Personality and Individual Differences 141 (2019) 252–257

Contents lists available at ScienceDirect

Personality and Individual Differences

journal homepage: www.elsevier.com/locate/paid

# Cyberbullying detection on twitter using Big Five and Dark Triad features

Vimala Balakrishnan[a,*], Shahzaib Khan[a], Terence Fernandez[b], Hamid R. Arabnia[c]

[a] Faculty of Computer Science and Information Technology, University of Malaya, 50603 Kuala Lumpur, Malaysia
[b] Spatial Media LLC, (Concise Version), GA, United States of America
[c] Department of Computer Science, Franklin College of Arts and Sciences, University of Georgia, Athens, GA, United States of America

ARTICLE INFO

*Keywords:*
Cyberbullying
Personality
Dark triad
Big five
Twitter
Random Forest

ABSTRACT

This paper presents a cyberbullying detection model based on user personality, determined by the Big Five and Dark Triad models. The model aims to recognize bullying patterns among Twitter communities, based on relationships between personality traits and cyberbullying. Random Forest, a well-known machine-learning algorithm was used for cyberbullying classification (i.e. aggressor, spammer, bully and normal), applied in conjunction with a baseline algorithm encompassing seven Twitter features (i.e. number of mentions, number of followers and following, popularity, favorite count, status count and number of hash tags). Findings indicate that factoring user's personality greatly improves cyberbullying detection mechanisms. Specifically, extraversion, agreeableness and neuroticism (Big Five), and psychopathy (Dark Triad) were found to be significant in detecting bullies, achieving up to 96% (precision) and 95% (recall). The emergence of significant personality traits in an experimental study supports existing empirical studies that show the relationships between personality traits and cyberbullying.

## 1. Introduction

Cyberbullying refers to "*any behavior performed through electronic media by individuals or groups of individuals that repeatedly communicates hostile or aggressive messages intended to inflict harm or discomfort on others*" (Tokunaga, 2010, p. 278). According to Willard (2007), cyberbullying may include flaming (a brief online fight using profanities and hostile languages), harassing (repeatedly sending offensive messages to someone), slandering (spreading malicious rumors), masquerading (pretending to be someone else) and exclusion (intentionally excluding a person from an online group).

### 1.1. Cyberbullying roles

Cyberbullying often involves multiple parties, most prominent being the bullies, victims, bully-victims and bystanders. Bullies perpetrate a bullying incident; victims are the ones bullied, and bystanders are witnesses to a bullying/victimization incident online. Bully-victim is a vicious cycle in which a victim turns into a bully, and vice-versa. There is a consensus across studies that cyberbullies fit the profile of being aggressive, manipulative, and exploitive, while victims are often associated with low self-esteem (Resett & Gamez-Guadix, 2017). Some studies differentiate between bullies and aggressors, whereby the latter refers to someone who engages in an offensive behavior once (Chatzakou et al., 2017).

The present study distinguishes four user roles from a Twitter study (Chatzakou et al., 2017), namely, bully (someone who posts multiple tweets/re-tweets with negative intentions, generally for the same topic in a repeated fashion), aggressor (someone who posts at least one tweet/re-tweet with a negative intention), spammer (someone who posts texts of advertising, marketing or other suspicious nature) and normal (someone who does not belong to any of the other roles).

### 1.2. Automatic cyberbullying detection

Considering detrimental consequences of cyberbullying ranging from psychological to emotional and physical harm (van Geel, Goemans, Toprak, & Vedder, 2017), various prevention and intervention strategies were recommended, including mechanisms on automatically detecting cyberbullying (Al-garadi, Varathan, & Ravana, 2016; Chatzakou et al., 2017). Automatic cyberbullying detection refers to identifying bullying patterns based on users' textual communication online, and is often addressed as a classification problem (bullying vs. non-bullying) (Salawu, He, & Lumsden, 2017).

Thus far, the majority of cyberbullying detection studies have used features such as content (abusive/cyberbully words), platform-features

* Corresponding author.
*E-mail addresses:* vimala.balakrishnan@um.edu.my (V. Balakrishnan), shahzaib198@gmail.com (S. Khan), conciseversion@gmail.com (T. Fernandez), hra@cs.uga.edu (H.R. Arabnia).

https://doi.org/10.1016/j.paid.2019.01.024
Received 1 September 2018; Received in revised form 12 January 2019; Accepted 15 January 2019
Available online 22 January 2019
0191-8869/ 

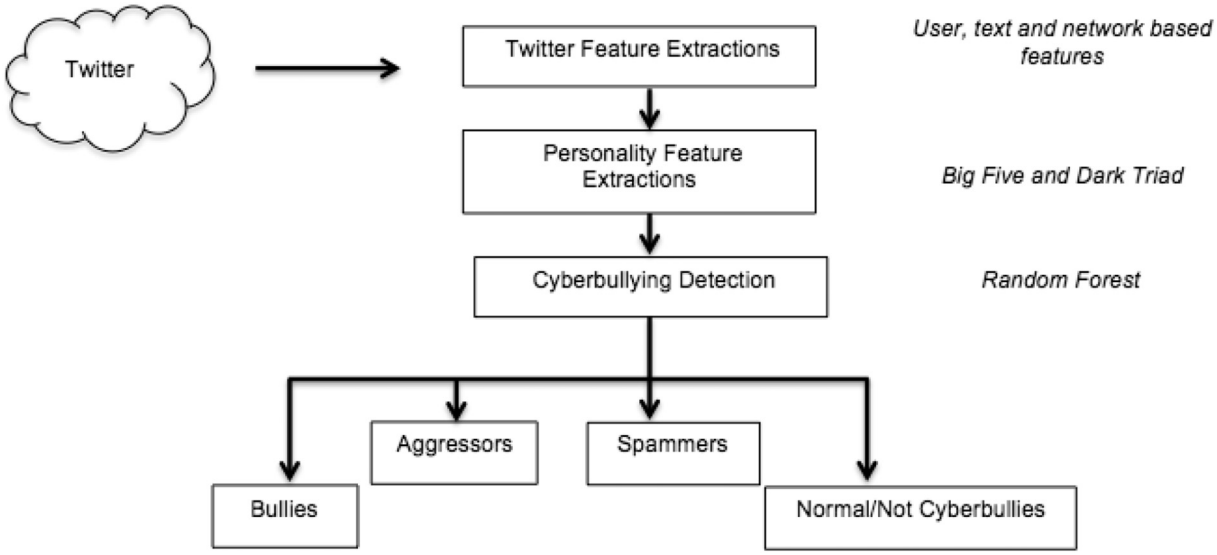


**Fig. 1.** The personality-based cyberbullying detection framework.

like time online and frequency of activities, and personal details such as gender and age (Al-garadi et al., 2016; Saravanaraj, Sheeba, & Devaneyan, 2016). Twitter-based studies found cyberbullying detection to improve when features such as number of tweets, user mentions, and number of followers-following were considered (Al-garadi et al., 2016; Chatzakou et al., 2017; Chen, Zhang, Chen, Xiang, & Zhou, 2015; Saravanaraj et al., 2016). The studies generally found bullies to use more hash tags and user mentions, post more tweets and have fewer friends. Power difference (calculated using user mentions and popularity) was noted between a bully and normal user, suggesting cyberbullying to be more impactful when the perpetrator is popular (Chatzakou et al., 2017). Therefore, the present study incorporates some of these Twitter features as the baseline model.

### 1.3. User personalities

Recently, cyberbullying studies have begun to explore the roles of user personalities on cyberbullying perpetration, with the majority focusing on the Big Five and Dark Triad models. Big Five, which encompasses five basic dimensions is among the most comprehensive and popular method to determine personality (McCrae & John, 1992). People who score high on *agreeableness* (kind, friendly, trusting and trustworthy) and *conscientiousness* (responsible, hardworking and diligent) generally avoid badmouthing (Stoughton, Thompson, & Meade, 2013) and attention seeking (Seidman, 2013). On the other hand, individuals who are *extraverted* (outgoing, gregarious and sociable), *neurotic* (depressed, fearful and anxious) and *open* (creative, perceptive, and thoughtful) tend to use technology more frequently to communicate (Marshall, Lefringhausen, & Ferenczi, 2015). Cyberbullying research using Big Five found agreeableness and conscientiousness relate negatively to cyberbullying perpetrations (Festl & Quandt, 2013; van Geel et al., 2017), whereas extraversion and neuroticism are positively correlated (Corcoran, Connolly, & O'Moore, 2012; Festl & Quandt, 2013). In comparing between cyberbullies and traditional bullies, Resett and Gamez-Guadix (2017) found cyberbullies scored low in neuroticism and high in agreeableness.

