Improving Cyberbullying Detection using Twitter Users' Psychological Features and Machine Learning

Vimala Balakrishnan Ph.D , Shahzaib Khan ,
Hamid R. Arabnia Ph.D

PII: S0167-4048(19)30247-0
DOI: https://doi.org/10.1016/j.cose.2019.101710
Reference: COSE 101710

To appear in: *Computers & Security*

Received date: 9 April 2019
Revised date: 27 November 2019
Accepted date: 31 December 2019

Please cite this article as: Vimala Balakrishnan Ph.D , Shahzaib Khan , Hamid R. Arabnia Ph.D , Improving Cyberbullying Detection using Twitter Users' Psychological Features and Machine Learning, *Computers & Security* (2019), doi: https://doi.org/10.1016/j.cose.2019.101710

Improving Cyberbullying Detection using Twitter Users' Psychological Features and Machine Learning

First and corresponding author:
Vimala Balakrishnan, Ph.D
Faculty of Computer Science and Information Technology, University of Malaya, 50603 Kuala Lumpur, Malaysia
Tel: +6 03 7967 6377
Email: vimala.balakrishnan@um.edu.my

Dr Vimala Balakrishnan is a Senior Lecturer, and Data Scientist affiliated with the Faculty of Computer Science and Information Technology, University of Malaya since 2010. She obtained her PhD. in the field of Ergonomics from Multimedia University. Dr Balakrishnan's main research interests are in data analytics and sentiment analysis, particularly related to social media. She has published approximately 47 articles in top indexed journals and also serves as an Associate Editor to the Malaysian Journal of Computer Science, and as an Associate Member for the Global Science and Technology Forum. She is also a fellow for the Leadership in Innovation program, a prestigious award by the Royal Academy of Engineering, UK, and a Fulbright Visiting Scholar 2018.

Second author
Shahzaib Khan,
Faculty of Computer Science and Information Technology, University of Malaya, 50603 Kuala Lumpur, Malaysia
Email: shahzaib198@gmail.com

Shahzaib Khan is a post-graduate student at the Faculty of Computer Science and Information Technology, University of Malaya. He recently worked and completed his Masters study on a thesis entitled multiple-feature enhanced cyberbullying detection model, which the current paper is based on. At the moment, Mr Khan is working on a data science related proposal to pursue his PhD program.

Third author
Hamid R. Arabnia, Ph.D
Department of Computer Science,
Franklin College of Arts and Sciences
University of Georgia, Athens, GA.
Email: hra@cs.uga.edu

Hamid R. Arabnia received a Ph.D. degree in Computer Science from the University of Kent (Canterbury, England) in 1987. He is currently a Professor of Computer Science at University of Georgia (Georgia, USA), where he has been since October 1987. His research interests include parallel and distributed processing techniques and algorithms, supercomputing, Big Data Analytics (in the context of scalable HPC), imaging science (image processing, computer vision, and computer graphics), and other compute intensive problems. Prof. Arabnia is Editor-in-Chief of The Journal of Supercomputing (one of the

oldest journals in Computer Science) published by Springer. He is also on the editorial and advisory boards of 40 other journals. Prof. Arabnia has published extensively in journals and refereed conference proceedings. He has about 200 peer-reviewed research publications as well as 250 edited research books in his areas of expertise.

Abstract

Empirical evidences linking users' psychological features such as personality traits and cybercrimes such as cyberbullying are many. This study deals with automatic cyberbullying detection mechanism tapping into Twitter users' psychological features including personalities, sentiment and emotion. User personalities were determined using Big Five and Dark Triad models, whereas machine learning classifiers namely, Naïve Bayes, Random Forest and J48 were used to classify the tweets into one of four categories: bully, aggressor, spammer and normal. The Twitter dataset contained 5453 tweets gathered using the hashtag #Gamergate, and manually annotated by human experts. Selected Twitter-based features namely text, user and network-based features were used as the baseline algorithm. Results show that cyberbullying detection improved when personalities and sentiments were used, however, a similar effect was not observed for emotion. A further analysis on the personalities revealed extraversion, agreeableness, neuroticism and psychopathy to have greater impacts in detecting online bullying compared to other traits. Key features were identified using the dimension reduction technique, and integrated into a single model, which produced the best detection accuracy. The paper describes suggestions and recommendations as to how the findings can be applied to mitigate cyberbullying.

Keywords – cyberbullying, personality, emotion, sentiment, Twitter, machine learning

1. Introduction

Cyberbullying is a deliberate and repetitive act to harm or humiliate someone using information and communication technologies such as mobile phones, e-mails and social

media (Hinduja and Patchin, 2010; Kowalski et al., 2012; Balakrishnan, 2015). It is often categorized into various forms, such as cyber harassment (i.e. repetitively harassing and threatening someone), denigration/slandering (i.e. sharing false information about someone), flaming (i.e. brief insulting online interactions) and happy slapping (i.e. recording a session while a person is being bullied for circulation purpose), among others (Kowalski et al., 2012; Marcum et al., 2012). Impacts of cyberbullying are detrimental in nature, ranging from emotional (anger, fear, self-blame etc.) to psychological (low self-esteem, depression, suicidal etc.) and physical (loss of sleep, headache, eating disorder etc.) (Bottino et al., 2015; Hemphill et al., 2015; Kowalski et al., 2014; Chu et al., 2018).

Despite the various prevention and intervention strategies, cyberbullying perpetration has not decreased in the last one decade (Hinduja and Patchin, 2015). Recent studies have looked into automatically detecting cyberbullying incidents, for instance, an affect analysis based on a lexicon and Support Vector Machine (SVM) was found to be effective in detecting cyberbullying, however the accuracy decreased when the size of the data increased, suggesting that SVM may not be ideal in dealing with frequent language ambiguities typical for cyberbullying (Ptaszynski et al., 2010). Murnion et al. (2018) automatically collected data from an in-game chat (World of Tanks) and found cyberbullying to be a learned behavior (i.e. new players are less likely to engage in cyberbullying).

Given the difficulty of detecting cyberbullying (i.e. nature of the phenomenon itself whereby a malicious/negative content had to be shared repetitively) compared to simpler types of unwanted content such as racism (e.g. specifically targeting a nationality/race/ethnicity in a negative manner), misogyny (e.g. cuss words or hashtags (#WomenSuck) referring to women, or explicit declaration to be misogynistic in profile/biography) or spam (e.g. strings

of gibberish characters such as http://$55o*(ghj), studies have begun exploring additional information about victims and bullies. For example, instead of only using words and emoticons expressing insults and profanity, machine learning models such as Naïve-Bayes and Decision Tree can also take into account gender, personality, user's membership duration activity, number of friends/followers etc. to detect cyberbullying (Al-garadi et al., 2016; Nahar et al., 2012; Navarro and Jasinski, 2012). In identifying bullies and aggressors using Twitter features, Chatzakou et al. (2017a) found bullies post less, participate in fewer online communities, and are less popular than normal users. Features such as "likes" and comments on Instagram (Hosseinmardi et al., 2015), and sentiment and emotion analyses were found promising in cyberbullying detection too (Xu et al., 2012; Patch, 2015; Murnion et al., 2018).

Studies employing automatic detection mechanisms for cyberbullying are still in its infancy. Although sentiments have been shown to have positive impact in other domains such as products and services (Gupta et al., 2015; Li, 2017), political analysis (Smailović et al., 2015; Mohammad et al., 2016) and epidemic breakouts (Almazidy et al., 2016; Missier et al., 2017), studies incorporating user sentiments and emotion in cyberbullying detection are completely lacking. Moreover, user personalities have not been fully investigated, with existing studies mostly being empirical in nature (Resett and Gamez-Guadix, 2017; van Geel et al., 2017; Goodboy and Martin, 2015). To address these visible gaps, this study presents a novel cyberbullying detection model that incorporates user personalities, sentiment and emotion to aid cyberbullying detection on Twitter. To be specific, it draws on the existing empirical evidences on user personalities and cyberbullying perpetration by building a detection model, and further enhanced it using users' sentiment and emotion.

The remaining structure of the paper is as follows: Section 2 presents the background details on the issue itself, that is, cyberbullying, followed by an elaboration on the various features used in the study. The research methodology including the data collection, detection mechanisms and evaluations are presented in Section 3, followed by the results and discussion in Section 4. The paper is concluded in Section 5.

