The Journal of Social Media in Society
Fall 2022, Vol. 11, No. 2, Page 209-235
thejsms.org

A Machine Learning and Qualitative Examination of Cyberbullying Disclosures on Twitter

Karla Dhungana Sainju[1*], Lisa Young[1], Akosua Kuffour[1], and Niti Mishra[2]
[1]Faculty of Social Science and Humanities, Ontario Tech University, Oshawa, ON, Canada
[2]Rotman School of Management, University of Toronto, Toronto, ON, Canada
*Corresponding Author: karla.dhungana-sainju@ontariotechu.ca, 905-721-8668 ext.5809

Although clear links exist between social networking sites (SNS) and cyberbullying, limited studies have examined the content of Twitter to better understand cyberbullying characteristics. This study provides one of the few explorations of cyberbullying using Twitter data. Using supervised machine learning, it analyzes the disclosure of cyberbullying episodes on Twitter to understand who is posting and why they are posting about cyberbullying. Additionally, a qualitative content analysis of 500 tweets provides further insights into the characteristics of cyberbullying episodes. The findings reveal that aside from serving as a medium for cyberbullying, Twitter is also a space for bystanders to engage in 'upstander' behavior and where victims make connections and receive validation. We also found that cyberbullying targets extend beyond the peer group; random strangers, celebrities, and entire groups are victimized on SNS platforms via multiple forms of cyberbullying. The paper discusses how SNS platforms can become a part of the fight against cyberbullying.

Keywords: cyberbullying, machine learning, qualitative content analysis, bullying roles, Twitter, social media

---

Bullying is a well-researched global issue (UNESCO, 2019). Traditional or face-to-face (F2F) bullying is often defined as aggressive and unwanted behavior characterized by an intent to harm, repetition, and an imbalance of power (Olweus, 1993). Similar to F2F bullying, cyberbullying is identified as intentional and repeated unwanted aggressive behavior; however, it is inflicted electronically through mediums such as computers, tablets, or cell phones (Hinduja & Patchin, 2015; Smith et al., 2008). Kowalski et al.'s (2014) meta-analysis of 131 cyberbullying studies and the growing body of cyberbullying studies that have been added since indicate that the cyberbullying literature is also strong and well versed. However, most prior studies of cyberbullying tend to rely on self-report data (Kowalski et al., 2014). While this approach has yielded a strong understanding of cyberbullying, it is often

limited to the perspectives of the bully and victim. It also poses several limitations including small sample sizes and issues around response and recall bias. Big data and in particular social media data, provides an innovative way to expand on our understanding of cyberbullying and allows us to address these limitations.

Social networking sites (SNS) such as Facebook, Twitter, and Instagram have all been reported as common spaces for cyberbullying perpetration (Newall, 2018; UNICEF, 2019; Whittaker & Kowalski, 2015). Researchers have explored the prevalence rates of cyberbullying on SNS platforms (Hamm et al., 2015; Gahagan et al., 2016), investigated the relationship between online activities and cyberbullying (Park et al., 2014), examined the association between problematic social media use, psychosocial factors, and cyberbullying (Kırcaburun et al., 2019), and studied the association between social media use, cyberbullying and mental health (Viner et al., 2019). However, for all the attention given to understanding the link between SNS and cyberbullying, surprisingly little has been done to study the digital content of the Twitter platform to understand bullying and cyberbullying. Twitter allows us to collect high-volume information about people's digital social interactions. The ability to collect semantically rich information from a geographically diverse sample in real-time and the opportunity to analyze behavior that is neither prompted nor solicited, suggests that Twitter SNS data could significantly advance our understanding of cyberbullying.

Launched in 2006, Twitter is a micro-blogging SNS that allows users to post short messages, or tweets, which are limited to 280 characters. With approximately 330 million active monthly users worldwide (Clement, 2019), Twitter provides the most comprehensive source of public conversation allowing users to follow any incident or topic in real-time. Additionally, unlike other SNS sites, the primarily public-facing nature of Twitter coupled with the accessibility of Twitter data makes it a timely and relevant social media source to examine the issue of cyberbullying. In this study, we provide one of the few explorations of cyberbullying through the lens of Twitter (Dhungana Sainju et al., 2021a, 2021b; Blanco et al., 2014; McHugh et al., 2019) and merge social science, computer science, and SNS data to enhance our understanding of cyberbullying. In addition to expanding the literature on cyberbullying by exploring Twitter data, the study also addresses the limitations associated with self-report data by utilizing data that is neither

prompted nor solicited. The use of Big Data also allows access to a larger sample size of Twitter users worldwide in real-time. For the first part of the study, 69,963 cyberbullying-related tweets and validated standard machine learning and language processing methods will be utilized to explore two questions related to cyberbullying experiences shared on Twitter: 1) who is posting about cyberbullying on Twitter? and 2) why are people posting about cyberbullying on Twitter? These questions are crucial to examine as they provide an exploration beyond the dyadic relationship of the victim and bully and examine additional bullying roles related to cyberbullying. To our knowledge, this study would be one of the first to utilize machine learning and Twitter data to explore different bullying roles and perspectives related specifically to cyberbullying. Secondly, a qualitative content analysis of 500 randomly selected tweets further explores the characteristics of cyberbullying discourse observed on Twitter. In all, this study contributes to the extant literature by providing one of the first studies to both quantitatively and qualitatively use Twitter data to identify key characteristics related to cyberbullying roles, the victim-bully relationship, and the various types of cyberbullying behavior.

LITERATURE REVIEW

Defining Cyberbullying

While the definition for both F2F and cyber bullying refers to the characteristics of intentional harm, repetition, and power imbalance, these characteristics may look different for each form of bullying. Often observed through physical strength, stature, or age in F2F bullying, researchers have suggested that in cyberbullying the imbalance of power could be characterized by the anonymity online spaces can afford cyberbullies (Butler et al., 2010) or those who are media savvy and able to exploit the power of technology (Dooley et al., 2009; Smith et al., 2008). Findings from Nocentini et al. (2010) and Dredge et al. (2014) further suggest that in cyber spaces, an act may not have to be repeated since a single post may be shared with many users and the publicity of the act is akin to repetition. Surveys of adolescent victims also indicate that their understanding of bullying online may include criteria not identified within the standard definitions of cyberbullying. Dredge et al. (2014) found that the most reported characteristic used by victims to define cyberbullying was whether the experience had a negative impact on them. Relatedly, Hellström et al.'s (2015) study respondents also indicated that

irrespective of the intent, the victim's experience of hurt and harm determined their interpretation of bullying.

Cyberbullying Roles

Studies have found correlations between offline and online roles; those who bully in F2F settings also tend to bully online (Kowalski et al., 2014) and victims of cyberbullying also report being victimized in person (Waasdorp & Bradshaw, 2015). Studies also indicate that most cyberbullies and victims are familiar with each other in real life (Dhungana Sainju et al., 2021b; Newall, 2018; Waasdorp & Bradshaw, 2015). However, given the broad scope of online platforms, studies have found that cyberbullies also target individuals beyond their peer group. Study participants admit to both harassing (Dowell et al., 2009) and being the victim of harassment (Waasdorp & Bradshaw, 2015) by strangers online. Pyzalski (2011) extended the victim-bully relationship beyond the peer group and introduced a typology based on the victim's identity; he found that while most young people cyberbullied individuals only known online, cyber aggression was also targeted towards random people, against groups, and celebrities.

Moreover, there are bullying roles beyond the bully and victim; additional bystander or participant roles that reinforce the bully, help the victim, or who alter their behavior according to the peer and social contexts have been identified (Lambe et al., 2019; Levy & Gumpel, 2018; Salmivalli et al., 1996; Wójcik & Flak, 2019). Findings from a meta-analysis examining the effects of school-based bullying prevention programs found that programs that viewed bullying as a group process and encouraged active and prosocial bystander behavior had an increased likelihood of bystander intervention (Polanin et al. 2012). This implies that while bystanders may be a part of the problem, they can also be a key part of the solution (Salmivalli, 2010).

Reasons for Posting and Sharing Online

The uses and gratifications theory explores why individuals use specific types of media and the needs associated with using them (Katz et al., 1973). Social media research based on this theory suggests that SNSs may fulfill several motives including the need to connect with others (Chen, 2011), navigate different social relationships (Marwick & Boyd, 2011), express thoughts and opinions, information sharing, and use social media to watch others (Whiting & Williams, 2013). Moreover, prior studies indicate that SNS platforms

such as Twitter can also serve as an important venue to stand up against online and offline bullying behavior, validate a bullying victim's experience, and support social movements and activism efforts (Dhungana Sainju et al., 2021a; Freelon et al., 2016; McHugh et al., 2019; Mundt et al., 2018).

Different Types of Cyberbullying

Prior research indicates that cyberbullying encompasses a range of behaviors. Exclusion and ostracism is the deliberate act of leaving someone out from a group (Willard, 2007). Outing or doxing happens when a bully shares personal or embarrassing information about someone without their consent to shame or publicly humiliate them (ETCB, n.d.; Willard, 2007). Masquerading occurs when a bully creates a fake profile or online persona to cyberbully someone (Securly, 2020). Harassment is a broader category of cyberbullying but refers to a repetitive pattern of sending offensive, rude and insulting messages (Securly, 2020; Willard, 2007). Flaming is characterized as an online fight exchanged via emails, messaging, or chat rooms (Willard, 2007). Lastly, trolling occurs when a bully intentionally tries to upset a victim and provoke a response by posting offensive or inflammatory remarks online (Securly, 2020).

Location of Cyberbullying

Cyberbullying can be perpetrated through various technological mediums such as SNS platforms, text messages, photos or video clips, phone calls, emails, chat rooms, instant messaging, and websites (Kowalski & Limber, 2007; Smith et al., 2008; Whittaker & Kowalski, 2015). Over time, research has shown the evolution and popularity of various platforms; Kowalski and Limber (2007) found instant messaging to be the most common venue for cyberbullying, Katzer et al. (2009) noted chat rooms as a popular venue for cyberbullying, and Whittaker and Kowalski (2015) reported Twitter and Facebook to be the most common online platforms for being cyberbullied. A more recent 2019 UNICEF poll indicated that almost three-quarters of youth reported that SNS platforms such as Facebook, Instagram, Snapchat, and Twitter were the most common venues for cyberbullying (UNICEF, 2019).

Cyberbullying Research Using Twitter Data

Despite the clear connection between Twitter and cyberbullying, limited empirical examinations exist using Twitter data. In the last decade, researchers primarily from the

computer science field have used machine learning methods to automate and detect aggression, hate speech, and cyberbullying within tweets (Blanco et al., 2014; Waseem & Hovy, 2016). A few more have utilized similar methodologies to identify and classify bullying participant roles within tweets and suggest that there are online disclosures about bullying from additional roles beyond the victim and bully (Bellmore et al. 2015; Chatzakou et al., 2017; Dhungana Sainju et al., 2021a; Xu et al., 2012). To our knowledge, only two studies have qualitatively examined bullying-related tweets. First, McHugh et al. (2019) utilized computer-assisted sentiment analysis to qualitatively examine 300 cyberbullying-related tweets. They found that most tweets were contributing to a negative atmosphere and were referring to known individuals and ongoing events. A second study by Dhungana Sainju et al. (2021b) qualitatively analyzed 780 bullying-related tweets to examine the characteristics related to perpetrators, targets, and helpers on Twitter. Their findings suggest that most bullying role players know each other, virtually or in real life, and they tweet about both current and past episodes of bullying. These studies all point to the efficacy of utilizing Twitter data to gain a better understanding of bullying characteristics on SNS. However, it also indicates that there are still unexplored aspects, with virtually no studies that examine the cyber victim and bully relationship or identify the different forms of cyberbullying using Twitter data.

Given what we know about bullying roles and how they can change according to the peer and social context ((Lambe et al., 2019; Levy & Gumpel, 2018; Salmivalli et al., 1996; Wójcik & Flak, 2019), the machine learning portion of the study aims to better understand how these roles translate specifically to cyberbullying and proposes the following hypotheses and research questions:

R1: Who is posting about cyberbullying on Twitter?

H1: Cyberbullying disclosures on Twitter will come from roles beyond the victim and the bully and will include additional participant roles such as those that reinforce the bully and those that help the victim.

Next, to expand on our understanding of cyberbullying roles and informed by the uses and gratifications theory (Katz et al., 1973), the machine learning models will also explore:

R2: Why are people posting about cyberbullying on Twitter?

H2: Twitter users will post about cyberbullying for varied reasons including to self-dislcose their own experiences, support bullying victims, validate victims' experiences, and call out bullying behavior.

The qualitative content analysis will also rely on prior literature to take an inductive approach to inform a directed content analysis and identify conceptual categories related to cyberbullying. The victim-bully relationship will be examined using a modified version of Pyzalski's (2011; 2012) victim-perpetrator typology to explore:

R3: What is the victim-bully relationship?

The remaining qualitative analysis will be guided by the following broad questions:

R4: Where does the cyberbullying behavior take place?

R5: What type of cyberbullying is being carried out or mentioned?

R6: When is the timeframe of the cyberbullying incident?

METHODS

Data Collection

The data for this study was collected as part of a larger data collection (see Dhungana Sainju et al., 2021a) via Twitter's streaming Application Program Interface (API), a free and automated retrieval service that allows access to up to 1 percent of the population's tweets. Tweets were collected using a list of primary keywords "*bullied, bully, bullying, cyberbullied, cyberbully, and cyberbullying*." Understanding that this would generate a large number of tweets, some of which may not be referring to a discrete bullying episode, the tweets were further filtered using secondary keywords. Using a conceptual concept analysis, where the number of times a word appears is quantified (Christie, 2007), previously conducted studies were scanned to examine middle school and high school students' written descriptions of bullying, and the following secondary keywords were identified "*mean, force, forced, text, texted, online, laugh, laughing, exclude, excluded, exclusion, destroy, destroyed, force, forced, spread, rumor, rumour, embarrass, embarrassing, embarrassed, repeatedly, repeat, mock, mocked, mocking, tease, teasing, teased, ignore, ignored, ignoring, hitting, and hit."* A previous study by Bellmore et al. (2015) utilized a similar approach, however, their study only used a limited number of secondary keywords that were only related to F2F bullying. To identify secondary words

related to cyberbullying, a content analysis was conducted on 10- to 18-year-old youth's interpretation of cyberbullying, which identified the following secondary keywords "*gossip, gossiped, manipulate, manipulating, manipulated, mislead, misleading, humiliating, humiliated, revenge, insult, insulted, anonymous, group text, and group chat."* Lastly, to further expand on the secondary keywords identified through the process noted above, the study authors also added the following secondary keywords "*isolate, isolated, social media, rejection, reject, rejected, aggressive, intimidate, intimidated, jealous, assault, harass, shove, shoved, pretend, pretended, scare, scared, shun, shunned, target, targeted, beat, insult, insulted, threat, threatened, bash, bashing, degrade, degrading, perpetrated, perpetrator, defending, stressful, bystander, coercion, suicide, Facebook, Snapchat, Instagram, and WhatsApp*."

Tweets were collected between August 7, 2019, and March 31, 2020. Following the methodology utilized by Bellmore et al. (2015) and Xu et al. (2012), only tweets that matched both a primary and secondary keyword were included. This was done to identify an "enriched dataset" where tweets that referred to a discrete bullying episode were more likely to be identified (Bellmore et al., 2015; Xu et al., 2012). Then, additional data processing steps were conducted to clean the dataset and remove spam accounts. Tweets that were re-tweets, non-English tweets, tweets that only included a URL, and tweets with six or more hashtags were removed. At the time of data collection, there were also a large number of tweets that mentioned the former U.S. President Donald Trump along with the word '*bully.*' According to McIntire et al. (2019), in 2019 Mr. Trump was tagged approximately 1000 times per minute on Twitter. Thus, to clean the dataset of any political references or debates, tweets that contained the terms "*realdonaldtrump, trump, whitehouse, white house, potus, flotus, president, @realdonaldtrump, @whitehouse, @potus, @flotus*" were also removed. After the keyword filtering and data processing, a total of 847,548 tweets were retained. Next, the tweets were tokenized, and using NLTK's part-of-speech (POS) each token was tagged to identify the lexical category of the tweet (Bird et al., 2019). Additionally, URLs and user mentions were replaced with placeholders, hashtags were converted to a single token, and each token and its tag were lemmatized using the NLTK's WordNet Lemmatizer (NLTK Project, 2020). Lastly, unigram (one token) and bigram (two consecutive tokens) were used to transform the tweets into a TF-

IDF matrix (Pedregosa et al., 2011) which was used for the machine learning model explained below.