Studies exploring darker traits of humans have focused on the Dark Triad model, which has three distinct, yet undesirable personality traits, namely, Machiavellianism (lacking empathy, engaging in impulsive and thrill-seeking behaviors), narcissism (feeling superior, grandiose, and entitled), and psychopathy (strategically manipulating others, callous and fearless) (Paulhus & Williams, 2002). Empirical studies found all three traits to be positively related to cyberbullying behavior, with psychopathy emerging as the strongest predictor (Gibb & Devereux, 2014; Goodboy & Martin, 2015). Ang, Tan, and Mansor (2011) particularly found narcissism to be linked with cyberbullying whereas Pabian, De Backer, and Vandebosch (2015) found psychopathy relating to cyber-aggression.

Despite these evidences, to the best of our knowledge, no studies have incorporated users' personalities in automatically detecting cyberbullying (Salawu et al., 2017). The extant of the literature revealed the empirical findings from survey-based studies, attempting to examine relationships between personality traits and cyberbullying perpetration. Our study differs, in the sense that the aim is to investigate whether users' personalities can be used collectively and individually to automatically detect cyberbullying based on their textual communication using artificial intelligence (i.e. Random Forest).

Social media platforms such as Twitter and Facebook have become an integral part of human lives, where posts and comments reveal personal information, including insights into user personalities (Pratama & Sarno, 2015). Therefore, user personalities analyzed in conjunction with other features such as platform-features, emotions etc. can help improve cyberbullying detection mechanisms, particularly on social media where most cyberbullying takes place. For instance, Twitter is listed among the top five platforms where the maximum percentage of users experience cyberbullying (Ditch the Label, 2017), along with Instagram and Facebook. By applying artificial intelligence that manipulate users' personalities, existing strategies such as timeout that ban users from using abusive language on Twitter can be further improved.

To fill the above-mentioned gaps, the study devised a cyberbullying detection model based on users' personalities determined by the Big Five and Dark Triad models, since most empirical findings on cyberbullying were based on these (Festl & Quandt, 2013; Goodboy & Martin, 2015; van Geel et al., 2017). Twitter is the social media platform of choice for the study. Knowledge of each user traits can help distinguish between individuals with tendencies to engage in cyberbullying and those who don't, will enable a more effective detection mechanism as opposed to identifying them solely based on the use of abusive words, or platform-features.

## 2. Methodology

Fig. 1 illustrates an overview of the cyberbullying detection framework, including the feature extraction and cyberbullying detection modules, along with their roles. Table 1 provides operational

**Table 1**
Operational definition for Twitter and personality features.

| Features | Definition | Source |
|---|---|---|
| Number of user mentions | The number of times other specific users (e.g. @Bob) are mentioned, e.g. 10 | Twitter |
| Number of hash tags | Frequency of hash tags (#bully, #joy etc.) used, e.g. 10 | Twitter Developer (2018) |
| Number of followers | The number of followers a user has, e.g. 100 | |
| Number of following/friends | The number of users an individual is following, e.g. 100 | |
| Popularity | The ratio of following-followers | |
| Favorite count | The number of tweets a user has liked, e.g. 100 | |
| Status count | The number of tweets and re-tweets a user has issued, e.g. 100 | |
| Extraversion | Tendency to be outgoing, gregarious, assertive, active, and interested in other people | Big Five Traits - McCrae and John (1992) |
| Openness | Tendency to be creative, perceptive, thoughtful, and curious | |
| Conscientiousness | Tendency to be responsible, hardworking and organized | |
| Agreeableness | Tendency to be kind, friendly, gentle, trusting and trustworthy | |
| Neuroticism | Tendency to be depressed, fearful, anxious and sensitive to threats | |
| Machiavellianism | Tendency to lack empathy and engage in impulsive and thrill-seeking behaviour | Dark Triad |
| Psychopathy | Tendency to strategically manipulate others, callous, unemotional and fearless | Paulhus and Williams (2002) |
| Narcissism | Tendency to feel superior, grandiose and entitled | |

definitions for all extracted features.

### 2.1. Cyberbullying dataset

The cyberbullying dataset was obtained from Chatzakou et al. (2017), along with the manually annotated data. The data were crawled from Twitter, using #GamerGate as the hash tag, containing tweets and other metadata such as creation time, followers, friends etc. Gamergate (i.e. video game) is a well-documented large-scale repository of bullying/aggressive behavior instances, including extreme cases such as threats of murder and rape (Massanari, 2015). The dataset contains a total of 9484 tweets, labeled as spammers (N = 3208; % = 33.8), aggressors (N = 336; % = 3.5), bullies (N = 528; % = 5.6) and normal (N = 5608; % = 59%).

