the results. Further research could focus on combining multiple personality assessments such as self-reports, interview-led reports and observer reports to reduce the sensitivity to measurement errors.

VII. Conclusions

This study highlights that there are relationships between Twitter activity, Dark Triad and Big Five personality traits, but that the practical performance of machine prediction is currently poor when applied directly to an individual. Our results demonstrate that while our models display a high degree of accuracy, defined as (TP + TN) / (TP + TN + FP + FN); the TPR and TNR remain poor, highlighting a need for greater focus on evaluation criteria in future studies.

As research in this field matures, behavioural research with social media will likely offer an important insight into the levels and variability of anti-social behaviours within and between social groups. While mainstream media reports may focus on individual level personality prediction, this field of research may be of greater use in examining changes in society over time.

Results also indicate that while users may be careful about the content they post on Twitter, the words they use may reveal more about their personalities than they would wish. This points to critical questions around the possible need for regulatory controls and/or raising awareness amongst users in order to prevent the misuse of information derived from Twitter and other online social network activity.

Acknowledgement

The winning Kaggle models were created by Yukihiro Tagami, Yahoo Corporation Japan, (psychopathy) and Willie Liao, University of California, Berkeley, (narcissism, Machiavellianism and Big Five traits).

References

[1] "Diagnostic and Statistical Manual of Mental Disorders (4th ed., text rev.)", Washington, DC: American Psychiatric Association, 2000.

To appear in conference proceedings at the IEEE 11th International Conference on Machine Learning and Applications ICMLA 2012
Please cite as C. Sumner, A. Byers, R. Boochever and G. J. Park. "Predicting Dark Triad Personality Traits from Twitter usage and a linguistic analysis of Tweets", unpublished.

[2] “Daily News and Analysis.”, [Online]. Available: http://www.dnaindia.com/world/report_can-twitter-help-expose-psychopath-killers-traits_1598342. 13 October 2011 [Accessed 7 June 2012].

[3] “Jobvite Social Recruiting Survey,” [Online]. Available: http://recruiting.jobvite.com/resources/social-recruiting-survey.php. 9 July 2012. [Accessed 18 July 2012].

[4] "Kaggle" [Online]. Available: http://www.kaggle.com/. [Accessed 2 6 2012].

[5] “Personality Prediction Based on Twitter Stream.” [Online]. Available: https://www.kaggle.com/c/twitter-personality-prediction. 8 May 2012 [Accessed 1 August 2012].

[6] “Psychopathy Prediction Based on Twitter Usage.” [Online]. Available: https://www.kaggle.com/c/twitter-psychopathy-prediction. 14 May 2012 [Accessed 1 August 2012].

[7] “Twitter reaches half a billion accounts.” [Online]. Available: http://semiocast.com/publications/2012_07_30_Twitter_reaches_half_a_billion_accounts_140m_in_the_US. 30 July 2012 [Accessed 10 August 2012].

[8] M. D. Back, "The predictive validity, judgeability and consequential outcomes of personality: A multisource approach," proc. European Association of Personality Psychology, Trieste, 2012.

[9] C. R. Boddy, “The Corporate Psychopaths Theory of the Global Financial Crisis,” Journal of Business Ethics, vol. 102, no. 2, pp. 255-259, 2011.

[10] R. Boochever, "Psychopaths Online: Modeling Psychopathy in Social Media Discourse", 2012. [Online]. Available http://hdl.handle.net/1813/29536

[11] D. Boyd, “Bibliography of Research on Twitter & Microblogging.” Internet: http://www.danah.org/researchBibs/twitter.php. [Accessed 6 June 2012].

[12] L. E. Buffardi and W. K. Campbell, “Narcissism and social networking web sites,” Personality and Social Psychology Bulletin, pp. 1303-1314, 2008.

[13] D. Freedman, “Premature Reliance on the Psychopathy Checklist-Revised in Violence Risk and Threat Assessment”, Journal of Threat Assessment, 2002.

[14] D. Funder, “On the accuracy of personality judgement: A realistic approach,” Psychological Review, pp. 652-670, 1995.

[15] J. Golbeck, C. Robles, M. Edmondson and K. Turner, “Predicting Personality from Twitter,” in IEEE International Conference on Privacy, Security, Risk and Trust and IEEE International Conference on Social Computing, Boston, Massachusetts, USA, 2011.

[16] J. Golbeck, C. Robles and K. Turner, “Predicting Personality with Social Media,” in Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems, Vancouver, BC, Canada, 2011.

[17] S. D. Gosling, S. Gaddis and S. Vazire, “Personality impressions based on Facebook profiles,” in International conference on Weblogs and Social Media, Boulder, CO, USA, 2007.

[18] S. D. Gosling, R. P. J. Renfrow and W.B. Swann Jnr, “A Very Brief Measure of the Big Five Personality Domains,” Journal of Research in Personality, pp. 504-528, 2003.

[19] R.M. Groves "Survey Errors and Survey Costs.", New York, John Wiley and Sons, 2004

[20] A.S. Gullhaugen and J.A. Nøttestad, "Looking for the Hannibal Behind the Cannibal: Current Status of Case Research", International Journal of Offender Therapy and Comparative Criminology, pp. 350-369, 2011

[21] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I. Witten, “The WEKA Data Mining Software: An Update,” SIGKDD Explorations, pp. Vol. 11, no. 1, pp 10-18, 2009.

[22] J. T. Hancock, M.T. Woodworth and S. Porter, “Hungry like the wolf: A word-pattern analysis of the language of psychopaths,” Legal and Criminological Psychology, pp. doi: 10.1111/j.2044-8333.2011.02025.x, 2011.

[23] R. D. Hare, “Psychopathy: A clinical and forensic overview”, Psychiatric Clinics of North America, pp. 709-724, 2006.

[24] S. D. Hart, "The role of psychopathy in assessing risk for violence: Conceptual and methodological issues", Legal and Criminological Psychology, pp. 121-137, 1998

[25] P. K. Jonason, N. P. Li and E. A. Teicher, “Who is James Bond?: The Dark Triad and an Agentic Social Style,” Individual Differences Research, vol. 8, no. 2, pp. 111-120, 2010.

[26] D. H. Kluemper, P. A. Rosen and K. W. Mossholder, “Social Networking Websites, Personality Ratings, and the Organizational Context: More Than Meets the Eye?”, Journal of Applied Social Psychology, pp. 1143–1172, 2012.

[27] R. Oliveria-Souza et al., "Psychopathy as a disorder of the moral brain: Fronto-temporo-limbic grey matter reductions demonstrated by voxel based morphonmetry", NeuroImage, 40, pp 1202-1213, 2008

[28] D.L.Paulhus and K. M. Williams, “The Dark Triad of personality: Narcissism, Machiavellianism, and psychopathy,” Journal of Research in Personality, pp. 556–563, 2002.

[29] D.L. Paulhus and D.N. Jones, "Introducing a short measure of the Dark Triad.". Poster presented at the meeting of the Society for Personality and Social Psychology, San Antonio, USA. 2011.

[30] J. W. Pennebaker, R. E. Boot and M. E. Francis, (2007). "Linguistic inquiry and word count: LIWC2007 - Operator's manual." Austin, TX: LIWC.net

[31] D. Quercia, M. Kosinki, D. Stillwell and J. Crowcroft, “Our Twitter Profiles, Our Selves: Predicting Personality with Twitter,” proc. 3rd IEEE International Conference on Social Computing, Boston, Massachusetts, USA, 2011.

[32] W. H. Reid, D. Dorr, J. I. Walker and J. W. Bonner III, "Unmasking the psychopath: Antisocial Personality and Related Syndromes.", New York: Norton, 1986.

[33] E. Straub, “The Roots of Evil: Social Conditions, Culture, Personality, and Basic Human Needs,” Personal and Social Psychology Review, pp. 179-192, 1999.

[34] R. Wald, T.M. Khoshgoftaar, A. Napolitano and C. Sumner “Using Twitter Content to Predict Psychopathy,” Unpublished

[35] X. Wang, M. S. Gerber and D. E. Brown, “Automatic Crime Prediction Using Events Extracted from Twitter Posts,” proc. International Conference on Social Computing, Behavioral-Cultural Modeling, & Prediction, Hyattsville, Maryland, USA, 2012.

[36] R. E. Wilson, S. D. Gosling and L. T. Graham, “A Review of Facebook Research in the Social Sciences,” Perspectives on Psychological Sciences, pp. 203-220, 2012.

To appear in conference proceedings at the IEEE 11th International Conference on Machine Learning and Applications ICMLA 2012
Please cite as C. Sumner, A. Byers, R. Boochever and G. J. Park. “Predicting Dark Triad Personality Traits from Twitter usage and a linguistic analysis of Tweets”, unpublished.