2.0 Background

2.1 Cyberbullying – a social media problem

The use of information and communication technologies, particularly social media has revolutionized the manner in which people communicate and form relationships with one another, with statistics around the world indicating a high prevalence rate of social media applications. For instance, according to the recent report by Pew Research Center (2018), Instagram (75%) and Snapchat (73%) were found to be most popular among those between 18 and 24 years old whereas Facebook and YouTube were more popular among those older than 50 (i.e. 68%). This unfortunately, provides an avenue for anti-social behaviors such as misogyny (Anzovino et al., 2018; Liu et al., 2018), sexual predation (Bogdanova et al., 2014), sexism (Frenda et al., 2019) and cyberbullying perpetration (Kowalski et al., 2019; Chatzakou et al., 2017a, b; Hosseinmardi et al., 2015).

Facebook, for instance, is one of the most popular social media platforms that allows its users to create their own profiles, upload their photos and videos, and send messages (both private and public). It has a wide reach, as any comments or posts can reach thousands of people, especially through “liking” and “sharing” mechanisms, and thus allowing cyberbullies to distribute nasty or unwanted information about their victims easily (Choo, 2016). Instagram allows its users to share photos and videos, to ‘follow’ others and support Stories. Like

Facebook, it is also easy for one to set up new, anonymous profiles for cyberbullying perpetration. The velocity and size of the distribution mechanism allow hostile comments or humiliating images to go viral within hours (Hosseinmardi et al., 2015).

Cyberbullying victimization, particularly involving young people has received an increasing level of scrutiny. For instance, Ask.fm (a platform that allows one to ask each other questions anonymously) had to launch new safety efforts when teenagers were bullied on their platform, resulting in several suicides. Likewise, Instagram recently introduced shadow banning online abusers (i.e. restricting a bully from posting or commenting on a post) as a mechanism to combat cyberbullying (LiveMint, 2019).

Twitter on the other hand, is listed as one of the top five social media platforms where the maximum percentage of users experience cyberbullying (turbofuture.com, 2019). It enables a user to send a message of 280-characters, with more than 330 million active users at present (Statista, 2018). Studies on cyberbullying and Twitter often reported extensive cases of the phenomenon, with the potential for serious, deleterious consequences for its victims (Chatzakou et al., 2017a; Balakrishnan et al., 2019; Sterner, 2017). Several measures have been taken by Twitter to mitigate cyberbullying, such as filtering unwanted messages from users without a profile picture, and enabling a time-out feature that bans users using abusive language, among others. Despite these positive attempts, the platform is not completely immune from cyberbullying (Bernazzani, 2017; Twitter, 2019).

2.2 Features

This section specifically focuses on the features incorporated in our cyberbullying detection model. It encompasses user personalities focusing on Big Five and Dark Triad models, sentiment, emotion and Twitter-based features.

2.2.1 User personalities

One of the most comprehensive and popular method to determine personality is based on the Big Five model (McCrae and John, 1992; John and Srivastava, 1999), which is a hierarchical organization of personality traits in terms of five basic dimensions/facets:

i. *Extraversion* - the tendency of being outgoing, sociable, to be interested in other people, assertive, active, paying more attention to external events and excitement seeking
ii. *Agreeableness* - the tendency to be kind, friendly, gentle, getting along with others and being warm to other people
iii. *Conscientiousness* – it presents how much a person pays attention to others when making decisions
iv. *Neuroticism* - the tendency to be depressed, fearful and moody
v. *Openness* - the tendency to be creative, perceptive, thoughtful, broad-minded, and willing to make adjustments in activities in accordance with new ideas.

Empirical evidences exist linking Big Five traits with cyberbullying, both in terms of perpetration and victimization. For instance, cyberbullying perpetration is linked to individuals who score low on agreeableness and conscientiousness, and high on extraversion and neuroticism (Alonso and Romero, 2017; van Geel et al., 2017; Festl and Quandt, 2013;

Wong and McBride, 2018). As for cybervictimization, victims tend to score higher on neuroticism and openness (Alonso and Romero, 2017).

Another personality model that is gaining momentum in cyberbullying studies is Dark Triad, which focuses more on the darker traits of user personalities. It refers to three distinct, yet undesirable (to other individuals) traits, namely, Machiavellianism (i.e. tendency to lack empathy and engage in impulsive and thrill-seeking behavior), psychopathy (i.e. tendency to strategically manipulate others) and narcissism (i.e. tendency to feel superior, grandiose, and entitled) (Paulhus and Williams, 2002). Empirical studies on Dark Triad and cyberbullying have consistently found all three traits to be positively related to cyberbullying perpetration, with psychopathy emerging as the strongest predictor (Goodboy and Martin, 2015).

2.2.2. Sentiment

Sentiments are the thoughts or opinions provoked due to the feelings attached with something, often categorized as positive, neutral or negative (Zhao et al., 2016). The process in which the unstructured data are computationally processed is referred to as sentiment analysis, and can be categorized into machine learning approach, lexicon-based approach and hybrid approach (Medhat et al., 2014). The machine learning approach employs algorithms such as Naïve Bayes, SVM, Decision Tree, etc., whereas the lexicon-based approach is dependent on sentiment lexicons (i.e. dictionary of opinion words and phrases with the assigned polarities and intensities) for gauging the sentiment of a text. In the context of cyberbullying, sentiment has been used to distinguish between bullies, victims and non-bullies (Dani et al., 2017; Nahar et al., 2012; Xu et al., 2012). For instance, Xu et al. (2012) identified cyberbullying victims using their sentiment scores as victims usually experience negative emotions such as depression, anxiety and loneliness.

2.2.3 Emotion

Unlike sentiment analysis, emotion analysis detects types of feelings through the expression of texts, such as anger, disgust, fear, happiness, sadness, and surprise. Three common methods exist in textual emotion detection, namely, keyword-based (i.e. uses synonyms and antonyms from dictionaries), learning-based (based on previously trained classifiers) and hybrid (combination of keyword and learning methods) (Ramalingam et al., 2018). Emotion analysis have been applied in various domains, such as novels (Mohammad, 2012) and suicide notes (Pestian et al., 2010; Vioulès et al., 2018). Xu et al. (2012) identified seven emotions that are common in cyberbullying behaviors, that is, anger, embarrassment, empathy, fear, pride, relief and sadness. To be specific, the authors found accusers to express more fear but less anger, compared to reporters and victims who seemed to experience more sadness and relief. Patch (2015) examined the presence of anger, sadness and fear in cyberbullying, however, no significant effect was reported despite an improvement on the overall accuracy in determining aggressive behaviors.

2.2.4 Twitter Features

The majority of the literature on cyberbullying detection have focused on Twitter features, such as user features (e.g. time when account was created, verified account, etc.), and text/content features (e.g. number of hashtags, number of symbols, number of user mentions, etc.). According to Chatzakou et al. (2017a), bullies use more hashtags and post more tweets when they become active. In addition, network features such as number of followers and following (bullies were reported to have fewer friends), power difference (more impact on victim when bullying done by someone well-known), communities (bullies tend not to cluster compared to normal users) have been shown to be effective in detecting aggressive behaviors online (Al-garadi et al., 2016; Chatzakou et al., 2017a; Navarro and Jasinski, 2012). Based on

these evidences, some of the Twitter features are selected and used as the baseline in the present study.

In general, literature review shows majority of the cyberbullying detection studies proposed techniques based on offensive keywords detected using features such as Bag of Words (BoW), number of abusive words, punctuation and lexical dictionaries, etc., however, the accuracy of these techniques remains limited (Saravanaraj et al., 2016; Hosseinmardi et al., 2015; Kayes et al., 2015). Also, despite the empirical evidences on the relationship(s) between personality and cyberbullying, to the best of our knowledge no studies have used such details in automatically detecting cyberbullying, particularly in using personality theories (i.e. Big Five and Dark Triad). Finally, attempts in investigating the impact and use of emotion and sentiment on cyberbullying detection are lacking, though promising results have been reported (Xu et al., 2012; Dani et al., 2017; Nahar et al., 2012).

The present study is the first to integrate these various aspects, especially users' psychological features to automatically detect cyberbullying. Based on the literature and empirical evidences, we hypothesize that the integration of these features may result in a more effective cyberbullying detection model.

Cyberbullying detection studies thus far, have mainly focused on the use of machine learning algorithms, such as Naïve Bayes, SVM and Decision Trees. The following sub-section elaborates on this.

2.2.5 Cyberbullying detection and machine learning

Machine learning, an application of artificial intelligence provides systems the ability to automatically learn and improve from experience without being explicitly programmed, often differentiated as supervised, unsupervised or semi-supervised algorithms (Jiang et al., 2017). The supervised algorithms take a set of training instances to build a model that generates a desired prediction for an unseen instance (i.e. based on labeled/annotated data), whereas unsupervised algorithms do not depend on labeled data, and thus often used for clustering problems (Jiang et al., 2017; Medhat et al., 2014). As cyberbullying is deemed to be a classification problem (i.e. categorizing an instance as bully or non-bully), the supervised learning algorithms were adopted in the present study.

Studies on cyberbullying detection are mainly based on supervised algorithms such as Naïve Bayes, SVM, Decision Trees (J48) and Random Forest, often with performance comparisons made between several of these classifiers (Chatzakou et al., 2017a,b; Huang et al., 2014). Naïve Bayes is a Bayesian theorem algorithm and is well-known for its ability to classify texts based on a probability (i.e. outcomes are based on the highest probability). It is therefore, well-suited for real-time predictions, text classifications and recommendation systems (Patel, 2017; Ting et al., 2011). It assumes independence between predictors, that is, the presence/absence of a feature is unrelated to the presence/absence of any other feature. Therefore, in the context of tweets, each word or feature is considered as a unique variable by Naive Bayes to determine the probability of that word/feature. For instance, Saravanaraj et al. (2016) proposed a model for detecting cyberbullying using Naïve Bayes, whereby the presence of an offensive word indicates cyberbullying, and the absence indicates otherwise. The authors however, did not evaluate their proposed model, but other similar studies such as Dinakar et al. (2011) reported an overall accuracy of 63% using Naïve Bayes to detect

cyberbullying based on YouTube comments. The present study adopted a similar approach whereby the presence of specific features (or combination of features) (e.g. high number of followers-following or a negative personality) may result in a tweet to be classified as a bully.

J48 is a popular decision tree algorithm, which uses the depth-first strategy that considers all the possible tests to split the dataset before one with the highest information gain is selected (Salzberg, 1994; Shahraki et al., 2015). The trees contain several nodes, that is, root (main node, no incoming edges), internal (with incoming and outgoing edges) and leaf (no outgoing edges). Both the root and internal nodes correspond to each feature tested whereas the leaf node is the final classification. Therefore, in the context of cyberbullying, features such as number of followers, popularity, positive sentiment can be used as the root or internal nodes, whereas bully or not-bully will be the corresponding leaf node. J48 is generally easy to use and relatively fast, however the preparation of large decision trees (i.e. large datasets with many features) are complex and time-consuming (Zhao and Zhang, 2008). Huang et al. (2014) explored the social network graphs features, namely the relationships between users and related features (e.g. number of friends), and network embeddedness (i.e. relationship between users) etc. using J48, with results indicating an accuracy of 62.8%.

Finally, Random Forest is an ensemble learning algorithm that generates multiple small decision trees (or forests) from random subsets of features, each capturing different trends in the data (Saravanaraj et al., 2016). This enables the algorithm to create classifiers for large datasets with many features efficiently. It is by far the most popular technique used by several cyberbullying detection studies (Al-garadi et al., 2016; Chatzakou et al., 2017a; Saravanaraj et al., 2016) due to its performance in dealing with multiple features which may be correlated. Chatzakou et al. (2017a) for example, used Twitter-based features to distinguish between

bullies and aggressors using several classifiers, with results indicating Random Forest to be the best (i.e. accuracy of 73.45%) compared to other tree classifiers, namely J48, Least Absolute Deviation Tree, Logistic Model Tree and Naïve Bayes Tree.

3. Methodology

3.1 Multiple-feature cyberbullying detection model

The cyberbullying detection model encompassing the various features is as illustrated in Figure 1. The overall model has three main stages, namely, Twitter data collection, feature extractions (i.e. Twitter-based, personalities, sentiment and emotion) and cyberbullying detection and classification.

3.1.1 Twitter dataset

The annotated cyberbullying dataset was obtained from Chatzakou et al. (2017a). To be specific, we were provided with 9484 annotated tweet IDs, which were extracted using #GamerGate as the hashtag. We used these tweet IDs to extract their respective tweets. The Gamergate (i.e. online video game) controversy is one of the most well documented, large-scale instances of bullying/aggressive behavior (Massanari, 2015). It originated from alleged improprieties in video game journalism, which quickly grew into a larger campaign centered around sexism and social justice (Chatzakou et al., 2017b). Extreme cases of cyberbullying and aggressive behavior were associated with Gamergate including direct threats of rape and murder. Chatzakou et al. (2017a) in fact, used the hashtag #GamerGate as a seed for snowball sampling of other hashtags (e.g. #IStandWithHateSpeech, #KillAllNiggers) which are also likely to be associated with cyberbullying and aggressive behavior.

Chatzakou et al. (2017a) accomplished the annotation via crowdsourcing (i.e. CrowdFlower), in which 834 contributors were recruited from various countries including the United States, Venezuela and Russia. Each batch of tweets was labeled by five different annotators and the final label was determined based on a majority vote. Instructions and definitions were provided to the annotators, particularly in reference to the three classes of labeling as follows:

Bully – someone who posts multiple tweets or re-tweets with negative intentions, generally for the same topic and in a repeated fashion

Aggressor - someone who posts at least one tweet or re-tweet to harm/insult other users (i.e. usually a one-off event)

Spammer - someone who posts texts of advertising/marketing or other suspicious nature, such as phishing attempts

This resulted in an annotated dataset containing 9484 tweets, with 4.5% of users labeled as bullies, 31.8% as spammers, 3.4% as aggressors and 60.3% as normal. The authors reported an average annotation accuracy of 66.5% (i.e. 83.75% for spam, 53.56% for bully, and 61.31% for aggressive cases) (Chatzakou et al., 2017a).

Pre-processing was then administered in the present study, whereby non-English tweets were removed along with those containing only special characters such as numbers, punctuations and stop words. Additionally, user profiles were examined to ensure only active user accounts were included. This resulted in the removal of profiles containing no data (i.e. previously marked as spam or bully/offensive by Twitter, and thus has been suspended or deleted). The final dataset contained 5453 tweets (i.e. normal = 3510, 64.3%; spammer = 1489, 27.3%; aggressor = 173, 3.1%; bully = 281, 5.1%). Abusive users (i.e. bullies and

aggressors) make up about 8% of the dataset, mirroring observations from previous studies including those of Chatzakou et al. (2017a) and Kayes et al. (2015).

3.1.2 Feature extractions

Twitter features that were shown to have improved cyberbullying detection in previous studies (Chatzakou et al., 2017a; Al-garadi et al., 2016; Saravanaraj et al., 2016) were extracted using Twitter API, whereas additional coding was done to extract features such as number of lower and upper case characters (i.e. content features). The features extracted were as follows:

- *Text/content features* – Features including number of characters, upper case characters, lower case characters, hashtags, symbols, user mentions, URLS and media.
- *User features* – These refer to features extracted from each individual's profile such as age of account, account verified, status count (i.e. number of tweets (including re-tweets) posted by the user), list count (i.e. number of public lists a user is a member of), and user favorite count (number of tweets a user has liked in his account's lifetime) (Chatzakou et al., 2017a)
- *Network features* – Features which measure the sociability of a user in their respective platforms such as the number of followers and following, and popularity (followers-following ratio)

The personality traits based on Big Five were determined using the IBM Watson's Personality Insight API, a tool that uses both linguistic and data analytics to predict an individual's personality. The API provides the value for each of the five dominant dimensions (i.e. extraversion, agreeableness, conscientiousness, openness and neuroticism),

along with six facets (sub-traits/extended traits) that further characterize an individual according to the dimension (IBM, 2018). The present study mapped the Big Five traits with Dark Triad based on the relationships reported in previous studies. For example, most studies found Dark Triad traits to be negatively correlated with conscientiousness and agreeableness (Ardic and Ozsoy, 2016; Paulhus and Willams, 2002; Douglas et al., 2012), whilst psychopathy to be positively related to extraversion (Douglas et al., 2012). The mappings are shown in Table 1, whereby a positive sign indicates a direct relationship whereas a negative sign indicates an inverse relationship. For example, a person who scores high on narcissism will tend to score high on openness as well.

Table 1: Mapping between Big Five and Dark Triad

		BIG FIVE PERSONALITY TRAITS				
		Openness	Conscientiousness	Extraversion	Agreeableness	Neuroticism
DARK TRIAD	N	+		+		
	M		-		-	
	P					+

Note. M: Machiavellianism, N: Narcissism, P: Psychopathy; +: positive correlation; - : negative correlation

On the other hand, the need for reliable sentiment analysis has led to an increase in the availability of sentiment analysis tools such as SentiStrength[1], Indico API[2] etc. Indico API is an advanced machine-learning, and a freely available application programming interface that supports various analyses including text, sentiment, image and emotion. For emotion analysis, the API is capable of detecting primary emotions (i.e. the body's basic emotional response that are usually easy to identify due to their strong nature, such as anger, fear etc.)

[1] http://sentistrength.wlv.ac.uk/
[2] https://indico.io/docs/emotion

(TenHouten, 2016). Indico API detects five of these, namely, anger, fear, joy, sadness and surprise.

Indico API was used for both emotion and sentiment analyses considering its ease of use and deemed appropriate as we were dealing with short texts (i.e. Twitter with limited number of characters). It has been successfully used in studies related to both emotion (Spitale et al., 2019) and sentiment analyses (Denis, 2017). The former for example, used the API to detect emotions based on a dataset comprising of original utterances of 142 Italian children aged 4 to 12 years.

For sentiments, Indico API estimates the positive and negative sentiment (on a [0, 1] scale) in short texts. A score value of more than 0.5 indicates a positive sentiment while a score lesser than 0.5 is a negative sentiment. A value of 0.5 indicates a neutral sentiment. For example, for a post such as:

This is a war on women in gaming waged by a group of sexist monsters. If you are not a horrible human being, get out of #gamergate

the score observed was 0.025, indicating a negative sentiment.

Figure 2 illustrates the sentiment distribution for the four categories of the users in the dataset, clearly showing the negative sentiments to be the highest among the bullies and aggressors, compared to spammers and normal users (dotted circle in Figure 2).

Indico API returns a set of five emotions (i.e. anger, fear, joy, sadness, surprise) for a post, with the highest value indicating the strongest emotion. For example, using the same example above, Indico API returns:

anger = 0.26; joy = 0.08; fear = 0.24; sadness = 0.35; and surprise = 0.04, indicating a strong emotion for anger, fear and sadness.

3.1.3 Cyberbullying detection and classification

This section presents our efforts to model cyberbullying using the multiple features extracted. The model was executed using WEKA 3.8 (i.e. an open source tool), with a 10-fold cross validation. Three machine learning algorithms were used, namely, Random Forest, Naïve Bayes and J48 (described in Section 2.2.5). Although Naïve Bayes has been shown to have a higher success rate compared to other algorithms (Kayes et al., 2015; Saravanaraj et al., 2016), however, during our preliminary experimental analysis, it performed poorly (i.e. average 62%) compared to Random Forest and J48 (i.e. above 90%), hence the algorithm was excluded from further analysis. The classifiers were trained using the manually annotated data from Chatzakou et al. (2017a). Finally, the cyberbullying detection model classifies the tweets into one of four user categories/roles as defined in Section 3.1.1, namely, bully, aggressor, spammer or normal (i.e. tweets that do not fit into any other classes).

3.2 Experiments

Several experiments were conducted to determine the effectiveness of the detection model based on the various features/aspects, namely, personality, sentiment and emotion. The aspects were evaluated both individually and jointly with other aspects, and executed using

Random Forest, J48 and Naïve Bayes based on the four roles: bully, aggressor, spammer and normal. The distinct models are as follows:

- *Baseline:* detection model using only the Twitter features
- *Baseline + Personalities:* the model above, combined with all the traits from Big Five and Dark Triad
- *Baseline + Personalities + Sentiment:* the model above, combined with sentiment analysis
- *Baseline + Personalities + Sentiment + Emotion:* the model above, combined with emotion analysis
- *Baseline + Sentiment:* baseline model combined with sentiment analysis
- *Baseline + Sentiment + Emotion*: the model above, combined with emotion analysis
- *Baseline + Key Features:* the baseline model coupled with significant features

3.3 Evaluation metrics

As the cyberbullying dataset used in this study is imbalanced (Section 3.1.1), two strategies were adopted to address this issue[3].. First, several machine learning algorithms, namely, Naïve Bayes, Random Forest and J48 were administered instead of using a single algorithm. Also, decision trees have been often reported to perform well on imbalanced datasets (Krawczyk et al., 2014), therefore, the selection of the algorithms are deemed to be appropriate. The second strategy adopted was to evaluate the model's effectiveness using various performance metrics (Dani et al., 2017). To be precise, standard metrics including F-measure (i.e. harmonious mean of precision and recall), accuracy (i.e. total correctly classified tweets normalized by the total number of tweets) and area under the receiver-operating characteristic (ROC) curve (AUC) were used. ROC curves are used to test the

[3] https://machinelearningmastery.com/tactics-to-combat-imbalanced-classes-in-your-machine-learning-dataset/

classifiers on different points to obtain true positive (i.e. recall) against false positive rates indicating the sensitivity of the model, and the resulting area under curve (AUC) is the probability of a classifier correctly ranking a random positive case higher than a random negative case (Davis and Goadrich, 2006). AUC and F-measure were particularly selected, as they are considered more robust when class imbalance exists (e.g. such as in this study where the four user roles are not balanced) (Dani et al., 2017).

Additionally, the overall kappa (compares an observed accuracy with an expected accuracy) and the root mean squared error (RMSE) that measures the differences between the predicted and observed values are provided as well. The confusion matrix, which is a breakdown of predictions showing correct predictions and the types of incorrect predictions made are also presented where applicable. Two sample *t*-tests were performed to examine if the differences between the models are significant, where appropriate using an alpha value of 0.05. The *t*-tests were carried out using Statistical Program for the Social Sciences (SPSS) version 25 (IBM, 2019).

4 Results and Discussion

Both Random Forest and J48 performed well for the cyberbullying classifications, although J48 performed slightly better (no significant difference). As the intent of the study is not in determining the best algorithm, therefore, only results for J48 are presented in this section.

Table 2: The effectiveness of the cyberbullying detection models

Models	Accuracy	AUC	F - score	Kappa	RMSE
Baseline	88.32	0.90	0.83	0.809	0.192
Baseline + Personalities	91.23	0.95	0.91	0.833	0.176
Baseline + Personalities + Sentiment*	**91.88**	**0.97**	**0.92**	**0.840**	**0.178**
Baseline + Personalities + Sentiment + Emotion	91.12	0.94	0.91	0.825	0.193
Baseline + Sentiment	90.63	0.95	0.90	0.816	0.194
Baseline + Sentiment + Emotion	89.95	0.94	0.89	0.811	0.207

*: Significant differences only noted with baseline

Table 2 indicates all the models to outperform the baseline algorithm, hence showing that psychological features can be effectively used to improve cyberbullying detection mechanisms. Generally, the combination of personalities and sentiment produced the best results across all the metrics, with an accuracy of 91.88%, a weighted AUC of 0.97, F-score of 92%, kappa of 0.840 and RMSE of 0.178. Table 3 below shows the breakdown of the classifications based on the four categories, with results indicating highest recall for the same model.

Table 3: Classifications based on each of the four user categories

Setup	Correct Count				Correctly Classified (Recall)			
	A	B	N	S	A	B	N	S
Baseline	82	233	3318	1292	47.40	82.92	94.53	86.77
Baseline + Personality	83	235	3341	1324	47.98	83.63	95.19	88.92
Baseline + Personality + Sentiment	**87**	**243**	**3345**	**1343**	**50.29**	**86.48**	**95.30**	**90.19**
Baseline + Personality + Sentiment + Emotion	79	215	3345	1330	45.66	76.51	95.30	89.32
Baseline + Sentiment	84	227	3319	1312	48.55	80.78	94.56	88.11
Baseline + Sentiment + Emotion	76	208	3312	1309	43.93	74.02	94.36	87.91

Actual count: (A) Aggressor = 173; (B) Bully = 281; (N) Normal = 3510; (S) Spammer = 1489;

Results from both Table 2 and 3 are in accordance with empirical evidences that have shown personalities (Alonso and Romero, 2017; van Geel et al., 2017; Festl and Quandt, 2013) and sentiments (Xu et al., 2012; Dinakar et al., 2011; Nahar et al., 2014) to be linked with cyberbullying perpetration. Although a direct comparison is not possible with other cyberbullying detection studies due to the nature of the dataset used, different classification algorithms and analysis mechanisms, our findings generally indicate a higher effectiveness in cyberbullying detection (e.g. our AUC of 0.970 as opposed to 0.943 in Al-garadi et al., 2016; 0.817 in Dani et al., 2017; and 0.815 in Chatzakou et al., 2017a). We conclude that this is probably due to the inclusion of users' psychological features, namely their personalities and sentiment which were not investigated in any of these mentioned studies.

As personalities were found to improve cyberbullying detection, a further analysis was administered whereby each individual personality traits was examined for its impact on cyberbullying detection. Table 4 provides the accuracy and F-scores for the individual models.

Table 4: Cyberbullying detection effectiveness for specific personality traits

Models	Accuracy	F-measure
Baseline + Openness	90.1	0.901
Baseline + Conscientiousness	90.1	0.901
*Baseline + Extraversion**	*91.1*	*0.911*
*Baseline + Agreeableness**	*91.0*	*0.910*
*Baseline + Neuroticism**	*91.1*	*0.911*
Baseline + Narcissism	90.8	0.908

Baseline + Machiavellianism	90.5	0.905
*Baseline + Psychopathy***	*91.7*	*0.918*

Note: * - Significantly different with Openness and Conscientiousness;

** - Significantly different with Narcissism and Machiavellianism

Higher scores for accuracy and F-scores for extraversion, agreeableness, and neuroticism (Big Five) and psychopathy (Dark Triad) indicate that these traits have greater impacts on cyberbullying detection. T-tests indicate significant differences between extraversion, agreeableness and neuroticism, with openness and conscientiousness (i.e. $p < 0.05$). As for Dark Triad, psychopathy was found to be significantly different with Machiavellianism and narcissism ($p < 0.001$). Similar outcomes were reflected in previous empirical studies in which the traits were found to have significant relations with cyberbullying perpetrations (van Geel et al., 2017). For example, extraverted people have a higher tendency to engage in cyberbullying perpetration to increase their social status (van Geel et al., 2017), and they communicate and use social media more compared to those who score low on extraversion (Marshall et al., 2015).

The inclusion of emotion however, had no positive impact on the detection model's performance (see Table 2). In fact, emotion resulted in lower accuracies in both Baseline + Personality + Sentiment + Emotion (i.e. 91.12%) and Baseline + Sentiment + Emotion (i.e. 89.95%), compared to those without and thus indicating no significant effect of emotion in cyberbullying detection, an observation that was reflected in Patch (2015). This could be attributed to the nature of the dataset, for example, most often negative emotions such as angry, fear, embarrassment etc. are related to cyber victims (Balakrishnan, 2018; Xu et al., 2012; Gan et al., 2014; Kokkinos et al., 2014), although there is a tendency among bullies to exhibit these emotions to a certain extent (Balakrishnan, 2018; Schenk et al., 2013). The

dataset lacked tweets related to victims; hence this may have affected the impact of emotion on the detection mechanism.

Looking at the best performing model (i.e. Baseline + Personality + Sentiment), we wanted to identify the specific features that may have contributed to the cyberbullying detection. For this reason, data dimensionality reduction technique was applied, particularly wrapper feature selection method. The wrapper method basically uses a predetermined learning algorithm (e.g. K-Means, Affinity Propagation, Greedy algorithm etc.) to prepare, evaluate and select the best features. The study used the wrapper method due to its high accuracy and ability to consider interactions between features and predictive models (Jindal and Kumar, 2017). To identify the top 10 key features, the greedy algorithm based on the best-first search were administered. This technique basically lists the best features first (or deletes the worst feature first) in each round (Hall et al., 2009; Jindal and Kumar, 2017; Panthong and Srivihok, 2015).

The 10 key features produced were number of followers, following, popularity, user favorite count and status count (i.e. Twitter features), extraversion, agreeableness and neuroticism (Big Five), psychopathy (Dark Triad) and sentiment. These key features were integrated into a single model, and compared against the best performing model in Table 2. Figure 3 clearly indicates that when key features are used, the performance of the cyberbullying detection model is further improved.

The finding indicates that although multiple features can be used to enhance cyberbullying detection, specific features play more profound roles in the process of detecting bullying patterns online. Table 5 depicts the breakdown of the classification for the key feature model,

indicating the model performed the best in detecting bullies online (i.e. 92.88%) compared to the Baseline + Personality + Sentiment model.

Table 5: Classification breakdown for Key Features versus Baseline + Personality + Sentiment

Setup	Correct Count				Correctly Classified (Recall)			
	A	B	N	S	A	B	N	S
Baseline + Personality + Sentiment	87	**243**	3345	1343	50.29	**86.48**	95.30	90.19
Baseline + Key Features	49	**261**	3379	1343	28.32	**92.88**	96.27	90.19

Actual count: (A) Aggressor = 173; (B) Bully = 281; (N) Normal = 3510; (S) Spammer = 1489;

The present study showed that not only personalities and sentiment can be effectively used to detect cyberbullying, but focusing on specific features further improves the detection process. The findings add support to existing empirical evidences linking specific personalities, particularly extraversion, agreeableness, neuroticism and psychopathy to cyberbullying perpetration, whereby these key traits had been shown to significantly improve online bullying detection. The top Twitter features extracted by the dimension reduction technique, namely, number of followers, following, popularity, user favorite count and status count belong to the user and network-based features, suggesting that activities and connectivity of a user in the network play important roles in identifying the bullies and non-bullies, a phenomenon observed in Al-garadi et al. (2016) and Chatzakou et al. (2017a).

5. Conclusion, Limitation and Future Directions

User features such as personalities, sentiment and emotion were used to improve cyberbullying detection using a dataset consisting of 5453 tweets. The execution of J48 revealed overall cyberbullying detection notably improved when user personalities and sentiments were used, with an accuracy of 91.88% and a weighted AUC of 0.97. Ten key

features were identified and further integrated into a single model, and the effectiveness of the detection improved to an accuracy of 92.88%. Emotion however, was not found to contribute positively to recognizing bullying patterns online.

In short, the study used empirical evidences for user personalities, and combined them with other pertinent features such as emotion and sentiment for cyberbullying detection. Results indicate that specific key personalities, particularly extraversion, agreeableness, neuroticism and psychopathy have greater impacts on cyberbullying perpetration compared to other traits. Knowing individual traits based on users' online communication styles (i.e. regardless of social media platforms used) 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. 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.

The study also found specific features such as popularity, number of followers and following to have greater impacts in identifying cyberbullying. With such findings, 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. Although the features identified in this study are Twitter-specific, most of these features (and others) are available and supported in a majority of the popular social media platforms. For instance, Facebook has number of friends, number of followers, user profiles, likes, shares and reactions whereas Instagram has number of followers-following, likes, hashtags, favorite counts and user profiles, among others. In fact,

Hosseinmardi et al. (2015) found Instagram communication sessions with cyberbullying and cyber-aggression incidents to have lower number of likes, but a higher number of followers. Therefore, similar techniques such as the dimension reduction applied in the present study can be adopted to identify top features in other social media platforms for detecting anti-social behaviors.

The study nevertheless had a few limitations. First, the dataset is limited both in terms of its size and the tweet categories (i.e. roles). The pre-processing resulted in an imbalanced dataset containing 5453 tweets with the majority being normal and spams (see Section 3.1.1). Although the breakdown is similar with previous studies (Chatzakou et al., 2017a; Kayes et al., 2015), we believe this may have contributed to the high accuracy and F-score for our proposed model. The present study addressed this issue by assessing the cyberbullying detection models using various metrics, including AUC, kappa and the confusion matrix, however, future studies could look into using a more balanced cyberbullying dataset, or generate synthetic samples using algorithms such as Synthetic Minority Over-sampling Technique (SMOTE[4]) (Chawla et al., 2002).

Second, the dataset contained no tweets pertaining to victims and bystanders. This is deemed important because most studies have consistently linked higher emotional responses such as anxiety, embarrassment and suicidal ideation with cyber-victims (Balakrishnan, 2018; Hinduja and Patchin, 2010). In fact, this could be one of the reasons as to why emotion was not found to be significant in the present study. Although our finding is similar with Patch (2015), however, emotion has been shown to have a significant impact in detecting other anti-social behavioral issues such as cyberpedophilia (Bogdanova et al., 2014) and

[4] https://machinelearningmastery.com/tactics-to-combat-imbalanced-classes-in-your-machine-learning-dataset/

irony/sarcasm/satire (Hernández et al., 2016; Thu and Aung, 2018; Ciucci and Baroncelli, 2014), and thus further support our recommendation for future studies to explore this particular aspect in cyberbullying.