### 2.2. Feature extractions

Seven basic Twitter features as defined in Table 1 were extracted using Twitter streaming API, and used as the baseline algorithm. These were identified based on previous cyberbullying detection studies (Al-garadi et al., 2016; Chatzakou et al., 2017).

The Big Five personality traits were determined using IBM Watson's Personality Insights API, a tool that incorporates linguistic and data analytics to predict an individual's personality. The API predicts each user's personality trait by examining tweets against the user. It provides the value for each of the five dominant dimensions of Big Five, along with six facets (sub-traits) that further characterize an individual according to the dimension (IBM, 2018). These values are within a range of 0–1, with a higher score indicating a higher inclination toward the trait. For instance, the personality score for User A may be: 0.56 (Extraversion), 0.89 (Neuroticism), 0.11 (Openness) and so on. Looking at the highest score User A will be categorized as neurotic. The Personality Insights API is an established tool, tested and freely provided by IBM Watson, and used in other scholarly studies (Mostafa, Crick, Calderon, & Oatley, 2016; Paruma-Pabón, González, Aponte, Camargo, & Restrepo-Calle, 2016).

To further incorporate the darker traits into the detection model, the study examined the relationships between Big Five and Dark Triad. Several empirical studies based on non-clinical measures have examined the similarities and differences between Big Five and Dark Triad models (Douglas, Bore, & Munro, 2012; Jakobwitz & Egan, 2006; Paulhus & Williams, 2002), with the majority showing most dark traits to be in negative associations with agreeableness and conscientiousness (Table 2).

Based on the relationships in Table 2, a condensed version was proposed as depicted in Table 3. The mapping between the models was done based on the majority of the occurrences, for example, most of the studies found positive correlations between extraversion and narcissism (Ardic & Ozsoy, 2016; Douglas et al., 2012; Paulhus & Williams, 2002), hence a positive correlation was marked for the respective traits (Table 3).

Mathematically, the Dark Triad classification for the personality was based on a cut-off point of 0.5 for Big Five (Big Five scores range between 0 and 1). Therefore, User A with a score of 0.89 (i.e. > 0.5, neuroticism), will be classified as Psychopathy (as per Table 3), with a positive correlation.

### 2.3. Cyberbullying detection

The cyberbullying detection was performed using Random Forest, a machine learning technique involving the generation of multiple decision trees, each providing a unique classifier. Random Forest was selected due to its popularity in cyberbullying detection studies (Al-garadi et al., 2016; Chatzakou et al., 2017; Saravanaraj et al., 2016). The final user classification was determined based on the majority vote (Saravanaraj et al., 2016).

### 2.4. Evaluation and experiment

#### 2.4.1. Evaluation metrics

Standard metrics, namely, precision (false positives), recall (false negatives), and F-measure were used to evaluate the effectiveness of the cyberbullying detection model. Precision is the ratio between the true positive (correct predictions) and the total predictions (Powers, 2011). By way of an example, if a model classifies ten items, out of which one of it is wrong, then the precision of the model will be 9/10 (90%). Mathematically, it can be defined as:

$$P = \frac{T_p}{T_p + F_p}$$

where $P$ = precision, $T_p$ = true positive, $F_p$ = false positive.

Recall is the ratio of the correct predictions and the total number of correct items in the set (Powers, 2011). For example, if a model classifies ten items, out of which 5 items were correct, the recall is 5/10 (50%). Mathematically, it can be defined as:

$$R = \frac{T_p}{T_p + F_n}$$

where R = recall, $T_p$ = true positive, $F_n$ = false negative.

The F-measure is the weighted harmonic mean of precision and recall, and is defined as:

$$F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

**Table 2**
Correlation between Dark Triad and the Big Five model.

| | | Big Five personality traits | | | | | Source |
|---|---|---|---|---|---|---|---|
| | | Openness | Conscientiousness | Extraversion | Agreeableness | Neuroticism | |
| Dark Triad | M | | | | | + | Ardic and Ozsoy (2016) |
| | N | + | | + | − | + | |
| | P | | | | − | + | |
| | M | | − | | − | | Paulhus and Williams (2002) |
| | N | + | | + | − | | |
| | P | + | − | + | − | − | |
| | M | | − | | − | + | Jakobwitz and Egan (2006) |
| | N | | − | | − | | |
| | P-Psy | | | | − | + | |
| | S-Psy | | − | | − | + | |
| | M | | − | | − | + | Douglas et al. (2012) |
| | N | | − | + | − | | |
| | P-Psy | − | − | | − | | |
| | S-Psy | | − | − | − | + | |

M: Machiavellianism, N: Narcissism, P: Psychopathy, P-Psy: Primary psychopathy, S-Psy: Secondary psychopathy, P-Psy and S-Psy are the two dimensions of P. + − positive correlation; −: negative correlation.