Identifying Cyberbullying-Related Tweets

Subsequently, using the dataset with primary and secondary keyword-matched tweets, we set to classify whether a tweet was a bullying trace. Guided by methodologies used in Xu et al. (2012) and Bellmore et al. (2015), any tweets where the author referred to a discrete bullying episode, where they were participating in or reporting bullying behavior, was classified as a bullying trace. Similar to Xu et al. (2012) and Bellmore et al. (2015), tweets were also taken at face value and did not follow the traditional definition of bullying, which tends to include features such as an imbalance of power and repetition (Olweus, 1993). Two of the study authors coded 7,868 randomly selected tweets and based on 1000 coded tweets, a percentage agreement of .79 and Cohen's kappa of κ =.59 was identified. The same set of human coders identified 3,096 or 39.35% of the randomly selected tweets as bullying traces and 4,772 or 60.65% were labeled as non-bullying traces. Following, 80% of the coded tweets were divided into the training set and the remaining 20% was put aside as the test dataset to test the skill of the final model.

Then using TF-IDF-based natural language processing methods and logistic regression and support vector machines (SVM) machine learning algorithms were used to classify the tweets and each model was trained using a stratified 12-fold cross-validation. The final model for each classification task was chosen based on the combination of the average F1 score and accuracy on the test set. This resulted in a 70% accuracy relative to the human coding. Relative to the baseline or the naïve model predicting only the majority class, or the non-bullying traces, which had an accuracy rate of 61%; this represents a moderate increase in skill. The final best model was then used to classify the remaining 847,548 unlabeled tweets. The machine learning models and natural language processing methods identified 28.58% or 240,018 of those tweets as bullying traces.

After bullying and non-bullying related tweets were categorized, we set to further classify the set of bullying traces to identify the form of bullying mentioned within the tweet. The two human coders labeled the 3,096 tweets identified as bullying traces to identify whether a tweet was related to general, verbal, physical, or cyberbullying. Tweets, where electronic forms of bullying were identified or mentioned, were labeled as

cyberbullying-related tweets. Similar to identifying bullying traces, the machine learning model was trained and tested to identify this classification. The human coders identified 1,060 or 34.30% of the labeled tweets as cyberbullying and the machine learning models trained on human coded tweets predicted 69,963 or 29.15% of unlabeled tweets as cyberbullying. This ML model attained 78% accuracy relative to the human coding and represented a significant increase in skill compared to the naïve or baseline model which had a 34% accuracy rate. Relatedly, Bellmore et al., (2015), whose methodology guides this study, also report similar predictive skills in the .70 range. This set of 69,963 cyberbullying-related tweets serves as the data for the current study.

Analytic Approach and Research Questions

Supervised Machine Learning. There are two parts to this study's analytic approach. To begin, using supervised machine learning methods and the set of 69,963 cyberbullying-related tweets; we set out to ask two research questions. First, research question 1 explored: who is posting about cyberbullying on Twitter? Informed by bullying roles identified in prior studies (Bellmore et al., 2015; Salmivalli et al. 1996; Xu et al., 2012), the role of the tweet author was classified into one of six categories. A "*victim*" was someone who referred to being bullied in the past or an ongoing event. A "*bully*" engaged in bullying behavior in the tweet or admitted to past bullying. A "*reporter*" was not involved in the bullying but shared information about a bullying episode. A "*defender*" took the victim's side and stood up against a bully. An "*assistant*" joined the bully and participated in the bullying. A "*reinforcer*" did not engage in the bullying, but their behavior reinforced and encouraged the bully's behavior. Finally, an "*accuser*" accused someone of bullying, and it is unclear whether they were a victim, defender, or other roles. Our dataset revealed a small number of assistants and reinforcers, thus for the current analysis, these were combined into one group as "*other*." Research question 2 sought to examine why people posted about cyberbullying on Twitter. Following the methodology used by Bellmore et al. (2015), several reasons were identified. A "*report*" was where a tweet author shared information about a bullying episode they saw in the news or were personally familiar with. A "*self-disclosure*" is one where the author revealed themselves as the bully, victim, assistant, or reinforcer. In a "*denial*" post, the author denied involvement in a bullying incident. Finally, in an "*accusation*" post the author accused

someone of being a bully. The 1,060 tweets labeled as cyberbullying were coded by the same two human coders and each tweet was categorized to identify the author's role and the reason for posting.

Qualitative Analysis. Secondly, a qualitative content analysis was conducted on a subset of tweets to explore the characteristics of cyberbullying discourse observed on Twitter. Using a directed content analysis approach and guided by research questions 3-6, the study authors began with decontextualisation, or open coding, which represented the first step of the process (Bengtsson, 2016). Tweets were taken at face value and, the context of each tweet was examined by the study authors to identify key themes and create an initial coding scheme. This process revealed conceptual categories related to the victim-bully relationship, forms of cyberbullying behavior being carried out or mentioned, the location of the cyberbullying behavior, and the timeframe of the episode. Then, the two human coders who labeled the tweets for the machine learning models independently coded a set of 25 randomly selected tweets to determine the level of agreement between coders. The labels were reviewed and discussed at length and new codes were added as needed. Two additional rounds of coding were conducted with 25 new randomly selected tweets per round until the agreement on all categories was 80% or higher. During the recontextualization phase (Bengtsson, 2016), the authors checked to make sure that the coding system was exhaustive and reflected the goals of the study. Once the coding categories were finalized, the qualitative software program Dedoose (Dedoose, 2018) was used for the categorization process and each of the coders labeled a new set of 250 tweets each, for a total of 500 randomly selected cyberbullying-related tweets. After all the tweets were labeled, the coded tweets were reviewed to ensure agreement with the coding. See Table 1 for a full list of the categories and coding scheme.

Table 1
Qualitative categories and coding scheme

Category	Codes and Definitions
Who: Victim-Bully Relationship	*Cyber aggression against peers*: The victim and bully know each other and are from the same offline/online group *Cyber aggression against the vulnerable:* The victims are vulnerable groups such as the homeless or alcoholics who may not be aware of the victimization *Cyber aggression against groups*: The victims are groups rather than a specific individual. For example, fans of a sports team or musical artist, ethnic or religious groups *Random cyber aggression*: Aggression perpetrated against anonymous victims not known to the bully *Cyber aggression against celebrities or public figures*: Aggression targeted towards celebrities or public figures not personally known to the bully
Where: Location of Cyberbullying Behavior	*Online (general)* *Twitter* *Instagram* *Facebook* *Snapchat or other group chat* *YouTube* *Gaming site* *Offline but being discussed and continued on Twitter* *Offline and online* *Not known*
What: Type of Cyberbullying	*Exclusion and ostracism:* Deliberate act of leaving someone out from a group *Outing or doxing:* sharing personal or embarrassing information without someone's consent to deliberately humiliate them *Masquerading:* Creating a made-up profile or identity online with the sole purpose of cyberbullying *Harassment*: A broad category referring to a repetitive and sustained pattern of sending rude, harassing messages online *Flaming*: An online fight exchanged via emails, messaging, or chat rooms *Trolling*: Purposefully trying to upset a victim and provoke a response by posting offensive or inflammatory information online *Not known*
When: Timeframe of Cyberbullying	*Ongoing:* Tweet referring to cyberbullying that is currently happening *Past:* Tweet refers to a cyberbullying incident that has already occurred *Not known*: Tweet does not include a time frame

RESULTS

Supervised Machine Learning

R1: Who is Posting About Cyberbullying on Twitter? The supervised machine learning model classified 69,963 cyberbullying-related tweets to identify the author of the tweet using the categories noted above in the methods section. Similar to identifying the bullying trace, to identify the tweet author the machine learning models were trained and tested using human-coded tweets. Varied roles including those that supported the victim and reinforced the bully were found to support hypothesis 1. As seen in Figure 1, the most common role predicted was a defender (36% or 24,793 tweets) which was someone who was taking the victim's side and standing up against the bully. This was followed by a reporter (30% or 21,227 tweets) who was referring to either a bullying event in the news or an event they were personally familiar with. Victims accounted for a little more than a quarter (26% or 18,726 tweets) of tweet authors, sharing about their own bullying victimization experiences. Accusers made up 7% (4,968 tweets) of the tweet authors, directly accusing someone of bullying in their post. Lastly, bullies (0.27% or 195 tweets) and others (0.07% or 54 tweets) (assistants and reinforcers) made up the smallest category, representing less than half a percentage point each.