Moreover, empirical studies often report bystanders to be the largest group in cyberbullying prevalence, usually categorized as assistants (i.e. one who encourages the bully or joins in the bullying) and defenders (i.e. one who defends the victim), hence further reiterating our point in using a larger dataset encompassing both victims and bystanders. It will also be beneficial to compare the performance of the proposed model using several other Twitter datasets so as to avoid biasness (i.e. by using other relevant hashtags). Furthermore, the detection model can be adapted and reproduced to analyse textual data from other social media platforms including Instagram or YouTube, focusing specifically on additional features that are not available on Twitter. For example, Twitter does not require a user to disclose their gender during registration unlike Facebook and Instagram. Empirical evidences show mixed results in the role of gender in cyberbullying (Balakrishnan, 2015; Festl and Quandt, 2013; Gan et al., 2014), therefore it would be interesting to adapt the present model to include gender to examine its effectiveness in cyberbullying detection.

Finally, as stated in Section 1, cyberbullying takes various forms including flaming, denigration, impersonation and harassment, among others. The present study did not differentiate between these cyberbullying types, unlike some studies in misogyny detection for example, whereby the authors categorized tweets into objectification, sexual harassment, discredit etc. (Anzovino et al., 2019). Therefore, it would be interesting to extend and examine if the proposed model can perform fine-grained cyberbullying classifications.

Acknowledgement

The study is partly supported by the Fulbright Scholar Program 2018. The authors express their gratitude to Mr. Despoina Chatzakou for kindly providing the Twitter IDs related to the dataset used in the present study.

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

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. https://doi.org/10.1016/j.chb.2016.05.051

Almazidy, A., Althani, H., & Mohammed, M. (2016). Towards a disease outbreak notification framework using Twitter mining for smart home dashboards. Procedia Computer Science, 82, 132-134. https://doi.org/10.1016/j.procs.2016.04.019

Alonso, C., & Romero, E. (2017). Aggressors and victims in bullying and cyberbullying: A study of personality profiles using the Five-Factor Model. The Spanish Journal of Psychology, 20, e76. https://doi.org/10.1017/sjp.2017.73

Anzovino M., Fersini E., and Rosso P. (2018). Automatic identification and classification of misogynistic language on Twitter. In 23rd International Conference on Applications of Natural Language to Information Systems, Springer International Publishing, Paris, France, 57-64.

Ardic, K., & Ozsoy, E. (2016) Examining the relationship between the Dark Triad traits and Big Five personality dimensions. In Proceedings of the Fifth European Academic Research Conference on Global Business, Economics, Finance and Banking (EAR16Turkey Conference), Istanbul-Turkey, pp. 1 - 9

Balakrishnan, V. (2015). Cyberbullying among young adults in Malaysia: The roles of gender, age and Internet frequency. Computers in Human Behavior, 46, 149-157. https://doi.org/10.1016/j.chb.2015.01.021

Balakrishnan, V. (2018). Actions, emotional reactions and cyberbullying – From the lens of bullies, victims, bully-victims and bystanders among Malaysian young adults. Telematics and Informatics, 35, 1190-1200. https://doi.org/10.1016/j.tele.2018.02.002

Balakrishnan, V., Khan, S., Fernandez, T., & Arabnia, H. R. (2019). Cyberbullying detection on Twitter using Big Five and Dark Triad features. Personality and Individual Differences, 141, 252-257. https://doi.org/10.1016/j.paid.2019.01.024

Bernazzani, S. (2017). How Twitter is fighting harassment and cyberbullying. Retrieved from https://blog.hubspot.com/marketing/twitter-harassment-cyberbullying (accessed 25 July 2019).

Bogdanova D., Rosso P., & Solorio T. (2014) Exploring high-level features for detecting cyberpedophilia. Computer Speech & Language, 28, 108-120

Bottino, S. M. B., Bottino, C., Regina, C. G., Correia, A. V. L., & Ribeiro, W. S. (2015). Cyberbullying and adolescent mental health: Systematic review. Cadernos de saude Publica, 31, 463-475. https://doi.org/10.1590/0102-311X00036114

Chatzakou, D., Kourtellis, N., Blackburn, J., De Cristofaro, E., Stringhini, G., & Vakali, A. (2017a). Mean birds: Detecting aggression and bullying on Twitter. In Proceedings of the 2017 ACM on Web Science Conference, Troy, New York, USA, pp. 13 - 22

Chatzakou, D., Kourtellis, N., Blackburn, J., De Cristofaro, E., Stringhini, G., & Vakali, A. (2017b). Measuring #Gamergate: A tale of hate, sexism, and bullying. In Proceedings of the 26th International Conference on World Wide Web Companion, pp. 1285-1290

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321-357. https://doi.org/10.1613/jair.953

Choo, M. S. (2016). Cyberbullying on Facebook and psychosocial adjustment in Malaysian adolescents (Doctor Of Philosophy Thesis). University of Hawai‘i, Mānoa, Hawaii.

Chu, X.-W., Fan, C.-Y., Liu, Q.-Q., & Zhou, Z.-K. (2018). Cyberbullying victimization and symptoms of depression and anxiety among Chinese adolescents: Examining hopelessness

as a mediator and self-compassion as a moderator. Computers in Human Behavior, 86, 377-386. https://doi.org/10.1016/j.chb.2018.04.039

Ciucci, E., & Baroncelli, A. (2014). Emotion-related personality traits and peer social standing: Unique and interactive effects in cyberbullying behaviors. Cyberpsychology, Behavior, and Social Networking, 17, 584-590. https://doi.org/10.1089/cyber.2014.0020

Costa Jr, P. T., & McCrae, R. R. (1992). Four ways five factors are basic. Personality and Individual Differences, 13, 653-665. https://doi.org/10.1016/0191-8869(92)90236-I

Dani, H., Li, J., & Liu, H. (2017). Sentiment informed cyberbullying detection in social media. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Cham, pp. 52-67.

Davis, J., & Goadrich, M. (2006). The relationship between precision-recall and ROC curves. In Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, Pennsylvania, USA, 233-240.

Denis, J. (2017). How do people use Facebook? A "comment" on modern social media interaction. Retrieved from http://www.jelanidenis.com/documents/facebook_report.pdf (accessed 25 May 2018).

Dinakar, K., Reichart, R., & Lieberman, H. (2011). Modeling the detection of textual cyberbullying. In Proceedings of the IEEE International Fifth International AAAI Conference on Weblogs and Social Media (SWM'11), Barcelona, Spain, pp. 11-17

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. https://doi.org/10.4236/psych.2012.33033

Dulovics, M., & Kamenská, J. (2017). Analysis of cyberbullying forms by aggressors in elementary and secondary schools. The New Educational Review, 49(3), 126-137. https://doi.org/10.15804/tner.2017.49.3.10

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, 101–126. https://doi.org/10.1111/j.1468-2958.2012.01442.x

Frenda, S., Ghanem, B., Montes-y-Gómez, M., & Rosso, P. (2019). Online hate speech against women: Automatic identification of misogyny and sexism on Twitter. Journal of Intelligent & Fuzzy Systems, 36(5), 4743-4752. https://doi.org/10.3233/JIFS-179023

Gan S. S., Zhong, C., Das, S., Gan J. S., Willis, S., & Tully, E. (2014). The prevalence of bullying and cyberbullying in high school: A 2011 survey. International Journal of Adolescent Medicine and Health, 26, 27-31. https://doi.org/10.1515/ijamh-2012-0106

Goodboy, A. K., & Martin, M. M. (2015). The personality profile of a cyberbully: Examining the Dark Triad. Computers in Human Behavior, 49, 1-4. https://doi.org/10.1016/j.chb.2015.02.052

Gupta, P., Kumar, S., & Jaidka, K. (2015). Summarizing customer reviews through aspects and contexts. In Proceedings of 16th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing), Cairo, Egypt, pp. 241-256.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: An update. ACM SIGKDD Explorations Newsletter, 11(1), 10-18. https://doi.org/10.1145/1656274.1656278

Hemphill, S.A., Kotevski, A., Heerde, J.A. (2015). Longitudinal associations between cyber-bullying perpetration and victimization and problem behavior and mental health problems in young Australians, International Journal of Public Health, 60, 227–237. https://doi.org/10.1007/s00038-014-0644-9

Hernández I., Patti V., Rosso P. (2016). Irony detection in Twitter: The role of affective content. ACM Transactions on Internet Technology, 16, 1-24

Hinduja, S. & Patchin, J. W. (2015). Bullying beyond the schoolyard: Preventing and responding to cyberbullying (2nd edition). Thousand Oaks, CA: Sage Publications.

Hinduja, S., & Patchin, J. W. (2010). Bullying, cyberbullying, and suicide. Archives of Suicide Research, 14(3), 206-221. https://doi.org/10.1080/13811118.2010.494133

Hosseinmardi, H., Mattson, S. A., Rafiq, R. I., Han, R., Lv, Q., & Mishra, S. (2015). Analysing labeled cyberbullying incidents on the Instagram social network. In Proceedings of the 7th International Conference, SocInfo 2015, Beijing, China, pp. 49-66.

Huang, Q., Singh, V. K., & Atrey, P. K. (2014). Cyber bullying detection using social and textual analysis. In Proceedings of the 3rd International Workshop on Socially Aware Multimedia, Orlando, Florida, USA, pp. 3-6.

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 18 June, 2019)

IBM. (2019). SPSS Software. Retrieved from https://www.ibm.com/analytics/spss-statistics-software (accessed 24 November 2019).

Jiang, C., Zhang, H., Ren, Y., Han, Z., Chen, K.-C., & Hanzo, L. (2017). Machine learning paradigms for next-generation wireless networks. IEEE Wireless Communications, 24(2), 98-105. https://doi.org/10.1109/MWC.2016.1500356WC

Jindal, P., & Kumar, D. (2017). A review on dimensionality reduction techniques. International Journal of Computer Applications, 173(2), 42-46. https://doi.org/10.5120/ijca2017915260

John, O. P., & Srivastava, S. (1999). The Big Five trait taxonomy: History, measurement, and theoretical perspectives. Handbook of Personality: Theory and Research, 2, 102-138.

Kayes, I., Kourtellis, N., Quercia, D., Iamnitchi, A., & Bonchi, F. (2015). The social world of content abusers in community question answering. In Proceedings of the 24th International Conference on World Wide Web, Florence, Italy, pp. 570-580.

Kokkinos, C. M., Antoniadou, N., & Markos, A. (2014). Cyber-bullying: An investigation of the psychological profile of university student participants. Journal of Applied Developmental Psychology, 35(3), 204-214. https://doi.org/10.1016/j.appdev.2014.04.001

Kowalski, R. M., Giumetti, G. W., Schroeder, A. N., & Lattanner, M. R. (2014). Bullying in the digital age: A critical review and meta-analysis of cyberbullying research among youth. Psychological Bulletin, 140(4), 1073 - 1137. http://dx.doi.org/10.1037/a0035618

Kowalski, R. M., Limber, S. P., & Agatston, P. W. (2012). Cyberbullying: Bullying in the digital age: John Wiley & Sons.

Kowalski, R. M., Limber, S. P., & McCord, A. (2019). A developmental approach to cyberbullying: Prevalence and protective factors. Aggression and Violent Behavior, 45, 20-32.

Krawczyk, B., Woźniak, M., & Schaefer, G. (2014). Cost-sensitive decision tree ensembles for effective imbalanced classification. Applied Soft Computing, 14, 554-562. https://doi.org/10.1016/j.asoc.2013.08.014

Lee, K., Mahmud, J., Chen, J., Zhou, M., & Nichols, J. (2014). Who will retweet this?: Automatically identifying and engaging strangers on Twitter to spread information. In Proceedings of the 19th International Conference on Intelligent User Interfaces, Haifa, Israel, pp. 247-256.

Li, G. (2017). Application of sentiment analysis: Assessing the reliability and validity of the global airlines rating program (Bachelor Thesis). University of Twente, Enschede, Netherlands.

Liu, B. (2012) Sentiment analysis and Opinion mining, Morgan & Claypool Publishers

Liu, H. Chiroma, F & Cocea, M. (2018) Identification and classification of misogynous tweets using multi-classifier fusion, In Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018), Sevilla, Spain, pp. 268-273.

LiveMint. (2019). Instagram is taking cyberbullying seriously, introduces 'shadow ban'. Retrieved from https://www.livemint.com/technology/tech-news/instagram-is-taking-cyberbullying-seriously-introduces-shadow-ban-1562648818682.html (accessed 25 July 2018).

Marcum, C. D., Higgins, G. E., Freiburger, T. L., & Ricketts, M. L. (2012). Battle of the sexes: An examination of male and female cyber bullying. International Journal of Cyber Criminology, 6(1), 904-911.

Marshall, T. C., 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. https://doi.org/10.1016/j.paid.2015.04.039

Martin, R. A., Lastuk, J. M., Jeffery, J., Vernon, P. A., & Veselka, L. (2012). Relationships between the Dark Triad and humor styles: A replication and extension. Personality and Individual Differences, 52(2), 178-182. https://doi.org/10.1016/j.paid.2011.10.010

Massanari, A. (2015). #Gamergate and the Fappening: How Reddit's algorithm, governance, and culture support toxic technocultures. New Media & Society, 19(3), 329-346. https://doi.org/10.1177/1461444815608807

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journal, 5(4), 1093-1113. https://doi.org/10.1016/j.asej.2014.04.011

Missier, P., McClean, C., Carlton, J., Cedrim, D., Silva, L., Garcia, A., & Romanovsky, A. (2017). Recruiting from the network: Discovering Twitter users who can help combat Zika epidemics, International Conference on Web Engineering, Springer, Cham, pp. 437-445.

Mohammad, S. M. (2012). From once upon a time to happily ever after: Tracking emotions in mail and books. Decision Support Systems, 53(4), 730-741.

Mohammad, S. M. (2016). Sentiment analysis: Detecting valence, emotions, and other affectual states from text, Emotion Measurement, 201-237. https://doi.org/10.1016/B978-0-08-100508-8.00009-6

Murnion, S., Buchanan, W. J., Smales, A., & Russell, G. (2018). Machine learning and semantic analysis of in-game chat for cyberbullying. Computers & Security, 76, 197-213. https://doi.org/10.1016/j.cose.2018.02.016

Nahar, V., Al-Maskari, S., Li, X. & Pang, C. (2014). Semi-supervised learning for cyberbullying detection, Social Networks, Databases Theory and Applications, 8506, 160-171.

Nahar, V., Unankard S., Li X., & Pang C. (2012) Sentiment analysis for effective detection of cyberbullying. In Sheng Q.Z., Wang G., Jensen C.S., Xu G. (eds) Web Technologies and Applications, Lecture Notes in Computer Science, 7235. Springer, Berlin, Heidelberg, pp. 767-774.

Navarro, J. N., & Jasinski, J. L. (2012). Going cyber: Using routine activities theory to predict cyberbullying experiences. Sociological Spectrum, 32(1), 81-94. https://doi.org/10.1080/02732173.2012.628560

Pabian, S., De Backer, C. J., & Vandebosch, H. (2015). Dark Triad personality traits and adolescent cyber-aggression, Personality and Individual Differences, 75, 41–46. https://doi.org/10.1016/j.paid.2014.11.015

Panthong, R., & Srivihok, A. (2015). Wrapper feature subset selection for dimension reduction based on ensemble learning algorithm. Procedia Computer Science, 72, 162-169. https://doi.org/10.1016/j.procs.2015.12.117

Patch, J. A. (2015). Detecting bullying on Twitter using emotion lexicons (Master of Science Thesis). University of Georgia, Athens, United States.

Patel, S. (2017). Supervised learning and Naive Bayes classification - Part 1 (Theory). Retrieved from https://medium.com/machine-learning-101/chapter-1-supervised-learning-and-naive-bayes-classification-part-1-theory-8b9e361897d5 (accessed 25 July 2018).

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. https://doi.org/10.1016/S0092-6566(02)00505-6

Pestian, J., Nasrallah, H., Matykiewicz, P., Bennett, A., & Leenaars, A. (2010). Suicide note classification using Natural Language Processing: A content analysis. Biomedical Informatics Insights, 3, 19–28. https://doi.org/10.4137%2FBII.S4706

Pew Research Center (2018) Share of U.S. adults using social media, including Facebook, is mostly unchanged since 2018, Retrieved from https://www.pewresearch.org/fact-tank/2019/04/10/share-of-u-s-adults-using-social-media-including-facebook-is-mostly-unchanged-since-2018/ (accessed 15 May 2019).