All three metrics are very useful in measuring the success of predictions especially when the classes are very imbalanced, as in the case of the current study where there is a wide distribution among the different user roles. Hence, all the aforementioned metrics were used to measure the performance of the detection model.

#### 2.4.2. Experiment setups

Experiments were conducted in several setups to determine the effectiveness of the model. These were executed using Weka® 3.8, an open source software, using the following setups:

- *Baseline:* detection model with the seven Twitter features
- *Baseline + Big Five + Dark Triad:* baseline model with all the personality traits
- *Baseline + Trait:* baseline model with each individual trait from Big Five and Dark Triad (e.g. Baseline + Openness)
- *Baseline + Key Traits:* baseline model with the significant personality traits identified from Baseline + Trait

To determine if the detection effectiveness differed significantly with one another, *t*-test was administered using SPSS 24, with a significant alpha value of 0.05.

## 3. Results

Table 4 depicts the cyberbullying detection effectiveness compared to the baseline model, and among the individual traits from Big Five and Dark Triad models.

A higher precision, recall and F-measure for all the models compared to the baseline indicate that users' personalities can be effectively used to improve detecting bullying patterns online. Generally, *t*-tests revealed all the models to have significantly outperformed the baseline ($p < 0.001$), hence showing that user personalities improve cyberbullying detection.

**Table 4**
Cyberbullying detection effectiveness based on precision, recall and F-measure.

| Models | Precision | Recall | F-measure |
|---|---|---|---|
| Baseline | 0.806 | 0.807 | 0.806 |
| Baseline + Big 5 + Dark Triad | 0.915 | 0.916 | 0.916 |
| Baseline + Openness | 0.901 | 0.901 | 0.901 |
| Baseline + Conscientiousness | 0.901 | 0.901 | 0.901 |
| *Baseline + Extraversion*[a] | *0.911* | *0.911* | *0.911* |
| *Baseline + Agreeableness*[a] | *0.910* | *0.911* | *0.910* |
| *Baseline + Neuroticism*[a] | *0.911* | *0.912* | *0.911* |
| Baseline + Narcissism | 0.908 | 0.910 | 0.908 |
| Baseline + Machiavellianism | 0.905 | 0.906 | 0.905 |
| *Baseline + Psychopathy*[b] | *0.917* | *0.918* | *0.918* |

[a] Significantly different with Openness and Conscientiousness;
[b] Significantly different with Narcissism and Machiavellianism.

Further experiments with singular personality traits produced higher accuracies for extraversion, agreeableness, and neuroticism (Big Five) and psychopathy (Dark Triad). Significant differences were noted between extraversion, agreeableness and neuroticism, with openness and conscientiousness (i.e. $p < 0.05$). As for Dark Triad, psychopathy performed significantly better than Machiavellianism and narcissism ($p < 0.001$). This shows that the four specific traits significantly contribute in cyberbullying detection compared to the other traits. These traits were aggregated into a single model, and its performance was compared with the baseline (Table 5). Cyberbullying detection effectiveness was found to significantly improve when the key personalities were used compared to the baseline (i.e. $p = 0.001$), and the rest of the models in Table 4.

## 4. Discussion and conclusion

Cyberbullying is a societal issue on a global scale. Considering that social media is now an unavoidable part of daily life, and cyberbullying,

**Table 3**
Mapping between Big Five and Dark Triad.

| | | Big Five personality traits | | | | |
|---|---|---|---|---|---|---|
| | | Openness | Conscientiousness | Extraversion | Agreeableness | Neuroticism |
| Dark Triad | N | + | | + | | |
| | M | | − | | − | |
| | P | | | | | + |

M: Machiavellianism, N: Narcissism, P: Psychopathy.

**Table 5**
Cyberbullying detection effectiveness for baseline versus key personality traits – precision, recall and F-measure.

| Models | Precision | Recall | F-measure |
|---|---|---|---|
| Baseline | 0.806 | 0.807 | 0.806 |
| *Baseline + Key Traits*[a] | *0.960* | *0.952* | *0.929* |

[a] Big Five – Extraversion, Agreeableness & Neuroticism, Dark Triad – Psychopathy.

a phenomenon proven to be an endemic, automatic cyberbullying detection is not only timely, but crucially necessary. This study improved the effectiveness of cyberbullying detection mechanism (vs. model using only Twitter features) using users' personalities determined by the Big Five and Dark Triad models. Random Forest was used to classify users into one of four roles, namely, bully, aggressor, spammer or normal, based on 9484 tweets.

Significant improvements were apparent for cyberbullying detection when user personalities were applied, with a precision of 91.5%, and a recall of 91.6%. This is consistent with empirical studies that have shown links between an individual's personality with both cyberbullying (Resett & Gamez-Guadix, 2017; van Geel et al., 2017) and traditional bullying (Festl & Quandt, 2013). Personality is a mental function that distinguishes one person from another, and its ability in predicting consequential outcomes (e.g. psychological well-being) has resulted in many studies investigating and manipulating users' personalities, including those who automatically predict one's personality based on his/her online communication style. Social media users are known to be motivated in greater self-disclosure, and tend to freely express their opinions and views concerning their perceptions and concerns, and thus making platforms such as Twitter ideal to examine human psychological profiles, particularly their personalities to help detect anti-social behaviors such as cyberbullying. Therefore, with findings that support deeper analytics of users' personalities to improve cyberbullying detection, social media platforms can play a bigger role in the line of defense, such as applying improved artificial intelligence, and designing supporting digital tools in the fight against cyberbullying.

Cyberbullying detection improved when all the personalities were incorporated, however, significant differences were noted for extraversion, agreeableness, neuroticism (Big Five) and psychopathy (Dark Triad). The singular model, based solely on these key personalities, was found to have the highest effectiveness score for cyberbullying detection (i.e. precision = 96%; recall = 95%). Our findings echo previous empirical results in which extraversion, agreeableness and neuroticism showed significant relations to cyberbullying perpetrations (Festl & Quandt, 2013; Resett & Gamez-Guadix, 2017; van Geel et al., 2017). Extraverted people for example, have the tendency to communicate more frequently using social media tools (Marshall et al., 2015), hence, there is greater likelihood for them to engage in cyberbullying perpetration compared to those who score low on extraversion. In fact, extraverted people have been known to engage in cyberbullying perpetration to increase their social status (van Geel et al., 2017). Similarly, cyberbullying perpetration has been linked with individuals who scored high on neuroticism, and low on agreeableness (Mitsopoulou & Giovazolias, 2015; van Geel et al., 2017). Individuals with high agreeableness tend to be gentle, trusting and altruistic, which likely inhibits them from engaging in harmful behaviors such as cyberbullying.

The emergence of psychopathy as a significant predictor for cyberbullying, compared to Machiavellianism and narcissism is also in line with empirical studies (Gibb & Devereux, 2014:Goodboy & Martin, 2015; van Geel et al., 2017). Unlike Machiavellians who are more likely to harm others if the perceived benefits are high and the personal risk is low, and the narcissists who tend to harm others when their sense of self feels threatened, psychopaths are predatory, callous and fearless (Paulhus & Williams, 2002). The predatory nature of psychopaths for example, may drive these individuals to seek potential victims to inflict emotional and psychological harm, as they also have a complete disregard for the distress they cause others. This probably explains why the trait has been consistently found to predict anti-social behaviors, including cyberbullying (Goodboy & Martin, 2015; van Geel et al., 2017). Therefore, being able to determine one's personality online and using the said information to detect not only cyberbullying, but other unwanted behaviors such as trolling is an added layer in innovation progress toward mitigating any kind of harmful behavioral patterns online.

