Abstract

Cyberbullying has become increasingly prevalent, particularly on social media. There has also been a steady rise in cyberbullying research across a range of disciplines. Much of the empirical work from computer science has focused on developing machine learning models for cyberbullying detection. Whereas machine learning cyberbullying detection models can be improved by drawing on psychological theories and perspectives, there is also tremendous potential for machine learning models to contribute to a better understanding of psychological aspects of cyberbullying. In this paper, we discuss how machine learning models can yield novel insights about the nature and defining characteristics of cyberbullying and how machine learning approaches can be applied to help clinicians, families, and communities reduce cyberbullying. Specifically, we discuss the potential for machine learning models to shed light on the repetitive nature of cyberbullying, the imbalance of power between cyberbullies and their victims, and causal mechanisms that give rise to cyberbullying. We orient our discussion on emerging and future research directions, as well as the practical implications of machine learning cyberbullying detection models.

Keywords: cyberbullying, machine learning, psychology, computer science, interdisciplinary research, social media

Harnessing the Power of Interdisciplinary Research with Psychology-Informed Cyberbullying Detection Models

Cyberbullying, bullying that occurs through electronic media, has become increasingly prevalent (Anderson, 2018; Kowalski, Toth, & Morgan, 2018; Wang et al., 2019), in part due to the widespread use of social media (Anderson & Jiang, 2018; Lenhart et al., 2010; Perrin, 2015). In fact, social media is now the most common venue for cyberbullying (Duggan, 2017; see also Bischoff, 2019). Across disciplines, research on cyberbullying and related phenomena has also increased steadily. In computer science, a major empirical focus has been on developing automated models for detecting instances of cyberbullying (Al-Garadi et al., 2019; Muneer & Fati, 2020; Rosa et al., 2019; Salawu, He, & Lumsden, 2020). These models primarily employ machine learning, a form of artificial intelligence whereby a system learns, or improves, on an automated task through experience (Jordan & Mitchell, 2015). With cyberbullying detection, the automated task typically involves the classification of individual items in large sets of data as cyberbullying or not (see Salawu et al., 2020, for a review). Several models, for example, implement algorithms that use textual features (e.g., the words in a social media post) to classify an instance (e.g., a social media comment) as cyberbullying or not (Dinakar, Reichart, & Lieberman, 2011; Van Hee et al., 2018; Zhao & Mao, 2016).

Machine learning approaches for detecting cyberbullying hold particular promise—not only for identifying cyberbullying in real world data, but also for bridging the largely disparate efforts of cyberbullying researchers across disciplines. For instance, there is emerging evidence that machine learning cyberbullying detection models can be improved by drawing on psychological research. To illustrate, recent approaches have used information about social

media users' personalities (e.g., Balakrishnan et al., 2019; Zarnoufi & Abik, 2019), emotional content in social media posts (e.g., Dani et al., 2017), and aspects of social influence (Cheng et al., 2019b; Squicciarini et al., 2015) to improve the accuracy of cyberbullying detection models. These models have received relatively little attention outside of computer science, however, and cyberbullying detection frameworks that draw explicitly on psychological research represent a departure from the norm. Furthermore, existing divides have led researchers within the social sciences to overlook the tremendous potential for machine learning models to inform a better understanding of psychological aspects of cyberbullying.

The goal of this paper is to initiate a conversation about future interdisciplinary work that addresses two core questions: How can machine learning models yield novel insights about the nature and defining characteristics of cyberbullying? And, how can machine learning approaches be applied to help clinicians, families, and communities reduce cyberbullying? In Section 1, we discuss three specific ways that machine learning models for detecting cyberbullying on social media, in particular, can offer new insights on crucial yet understudied aspects of cyberbullying. These include the temporal properties and repetitive nature of cyberbullying, the imbalance of power between cyberbullies and their victims, and causal mechanisms that give rise to cyberbullying. In Section 2, we discuss key clinical and practical implications of machine learning cyberbullying detection models, with an emphasis on how machine learning approaches can strengthen efforts to prevent cyberbullying and its detrimental impact. Rather than provide a comprehensive review of the relevant literatures, our aim is to highlight vital directions for future work that uses machine learning to better understand cyberbullying and how to combat it.

I. New Insights on the Nature and Causes of Cyberbullying

Despite inconsistency and a lack of conceptual clarity in how cyberbullying is defined by researchers (cf., Kowalski et al., 2014; Langos, 2012; Rosa et al., 2018), most definitions reflect a consensus that cyberbullying entails intentionally harmful (i.e., aggressive or hostile) behavior that occurs *repeatedly* over time via electronic media (e.g., Hinduja & Patchin, 2020). Many definitions also identify or imply a power imbalance between cyberbullies and their victims (e.g., victims' inability to defend themselves) as an additional component of cyberbullying (Langos, 2012). With the work of Ziems and colleagues (2020) as a notable exception, discussed in greater detail below, few machine learning cyberbullying detection models have taken these definitional criteria into account. That is, of the numerous automated detection models published to date, only a small minority have included information about the specific cyberbullying criteria that informed how the models or datasets were built or how human annotators were instructed to differentiate cyberbullying from normal (i.e., non-bullying) content in training data (Rosa et al., 2018). Moreover, the vast majority of machine learning cyberbullying detection models rely on textual features and/or sentiment analysis (Rosa et al., 2018; Salawu et al., 2020), with a predominant focus on the intentional harm aspect of cyberbullying behavior.

Yet, machine learning frameworks for detecting cyberbullying on social media are uniquely suited for building a better understanding of the elements of repetition and power imbalance. This stems in large part from the hierarchical structure of social media data—i.e., individual comments are made up of words and sessions are made up of comments, images/videos, and social content that occur over time. Cyberbullying detection models that account for and leverage the hierarchical structure of social media data can yield important insights about temporal properties of cyberbullying and power differentials among users within a session, in particular (Cheng et al., 2020).

Temporal dynamics of cyberbullying

Machine learning models that incorporate temporal information can more effectively capture the hierarchical structure of a social media session and, by doing so, can increase the accuracy of cyberbullying detection. They also have unique benefits for investigating temporal patterns in cyberbullying interactions as they unfold over time within a session. This is evidenced, for instance, by cyberbullying detection models that employ hierarchical attention networks (e.g., Cheng et al., 2019a, 2020)—a technique that first constructs a representation of social media comments, aggregates them into a session representation, and then assigns different weights (i.e. "attention") to certain words and comments (Yang et al., 2016).

To illustrate, Cheng and colleagues (see Cheng et al., 2020) used the following information to build a representation of each social media session in an Instagram dataset: (1) the words comprising a comment or caption, (2) weights reflecting degree of relevance of each word to cyberbullying, (3) comments and associated weights reflecting the relevance of each comment to cyberbullying, (4) the timestamp and social content (i.e., number of 'likes' and shares) for each comment, and (5) the image/video shared in the initial post. They then used either time interval prediction (Cheng et al., 2019a) or temporal encoding (Cheng et al., 2020) to model the temporal ordering of comments (Cheng et al., 2020).

Two recent papers that have used the hierarchical structure of social media sessions to explore temporal patterns in cyberbullying offer early insights. Soni and Singh (2018) compared the temporal properties of Instagram social media sessions—where each session consisted of an initial post and all subsequent comments—that had been previously labeled, at the session-level, as cyberbullying or not (Hosseinmardi et al., 2016). They found that sessions determined (by human annotators) to constitute cyberbullying had, on average, a longer interval of time between

the initial post and the first comment, smaller intervals of time between all subsequent comments, and higher levels of activity than sessions without cyberbullying. Using the same Instagram data, Gupta and colleagues (2020) manually labeled each comment within each session of the same Instagram data as cyberbullying or not. They found that in sessions labeled holistically as cyberbullying, the first comments labeled as cyberbullying occurred within the first hours of the session; that is, in roughly 50% of cyberbullying sessions, the first cyberbullying comment occurred within the first hour of the session and in roughly 75% of the sessions, initial cyberbullying comments occurred within the first five hours of the session. Burst analyses also indicated that activity in cyberbullying sessions tended to peak in the first hour of a session, highlighting that within individual social media sessions, the repetitive nature of cyberbullying manifests early on. A practical implication is that efforts to identify and curtail cyberbullying can be tailored to focus more heavily on messages that occur earlier in a social media session, and social scientists may gain especially helpful insights by examining the initial exchanges within a session.

Together, these findings represent meaningful initial steps but the potential for machine learning models to help researchers understand how cyberbullying occurs repetitively over time remains mostly untapped. A number of open questions that have yet to be explored include what degree of repetition (e.g., number of successive cyberbullying comments) characterizes cyberbullying sessions that are more (versus less) severe, whether the psychological harm induced by cyberbullying messages increases over time or with greater repetition (or potentially plateaus or diminishes after reaching a tipping point), whether greater similarity or variability in successive cyberbullying comments is more detrimental, and how different forms of repetition (e.g., reposting/sharing cyberbullying content) can impact the nature of cyberbullying

interactions. Needless to say, cyberbullying researchers across disciplines will benefit from future work that identifies and seeks to understand patterns of repetition in cyberbullying.

Power imbalances among users

Relatively little is known about the dynamics of power imbalances within cyberbullying interactions. This is exacerbated by the anonymity that frequently characterizes online environments (Sarda et al., 2019) and the complexity in how power can be operationalized in online interactions (Langos, 2012). In light of these challenges, it is perhaps not surprising that few machine learning cyberbullying detection approaches have incorporated power or status differences between users into their models.

Among existing cyberbullying detection frameworks, Squicciarini and colleagues (2015) included a pairwise influence component to model the directional influence of a cyberbully on other users, reflected in the increased likelihood that a user who observes the cyberbully will subsequently bully others. Cheng and colleagues (2019b) incorporated a peer-influence component that added predictive value to a global cyberbullying detection model by leveraging between-user similarities in language use. This component was based on psychological research that has identified patterns of similarity in bullying behavior and victimization within child and adolescent peer groups (Espelage et al., 2003; Festl, Scharkow, & Quandt, 2013; see also Hinduja & Patchin, 2013). Notably, neither of these approaches directly models the imbalance of power that may exist between cyberbullies and their victims.

Within face-to-face bullying contexts, power and social status are often reflected in physical stature or strength or sociometric (e.g., peer nomination) indicators of social status (e.g., popularity) (see Nelson et al., 2019). Online environments necessitate different metrics of power and status that can offer innovative insights on power differentials in cyberbullying on social

media. For instance, the size of a user's social network—reflected in their number of "friends" or followers or the frequency with which their social media content is "liked" or shared by others—and even a user's privacy settings (i.e., the extent to which their social media content is accessible by others; cf., Kasper, 2007) provide ways to conceptualize and quantify social status and power.

Indeed, two promising approaches for modeling power imbalance in cyberbullying interactions build on social network analysis (e.g., Huang et al., 2015; Ziems et al., 2020). First, Haung and colleagues (2015) used social network features to improve text-based automated cyberbullying detection. Specifically, they analyzed Twitter comments reflecting a direct interaction, or network path, between two users, indicated by the inclusion of @. Social network features, including the number of nodes (an indicator of the size, or number of users within one's network) and number of edges (an indicator of degree of connectedness within one's network), were extracted for both users based on their respective interaction histories. Cyberbullying detection that incorporated text analysis and users' social network features outperformed models based on text analysis alone.

Second, Ziems et al. (2020) used social network features, including neighborhood overlap and user-based features (e.g., number of friends and followers), to more directly model power imbalance between cyberbullying perpetrators and victims. Ziems and colleagues first asked human annotators to label a corpus of Twitter comments based on their subjective assessment of whether the author of the tweet was more powerful than the target, the target was more powerful than the author, or the author and target were equal in power. Next, the researchers calculated social network features that might be reflective of power for each author and target, including characteristics of neighborhood overlap (e.g., number of paths between the two users via

'following' and 'followed by' relations) and overall number of friends and followers. Their results from experiments using a series of machine learning detection models point to the potential utility of social network-based metrics of power imbalance.

Notwithstanding the work by Ziems and colleagues (2020), few existing machine learning frameworks incorporate quantitative metrics of power imbalance between users into cyberbullying detection tasks. Cyberbullying detection models that take indicators of users' power into account--by, for example, incorporating users' social network features during the model training phase--will be an invaluable tool for advancing the understanding of how cyberbullying interactions are shaped by and reflective of power differentials. Open research questions include the extent to which the identification of cyberbullying (by humans and automated models) is influenced by the imbalance of power between users, whether cyberbullies leverage their power in strategic ways (by, for example, selectively bullying users with lower embeddedness within a network), and whether larger power discrepancies are associated with greater psychological harm or characterize specific forms of cyberbullying.

Understanding causality in cyberbullying

Understanding the causal factors and mechanisms that give rise to cyberbullying has profound implications for efforts to identify and ultimately prevent cyberbullying. Although beneficial for investigating associations between cyberbullying and psychological factors, the predominantly cross-sectional correlational design of cyberbullying studies in psychology has severely limited researchers' ability to draw causal inferences. Whereas longitudinal studies of cyberbullying (e.g., Camerini et al., 2020; Marciano, Schulz, & Camerini, 2020; Zhang et al., 2020a; 2020b)—in which causal relations between variables measured at later time points can, under certain circumstances, be inferred from variables measured at preceding time points (see

Rutter, 1988; Selig & Little, 2012)—these designs are, for many cyberbullying researchers, prohibitively time-, labor-, and resource-intensive.

Given the inherently data-driven nature of machine learning, it is typically viewed as a tool for prediction rather than causal analysis. Indeed, differences in potential confounding variables across the real-world data sets tend to constrain the transportability--or generalizability--of machine learning models to new data, largely precluding causal analysis (Cheng, Guo, & Liu, 2019; Pearl & Bareinboim, 2011). Yet, innovative approaches for developing causality-powered machine learning models are being introduced that hold considerable promise for identifying psychological factors that are causally-related to cyberbullying behavior.

One such approach comes from the work of Cheng, Guo, and Liu (2019), who used a machine learning framework to examine causal relations between psychological predictors and cyberbullying detection in social media data. Their approach involved identifying psychological covariates of cyberbullying and potential confounders in data collected from Twitter and Formspring and then employing a de-confounding mechanism to isolate the causal effects of the psychological covariates on cyberbullying detection. Specifically, they extracted psychological variables from the data using the Linguistic Inquiry and Word Count (LIWC) tool; Pennebaker et al., 2001) including, for example, affective processes (e.g., text reflecting negative emotion), motivational processes (e.g., text pertaining to drives for affiliation, achievement, etc.), and biological processes (e.g., text pertaining to the body, health, etc.). Cheng and colleagues then identified potential confounding variables, defined as factors contributing to a spurious relation between a psychological covariate and cyberbullying detection, drawing on a phenomenon known as Simpson's paradox (see Pearl, 2000). To block the influence of the identified confounders, they developed a de-confounding mechanism (by disaggregating the data into

smaller, more homogeneous subgroups). Finally, they examined the cyberbullying detection performance of machine learning models with and without de-confounding. Not only did the de-confounding mechanism improve cyberbullying detection, it also increased the models' transportability—evidenced by the effectiveness of models trained on one data set (i.e., Twitter or Formspring data) and tested on the other.

Clearly, research on causality-powered cyberbullying detection models is incipient and a vital task will be communicating these and subsequent findings to those who may be less familiar with the specific computational and statistical techniques employed. Still, continued efforts in this direction can help equip psychologists with new tools for investigating causal relations among predictors and outcomes associated with cyberbullying. This may especially be the case for cyberbullying detection frameworks that take the hierarchical structure of social media data into account. For instance, models can be built to identify the content and characteristics (e.g., user information, number of likes/shares) of comments that immediately precede the first cyberbullying comment within a session. When paired with mechanisms that control for confounding bias, a deeper understanding of the affective and motivational antecedents that directly contribute to cyberbullying is possible.

II. Practical Applications of Machine Learning Cyberbullying Detection

One of the most compelling motivations for studying cyberbullying is to develop mechanisms and tools for preventing cyberbullying and its negative consequences. Not only can cyberbullying detection models facilitate a better understanding of cyberbullying, they also represent an essential step toward realizing the goal of prevention. Two open (and related) challenges are to build models that shed light on the severity of cyberbullying instances and models that help predict *which* instances of cyberbullying are most likely to result in

psychological harm. In a recent systematic review, Salawu and colleagues (2020) identified 46 published papers that introduced an automated cyberbullying detection framework. Of these papers, 34 (73.9%) used a binary cyberbullying classification task; only six (13.0%) used a task that provided some level of distinction in cyberbullying severity. For instance, Talpur and O'Sullivan (2020) used a multi-class categorization task to identify instances of cyberbullying that were low, moderate, or high in severity in Twitter data. Potha and Maragoudakis (2014) used time-series modeling to classify the severity of cyberbullying attacks in dialogue exchanged between cyber-predators and victims. Despite being outnumbered by binary classification models, models that identify varying degrees of cyberbullying (e.g., Aggarwal, Maurya, and Chaudhary, 2020; Potha & Maragoudakis, 2014; Talpur & O'Sullivan, 2020) have so far yielded promising results. Detection tasks that move beyond a dichotomous categorization will be instrumental for helping practitioners identify which cyberbullying instances are most strongly linked with negative psychological outcomes.

A related direction for future research with considerable implications for both policy and clinical practice involves the integration of automated models for detecting depression, suicidal thoughts, and behavior, and cyberbullying. Several machine learning-based approaches for detecting suicidal thoughts and behavior have been introduced in recent years (e.g., O'Dea et al., 2015; Just et al., 2017; Walsh, Ribiero, & Franklin, 2017; Walsh, Ribiero, & Franklin, 2018; see also Linthicum, Schafer, & Ribeiro, 2019). Whereas many of these models draw on similar machine learning techniques to those being developed in cyberbullying research, to our knowledge, there are no models that combine cyberbullying and suicide risk detection in a single framework.

More broadly, efforts that integrate cyberbullying detection with indicators of mental health may help increase the likelihood that cyberbullying victims and perpetrators receive psychological support and improve the quality of care that clinicians provide. Recent guidelines for clinicians working with youth who have experienced cyberbullying, for instance, recommend adopting a holistic approach whereby clients (i.e., youth), caregivers, and schools work with clinicians to develop a more comprehensive system of support (Byers et al., 2021). Efforts to identify cyberbullying risk and its psychological correlates at the family- and school-level, discussed in greater detail below, can thus improve the mental health outcomes of potential victims.

An especially promising and impactful direction for future work will be to incorporate machine learning cyberbullying detection models into practical tools for use in clinical, educational, community, and family contexts. One template for this comes from the work of Silva and colleagues (2018) to develop mobile applications that help parents identify and respond to changes in their children's cyberbullying risk. As one example, they developed a rule-based model for quantifying the relative likelihood that a teen is being bullied on social media. This identification model was then built into a mobile application that allows parents to track changes in their child's cyberbullying risk over time, better understand their child's risk factors for cyberbullying, and locate resources tailored to their child's unique circumstances. Although a detailed discussion of the app is beyond the scope of the present paper (see Silva et al., 2018), we believe it provides one blueprint for how technological tools can be developed from psychology-informed cyberbullying detection models with far-reaching practical benefits.

Technological tools aimed at helping parents identify and prevent instances of cyberbullying are currently an underutilized avenue for cyberbullying prevention within families.

Common strategies utilized by parents to reduce their childrens' cyberbullying risk include monitoring technology use (Ghosh, Badillo-Urquiola, & Wisniewski, 2018; Mesch, 2009) and communicating safe technology practices (Padilla-Walker et al., 2019) (see Hutson, Kelly, & Mitello, 2018, for a review of cyberbullying interventions for families). In a recent qualitative study involving a series of focus groups with parents and guardians of children in fourth through sixth grade, Helfrich and colleagues (2020) found that parent-child communication about online behavior was the main strategy used by parents to mitigate cyberbullying risk. Some parents also reported engaging in active monitoring (i.e., using and exploring online media together) and restrictive monitoring (i.e., preventing or blocking access to technology or certain websites), but expressed frustration with these methods due to their limited understanding of how to utilize online parental controls. These findings highlight an area in which alternative technological tools, including apps that implement automated cyberbullying risk detection, may be especially beneficial. Although there are some software and mobile applications geared toward bullying prevention for parents (see Topcu-Uzer & Tanrikulu, 2018)—only a small handful are specific to cyberbullying and few use machine learning-based techniques that are informed by and have the ability to inform psychological research.

Whereas anti-bullying and cyberbullying interventions at the level of schools have been more common, most have been aimed at increasing knowledge and awareness of cyberbullying and its effects and on fostering social skills of children and teens such as empathy, prosocial motivations, and adaptive coping (see Gaffney et al., 2019 and Lancaster, 2018, for reviews). One technological tool that has yielded encouraging results within school settings are serious video games—i.e., video games designed for purposes other than entertainment—to reduce cyberbullying (Calvo-Morata et al., 2020). The KiVa antibullying program in Finland (Herkama

& Salmivalli, 2017; Salmivalli & Poskiparta, 2012), for example, uses a serious game to teach students about group-related aspects of cyberbullying. The emphasis of the broader KiVa program is on developing the skills youth need to support peers who may be experiencing cyberbullying victimization. We are unaware of any existing school- or community-level interventions that use machine learning models to detect cyberbullying or identify cyberbullying risk. Machine learning-based tools may be particularly beneficial for identifying, preventing, and understanding cyberbullying that occurs in online learning and classroom management platforms. Instances of cyberbullying via school-related online platforms may become increasingly prevalent and important to detect as shifts to remote learning occur throughout the COVID-19 pandemic.

Concluding Remarks

As both social media use and cyberbullying become more pervasive across increasingly broad segments of the population, interdisciplinary collaborations between computer science and psychology hold considerable promise for cyberbullying research. The core aim of this paper was to highlight how machine learning frameworks for detecting cyberbullying in social media data, in particular, can facilitate a deeper understanding of psychological aspects of cyberbullying and be applied in ways that help combat cyberbullying. Our hope is that these insights from emerging and future research directions will inspire an ongoing synergy among cyberbullying researchers across disciplines.

Conflict of Interest Statement

On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

Aggarwal, A., Maurya, K., & Chaudhary, A. (2020). Comparative study for predicting the severity of cyberbullying across multiple social media platforms. *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)*, 871–877. https://doi.org/10.1109/ICICCS48265.2020.9121046.

Al-Garadi, M. A., Hussain, M. R., Khan, N., Murtaza, G., Nweke, H. F., