

Predicting Author Profiles From Online Abuse Directed at Public Figures

Isabelle van der Vegt¹, Bennett Kleinberg^{1, 2}, and Paul Gill¹

¹ Department of Security and Crime Science, University College London

² Department of Methodology and Statistics, Tilburg University



The problem of online threats and abuse directed at public figures could potentially be mitigated with a computational approach, where sources of abusive language are better understood or identified through author profiling. However, abusive language constitutes a specific domain of language that is untested on whether differences emerge based on personality, age, or gender of text authors. The present study presents a unique data set of 789 abusive messages directed at politicians. It examines statistical relationships between author demographics of text authors and (abusive) language, then uses a machine learning approach to predict personality, age, and gender based on language in the texts. Results showed that (a) personality traits could be determined within 10% of their actual value, (b) age was determined with an error margin of 10 years, and (c) gender was classified correctly in 70% of the cases. Even though we found statistically significant relationships between language use and demographics, prediction performance was poor when compared to previous research on author profiling. Therefore, we suggest that further research is needed before author profiling systems can be of significant value within the context of abusive language and threat assessment.

Public Significance Statement

We test the feasibility of determining author characteristics (personality, age, and gender) through language, focussing specifically on threats and abuse directed at public figures. Results show that there are indeed statistical relationships between author characteristics and language use in abusive texts. However, machine learning prediction of author traits through language remains error prone.

Keywords: author profiling, abusive language, threat assessment, threats to public figures, machine learning


Supplemental materials: <https://doi.org/10.1037/tam0000172.supp>


This article was published Online First January 27, 2022.
Isabelle van der Vegt  <https://orcid.org/0000-0001-6448-3388>

Bennett Kleinberg  <https://orcid.org/0000-0003-1658-9086>

Open Science Framework repository: <https://osf.io/ag8hu/>

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant Agreement No. 758834).

 The data are available at https://osf.io/ag8hu/?view_only=f6b8ee9ece1a41fd964f80ef0b06aa3f

 The experiment materials are available at https://osf.io/ag8hu/?view_only=f6b8ee9ece1a41fd964f80ef0b06aa3f

Correspondence concerning this article should be addressed to Isabelle van der Vegt, Department of Security and Crime Science, University College London, 35 Tavistock Square, London WC1H 9EZ, United Kingdom. Email: isabelle.vandervegt@ucl.ac.uk

In June 2016, a far-right terrorist murdered Labour Party MP Jo Cox during the European Union referendum campaign in the United Kingdom (Cobain & Taylor, 2016). Prior to the U.K. elections in December 2019, a record number of female MPs stood down citing the constant abuse and threats they endure (Perraudin & Murphy, 2019). Violent threats to politicians and public figures remain a serious problem, in particular, due to the rise of threats communicated over the internet. Computational linguistics can play a key role in better understanding and mitigating this social phenomenon.

In recent years, studies increasingly attempted to understand and detect abusive language and hate speech. These include studying abusive posts on social media, comment sections and forums (Nobata et al., 2016; Waseem & Hovy, 2016), or online extremist language use (Figea et al., 2016; Scrivens et al., 2018). Some studies aim to “profile” the authors of text, examining language use to estimate, for example, the age, gender, and personality of the author. For instance, measures of personality through language are utilized in a tool aimed at assessing risk of violence in written communication (Akrami et al., 2018). While such an approach may be of particular interest to threat assessment practitioners and law enforcement agencies to triage online threats, we argue that author profiling in this domain requires further testing before it can be successfully deployed in practice.

Although a large body of wider research examined the relationship between writing and personality (Pennebaker & King, 1999), age (Pennebaker & Stone, 2003), and gender (Newman et al., 2008), the majority have obtained small effects (Azucar et al., 2018; Qiu et al., 2012). A few have used linguistic variables to predict author characteristics, but classification accuracy varies widely from, for example, 45%-92% (personality; Argamon et al., 2005; gender; Burger et al., 2011; age; Nguyen et al., 2013; personality; Preotiuc-Pietro et al., 2016). However, it has yet to be examined whether there is a relationship between personality, age, and gender when individuals write abusive text. For instance, do highly extraverted or narcissistic individuals write an abusive message differently than people who score low on these traits? Do men use more violent language than women? Whether aspects of abusive language can be used to infer one’s personality, age, and gender also remains untested. These endeavors will be especially important in the context of assessing violent threats directed at public figures, or other threat assessment

purposes where a computational approach may increase insight and reduce human workload.

The present study presents an experiment in which participants write a neutral, nonoffensive text, as well as an abusive text directed at a politician. Our aim is to (a) examine relationships between author characteristics (personality, age, and gender) and neutral and abusive language, and (b) predict author profiles based on the linguistic characteristics of neutral and abusive texts. In our view, this is an important endeavor because author profiling is gaining traction within the field of violence research (Kop et al., 2019; Neuman et al., 2015) and threat assessment tools (Akrami et al., 2018) developed for possible use in practice. Therefore, it is crucial to test whether author profiling approaches can indeed be generalized to the domain of abusive language where the feasibility of author profiling thus far is unknown. Before discussing previous research on author profiling, we examine the issue of harassment and threats directed at public figures.

On- and Offline Threats to Public Figures

Politicians and other public figures are at an increased risk of threats of violence, stalking, and harassment as a result of their visibility. This problem has been identified in a large number of countries. For instance, of 239 surveyed MPs in the United Kingdom, a majority reported being the victim of either intrusive and aggressive behaviors (81%) or stalking and harassment (53%; James et al., 2016). Similar figures emerged for politicians in New Zealand, where 87% of politicians experienced unwanted harassment, and 50% of MPs reported being approached by their harassers (Every-Palmer et al., 2015). In Norway, 82% of politicians experienced unwanted behavior or threats ((Bjørge & Silkoset, 2018).

Several different samples of public figure attackers identify high rates of mental health disorders. In a study of nonterrorist attackers on politicians in Western Europe between 1990 and 2004, almost half were psychotic (James et al., 2007). Those with mental disorders were responsible for the more serious and fatal attacks (James et al., 2007). Several countries set up Fixated Threat Assessment Centres (FTAC), where mental health professionals and police collaborate to assess and manage risks posed by individuals who have pathological fixations on politicians, royalty, or other public figures (James et al., 2010). In cases from the U.K. FTAC, 83.6% of individuals suffered serious mental disorders (James et al., 2009), while 70% of

cases in the Australian Queensland FTAC had a formal psychiatric diagnosis (Pathé et al., 2015). In an analysis of 4,387 cases of threatening contact to U.S. members of Congress, individuals who engaged in problematic approach behavior were significantly more likely to have a prior criminal record and display signs of serious mental illness (Scalora et al., 2002). Other research has pointed to an increased polarization of political debate, facilitated by online communication, leading to greater number of threats (Lelkes et al., 2017). In an overview and analysis of problematic approaches to public figures in the United States, the authors note that individuals who wrote to celebrities with “an excessive sense of self-importance or uniqueness (grandiosity or narcissism)” were more likely to approach (Dietz & Martell, 2010). The characteristics of public figures (politicians) themselves have also been reported to play a role, with (younger) females and ethnic minorities receiving more abuse and threats (Inter-Parliamentary Union, 2016).

Importantly, a large amount of abuse and threats to public figures now occurs online. In a study of 270,000 tweets directed at 573 U.K. MPs, 62% of MPs received at least one abusive tweet in a 2-month timeframe (Ward & McLoughlin, 2020). Recognition (measured by number of Twitter followers and mentions) was positively correlated with the amount of abuse. While male MPs received almost twice as many abusive tweets than female MPs (3% of male MPs received abuse vs. 1.7% in female MPs), female MPs were highly overrepresented in the group receiving hate speech tweets (86% of hate speech targets were female). Therefore, the authors suggest that the abuse MPs receive depends on gender and could potentially be viewed as more threatening for females (Ward & McLoughlin, 2020).

The nature of online communication has also been viewed as contributing element to online abuse and threats (Ward & McLoughlin, 2020). The relative ease and low cost of online communication have been raised as an important factor, while online anonymity is said to promote disinhibition (Rowe, 2015). As a result, individuals may express views and abuse without fearing sanction (Ward & McLoughlin, 2020). The anonymity of internet users makes it particularly difficult to gain information about the demographics of threateners, or even to identify possible suspects in a law enforcement context. Resultingly, author profiling based on language may provide a possible solution. In the next sections, we describe previous research on the correlates between language use and author

characteristics, as well as previous attempts at predicting author characteristics from language.

Linguistic Correlates of Author Characteristics

Early studies using automated approaches to studying language departed from the assumption that linguistic content and style differ between individuals (Pennebaker & King, 1999). Specific traits such as the Big Five personality traits (Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism) were correlated with certain linguistic characteristics, such as the use of negative emotion words, negations, and present tense verbs (Pennebaker & King, 1999). Linguistic variables were measured with Linguistic Inquiry and Word Count software, which can be used to measure descriptive categories (e.g., words per sentence, words longer than six characters), grammatical categories (e.g., pronouns, articles), psychological concepts and processes (e.g., power, positive emotion), personal concern categories (e.g., family, money), informal language (e.g., swearing, filler words), and punctuation (e.g., periods, commas) in text (Pennebaker et al., 2015). Measurements are based on word count, in that the LIWC reports the number (proportion) of words found in a document that relate to each category. For the linguistic assessment of personality, the LIWC was applied to a sample of psychology students’ writing samples ($N = 1,203$), who wrote a “stream-of-consciousness (SOC)” essay describing current thoughts, feelings, and sensations (Pennebaker & King, 1999). Results showed small positive correlations between neuroticism and negative emotion words ($r = 0.16$), and a positive correlation between positive emotion words ($r = 0.15$), social references ($r = 0.12$), and extraversion. Other endeavors showed correlations of $r = 0.23$ between personality traits and LIWC categories (Hirsh & Peterson, 2009).

A possible effect of age on language has also been examined. In a large-scale study, references to the self and others decreased with age, as well as an increase of present- and future-tense over past-tense verbs with age (Pennebaker & Stone, 2003). Aging was also associated with an increase in positive emotion words ($r = 0.05$) and a decrease in negative emotion words ($r = -0.04$; Pennebaker & Stone, 2003). Gender differences in language emerged in a study of 14,324 text samples including stream-of-consciousness essays (Newman et al., 2008). Women more

often used LIWC dimensions such as pronouns ($r = 0.18$) and social words ($r = 0.10$).

Predicting Author Characteristics From Language

Besides the study of linguistic correlates of author profiles, linguistic information has also been used to predict personality traits, age, and gender using a machine learning approach. In one prediction example, participants completed a personality questionnaire and wrote stream-of-consciousness essays (i.e., expressing their current thoughts and feelings), after which the traits neuroticism and extraversion were predicted (Argamon et al., 2005). A binary classification task was performed, where participants were either high (top third) or low (bottom third) scoring on the traits. Various psycholinguistic measures (such as the LIWC) were used as features, and the average classification accuracy was 58% (Argamon et al., 2005). In a similar effort, n -grams (i.e., word occurrences) were used as features to predict Big Five scores in several binary and multiclass prediction tasks (Oberlander & Nowson, 2006). Accuracies ranged from 45% to 100% depending on the task, personality trait, and feature set (Oberlander & Nowson, 2006).

Importantly, personality traits are considered more accurately conceptualized as continuous constructs rather than as binary or categorical variables (Haslam et al., 2012). Some prediction efforts estimated traits on a continuous scale, using a regression approach. This has, for example, been done for Big Five personality impressions (i.e., third-person annotations) of YouTube vlogger videos using the LIWC (Farnadi et al., 2014). The best performance was achieved for conscientiousness (RMSE = 0.64 on a scale of 1-7, $R^2 = 0.18$). Another study predicted Dark Triad traits (narcissism, Machiavellianism, and psychopathy) from Twitter data including unigrams, LIWC categories, and profile picture features, with ground truth established through a self-report survey (Preotiuc-Pietro et al., 2016). The best model showed a correlation of 0.25 between predicted and observed values (Preotiuc-Pietro et al., 2016). In another study, both regression and classification tasks were used for Big Five and Dark Triad prediction with LIWC measures of Twitter profiles as features (Sumner et al., 2012). Prediction performance was poor for both tasks, even though the authors identified correlations between personality traits and LIWC categories in the Twitter data (Sumner et al., 2012).

Various studies also worked on predicting age and gender. In the PAN¹ 2016 shared task on this topic, the best performance for predicting five age classes was 58.97% using stylistic features and vector representations of terms and documents (Rangel et al., 2016). Gender was correctly classified 75.64% of the time using stylometric features (e.g., pronouns and adjectives) and n -grams (Rangel et al., 2016). Age has also been predicted on a continuous scale using unigrams, with a mean absolute error; MAE of approximately 4 years (Nguyen et al., 2013). Furthermore, gender classification on Twitter using n -grams achieved 91.80% accuracy when all tweets from a profile were used (Burger et al., 2011).

Author Profiling Grievance-Fueled Communications

Importantly, author profiling is also gaining traction within violence threat assessment, for example, when the source of an abusive, threatening, or extremist text posted online needs determining. The Profile Risk Assessment Tool (PRAT), which is intended for risk assessment of violent written communications, constructs a personality profile of a text author (Akrami et al., 2018). The profiles are constructed by means of IBM Watson Personality Insights, which predicts Big Five traits with models trained on word embeddings (i.e., words represented by vectors of other semantically close words) from a large data set for which personality traits of text authors were known. IBM Personality Insights has also been used to study the texts of “pseudocommando mass murderers,” defined as individuals who “are obsessed with weapons and meticulously plan their attack” (Kop et al., 2019). Personality traits measured in the mass murderer texts were compared to population medians, with the former scoring higher on openness, but lower on extraversion and agreeableness (Kop et al., 2019). In a study on profiling the texts of school shooters, personality profiles were constructed by means of semantic vector representations of text (Neuman et al., 2015). Distances were computed between vectors for each school shooter text and vectors representing traits such as narcissism, but also for disorders such as paranoid personality disorder, schizotypal personality disorders, and depression. The same was done for control samples of neutral writing. After ranking all

¹ Plagiarism analysis, Authorship identification, and Near-duplicate detection: <https://pan.webis.de/>

texts on these measures, school shooter texts could be identified by examining 3% of the entire corpus (Neuman et al., 2015).

A notable approach to linguistically studying grievance-fuelled communications is the Grievance Dictionary, a psycholinguistic dictionary similar to the LIWC, but specifically aimed at grievance-fuelled language (van der Vegt et al., 2021). The dictionary can be used to measure concepts relevant to threat assessment in text, such as categories relating to violence, hate, paranoia, and weaponry. Although the Grievance Dictionary has previously been used to distinguish between violent and non-violent writing samples (van der Vegt et al., 2021), it has yet to be used for the specific purpose of author profiling. The present study will explore the use of both the Grievance Dictionary and the LIWC for author profiling of abusive texts.

The Present Study

Since author profiling is increasingly applied within the domain of understanding (potentially) violent individuals and threat assessment, we recognize the importance of testing (a) whether there are statistically significant relationships between author characteristics (personality, age, and gender) and abusive language use, (b) whether author characteristics can indeed be predicted from abusive texts. We focus on personality due to its increased popularity in violence and threat assessment research (Akrami et al., 2018; Kop et al., 2019; Neuman et al., 2015), whereas age and gender may be of particular interest in practice to determine the source of an anonymous threatening communication.

Method

Transparency Statement

Data, code, and supplemental materials are publicly available on the Open Science Framework: <https://osf.io/ag8hu/>.

Sample

Eight hundred participants were recruited through the online crowdsourcing platform Prolific Academic. Only adult U.K. citizens with English as their first language were eligible. Participants who failed the attention checks² were excluded, resulting in a sample of 789.

Procedure

The study procedure was approved by the local departmental ethics committee. Participants wrote both a SOC essay about current thoughts and feelings, and an abusive text directed at a politician. Each task lasted for at least 3 min and participants had to write at least 100 words. For the abusive writing task, participants rated eight U.K. politicians from most to least favorite, then were assigned to write about their negative thoughts and feelings about their least favorite politician. They were told they could be as insulting, abusive, and offensive as they wanted. Finally, the participants completed two personality questionnaires and were asked about their gender and age. Data were collected between 22 and 23 October 2019.

Personality Measures

In order to assess personality, two tests were used. The HEXACO-60 (Ashton & Lee, 2009) measures honesty–humility, emotionality, extraversion, agreeableness versus anger, conscientiousness, and openness to experience, on a scale from 1 = *strongly disagree* to 5 = *strongly agree*, with 10 questions per trait (i.e., resulting in a score between 1 and 50 per person). The Short Dark Triad (SD3; Jones & Paulhus, 2014) measures Machiavellianism, narcissism, and psychopathy on a Likert scale from 1 = *strongly disagree* to 7 = *strongly agree*, with nine questions per trait (i.e., a score of 1–63 per person).

Writing Examples

Below, we provide a writing example (original wording, anonymization added) for both the SOC and abusive writing tasks.

Stream-of-Consciousness

I feel content and I am reasonably happy at this present moment in time. It may be a challenging few months for me and I am looking forward to the time ahead. Some times I do feel at times that things get on top of me and find it hard to get going in the morning. I think that the future is bright for me and I fight on with perseverance and determination even though I have had some setbacks. I overall feel more confident and determined than ever even though at times I doubt myself for a brief moment.

² Two questions asking participants to select a specific response (e.g., *strongly disagree*) to continue.

Abusive Writing

[POLITICIAN] you are a liar, a cheat, an abhorrent person, your arrogance is beyond repair, you are determined to drag the country into the gutter, you are a complete shit with total disregard for women, I hope you die in regret of what you have dragged our country into, we are now the laughing stock of [redacted], I hope you rot, shame on you, you are possibly the worst politician that we have ever had, you deserve a long and hard punishment for what you've done, you utter prick, please rot in hell for a long long time I hope.

Statistical Tests

Prior to performing the prediction tasks, we tested to determine the presence of any statistically significant relationships between author characteristics (personality, age, and gender) and LIWC2015 variables (Pennebaker et al., 2015) as well as Grievance Dictionary categories (van der Vegt et al., 2021). We computed correlations for personality traits, applying a Bonferroni-corrected threshold of $0.05/(89 * 9) = 0.000062$ for 89 LIWC categories and nine personality traits, and $0.05/(22 * 9) = 0.00025$ for 22 Grievance Dictionary categories and nine personality traits. A Bonferroni correction accounts for the possibility of inflated false positives as a result of conducting multiple tests (for each linguistic category and personality trait).

Multivariate regression assessed the effect of age (and quadratic age, here: the absolute difference from age 40) on all LIWC2015 and Grievance Dictionary categories, while controlling for gender, following Pennebaker and Stone (2003). To examine gender and language, we assessed whether there is a multivariate effect of gender in a MANOVA for all LIWC2015 categories following Newman et al. (2008). We did the same for Grievance Dictionary categories. For both analyses, we report Pillai's Trace, a test statistic (ranging between 0 and 1) that increases if the (gender) effects are contributing more to the model. Thereafter, we conducted univariate ANOVAs to demonstrate the direction and magnitude (reported using Cohen's d effect size, Cohen, 1988) of gender differences in LIWC and Grievance Dictionary categories.

Prediction Tests

All prediction and classification tests below were performed for SOC and abusive writing separately. In addition to the LIWC and Grievance Dictionary measures of linguistic content,

we also examined prediction performance for stylistic features (e.g., grammatical categories in the LIWC, parts-of-speech, number of words). For each machine learning task, we tested each of the following feature sets:

1. Number of words (baseline model)
2. Stemmed uni- and bigrams: frequencies of single words (e.g., "kingdom") and word pairs (e.g., "united kingdom")
3. Parts-of-speech (universal POS tags from the spacyr R package: (Benoit & Matsuo, 2020): frequencies of grammatical categories such as nouns, verbs, and pronouns.
4. All 89 LIWC2015 categories (Pennebaker et al., 2015). In the abusive writing condition, we also include the proportion of abusive language³ words in this feature set.
5. All 22 Grievance Dictionary categories (van der Vegt et al., 2021)
6. Composite feature set: all of the above features.
7. Filtered feature set: a selection of features from the composite feature set, filtered using a General Additive Model (Chouldechova & Hastie, 2015), and included if there is a functional relationship ($p < .05$) between the feature and outcome variable, during 10 resampling iterations (Kuhn, 2010).
8. Pretrained word embeddings, using the GloVe 6B corpus (Pennington et al., 2014): each word is represented as a vector of the cosine distance with 100 semantically similar words from the corpus. These measures are then averaged in order to represent each text as a function of 100 distances.
9. Pretrained Bidirectional Encoder Representations from Transformers (BERT) language model (base uncased model with 12 layers and 768 hidden nodes): similarly represents words as a vector, but takes into account contextual relations between words through bidirectional training (Devlin et al., 2019).

All tasks were performed with a 10-fold cross-validation on the training set. The training set consisted of 80% of the data, and the remaining 20% of the sample was used as a hold-out test

³ A composite measure of abusive language following Kleinberg et al. (2021), measuring profane and racist language from various dictionaries.

set. Allocation to the training or hold-out test set occurred randomly. The model was subsequently trained 10 times on 10 different random samples from the training data. Then, the optimal model was chosen to perform test set predictions on the test set (the held out 20% of data). We could then evaluate the prediction performance by comparing the predictions for the test set to the actual observed values of this sample. The prediction analysis included the following steps:

- Predicting the HEXACO and Dark Triad traits in isolation on a continuous scale (a regression model using a Support Vector Machine (SVM) algorithm). We report the Mean Absolute Percentage Error (MAPE), which represents the average prediction error across all iterations in proportional form. For instance, a MAPE of 10% on a scale of 1-100 means predictions by the algorithm were 10% (=10 points) off on average.
- Predicting partitioned personality traits (binary classification with a Naïve Bayes algorithm). Following Celli et al. (2013), we performed a median split on each personality trait. We report classification accuracy, a measure representing the number of correct classifications divided by all classifications performed.
- Predicting author age (regression with an SVM algorithm). Performance metrics reported are MAE and MAPE.
- Predicting author gender (male or female; binary classification with a Naïve Bayes classifier). Again, we report classification accuracy.

Results

Descriptive Statistics

Mean age of the participants was 37 years ($SD = 12.73$; 63.75% female). The average word count for SOC writing was 120.51 words, and 120.62 for abusive writing, with no significant order effect found for word count. We observed differences between SOC and abusive writing (i.e., manipulation check) for 60 out of 89 LIWC categories (adjusted p value of 0.05/89 LIWC categories). Furthermore, the average number of abusive words in abusive writing was 4.03, with a mean of 2.05 in SOC writing,

representing a difference of $t(788) = 16.992$, $p < .001$, Cohen's $d = 0.60$. The order in which participants wrote texts did not affect the number of abusive words written in the abusive text, $t(781.88) = -1.67$, $p > .05$. Participants, who wrote the SOC essay after the abusive text, used somewhat more abusive words, $t(745.86) = 4.12$, $p < .001$, albeit with a small effect size $d = 0.29$.

Personality

Correlations

In Table 1, we present significant correlations between HEXACO and Dark Triad traits with LIWC2015 ($p < .000062$) and Grievance Dictionary ($p < .00025$) variables. Note that no significant correlations were found for honesty, agreeableness, conscientiousness, narcissism, and Machiavellianism with any of the traits and in neither of the writing conditions. In short, for SOC writing we found significant relationships for only three out of nine personality traits (i.e., emotionality, extraversion, and openness), and 11 out of 89 LIWC categories (i.e., personal pronouns, first-person singular, negative emotion, anxiety, tone, negation, cognitive processes, differentiation, seeing, leisure, commas) and 5 out of 22 Grievance Dictionary categories (i.e., desperation, grievance, loneliness, paranoia, and suicide). For abusive writing, we saw effects for four out of nine traits (i.e., emotionality, extraversion, openness, and psychopathy) with 7 out of 89 LIWC (i.e., function words, pronouns, verbs, cognitive processes, comma, sexual words, and informal language) categories and 3 out of 22 Grievance Dictionary categories (i.e., hate, murder, and violence). The effects ranged between $r = 0.14$ and $r = 0.24$ for SOC writing, and $r = 0.14$ to $r = 0.20$ for abusive writing.

Prediction

Next, we report personality prediction performance for SOC (Table 2) and abusive writing (Table 3). On average, honesty, emotionality, extraversion, agreeableness, conscientiousness, and openness (i.e., HEXACO traits) were predicted in SOC writing with an error margin of 9.62 points on a scale from 1 to 50 (MAPE = 19.24%), and 9.46 points for abusive writing (MAPE = 18.93%). The lowest average error in SOC writing was 7.60 points (MAPE = 15.20%)

Table 1
Correlations LIWC and Grievance Dictionary With Personality Traits

Stream of consciousness			Abusive writing		
Dictionary	Category	r (R^2)	Dictionary	Category	r (R^2)
Emotionality			Emotionality		
LIWC	Per. pronouns	0.19 (0.04)	LIWC	Function words	0.15 (0.02)
LIWC	1st pers. sing.	0.20 (0.04)	LIWC	Pronouns	0.17 (0.03)
LIWC	Neg. emotion	0.14 (0.02)	LIWC	Verbs	0.15 (0.02)
LIWC	Anxiety	0.18 (0.03)	Extraversion		
GD	Desperation	0.24 (0.06)	GD	Hate	0.14 (0.02)
GD	Grievance	0.14 (0.02)	Openness		
GD	Loneliness	0.15 (0.02)	LIWC	Verbs	-0.15 (0.02)
GD	Paranoia	0.15 (0.02)	LIWC	Cogn. processes	-0.15 (0.02)
GD	Suicide	0.14 (0.02)	LIWC	Comma	0.18 (0.03)
Extraversion			GD	Murder	0.20 (0.04)
LIWC	Tone	0.15 (0.02)	GD	Violence	0.17 (0.03)
LIWC	Negation	-0.15 (0.02)	Psychopathy		
LIWC	Cogn. processes	-0.16 (0.03)	LIWC	Sexual words	0.15 (0.02)
LIWC	Differentiation	-0.16 (0.03)		Informal language	0.15 (0.02)
LIWC	Seeing	0.14 (0.02)			
LIWC	Leisure	0.15 (0.02)			
Openness					
LIWC	Commas	0.19 (0.04)			

Note. LIWC = Linguistic Inquiry and Word Count; GD = Grievance dictionary.

for predicting conscientiousness, with equal performance using the baseline model, parts-of-speech, the Grievance Dictionary, the filtered feature set, or word embeddings. For abusive writing, this was the case for conscientiousness using the filtered feature set (average error 7.20 points, MAPE = 14.40%).

For Dark Triad predictions, the average error rate was 17.40 points on a scale of 1–63 (MAPE = 27.61%) for SOC writing and 17.07 points (MAPE = 27.10%) for abusive writing. The best performance in SOC writing was obtained

for predicting Machiavellianism, using either the baseline model or the Grievance Dictionary (MAPE = 17.70%). In abusive writing, Machiavellianism was best predicted using word embeddings (MAPE = 17.60%). Importantly, a baseline model using only number of words often outperformed other feature sets. In both conditions, n -grams, parts-of-speech, LIWC, the composite and filtered feature sets, and the BERT language model did not perform best for any of the traits.

We also performed binary classifications for each personality trait (based on median splits on

Table 2
SVM Prediction Performance for SOC Writing (Mean Absolute Percentage Error)

Model	HEXACO						Dark Triad		
	Hon.	Emot.	Extr.	Agr.	Consc.	Open.	Narc.	Mach.	Psych.
Baseline	18.9	16.3	21.3	18.8	15.2	17.4	29.6	17.7	30.3
n -grams	23.1	18.0	24.3	21.4	18.0	19.9	33.7	21.9	34.8
POS	19.1	16.4	21.6	18.8	15.2	17.3	30.5	18.0	29.7
LIWC	21.0	16.8	22.6	20.0	15.8	18.8	31.5	19.5	32.2
Grievance	19.0	16.2	21.3	18.8	15.3	17.1	29.6	17.7	30.2
Composite	23.8	22.3	25.1	23.1	20.2	22.5	38.2	23.6	37.2
Filtered	20.1	15.8	22.4	19.8	15.2	17.8	30.6	18.3	29.5
Embeddings	19.1	16.1	21.1	18.6	15.2	17.2	29.3	17.5	29.5
BERT	21.4	18.2	25.7	19.3	16.5	19.3	30.4	20.5	34.5

Note. POS = Parts-of-speech; LIWC = Linguistic Inquiry and Word Count; BERT = Bidirectional Encoder Representations from Transformers. Bold values represent the lowest MAPE across all traits and models.

Table 3*SVM Prediction Performance for Abusive Writing (Mean Absolute Percentage Error)*

Model	HEXACO						Dark Triad		
	Hon.	Emot.	Extr.	Agr.	Consc.	Open.	Narc.	Mach.	Psych.
Baseline	19.0	16.3	21.3	18.8	15.1	17.0	29.3	17.8	29.2
<i>n</i> -grams	21.1	18.2	27.0	19.7	17.3	19.9	32.9	21.2	32.2
POS	19.5	15.9	21.7	19.3	14.7	15.6	29.5	17.9	29.4
LIWC	19.9	16.6	23.2	19.0	16.3	17.1	31.2	18.9	30.7
Grievance	19.0	16.3	21.2	18.8	15.1	16.8	29.3	17.8	29.4
Composite	22.9	20.3	27.7	21.0	17.9	20.6	37.0	22.5	34.7
Filtered	19.8	16.2	22.6	19.7	14.4	17.1	30.8	19.4	31.1
Embeddings	19.0	16.0	21.3	18.5	15.1	16.2	29.2	17.6	27.7
BERT	19.79	19.44	22.96	19.60	17.58	19.22	34.50	19.89	30.73

Note. POS = Parts-of-speech; LIWC = Linguistic Inquiry and Word Count; BERT = Bidirectional Encoder Representations from Transformers. Bold values represent the lowest MAPE across all traits and models.

each trait), using the same features. In SOC writing (Table 4), the highest accuracy (0.62) was achieved for predicting openness (random baseline = 0.50) using BERT. For abusive writing (Table 5), the highest accuracies (0.62) were achieved in predicting openness using word embeddings. The baseline feature set was never the top performer in either prediction task.

Age

First, we tested for possible statistical relationships between age with LIWC and Grievance Dictionary categories. In both writing conditions, no significant effect of age or quadratic age (while controlling for gender) on any of the LIWC2015 categories was found (all $p > .00056$, α -level adjusted 89 LIWC categories) nor on any of the Grievance Dictionary categories ($p > .0028$, α -level adjusted for 22 Grievance Dictionary categories).

The results of the age prediction task are presented in Table 6, which shows that the different models predicted age with an average error of about 10 years. For the prediction of age in SOC writing, the best-performing model using the filtered feature set achieved a MAE of 9.15 years (MAPE = 24.61%). For abusive writing, best performance was achieved using word embeddings as features achieving a MAE of 10.01 years (MAPE = 27.04%).

Gender

We observed a significant multivariate effect of gender on LIWC2015 variables in SOC writing, Pillai's Trace = 0.30, $F(178, 1,398) = 1.37$, $p < .001$. Significant differences between genders ($p < .00056$), on individual LIWC categories were also found, where a positive Cohen's d value means the category was used more by women.

Table 4*Classification Results Stream-of-Consciousness Writing (Accuracy)*

Model	HEXACO						Dark Triad		
	Hon.	Emot.	Extr.	Agr.	Consc.	Open.	Narc.	Mach.	Psych.
Baseline	0.52	0.49	0.50	0.52	0.46	0.46	0.49	0.50	0.53
<i>n</i> -grams	0.54	0.46	0.58	0.48	0.52	0.52	0.49	0.49	0.49
POS	0.50	0.52	0.52	0.51	0.50	0.52	0.45	0.52	0.56
LIWC	0.57	0.48	0.56	0.62	0.50	0.53	0.47	0.52	0.55
Grievance	0.56	0.50	0.49	0.56	0.47	0.52	0.48	0.48	0.49
Composite	0.50	0.46	0.54	0.49	0.51	0.58	0.55	0.49	0.51
Filtered	0.54	0.55	0.54	0.50	0.51	0.52	0.55	0.52	0.50
Embeddings	0.58	0.51	0.55	0.55	0.48	0.60	0.41	0.58	0.48
BERT	0.52	0.52	0.49	0.61	0.58	0.63	0.50	0.51	0.46

Note. POS = Parts-of-speech; LIWC = Linguistic Inquiry and Word Count; BERT = Bidirectional Encoder Representations from Transformers. Bold values represent the highest accuracy across all traits and models.

Table 5
Classification Results Abusive Writing (Accuracy)

Model	HEXACO						Dark Triad		
	Hon.	Emot.	Extr.	Agr.	Consc.	Open.	Narc.	Mach.	Psych.
Baseline	0.52	0.54	0.54	0.52	0.52	0.56	0.52	0.53	0.49
<i>n</i> -grams	0.51	0.50	0.49	0.53	0.55	0.52	0.53	0.48	0.51
POS	0.53	0.54	0.53	0.52	0.52	0.57	0.48	0.48	0.52
LIWC	0.47	0.51	0.48	0.45	0.52	0.47	0.58	0.49	0.56
Grievance	0.52	0.49	0.52	0.52	0.52	0.61	0.51	0.51	0.51
Composite	0.55	0.50	0.48	0.54	0.56	0.60	0.53	0.48	0.47
Filtered	0.54	0.56	0.54	0.55	0.52	0.58	0.56	0.61	0.48
Embeddings	0.56	0.54	0.51	0.48	0.52	0.62	0.45	0.53	0.49
BERT	0.60	0.46	0.59	0.52	0.52	0.60	0.54	0.51	0.52

Note. POS = Parts-of-speech; LIWC = Linguistic Inquiry and Word Count; BERT = Bidirectional Encoder Representations from Transformers. Bold values represent the highest accuracy across all traits and models.

That is, men used more *analytical language* ($d = -0.34$), whereas women used more *pronouns* ($d = 0.27$), *personal pronouns* ($d = 0.30$), *first-person singular* ($d = 0.28$), *verbs* ($d = 0.35$), *discrepancies* ($d = 0.27$), *focus on the present* ($d = 0.26$), and *apostrophes* ($d = 0.28$). We also observed a significant multivariate effect of gender on all Grievance Dictionary categories, Pillai's Trace = 0.07, $F(22, 764) = 2.79, p < .001$. Significant differences between genders were found for the categories *desperation* ($d = 0.38$), *grievance* ($d = 0.24$), and *soldier* ($d = -0.30$).

For abusive writing we also found a multivariate effect on LIWC categories, Pillai's Trace = 0.32, $F(178, 1,398) = 1.47, p < .001$. Significant differences between genders ($p < .00056$) were found, with men using more *analytical language* ($d = -0.44$), *articles* ($d = -0.32$), and *sexual*

words ($d = -0.24$). In contrast, women used more *function words* ($d = 0.41$), *pronouns* ($d = 0.47$), *personal pronouns* ($d = 0.47$), *first-person singular* ($d = 0.31$), *auxiliary verbs* ($d = 0.33$), *verbs* ($d = 0.51$), *social words* ($d = 0.33$), *present focus words* ($d = 0.45$), and *apostrophes* ($d = 0.26$). We also observed a significant multivariate effect of gender on all Grievance Dictionary categories, Pillai's Trace = 0.07, $F(22, 764) = 2.79, p < .001$. Significant differences between genders were found for *desperation* ($d = 0.38$), *grievance* ($d = 0.24$), and *soldier* ($d = -0.30$).

Results for the gender classification task are presented in Table 7. For the prediction of gender in SOC writing, the highest accuracy of 0.64 was achieved using parts-of-speech as features. For abusive writing, best-performing prediction accuracy was 0.70, again using parts-of-speech. It must be noted that the proportion of males in the data set was 0.64; therefore, there is practically no improvement over a model which always predicts the majority class.

Table 6
Results Age Prediction (Mean Absolute Error)

Model	Stream-of-consciousness	Abusive writing
Baseline	10.10	10.23
<i>n</i> -grams	10.57	11.25
POS	9.29	10.04
LIWC	9.67	10.44
Grievance dictionary	10.21	10.22
Composite	11.11	12.28
Filtered	9.15	10.16
Embeddings	9.67	10.01
BERT	10.13	10.70

Note. POS = Parts-of-speech; LIWC = Linguistic Inquiry and Word Count; BERT = Bidirectional Encoder Representations from Transformers. Bold values represent the lowest mean absolute error across models.

Discussion

The present study examined the feasibility of author profiling through normal and abusive language, supplementing linguistic content with stylistic features of text. We looked at statistical relationships between linguistic variables and authors' personality traits and demographics (age, gender), and performed prediction experiments.

Statistical Relationships

First and foremost, some statistical relationships between (abusive) writing and author

Table 7
Results Gender Classification (Accuracy)

Model	Stream-of-consciousness Observed proportion of males: 0.64	Abusive writing
Baseline	0.62	0.59
<i>n</i> -grams	0.55	0.56
POS	0.64	0.70
LIWC	0.63	0.63
Grievance dictionary	0.54	0.54
Composite	0.58	0.63
Filtered	0.62	0.60
Embeddings	0.56	0.66
BERT	0.60	0.55

Note. POS = Parts-of-speech; LIWC = Linguistic Inquiry and Word Count; BERT = Bidirectional Encoder Representations from Transformers. Bold values represent the highest accuracy across models.

characteristics were observed. Language use in abusive texts was related to emotionality, openness, and psychopathy scores, whereas neutral writing showed relationships with emotionality, extraversion, and openness. We also observed gender differences in both types of text, but no significant effect of age on writing was found. Interestingly, our results seem to confirm that neutral and abusive writing are differently related to personality traits. Of particular interest is the fact that differences in language use based on differences in psychopathy can be measured in abusive writing, but did not emerge in neutral writing. Of further interest is the fact that differential gender differences emerged in abusive writing when compared to SOC writing with men using more sexual words (e.g., dick, whore, pervert), and women using more social words (e.g., mate, mother, together).

It is important to note that the majority of LIWC categories and personality traits did not seem to be related to abusive or neutral writing. We also observed fairly low correlations with personality traits, with an average of $r = 0.14$ for SOC writing, and $r = 0.12$ for abusive writing. These values are smaller than the average correlation of $r = 0.23$ (Hirsh & Peterson, 2009) or $r = 0.32$ (Azucar et al., 2018) found elsewhere. Results were also qualitatively different from previous research: We do not observe relationships between agreeableness and conscientiousness with any linguistic variable in either writing condition, whereas previous research reported such effects (Azucar et al., 2018; Qiu et al., 2012). These disparities are largely due to the

more stringent statistical criteria applied in the present study, but it can be argued that these corrections should have also been applied in previous studies in the first place. For instance, none of the correlations reported in Hirsh and Peterson (2009), a widely cited study on LIWC and personality traits, would have been considered statistically significant if corrections for the number of traits and LIWC categories had been performed⁴.

In some cases, the relationships that emerged between author traits and LIWC categories are seemingly straightforward to interpret. For example, it is perhaps not surprising that participants who scored higher on the trait Emotionality used more words from the emotional LIWC categories negative emotion and anxiety, as well as similar (negative) Grievance Dictionary categories such as desperation, grievance, loneliness, paranoia, and suicide. The positive correlation between Extraversion and “leisure” words could also have been anticipated, since it also replicates previous research (Nguyen et al., 2011). The result showing that individuals who scored higher on Psychopathy used more sexual words (in the abusive writing condition only) is interesting in light of previous research on the relationship between psychopathy and sexual deviance (Olver & Wong, 2006). For other relationships, particularly those with style categories, it is more difficult to explain why certain effects emerged (e.g., why higher openness was related to more use of commas or why high emotionality is related to more use of function words and pronouns). Of particular interest are the positive relationships between extraversion and hate, as well as those between high openness with murder and violence (all are Grievance Dictionary categories). These results suggest that extraverted and open individuals are more inclined to write more violent abuse (e.g., using words such as “bloodshed”, “fight”, “punch”). This effect has not previously been shown. However, it is important to replicate this study in future in order to test whether these relationships persist. This study served as an exploratory study assessing possible relationships with abusive writing. In future replication studies, direct hypotheses on these relationships can perhaps be tested.

⁴ The largest r in Hirsh and Peterson (2009) is 0.29 (for neuroticism and LIWC sadness), which equates to a $p = 0.0046$ (based on the reported $N = 94$), which is above the threshold of $p = 0.00026$ if corrections for five traits and 39 LIWC categories are applied.

It must also be noted that the small effects obtained in this study would only be of practical significance for the specific purpose of author profiling (e.g., to identify sources of threats), if the linguistic variables can also serve as features for predicting demographic traits. For example, when converting correlations for personality traits to explained variance (R^2), on average the significantly related LIWC categories would explain just 0.01% of the variance in each of the traits. This means that the vast majority of variance cannot be explained by the LIWC or Grievance Dictionary, and we must explore further explanatory variables. In the next section, we discuss our machine learning approach to author profiling.

Prediction Tasks

On average, the continuous prediction of personality traits was approximately 10% off in both neutral and abusive writing. Baseline models (using number of words) performed surprisingly well, whereas feature sets (such as the LIWC) that showed success in previous studies (Golbeck et al., 2011; Preotiuc-Pietro et al., 2016) performed poorly in the present study. When personality prediction was simplified into a binary classification task, accuracy was also markedly lower than in previous research (Argamon et al., 2005; Oberlander & Nowson, 2006). The statistical tests showed that the LIWC and Grievance Dictionary alone explain little variance in personality, and even when supplementing these measures with a mixture of additional variables (n -grams, parts-of-speech, embeddings, language models) we were not able to reach high regression or classification performance. Importantly, performance between writing conditions did not follow the same patterns, further illustrating the difference between abusive and neutral writing.

When predicting age, we observed an error margin of approximately 10 years in both conditions. This stands in stark contrast with previous research, which used the same or fewer features and achieved an error of 4 years (Nguyen et al., 2013), potentially because a larger amount of data (in terms of text and participants) was available. However, approximating someone's age based on their language to plus or minus 10 years may be helpful in a context where there is a wide range of possible ages.

Although we achieved an accuracy of gender classification of 70%, this is only marginally superior to a model which always predicts the majority class. Previous attempts achieved accuracy levels in the range of approximately 75% (with a 0.56 random baseline) to 92% (with a 0.55 random baseline) with similar feature sets as in the current work (Burger et al., 2011; Rangel et al., 2016). Again, even though we observed gender differences for various LIWC and Grievance Dictionary categories, these effects did not seem to transfer into high prediction performance.

There are several possible explanations for why the current results differ substantially from previous work on author profiling. First of all, our writing task involved instructed online writing, which is arguably different from handwritten SOC essays (Hirsh & Peterson, 2009; Pennebaker & King, 1999) or more natural, uninstructed social media posts on Twitter or Facebook (Azucar et al., 2018; Preotiuc-Pietro et al., 2016). In addition, the fact that participants were instructed to write abusive text when they normally may not be inclined to do so, may have lowered the external validity of the study. On the other hand, the highly anonymous nature of our task may have enabled some participants to be even more abusive than they would be in an online setting where messages can be traced back to a user profile. Finally, the number of words (120 on average) may have impacted on our ability to adequately predict author traits from language. Nevertheless, online writing is generally short in nature, and therefore testing the ability to make predictions on short texts seems especially relevant for applying these methods to online contexts.

Practical Significance

Whether the error rates for personality, age, and gender obtained in this study are problematic, is a matter of perspective. One could argue that a prediction of personality within 10% of the actual value is useful if a general profile of a text author is desired. The same holds for the prediction of age and gender. However, if such an author profiling system were deployed in a threat assessment or law enforcement context, where decisions based on such a system may have far-reaching consequences, these inaccuracies may be highly problematic. For example, an inaccurate profile may lead to the identification or arrest of an innocent individual, and vice versa, the true source of a threat may be missed. However, to

adequately evaluate the practical potential of an automatic system such as that utilized here, we would need to know what the “accuracy rates” of human judgment of author profiles are. If the accuracy of human judgment is lower or equivalent to an automatic system, the benefits of an automatic system (scalability, reliance on measurable features) may be preferable.

The results of this study illustrate another important point: Statistical significance does not equate to practical significance. Even though we observed significant statistical relationships between author demographics and language, these effects do not translate into accurate predictions, even when supplementing them with additional linguistic features. Increasingly, research focusing on violent individuals examines author characteristics through language, for example in terrorist manifestos and extremist forums (Akrami et al., 2018; Kop et al., 2019; Neuman et al., 2015). Oftentimes, these studies refer back to original research that has “established” a link between language and personality (Hirsh & Peterson, 2009; Pennebaker & King, 1999), assuming that this relationship generalizes to other types of language, such as that in violent or threatening texts.

The present study is the first to test this assumption in a context of abusive language, and found that these relationships are markedly different from neutral language, but of little importance in constructing accurate personality profiles. As such, our study suggests that the empirical body underpinning many studies on linguistic examinations of threats and terrorism may be weaker than how it is portrayed. While the present study demonstrates that such predictions are currently inaccurate for the type of (abusive) writing tasks performed here, further research is necessary to explore if indeed there are other conditions where predictions are more successful. One future avenue may include using nonlinguistic information (e.g., social media metadata) as additional features in prediction algorithms. Other author characteristics may also be considered for prediction, such as education level or language proficiency (e.g., whether English is the first language of the author). The focus on age and gender in this study is straightforward because of its relevance to (criminal) investigations, for example those involving threateners of public figures, whereas personality prediction was chosen due to its increased popularity in threat assessment

and offender profiling (Akrami et al., 2018; Neuman et al., 2015).

All in all, regardless of which author characteristics and language features are used, it remains important to realize that these predictions are highly complex. Therefore, it is crucial to consider the limitations (i.e., error margins) of these systems before they are implemented in practice.

Conclusion

The study was designed to test whether there are significant relationships between author personality, age, and gender and the way in which texts are written, with specific attention paid to abusive texts, particularly those directed at public figures. We then used linguistic features from the (abusive) texts to predict personality, age, and gender. Statistically significant relationships between author demographics and linguistic measures were found. For instance, individuals who scored high on extraversion and openness wrote more violently abusive texts. However, these statistical effects did not result in high prediction performance when compared to previous author profiling research. The results illustrate that statistical significance does not necessarily translate into practical significance. Therefore, we recommend that further research is conducted on author profiling in the threat assessment domain. In the meantime, we urge researchers and threat assessment practitioners to exercise caution in author profiling based on (abusive) language, specifically in contexts where potentially dangerous individuals are the subject of interest.

References

- Akrami, N., Shrestha, A., Berggren, M., Kaati, L., Obaidi, M., & Cohen, K. (2018). *Assessment of risk in written communication: Introducing the Profile Risk Assessment Tool (PRAT)*. EUROPOL. <https://www.europol.europa.eu/publications-documents/assessment-of-risk-in-written-communication>
- Argamon, S., Dhawle, S., Koppel, M., & Pennebaker, J. W. (2005, January). Lexical predictors of personality type. *Proceedings of joint annual meeting of the interface and the classification society of North America* (pp. 1–16). <https://doi.org/10.2105/AJPH.50.1.21>
- Ashton, M. C., & Lee, K. (2009). The HEXACO-60: A short measure of the major dimensions of personality. *Journal of Personality Assessment*, 91(4),

- 340–345. <https://www.tandfonline.com/doi/full/10.1080/00223890902935878>. <https://doi.org/10.1080/00223890902935878>
- Azucar, D., Marengo, D., & Settanni, M. (2018). Predicting the Big 5 personality traits from digital footprints on social media: A meta-analysis. *Personality and Individual Differences, 124*, 150–159. <https://doi.org/10.1016/j.paid.2017.12.018>
- Benoit, K., & Matsuo, A. (2020). *spacyr: An R wrapper for spaCy*.
- Bjørger, T., & Silkoset, E. (2018). *Threats and threatening approaches to politicians: A survey of Norwegian parliamentarians and cabinet ministers* (p. 56). Politihøgskolen.
- Burger, J. D., Henderson, J., Kim, G., & Zarrella, G. (2011). *Discriminating gender on Twitter* (p. 9). The MITRE Corporation.
- Celli, F., Pianesi, F., Stillwell, D., & Kosinski, M. (2013). Workshop on computational personality recognition: Shared task. *Proceedings of the International AAAI Conference on Web and Social Media, 7*(1), Article 4.
- Chouldechova, A., & Hastie, T. (2015). *Generalized additive model selection*. ArXiv:1506.03850 [Stat]. <http://arxiv.org/abs/1506.03850>
- Cobain, I., & Taylor, M. (2016, November 23). Far-right terrorist Thomas Mair jailed for life for Jo Cox murder. *The Guardian*. <https://www.theguardian.com/uk-news/2016/nov/23/thomas-mair-found-guilty-of-jo-cox-murder>
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Academic Press.
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of deep bidirectional transformers for language understanding*. ArXiv:1810.04805 [Cs]. <http://arxiv.org/abs/1810.04805>
- Dietz, P., & Martell, D. A. (2010). Commentary: Approaching and stalking public figures—a prerequisite to attack. *The Journal of the American Academy of Psychiatry and the Law, 38*, 341–348.
- Every-Palmer, S., Barry-Walsh, J., & Pathé, M. (2015). Harassment, stalking, threats and attacks targeting New Zealand politicians: A mental health issue. *The Australian and New Zealand Journal of Psychiatry, 49*(7), 634–641. <https://doi.org/10.1177/0004867415583700>
- Farnadi, G., Sushmita, S., Sitaraman, G., Ton, N., De Cock, M., & Davalos, S. (2014). *A multivariate regression approach to personality impression recognition of vloggers* [Conference session]. Proceedings of the 2014 ACM Multi Media on Workshop on Computational Personality Recognition - WCPR '14, Orlando, Florida, United States. <https://doi.org/10.1145/2659522.2659526>
- Figea, L., Kaati, L., & Scrivens, R. (2016). *Measuring online affects in a white supremacy forum* [Conference session]. IEEE International Conference on Intelligence and Security Informatics: Cybersecurity and Big Data, ISI 2016, Tucson, Arizona, United States. <https://doi.org/10.1109/ISI.2016.7745448>
- Golbeck, J., Robles, C., & Turner, K. (2011). *Predicting personality with social media* [Conference session]. Proceedings of the 2011 annual conference extended abstracts on human factors in computing systems - CHI EA '11, Vancouver, BC, Canada. <https://doi.org/10.1145/1979742.1979614>
- Haslam, N., Holland, E., & Kuppens, P. (2012). Categories versus dimensions in personality and psychopathology: A quantitative review of taxometric research. *Psychological Medicine, 42*(5), 903–920. <https://doi.org/10.1017/S0033291711001966>
- Hirsh, J. B., & Peterson, J. B. (2009). Personality and language use in self-narratives. *Journal of Research in Personality, 43*(3), 524–527. <https://doi.org/10.1016/j.jrp.2009.01.006>
- Inter-Parliamentary Union. (2016). *Sexism, harassment and violence against women parliamentarians*.
- James, D. V., Kerrigan, T. R., Forfar, R., Farnham, F. R., Preston, F. (2010). The fixated threat assessment centre: Preventing harm and facilitating care. *Journal of Forensic Psychiatry & Psychology, 21*(4), 521–536. <https://doi.org/10.1080/14789941003596981>
- James, D. V., Mullen, P. E., Meloy, J. R., Pathé, M. T., Farnham, F. R., Preston, L., & Darnley, B. (2007). The role of mental disorder in attacks on European politicians 1990–2004. *Acta Psychiatrica Scandinavica, 116*(5), 334–344. <https://doi.org/10.1111/j.1600-0447.2007.01077.x>
- James, D. V., Mullen, P. E., Pathé, M. T., Meloy, J. R., Preston, L. F., Darnley, B., & Farnham, F. R. (2009). Stalkers and harassers of royalty: The role of mental illness and motivation. *Psychological Medicine, 39*(9), 1479–1490. <https://doi.org/10.1017/S0033291709005443>
- James, D. V., Sukhwil, S., Farnham, F. R., Evans, J., Barrie, C., Taylor, A., & Wilson, S. P. (2016). Harassment and stalking of members of the United Kingdom Parliament: Associations and consequences. *Journal of Forensic Psychiatry & Psychology, 27*(3), 309–330. <https://doi.org/10.1080/14789949.2015.1124909>
- Jones, D. N., & Paulhus, D. L. (2014). Introducing the short Dark Triad (SD3): A brief measure of dark personality traits. *Assessment, 21*(1), 28–41. <https://doi.org/10.1177/1073191113514105>
- Kleinberg, B., van der Vegt, I., & Gill, P. (2021). The temporal evolution of a far-right forum. *Journal of Computational Social Science, 4*(1), 1–23. <https://doi.org/10.1007/s42001-020-00064-x>
- Kop, M., Read, P., & Walker, B. R. (2019). Pseudo-commando mass murderers: A big five personality profile using psycholinguistics. *Current Psychology*. Advance online publication. <https://doi.org/10.1007/s12144-019-00230-z>

- Kuhn, M. (2010). *Variable Selection Using The caret Package*.
- Lelkes, Y., Sood, G., & Iyengar, S. (2017). The Hostile audience: The effect of access to broadband internet on Partisan affect. *American Journal of Political Science*, 61(1), 5–20. <https://doi.org/10.1111/ajps.12237>
- Neuman, Y., Assaf, D., Cohen, Y., & Knoll, J. L. (2015). Profiling school shooters: Automatic text-based analysis. *Frontiers in Psychiatry*, 6(86), Article 86. <https://doi.org/10.3389/fpsy.2015.00086>
- Newman, M. L., Groom, C. J., Handelman, L. D., & Pennebaker, J. W. (2008). Gender differences in language use: An analysis of 14,000 text samples. *Discourse Processes*, 45(3), 211–236. <https://doi.org/10.1080/01638530802073712>
- Nguyen, D., Gravel, R., Trieschnigg, D., & Meder, T. (2013). *How old do you think i am?: A study of language and age in Twitter*, 10.
- Nguyen, T., Phung, D., Adams, B., & Venkatesh, S. (2011, July 5). *Towards discovery of influence and personality traits through social link prediction*. Fifth International AAAI Conference on Weblogs and Social Media. Fifth International AAAI Conference on Weblogs and Social Media. <https://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2772>
- Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., & Chang, Y. (2016). Abusive language detection in online user content. *Proceedings of the 25th international conference on world wide web* (pp. 145–153). ACM. <https://doi.org/10.1145/2872427.2883062>
- Oberlander, J., & Nowson, S. (2006, July). *Whose thumb is it anyway?: Classifying author personality from weblog text* [Conference session]. Proceedings of the COLING/ACL on Main Conference Poster Sessions, Sydney, Australia. <https://doi.org/10.1177/0266382105060607>
- Olver, M. E., & Wong, S. C. P. (2006). Psychopathy, sexual deviance, and recidivism among sex offenders. *Sexual Abuse*, 18(1), 65–82. <https://doi.org/10.1177/107906320601800105>
- Pathé, M. T., Lowry, T., Haworth, D. J., Webster, D. M., Mulder, M. J., Winterbourne, P., & Briggs, C. J. (2015). Assessing and managing the threat posed by fixated persons in Australia. *Journal of Forensic Psychiatry & Psychology*, 26(4), 425–438. <https://doi.org/10.1080/14789949.2015.1037332>
- Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of LIWC2015*. The University of Texas at Austin. <https://repositories.lib.utexas.edu/handle/2152/31333>
- Pennebaker, J. W., & King, L. A. (1999). Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology*, 77(6), 1296–1312. <https://doi.org/10.1037/0022-3514.77.6.1296>
- Pennebaker, J. W., & Stone, L. D. (2003). Words of wisdom: Language use over the life span. *Journal of Personality and Social Psychology*, 85(2), 291–301. <https://doi.org/10.1037/0022-3514.85.2.291>
- Pennington, J., Socher, R., & Manning, C. (2014). *Glove: Global vectors for word representation* [Conference session]. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar. <https://doi.org/10.3115/v1/D14-1162>
- Perraudin, F., & Murphy, S. (2019, October 31). Alarm over number of female MPs stepping down after abuse. *The Guardian*. <https://www.theguardian.com/politics/2019/oct/31/alarm-over-number-female-mps-stepping-down-after-abuse>
- Preotiuc-Pietro, D., Carpenter, J., Giorgi, S., & Ungar, L. (2016). *Studying the Dark Triad of personality through Twitter behavior* [Conference session]. Proceedings of the 25th ACM International on Conference on Information and Knowledge Management - CIKM '16, Indianapolis, Indiana, United States. <https://doi.org/10.1145/2983323.2983822>
- Qiu, L., Lin, H., Ramsay, J., & Yang, F. (2012). You are what you tweet: Personality expression and perception on Twitter. *Journal of Research in Personality*, 46(6), 710–718. <https://doi.org/10.1016/j.jrp.2012.08.008>
- Rangel, F., Rosso, P., Verhoeven, B., Daelemans, W., Potthast, M., & Stein, B. (2016). *Overview of the 4th author profiling task at PAN 2016: cross-genre evaluations*, 35.
- Rowe, I. (2015). Civility 2.0: A comparative analysis of incivility in online political discussion. *Information Communication and Society*, 18(2), 121–138. <https://doi.org/10.1080/1369118X.2014.940365>
- Scalora, M. J., Baumgartner, J. V., Zimmerman, W., Callaway, D., Hatch Maillette, M. A., Covell, C. N., Palarea, R. E., Krebs, J. A., & Washington, D. O. (2002). An epidemiological assessment of problematic contacts to members of Congress. *Journal of Forensic Sciences*, 47(6), 1360–1364. <https://doi.org/10.1520/JFS15575J>
- Scrivens, R., Davies, G., & Frank, R. (2018). Searching for signs of extremism on the web: An introduction to sentiment-based identification of radical authors. *Behavioral Sciences of Terrorism and Political Aggression*, 10(1), 39–59. <https://doi.org/10.1080/19434472.2016.1276612>
- Sumner, C., Byers, A., Boochever, R., & Park, G. J. (2012). *Predicting Dark Triad personality traits from Twitter usage and a linguistic analysis of tweets* [Conference session]. 11th International Conference on Machine Learning and Applications, Boca Raton, Florida, United States. <https://doi.org/10.1109/ICMLA.2012.218>
- van der Vegt, I., Mozes, M., Kleinberg, B., & Gill, P. (2021). *The Grievance Dictionary: Understanding*

- threatening language use. *Behavior Research Methods*, 53, 2105–2119. <https://doi.org/10.3758/s13428-021-01536-2>
- Ward, S., & McLoughlin, L. (2020). Turds, traitors and tossers: The abuse of UK MPs via Twitter. *Journal of Legislative Studies*, 26(1), 47–73. <https://doi.org/10.1080/13572334.2020.1730502>
- Waseem, Z., & Hovy, D. (2016). Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. *Proceedings of the NAACL student research workshop* (pp. 88–93). Association for Computational Linguistics. <https://doi.org/10.18653/v1/N16-2013>

Received December 16, 2020

Revision received November 11, 2021

Accepted November 19, 2021 ■

Correction to van der Vegt et al. (2022)

In the article “Assessment Procedures in Anonymously Written Threats of Harm and Violence” by Isabelle van der Vegt, Pippa Gregory, Bram B. van der Meer, Junyi Yang, Bennett Kleinberg, and Paul Gill (*Journal of Threat Assessment and Management*. Advance online publication. January 27, 2022. <https://doi.org/10.1037/tam0000168>), the funding statement was missing from the author note. All versions of this article have been corrected.

<https://doi.org/10.1037/tam0000176>