This study contributes to the body of evidence proving the relationship between user personalities and cyberbullying perpetration, particularly by showing that specific traits can be used effectively to detect online bullying patterns. The identification of the key personalities from Big Five and Dark Triad show that, although every one of these traits impact cyberbullying collectively, extraversion, agreeableness, neuroticism and psychopathy have bigger impacts on cyberbullying perpetration. In fact, integrated together, these four key personalities further improved the cyberbullying detection mechanism compared to the rest of the models. Being technological in nature, the detection algorithm can be adapted into existing online platforms such as Twitter, gaming websites or forums (e.g. Reddit) where massive textual communication takes place, including those that are abusive and offensive in nature. Alternatively, knowing individual traits based on their online communication styles provide an opportunity for a higher level of monitoring for those who score high on specific personalities, especially on the negative traits such as neuroticism and psychopathy. We note that although the idea is feasible, respective Internet organizations need to be cooperative in addressing the heinous act of cyberbullying. When applied in an educational setting such as communication platforms (forums, blogs) for universities and schools, early identification of negative personalities and cyberbullying detection may potentially help educators or counselors to focus on these individuals. For example, although individuals scoring high on psychopathic traits often require intensive professional treatment, an integrative and progressive approach targeting these individuals, consisting of education, technology and awareness that cyberbullying is detrimental and morally wrong may be deemed as useful practice in mitigating cyberbullying.

The study however, is not without its limitations. The dataset was limited to a specific community (i.e. GamerGate), with the majority of the tweets categorized as spam and normal. Although this is considered normal in any social media study, the low number of bullies compared to other roles could have affected the cyberbullying detection accuracy. Future studies can explore other (and probably bigger) datasets. Future studies could venture into detecting cyberbullying victims as well as bystanders. The latter role is especially important considering majority of studies have reported a huge number of bystanders playing major roles in either supporting or defending cyberbullying perpetrations (Balakrishnan, 2017). Finally, recent studies investigating anti-social behaviors found another dark trait, that is, sadism (tendency to take pleasure in the suffering of others) to predict bullying as a whole, and cyberbullying (van Geel et al., 2017). In such a case, studies may replicate the methodology of the present study by incorporating the Dark Tetrad (i.e. Dark Triad + sadism) model.

## Acknowledgement

The Fulbright Visiting Scholar Program 2018, awarded to the main author of this paper, supported this work. The authors would like to extend their gratitude to Mr. Despoina Chatzakou for sharing his cyberbullying dataset.

## References


Al-garadi, M. A., Varathan, K. D., & Ravana, S. D. (2016). Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network. *Computers in Human Behavior, 63*, 433–443.

Ang, R. P., Tan, K.-A., & Mansor, A. T. (2011). Normative beliefs about aggression as a mediator of narcissistic exploitativeness and cyberbullying. *Journal of Interpersonal Violence, 26*(13), 2619–2634.

Ardic, K., & Ozsoy, E. (2016). Examining the relationship between the dark triad traits and big five personality dimensions. *Proceedings of the fifth European academic research conference on global business, economics, finance and banking (EAR16Turkey Conference), Istanbul-Turkey*.

Balakrishnan, V. (2017). Unraveling the underlying factors SCulPT-ing cyberbullying behaviours among Malaysian young adults. *Computers in Human Behavior, 75*, 194–205.

Chatzakou, D., Kourtellis, N., Blackburn, J., De Cristofaro, E., Stringhini, G., & Vakali, A. (2017). Mean birds: Detecting aggression and bullying on Twitter. *Proceedings of the 2017 ACM on web science conference* (pp. 13–22). New York: USA.

Chen, C., Zhang, J., Chen, X., Xiang, Y., & Zhou, W. (2015). 6 million spam tweets: A large ground truth for timely Twitter spam detection. *Proceedings of the IEEE International Conference on Communications (ICC), London: UK, IEEE*.

Corcoran, L., Connolly, I., & O'Moore, M. (2012). Cyberbullying in Irish schools: An investigation of personality and self-concept. *The Irish Journal of Psychology, 33*(4), 153–165.

Douglas, H., Bore, M., & Munro, D. (2012). Distinguishing the dark triad: Evidence from the five-factor model and the Hogan development survey. *Psychology, 3*(3), 237–242.

Festl, R., & Quandt, T. (2013). Social relations and cyberbullying: The influence of individual and structural attributes on victimization and perpetration via the internet. *Human Communication Research, 39*(1), 101–126.

Gibb, Z. G., & Devereux, P. G. (2014). Who does that anyway? Predictors and personality correlates of cyberbullying in college. *Computers in Human Behavior, 38*, 8–16.

Goodboy, A. K., & Martin, M. M. (2015). The personality profile of a cyberbully: Examining the Dark Triad. *Computers in Human Behavior, 49*, 1–4.

Jakobwitz, S., & Egan, V. (2006). The dark triad and normal personality traits. *Personality and Individual Differences, 40*(2), 331–339.

Marshall, T. M., Lefringhausen, K., & Ferenczi, N. (2015). The Big Five, self-esteem, and narcissism as predictors of the topics people write about in Facebook status updates. *Personality and Individual Differences, 85*, 35–40.

Massanari, A. (2015). #Gamergate and the Fappening: How Reddit's algorithm, governance, and culture support toxic technocultures. *New Media & Society, 19*(3), 329–346.

McCrae, R. R., & John, O. P. (1992). An introduction to the five factor model and its applications. *Journal of Personality, 60*(2), 175–215.

Mitsopoulou, E., & Giovazolias, T. (2015). Personality traits, empathy and bullying behavior: A meta-analytic approach. *Aggression and Violent Behavior, 21*, 61–72.

Mostafa, M., Crick, T., Calderon, A. C., & Oatley, G. (2016). Incorporating emotion and personality-based analysis in user-centered modelling. In M. Bramer, & M. Petridis (Eds.). *Research and Development in Intelligent Systems XXXIII*. (pp. 383–389). Cham: Springer.

Pabian, S., De Backer, C. J., & Vandebosch, H. (2015). Dark Triad personality traits and adolescent cyber-aggression. *Personality and Individual Differences, 75*, 41–46.

Paruma-Pabón, O. H., González, F. A., Aponte, J., Camargo, J. E., & Restrepo-Calle, F. (2016). Finding relationships between socio-technical aspects and personality traits by mining developer e-mails. *Proceedings of the 9th international workshop on cooperative and human aspects of software engineering* (pp. 8–14). New York: ACM.

Paulhus, D. L., & Williams, K. M. (2002). The dark triad of personality: Narcissism, Machiavellianism, and psychopathy. *Journal of Research in Personality, 36*(6), 556–563.

Powers, D. M. W. (2011). Evaluation: From precision, recall and f-measure to roc., informedness, markedness & correlation. *Journal of Machine Learning Technologies, 2*(1), 37–63 (doi:citeulike-article-id:12882259).

Pratama, B. Y., & Sarno, R. (2015). *Personality classification based on Twitter text using Naive Bayes, KNN and SVM. Proceedings of the international conference on data and software engineering.* Indonesia: IEEE170–174.

Resett, S., & Gamez-Guadix, M. (2017). Traditional bullying and cyberbullying: Differences in emotional problems, and personality. Are cyberbullies more Machiavellians? *Journal of Adolescence, 61*, 113–116.

Salawu, S., He, Y., & Lumsden, J. (2017). Approaches to automated detection of Cyberbullying: A survey. *IEEE Transactions on Affective Computing* (Vol. early online, 10.10.2017, http://doi.ieeecomputersociety.org/10.1109/TAFFC.2017.2761757).

Saravanaraj, A., Sheeba, J., & Devaneyan, S. P. (2016). Automatic detection of Cyberbullying from Twitter. *International Journal of Computer Science and Information Technology & Security, 6*(6), 26–32.

Seidman, G. (2013). Self-presentation and belonging on Facebook: How personality influences social media use and motivations. *Personality and Individual Differences, 54*(3), 402–407.

Stoughton, J. W., Thompson, L. F., & Meade, A. W. (2013). Big five personality traits reflected in job applicants' social media postings. *Cyberpsychology, Behavior and Social Networking, 16*(11), 800–805.

Tokunaga, R. S. (2010). Following you home from school: A critical review and synthesis of research on cyberbullying victimization. *Computers in Human Behavior, 26*(3), 277–287.

van Geel, M., Goemans, A., Toprak, A., & Vedder, P. (2017). Which personality traits are related to traditional bullying and cyberbullying? A study with the Big Five, Dark Triad and sadism. *Personality and Individual Differences, 106*, 231–235.

Willard, N. E. (2007). *Cyberbullying and cyberthreats: Responding to the challenge of online social aggression, threats, and distress.* IL: Research Press.


## Web references


Ditch the Label (2017). The annual bullying survey 2017. Retrieved from https://www.ditchthelabel.org/wp-content/uploads/2017/07/The-Annual-Bullying-Survey-2017-1.pdf, Accessed date: 29 November 2018.

IBM (2018). Personality insights - API reference | IBM Watson Developer Cloud. Retrieved from https://www.ibm.com/watson/developercloud/personality-insights/api/v3/curl.html, Accessed date: 15 October 2018.

Twitter Developer (2018). Tweet objects. Retrieved from https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/user-object, Accessed date: 18 November 2018.