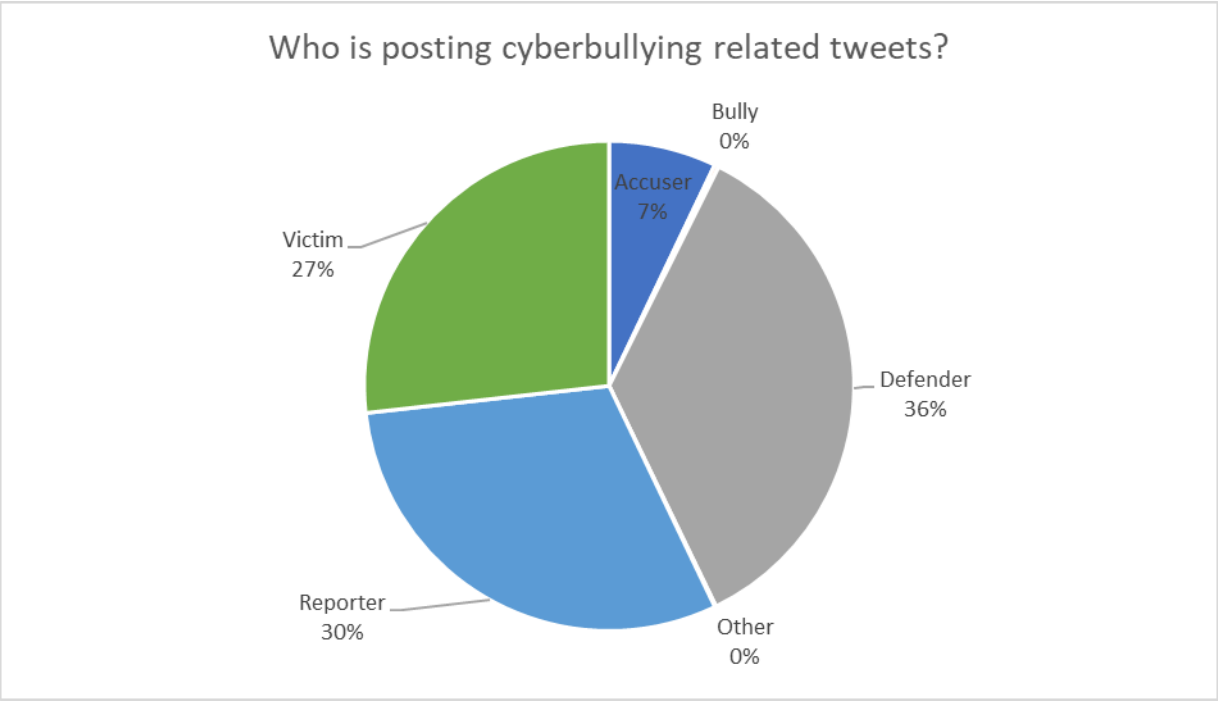

Figure 1. Who is posting cyberbullying-related tweets?

R2: Why are People Posting About Cyberbullying on Twitter? Next, we aimed to identify why someone was posting about cyberbullying on Twitter using the categories discussed in the methods section. Support was also found for hypothesis 2. Of the 69,963 tweets identified as cyberbullying-related tweets, the machine learning model found that the most common reason for posting was to report, with more than 4 in 10 (44.4% or 30,976) authors tweeting to share information about a cyberbullying episode. This was followed by accusations, which made up a little more than one-third (34.3% or 24,045) of the tweets. Finally, self-disclosures accounted for 21.3% (14,942) of the tweets, where the tweet author was disclosing their role in the cyberbullying episode. No denial posts were captured. Examining the convergence of the "who" and the "why" analysis reveals that the most common cyberbullying-related tweets were reporters reporting, defenders accusing, and victims' self-disclosing. As Figure 2 shows, reports were most common from reporters, followed by defenders and victims. Accusations were most likely to come from a defender, although a small number of accusers, victims, and reporters were also tweeting to accuse. Finally, a self-disclosure tweet was most likely to come from a victim.

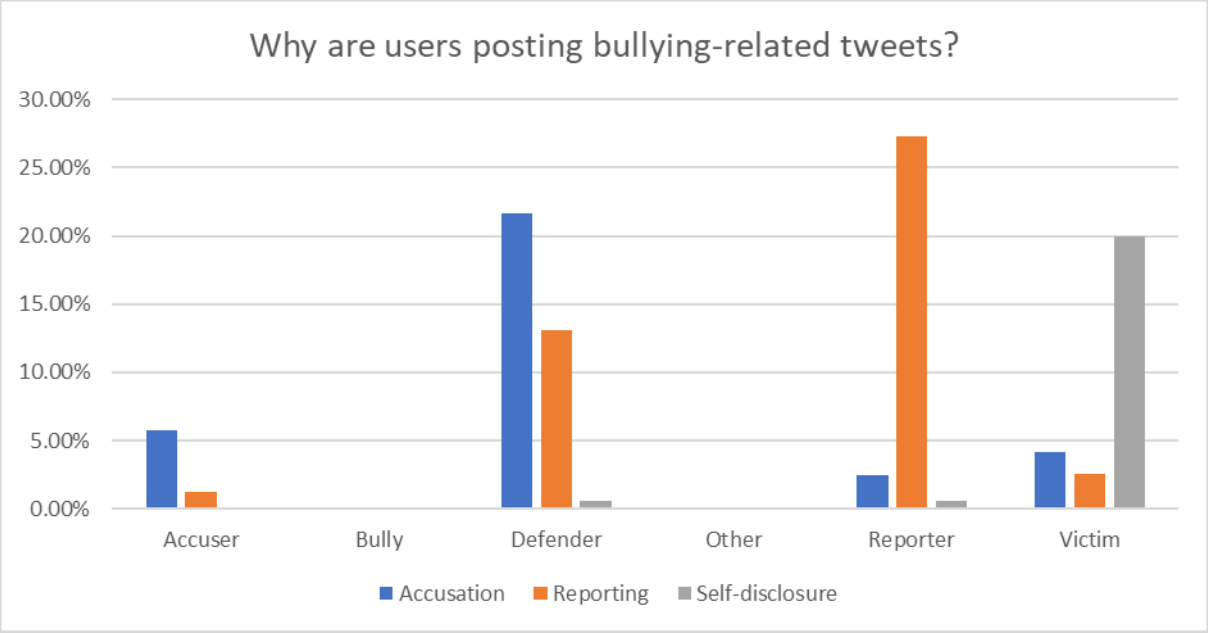

Figure 2. Who and why are people posting about cyberbullying on Twitter?

Qualitative Analysis

R3: Who: Victim-Bully Relationship. Based on a modified version of the victim-perpetrator typology introduced by Pyzalski (2011; 2012), tweets were coded to classify the victim-bully relationship. More than half (52.83%) of the cyberbullying-related tweets were coded as random acts of cyber aggression between a bully and a victim not personally known to them, as seen in this tweet "*People are dragging this stranger on the other side of the world, is she so offensive to you that you have to go out of your way to bully her online.*"[1] Almost a quarter (23.48%) were referring to cyber aggression among peers or known individuals from the same offline or online group. For example, "*I like to bully @user on snapchat.*" 11.94% of tweets suggested cyber aggression being targeted at celebrities or public figures, in tweets such as "*These conservatives rallying online to bully a 15 year old climate change activist is so sad"* and "*If I see another article from Piers Morgan about Harry and Megan I will report it since its clearly bullying and pathetic.*" About 1 in 10 (9.71%) of tweets were cyber aggression against entire groups rather than a specific individual, which could be seen through tweets such as "*I enjoy bullying boomers online*" and "*the amount of bullying and hate from our fandom for that one tweet makes me sick.*" Only 2.02% were tweets that implied cyber aggression against the vulnerable. For example, "*why are ya'll bullying a child who is already suicidal*," and *"@user is cyber bullying someone with a disability and thinks its funny, you don't deserve to be in our community.*"

R4: Where: Location of Cyberbullying Behavior. Each tweet was examined to determine if the author mentioned where the cyberbullying occurred. More than half (51.10%) of the cyberbullying-related tweets did not indicate a specific location and referred to online spaces in general. About 3 in 10 (29.17%) specifically mentioned Twitter or were engaging in cyberbullying on Twitter. In about 5.23% of tweets, there was not enough information to determine the location. Instagram was mentioned in 4.22% and Facebook in 3.42% of tweets. Offline behavior that was continued and discussed on Twitter was seen in 2.21% of tweets, while 1.81% of tweets reflected both online and offline bullying. Another 1.2% mentioned gaming sites. Finally, 0.8% of tweets noted

[1]To avoid traceability of an example tweet, all user names have been removed and tweets have been shortened or slightly paraphrased instead of including it verbatim.

cyberbullying occurring on Snapchat or other group chat and 0.8% mentioned cyberbullying on YouTube.

R5: What: Type of Cyberbullying. Tweets were also coded to identify the type of cyberbullying behavior. Almost 4 in 10 tweets did not have enough information or context for us to determine the type of cyberbullying behavior that was being displayed. The remainder of the tweets, however, reflected a range of cyberbullying behaviors. 28.05% of tweets emphasized harassment or a pattern of hurtful and harassing messages. 12.82% of tweets suggested flaming or an online fight. The behavior of trolling was seen in 11.42% of tweets. Outing or doxing was reflected in 2.80% of tweets and finally, only 1% of tweets revealed characteristics of exclusion.

R6: When: Timeframe of Cyberbullying. Finally, tweets were examined to determine the timing of the cyberbullying incident. 5 in 10 (51.81%) tweets were referring to an ongoing cyberbullying episode. A quarter (25.80%) were discussing a past cyberbullying event, whereas 21.57% did not have enough information to determine the time frame. A small percentage (0.80%) was referring to a future cyberbullying event. See Figure 3 for a full list of cyberbullying-related tweet characteristics.

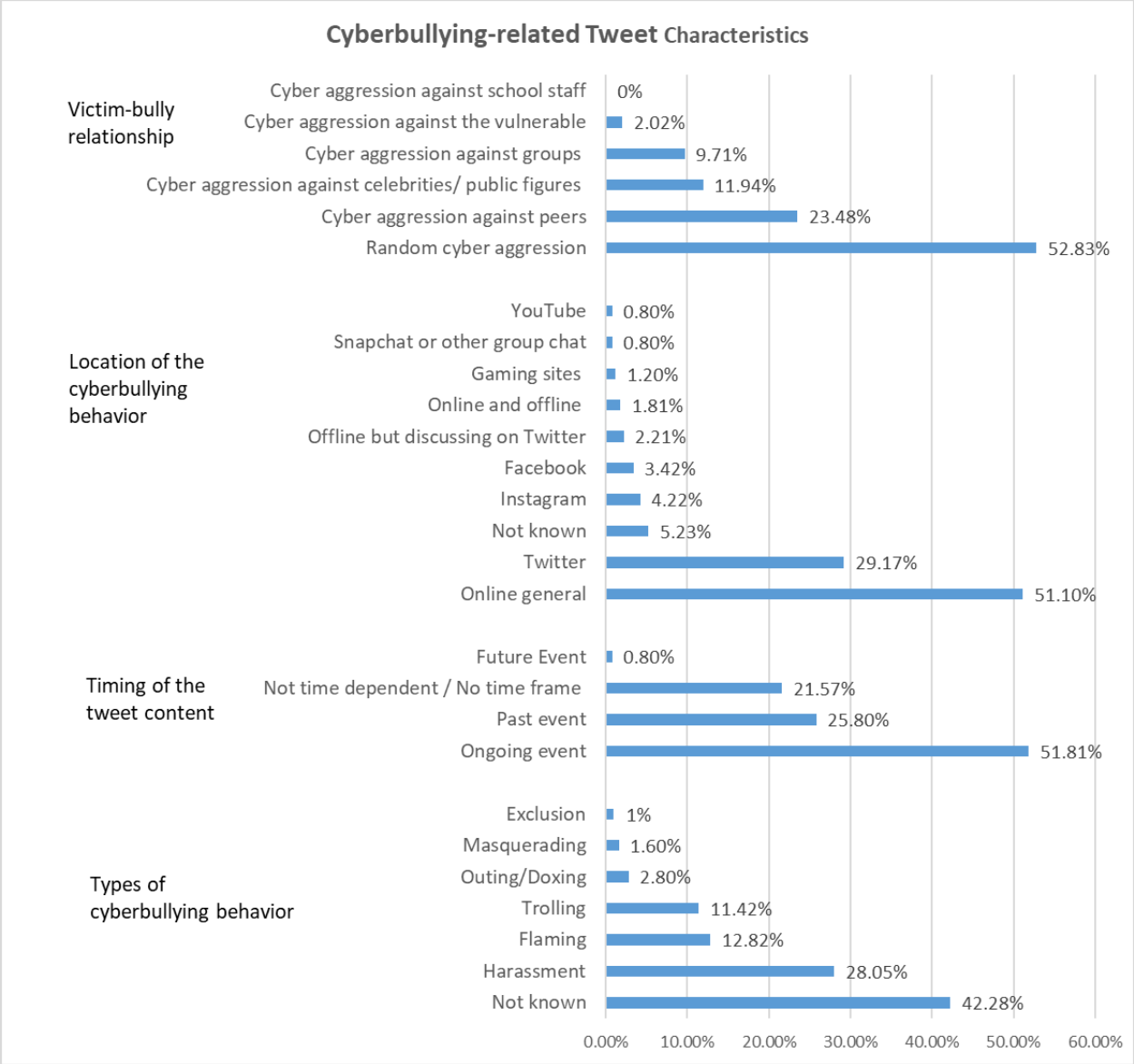

Figure 3. Cyberbullying-related tweet characteristics

DISCUSSION

This study is one of the first of its kind to utilize machine learning and qualitative content analysis to identify key characteristics related to cyberbullying discourse on Twitter. A key strength and contribution of the study was the opportunity to examine tweets at face value and to be able to analyze unprompted and unsolicited digital social interactions rather than self-reported conduct. Rather than relying on traditional definitions of cyberbullying, the study allowed us to analyze tweets where Twitter users posted first-hand experiences and interpretations of cyberbullying. The findings reveal a

diverse range of bullying roles and reasons for tweeting about cyberbullying incidents. Reporters were reporting about cyberbullying episodes, defenders were standing up for a victim and accusing a bully, and victims felt safe to self-disclose and share their own experiences. While a little less than a third of tweets referred to cyberbullying occurring on Twitter, the analysis revealed that users were also tweeting to mention cyberbullying happening on other SNS platforms and electronic mediums. Additionally, about half of the tweets were referring to ongoing incidents and a quarter was reflecting on a past episode. Over half of the tweets suggested random aggression against strangers; however, cyber aggression was also aimed at known peers, celebrities, specific groups, and the vulnerable. The analysis also found an assortment of cyberbullying behaviors being perpetrated or referred to in the tweets.

The tweets in the current study primarily captured the experience of being cyberbullied rather than cyberbullying itself. This can largely be attributed to the keyword selection as cyberbullies are not likely to use the primary keywords such as cyberbullied, cyberbully, or cyberbullying when engaging in cyberbullying behavior. Rather, our findings point to victims and their supporters actively using Twitter as a space to report, accuse, and self-disclose. Bellmore et al. (2015) and Dhungana Sainju et al. (2021a) also found similar results and hardly any bully authored tweets were captured in their analyses. This is not to say that cyberbullying does not occur on Twitter. Almost a third of tweets referred to bullying happening on Twitter which supports prior reports about Twitter being a venue for cyberbullying (Whittaker & Kowalski, 2015; Unicef, 2019). Taken with the finding that about half of the tweets were referring to ongoing events, and considering the uses and gratification perspective (Katz et al., 1973), it indicates that while Twitter users experience cyberbullying on the platform, it is also creating a safe space for victims, reporters, and defenders to make connections in real-time, receive validation for their experiences, and engage in prosocial bystander behavior. This emotional reciprocity, of making connections and responding empathetically to others, can have important implications. Sharing emotional and negative experiences can be therapeutic (Wagner et al., 2015), and receiving empathetic and validating responses can improve emotional regulation and reactivity (Shenk & Fruzzetti, 2011).

It appears that SNS and online platforms also create opportunities to cyberbully beyond one's peer group. The tweets referred to multiple SNS platforms besides Twitter, and potential targets of cyber aggression included both known and unknown individuals. These findings are in line with Pyzalski's (2012) survey of adolescents which found that they were most likely to perpetrate cyber aggression against people only known online. The study respondents also noted targeting groups of people, celebrities, and random or unknown individuals. Similarly, Whittaker and Kowalski (2015) also tested out a modified version of Pyzalski's typology and found that, when perpetrating cyber aggression, participants directed it most often towards random people and least often towards peers. When asked about witnessing cyber aggression, the respondents noted seeing the most comments targeted towards celebrities and the least comments targeted towards peers. This could imply that characteristics relevant to online spaces, including a measure of anonymity afforded to cyberbullies, increased accessibility to potential targets, the viral nature of online posts, and not being able to see the victim's reaction may embolden and encourage behavior beyond what would normally be done in person (Della Cioppa et al., 2015; Hinduja & Patchin, 2015; Tokunaga, 2010; Twyman et al., 2010).

Relatedly, our findings also point to the association between cyber aggression and celebrities. Public attention and media scrutiny have been part and parcel of celebrity lives. By putting themselves on social media platforms, celebrities are opening themselves up to engage in a form of parasocial relationship, associations that are often one-sided where fans develop a sense of intimacy and connection to the celebrity (Stever & Lawson, 2013). The ease and accessibility of Twitter mean that any user can tag, mention, or direct message a celebrity in an attempt to gain their attention. The attention, however, is not always positive. Many celebrities have been subject to online trolls, public shaming, or the increasing push to "cancel" celebrities in response to objectionable behavior or comments (BBC, n.d.; Romano, 2019). Indeed, in addition to our current findings, results from Pyzalski (2012), Whittaker and Kowalski (2015), McHugh et al. (2019), and Dhungana Sainju et al. (2021b) all indicate that SNS platforms are being used to perpetrate cyber aggression against celebrities. While one could argue that celebrity harassment comes with the territory, celebrities are human and the impacts are real. Celebrities such as Ed Sheeran, Zayn Malik, Demi Lovato, and Lizzo have all reported suffering from online

trolls and temporarily or permanently deleted their social media accounts (Vanderberg, 2020). In more extreme and tragic situations, it may have also been a contributing factor in the recent suicides of British TV personality Caroline Flack (Picheta, 2020), K-Pop star Sulli (McCurry, 2019), and the Japanese wrestler and Netflix reality show star Hana Kimura (BBC, 2020).

Lastly, while a majority of tweets did not contain enough information to determine the type of cyberbullying, the results found harassment to be the most common form, followed by flaming and trolling. Similar to our findings, Staude-Müller et al.'s (2012) online victimization survey revealed that verbal and sexual harassment was most common, whereas outing, exclusion, impersonation, and cyberstalking were less frequent. However, they also found that the less common forms were more emotionally distressing for victims. Likewise, Wolak et al.'s (2007) results indicate that a majority of online harassment was not distressing to targets, and incidents that involved repetition, asking for pictures, and involving online-only contacts who were 18-or older were more likely to be associated with distress. This suggests that not all forms of cyberbullying may trigger similar reactions among victims. El Asam and Samara (2016) also suggests that advances in technology and technological skills could change cyberbullying trends over time so it is crucial to continue to examine the types of cyberbullying and understand the consequences of differing cyberbullying behavior. Moreover, as researchers such as O'Sullivan and Flanagin (2003) and Wolak et al. (2007) aptly point out, clearer conceptualization and effective measures are needed to identify common cyberbullying behaviors such as harassment and flaming as current definitions vary significantly.

Study Implications

Our findings point to several important implications for researchers, educators, and policymakers. It is evident that analyzing social media data can complement and strengthen prior cyberbullying studies that utilize self-reported data. By analyzing digital social interactions, we discovered that bystanders on Twitter are frequently taking on the role of an 'upstander,' or someone who intervenes and takes action to stop or call out bullying (Padgett & Notar, 2013). Prior research tells us that upstanders are an integral part of stopping bullying; they are easier to influence than bullies (Salmivalli, 2014) and they have a high likelihood of quickly stopping the bullying behavior (Hawkins et al.,

2001; Polanin et al., 2012). We also found that Twitter encourages victims to speak out. Socially sharing our emotional experiences with others and in turn receiving validating responses have been found to increase positive emotional regulation (Shenk & Fruzzetti, 2011; Wagner et al., 2015). Given the global reach and popularity of Twitter, anti-bullying initiatives should consider how upstander behavior and empowering victims could be leveraged and promoted on the SNS platform. Our findings also point to the importance of further understanding and addressing aggression targeted toward celebrities and public figures. Future examinations should analyze how to discourage and combat online trolls and explore potential ways in which celebrities can serve as advocates and upstanders against cyberbullying behavior. Finally, our results indicate that additional empirical examinations are needed to understand the different types of cyberbullying behavior. Our study was limited to only examining the prevalence, however, future studies should utilize SNS data to better conceptualize the behavior to appropriately target the causes and consequences as well.

Limitations and Future Directions

This study adds an important contribution to the cyberbullying literature; however, there were some limitations. It is important to note that all machine learning models rely on the "ground truth" and need to have clear definitions to perform the classification tasks. The machine learning models for this study reported a moderate increase in skill compared to the naïve baseline models and the analysis was driven purely by keyword selection. While our keywords contained a more extensive keyword selection compared to similar prior studies (Bellmore et al., 2015; McHugh et al., 2019), as noted above it largely captured the experience of being bullied. Future studies should look to expand the keywords to capture both sides of the cyberbullying experience. Additionally, since no contextual or background information was available each tweet had to be coded at face value which meant that instances of sarcasm, harmless teasing, and more importantly, the intent of the author could not be fully captured. Analyzing qualitative data also runs the risk of the coder's cognitive biases influencing the coding process. Thus, all coded tweets were reviewed for agreement. No background or demographic information was available for the tweet authors, so we were unable to identify traits that were associated with a specific role or the likelihood of engaging in or experiencing a specific type of

cyberbullying behavior. Context also matters, and tweets often involve conversation threads where users can reply, retweet, and like a tweet. The current study only analyzed one tweet at a time, potentially missing out on valuable contextual information to situate the cyberbullying behavior. Future studies should attempt to analyze threads or conversations between users to get a fuller picture of cyberbullying discourse on Twitter. Given the complexity of human language and behavior, an increase in sample size would also improve the classification tasks allowing the machine learning models to access more examples to learn from and improve model accuracy and strengthen performance.

CONCLUSION

Given the proliferation of social media users worldwide, with one-in-three people in the world and more than two-thirds of all internet users using SNS (Ortiz-Ospina, 2019), our study points to the importance of analyzing direct digital interactions on SNS platforms such as Twitter rather than solely relying on self-report data to understand cyberbullying. The study is one of the first to quantitatively and qualitatively use Twitter data to explore cyberbullying characteristics. The objective of the study was to utilize machine learning methodologies and qualitative content analysis to identify key characteristics related to cyberbullying discourse on Twitter. Our findings confirm the utility and value of Twitter data to expand our understanding of cyberbullying roles, cyberbullying victims, cyberbullying locations, and cyberbullying behaviors. By exploring unpromoted and unsolicited disclosures of cyberbullying incidents, the study establishes that Twitter is not just a conduit for cyberbullying, it also serves as a space for reporters to report, defenders to accuse and victims' to self-disclose about past and current incidents. It is a dynamic space where strangers, celebrities, and everyday users can become targets of cyberbullying but can also become upstanders to intervene and curb cyberbullying behaviors. The machine learning methodology used in the current study can also be used to build improved classifiers for ML models and help refine the categories used to define cyberbullying and its characteristics. In all, the study provides promising implications of how Twitter and SNS platforms can become a part of the fight against cyberbullying.

References

BBC. (n.d.). 7 stars who have personal experiences of online bullying. *BBC Radio 1*. Retrieved from https://www.bbc.co.uk/programmes/articles/3QcD9W13Dr0bxmt4CMWVkGk/7-stars-who-have-personal-experiences-of-online-bullying

BBC. (May 23, 2020). Hana Kimura: Netflix star and Japanese wrestler dies at 22. Retrieved from https://www.bbc.com/news/world-asia-52782235

Bengtsson, M. (2016). How to plan and perform a qualitative study using content analysis. *NursingPlus Open*, 2, 8-14.

Blanco, K., Briceno, A. & Steele, A., Tapia, J., McKay, J., Towers, S. & Yong, K. (2014). The Dynamics of Offensive Messages in the World of Social Media: the Control of Cyberbullying on Twitter. *arXiv*:1408.0694 [cs.SI]

Bellmore, A., Calvin, A., Xu, J-M., & Zhu, X. (2015). The five W's of "bullying" on Twitter: Who, What, Why, Where and When. *Computers in Human Behavior,* 44, 305–314.

Bird, S., Klein, E., & Loper, E. (2019). Natural language Processing with Python: Analyzing text with the natural language toolkit. O'Reilly Media.

Butler, D., Kift, S., Campbell, M. (2009). Cyberbullying in schools and the law: Is there an effective means of addressing the power imbalance? *eLaw Journal*, 16(1), 84-114.

Chatzakou, D., Kourtellis, N., Blackburn, J., De Cristofaro, E., Stringhini, G., & Vakali, A. (2017). Mean Birds: Detecting Aggression and Bullying on Twitter. 13-22. https://doi.org/10.1145/3091478.3091487

Chen, G.M. (2011). Tweet this: A uses and gratification perspective on how active Twitter use gratifies a need to connect with others. *Computers in Human Behavior,* 27, 755-762.

Clement, J. (2019). Number of monthly active Twitter users worldwide from 1st quarter 2010 to first quarter 2019. *Statistica*. Retrieved from: https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/

Christie, C. (2007). Content analysis. In R. Baumeister, & K. Vohs (Eds.), *Encyclopedia of social psychology*. Thousand Oaks, CA: Sage.

Dedoose Version 8.0.35 web application for managing, analyzing, and presenting qualitative and mixed method research data (2018). Los Angeles, CA: SocioCultural Research Consultants, LLC www.dedoose.com.

Della Cioppa, V., O'Neil, A., & Craig, W. (2015). Learning from traditional bullying interventions: A review of research on cyberbullying and best practice, *Aggression and Violent Behavior*, 23, 61-68, https://doi.org/10.1016/j.avb.2015.05.009.

Dhungana Sainju, K., Mishra, N., Kuffour, A., Young, L. (2021a). Bullying discourse on Twitter: An examination of bully-related tweets using supervised machine learning. *Computers in Human Behavior*, 120, 106735.

Dhungana Sainju, K., Kuffour, A., Young, L., Mishra, N. (2021b). Bullying-related tweets: A qualitative examination of perpetrators, targets, and helpers. *International Journal of Bullying Prevention*, 4, 6–22.

Dooley, J.J., Pyzalski, J., & Cross, D. (2009). Cyberbullying versus face-to-face bullying: A theoretical and conceptual review. *Zeitschrift für Psychologie/Journal of Psychology*, 217(4), 182–188.

Dowell, E.B., Burgess, A.W., & Cavanaugh, D.J. (2009). Clustering of internet risk behaviors in a middle school student population. *Journal of School Health,* 11, 547–53.

Dredge, R., Gleeson, J., & Garcia, X. D. (2014). Cyberbullying in social networking sites: An adolescent victim's perspective. *Computers in Human Behavior*, 36, 13-20.

El Asam, A., & Samara, M. (2016) Cyberbullying and the law: a review of psychological and legal challenges. *Computers in Human Behavior*, 65, 127-141.

Freelon, D., McIlwain, C.D., Charlton, D., & Clark, M. (2016). Beyond the Hashtags: #Ferguson, #Blacklivesmatter, and the Online Struggle for Offline Justice; Center for Media & Social Impact, American University: Washington, DC, USA.

Gahagan, K., Vaterlaus, J.M., & Frost, L.R. (2016). College student cyberbullying on social networking sites: Conceptualization, prevalence, and perceived bystander responsibility. *Computers in Human Behavior*, 55(B), 1097-1105.

Hamm, M.P., Newton, A.S., Chisholm, A., Shulhan, J., Milne, A., Sundar, P., Ennis, H., Scott, S.D., Hartling, L. (2015). Prevalence and effect of cyberbullying on children and young people: A scoping review of social media studies. *JAMA Pediatrics*, 169(8), 770–777. doi:10.1001/jamapediatrics.2015.0944

Hawkins, D. L., Pepler, D. J., & Craig, W. M. (2001). Naturalistic observations of peer interventions in bullying. *Social Development*, 10(4), 512–527.

Hellström, L., Persson, L. & Hagquist, C. (2015). Understanding and defining bullying – adolescents' own views. *Archives of Public Health*.73, 4.

Hinduja, S. & Patchin, J. W. (2015). Bullying Beyond the Schoolyard: Preventing and Responding to Cyberbullying (2nd edition). Thousand Oaks, CA: Sage Publications.

Lambe, L. J., Della Cioppa, V., Hong, I. K., & Craig, W. M. (2019). Standing up to bullying: A social ecological review of peer defending in offline and online contexts. *Aggression and Violent Behavior*, 10.1016/j.avb.2018.05.007

Levy, M. & Gumpel, T. P. (2018). The interplay between bystanders' intervention styles: An examination of the "bullying circle" approach, *Journal of School Violence*, 17(3), 339-353.

Katz, E., Blumler, J. G., & Gurevitch, M. (1973). Uses and Gratifications Research. *The Public Opinion Quarterly*, 37(4), 509–523.

Katzer, C., Fetchenhauer, D., & Belschak, F. (2009). Cyberbullying: Who are the victims? A comparison of victimization in Internet chatrooms and victimization in school. *Journal of Media Psychology,* 21, 25–36. doi:10.1027/1864-1105.21.1.25

Kırcaburun, K., Kokkinos, C.M., Demetrovics, Z. Király, O., Griffiths, M.D., & Seda Çolak, T. (2019). Problematic online behaviors among adolescents and emerging adults: Associations between cyberbullying perpetration, problematic social Media Use, and psychosocial Factors. *International Journal of Mental Health and Addiction* 17, 891–908.

Kowalski, R. M., & Limber, S. P. (2007). Electronic bullying among middle school students. *Journal of Adolescent Health*, 41(6), S22–S30. doi:10.1016/j.jadohealth.2007.08.017

Kowalski, R., 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.

Marwick, A. C., & Boyd, D. (2011). I tweet honestly, I tweet passionately: Twitter users,

context collapse, and the imagined audience. *New Media & Society*, 13(1), 114–133.
McCurry, J. (2019). K-pop under scrutiny over 'toxic fandom' after death of Sulli. *The Guardian*. Retrieved from https://www.theguardian.com/music/2019/oct/18/k-pop-under-scrutiny-over-toxic-fandom-after-death-of-sulli
McHugh, M.C., Saperstein, S.L., & Gold, R.S. (2019). OMG U #Cyberbully! An exploration of public discourse about cyberbullying on Twitter. *Health Education & Behavior*, 46(1) 97–105.
McIntire, M., Yourish, K., & Buchanan, L. (2019). Trump's twitter feed: Conspiracymongers, racists, and spies. The New York Times. Retrieved from https://www.nytimes.com/interactive/2019/11/02/us/politics/trump-twitter-disinformation.html.
Mundt, M., Ross, K., & Burnett, C.M. Scaling Social Movements Through Social Media: The Case of Black Lives Matter. *Social Media + Society*, 4, 2056305118807911
Newall, M. (2018). Global Views on Cyberbullying. *Ipsos*. Retrieved from https://www.ipsos.com/en/global-views-cyberbullying
NLTK Project. (2020). NLTK 3.5 documentation. Source code for nltk.stem.wordnet. Retrieved from https://www.nltk.org/_modules/nltk/stem/wordnet.html.
Nocentini, A., Calmaestra, J., Schultze-Krumbholz, A., Scheithauer, H., Ortega, R., & Menesini, E. (2010). Cyberbullying: Labels, Behaviours and Definition in Three European Countries. *Australian Journal of Guidance & Counselling*, 20(2), 129–142.
Olweus, D. (1993). Bullying at School: What We Know and What We Can Do. Blackwell Publishers. Cambridge, MA.
Ortiz-Ospina, E. (2019). The rise of social media. Our World in Data. Retrieved from https://ourworldindata.org/rise-of-social-media
O'Sullivan, P. B., & Flanagin, A. J. (2003). Reconceptualizing 'flaming' and other problematic messages. *New Media & Society*, 5(1), 69–94.
Padgett, S., & Notar, C. E. (2013). Bystanders are the key to stopping bullying. *Universal Journal of Educational Research*, *1*(2), 33–41. doi:10.13189/ujer.2013.010201
Park, S., Na, E., & Kim, E. (2014). The relationship between online activities, netiquette and cyberbullying. *Children and Youth Services Review*, 42, 74-81.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
Picheta, R. (2020). Caroline Flack, 'Love Island,' and the industry of outrage surrounding the star's death. *CNN.* Retreived from https://www.cnn.com/2020/02/17/media/caroline-flack-death-reaction-scli-gbr-intl/index.html
Polanin, J. R., Espelage, D. L., & Pigott, T. D. (2012). A Meta-Analysis of School-Based Bullying Prevention Programs' Effects on Bystander Intervention Behavior. *School Psychology Review*, 41(1), 47–65.
Pyzalski, J. (2011). Electronic aggression among adolescents: An old house with a new façade (or even a number of houses). In Youth culture and net culture: Online social practices, ed. C. Hällgren, E. Dunkels and G.-M. Frånberg, 278–95. Hershey, PA: IGI Global.

Pyzalski, J. (2012). From cyberbullying to electronic aggression: typology of the phenomenon.*Emotional and Behavioural Difficulties*, 17(3–4), 305–317.

Romano, A. (2019). Why we can't stop fighting about cancel culture. *Vox.* Retrieved from https://www.vox.com/culture/2019/12/30/20879720/what-is-cancel-culture-explained-history-debate

Salmivalli, C., Lagerspetz, K. M. J., BjoÈrkqvist, K., OÈ sterman, K. & Kaukiainen, A. (1996). Bullying as a group process: participant roles and their relations to social status within the class. *Aggressive Behavior*, 22, 1-15.

Salmivalli, C. (2010). Bullying and the peer group: A review. *Aggression and Violent Behavior*, 15(2), 112-120, 10.1016/j.avb.2009.08.007

Salmivalli, C. (2014). Participant roles in bullying: How can peer bystanders be utilized in interventions? *Theory Into Practice*, 53(4), 286–292.

Securly. (2020). The 10 types of cyberbullying. Retrieved from https://blog.securly.com/2018/10/04/the-10-types-of-cyberbullying/

Shenk, C. E., & Fruzzetti, A. E. (2011). The impact of validating and invalidating responses on emotional reactivity. *Journal of Social and Clinical Psychology*, 30(2), 163-183.

Smith, P.K., Mahdavi, J., Carvalho, M., Fisher, S., Russell, S., & Tippett, N. (2008). Cyberbullying: Its nature and impact in secondary school pupils, *Journal of Child Psychology and Psychiatry*, 49 (4), 376-385

Staude-Müller, F., Hansen, B., & Voss, M. (2012). How stressful is online victimization? Effects of victim's personality and properties of the incident. *European Journal of Developmental Psychology*, 9(2), 260-274, DOI: 10.1080/17405629.2011.643170

Stever, G. S., & Lawson, K. (2013). Twitter as a way for celebrities to communicate with fans: Implications for the study of parasocial interaction. *North American Journal of Psychology*, 15(2), 339-354.

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.

Twyman, K., Saylor, C., Taylor, L.A., & Comeaux, C. (2010). Comparing children and adolescents engaged in cyberbullying to matched peers. *Cyberpsychology, Behavior, and Social Networking*, 13(2):195-199. doi:10.1089/cyber.2009.0137

UNESCO (2019). Behind the numbers: Ending school violence and bullying. United Nations Educational, Scientific and Cultural Organization. Paris, France.

UNICEF. (2019). UNICEF poll: More than a third of young people in 30 countries report being a victim of online bullying. Retrieved from https://www.unicef.org/press-releases/unicef-poll-more-third-young-people-30-countries-report-being-victim-online-bullying

Vanderberg, M. (2020). 18 celebrities who have quit social media and why. *Insider*. Retrieved from https://www.insider.com/celebrities-who-quit-social-media-twitter-2018-8

Viner, R.M., Gireesh, A., Stiglic, N., Hudson, L.D., Goddings, A-L., Ward, J.L., & Nicholls, D.E. (2019). Roles of cyberbullying, sleep, & physical activity in mediating the effects of social media use on mental health & wellbeing among young people in England: a secondary analysis of longitudinal data, *The Lancet Child & Adolescent Health*, 3(10), 685-696.

Waasdorp, T.E. & Bradshaw, C.P. (2015). The overlap between cyberbullying and traditional bullying. *Journal of Adolescent Health*, 56(5), 483-488.

Wagner, U., Galli, L., Schott, B.H., Wold, A., van der Schalk, J., Manstead, A.S.R., Scherer, K., & Walter, H. (2015). Beautiful friendship: Social sharing of emotions improves subjective feelings and activates the neural reward circuitry. *Social Cognitive and Affective Neuroscience*, 10(6), 801–808.

Wójcik, M., & Flak, W. (2019). Frenemy: A New Addition to the Bullying Circle*. Journal of Interpersonal Violence*. https://doi.org/10.1177/0886260519880168

Wolak, J., Mitchell, K. J., & Finkelhor, D. (2007). Does online harassment constitute bullying? An exploration of online harassment by known peers and online-only contacts. *The Journal of Adolescent Health* : official publication of the Society for Adolescent Medicine, 41(6 Suppl 1), S51–S58. https://doi.org/10.1016/j.jadohealth.2007.08.019

Waseem, Z. & Hovy, D. (2016). Hateful symbols or hateful people? Predictive features for hate speech detection on twitter. *Proceedings of North American Chapter of the Association for Computational Linguistics: Human Language Technologies 2016*, 88–93.

Whiting, A. & Williams, D. (2013). Why people use social media: A uses and gratifications approach. *Qualitative Market Research: An International Journal*, 16(4), 362-369.

Whittaker, E. & Kowalski, R.M. (2015). Cyberbullying via social media. *Journal of School Violence*, 14, 11-29.

Willard, N. (2007). Cyberbullying and cyberthreats. Effectively managing internet use risks in schools. Center for Safe and Responsible Use of the Internet. Retrieved from https://www.cforks.org/Downloads/cyber_bullying.pdf

Xu, J.-M., Jun, K.-S., Zhu, X., & Bellmore, A. (2012). Learning from bullying traces in social media. In Proceedings of the conference of the North American chapter of the association for computational linguistics: human language technologies (pp. 656–666). Montreal, Quebec, Canada: Association for Computational Linguistics. Retrieved from http://www.proceedings.com/15547.html.

Funding and Acknowledgements

The authors declare no funding sources or conflicts of interest. This study was supported by a grant to the first author from the Social Sciences and Humanities Research Council of Canada (SSHRC) through the SSHRC Institutional Grants Program 2019 for the University of Ontario Institute Of Technology.

The current study was reviewed and approved by the University of Ontario Institute of Technology's Research Ethics Board (REB) as a secondary use of data study (REB# 15,917). All procedures performed in the study were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.