Ptaszynski, M., Dybala, P., Matsuba, T., Masui, F. Rzepka, R., & Araki, K. (2010). Machine learning and affect analysis against cyberbullying, International Journal of Computational Linguistics Research, 1(3), 135-154

Ramalingam, Pandian, A., Abhijeet, J., & Nikhar, B. (2018). Emotion detection from text. Journal of Physics: Conference Series, 1000(1), 1-5. https://doi.org/10.1088/1742-6596/1000/1/012027

Rauthmann, J. F., & Kolar, G. P. (2012). How "dark" are the Dark Triad traits? Examining the perceived darkness of narcissism, Machiavellianism, and psychopathy. Personality and Individual Differences, 53(7), 884-889. https://doi.org/10.1016/j.paid.2012.06.020

Resett, S., & Gámez-Guadix, M. (2017). Traditional bullying and cyberbullying: Differences in emotional problems, and personality. Are cyberbullies more Machiavellians? Journal of Adolescence, 61, 113-116. https://doi.org/10.1016/j.adolescence.2017.09.013

Salzberg, S. L. (1994). C4.5: Programs for machine learning, Machine Learning, 16(3), 235-240. https://doi.org/10.1007/bf00993309

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), 2249-9555.

Schenk, A. M., Fremouw, W. J., & Keelan, C. M. (2013). Characteristics of college cyberbullies. Computers in Human Behavior, 29(6), 2320-2327. https://doi.org/10.1016/j.chb.2013.05.013

Schultze-Krumbholz, A., Jakel, A., Schultze, M. & Scheithauer, H. (2012) Emotional and behavioural problems in the context of cyberbullying: A longitudinal study among German adolescents, Emotional and Behavioural Difficulties 17, 329 – 345. https://doi.org/10.1080/13632752.2012.704317

Shahraki, N., Hossein, M., Torabi, Z. S., & Nabiollahi, A. (2015). Using J48 tree partitioning for scalable SVM in spam detection. Computer and Information Science, 8(2), 37. http://dx.doi.org/10.5539/cis.v8n2p37

Smailović, J., Kranjc, J., Grčar, M., Žnidaršič, M. & Mozetič, I. (2015). Monitoring the Twitter during the Bulgarian elections, In Proceedings of the 2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA), Paris, 2015, pp. 1-10.

Spitale, M., Catania, F., Cosentino, G., Gelsomini, M., & Garzotto, F. (2019). WIYE: Building a corpus of children's audio and video recordings with a story-based app. In Proceedings of the 24th International Conference on Intelligent User Interfaces: Companion, Marina del Ray, California, pp. 33-34

Statista. (2018). Number of monthly active Twitter users worldwide from 1st quarter 2010 to 1st quarter 2018 (in millions). Retrieved from https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/ (accessed 30 May 2018).

Sterner, G., & Felmlee, D. (2017). The social networks of cyberbullying on Twitter. International Journal of Technoethics, 8(2), 1-15. https://doi.org/10.4018/IJT.2017070101

TenHouten, W. D. (2016). Normlessness, anomie, and the emotions, Sociological Forum, 31, 465-486. https://doi.org/10.1111/socf.12253

Thu, P. P. & Aung, T. N. (2018) Implementation of emotional features on satire detection, International Journal of Networked and Distributed Computing, 6, 78 - 87.

Ting, S., Ip, W., & Tsang, A. H. (2011). Is Naive Bayes a good classifier for document classification. International Journal of Software Engineering and Its Applications, 5(3), 37-46.

turbofuture.com. (2019). Cyberbullying and social media. Retrieved from https://turbofuture.com/internet/Cyberbullying-and-Social-Media (accessed 21 May 2019).

Twitter. (2019). About online abuse. Retrieved from https://help.twitter.com/en/safety-and-security/cyber-bullying-and-online-abuse (accessed 25 July 2018).

van Geel, M., Goemans, A., Toprak, F., & 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. https://doi.org/10.1016/j.paid.2016.10.063

Vioulès, M. J., Moulahi, B., Azé, J., & Bringay, S. (2018). Detection of suicide-related posts in Twitter data streams. IBM Journal of Research and Development, 62(1), 7: 1-7: 12. https://doi.org/10.1147/JRD.2017.2768678

Wong, N., & McBride, C. (2018). Fun over conscience: Fun-seeking tendencies in cyberbullying perpetration. Computers in Human Behavior, 86, 319-329. https://doi.org/10.1016/j.chb.2018.05.009

Xu, J.-M., Zhu, X., & Bellmore, A. (2012). Fast learning for sentiment analysis on bullying. In Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining, Article 10, Beijing, China.

Zhao, J., Liu, K., & Xu, L. (2016). Sentiment analysis: Mining opinions, sentiments, and emotions. Computational Linguistics, 42(3), 595-598. https://doi.org/10.1162/COLI_r_00259

Zhao, Y., & Zhang, Y. (2008). Comparison of decision tree methods for finding active objects. Advances in Space Research, 41(12), 1955-1959. https://doi.org/10.1016/j.asr.2007.07.020

Figure 1: Multiple-feature cyberbullying detection model

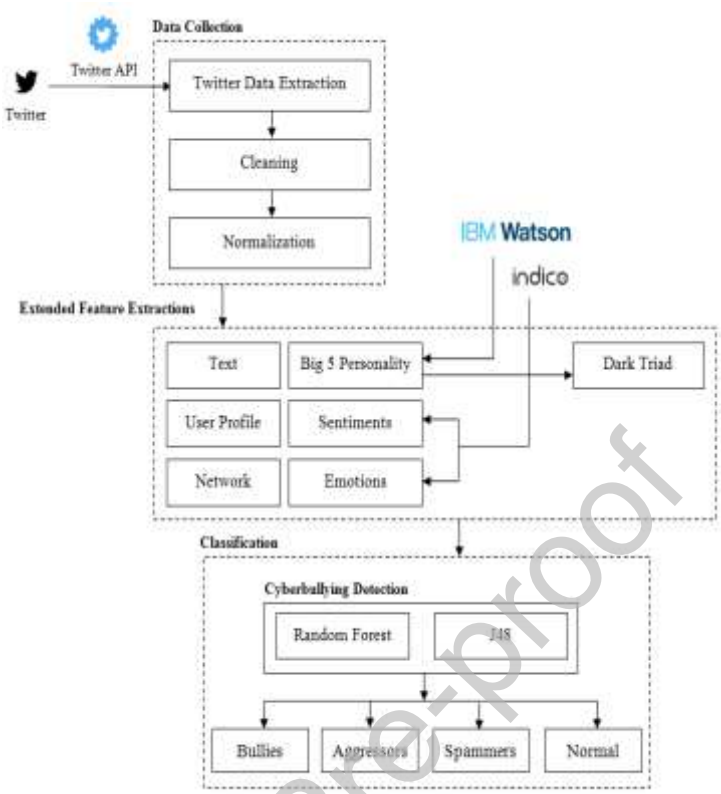

Figure 2: Sentiment distribution

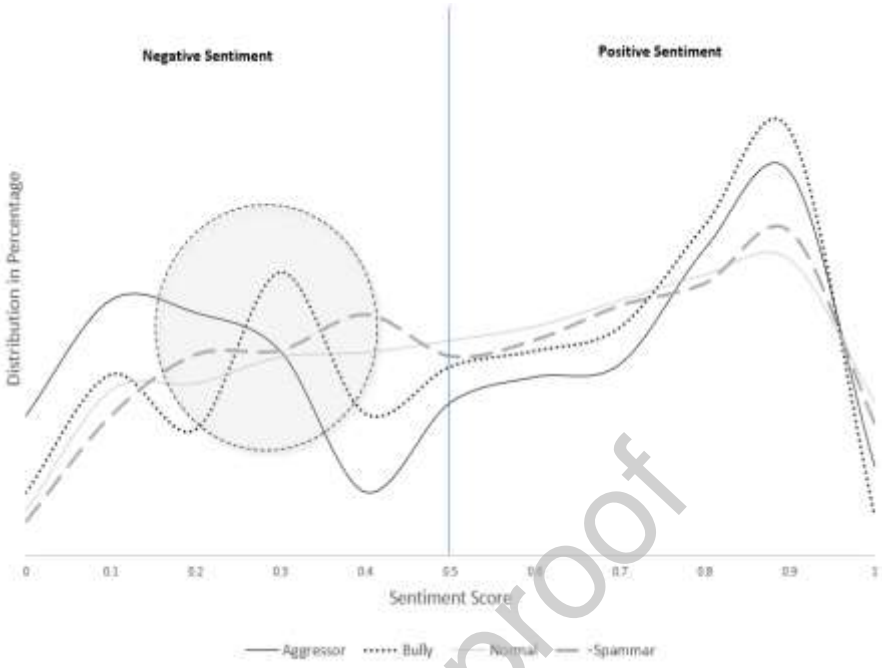

Figure 3: Key Features versus Baseline + Personality + Sentiment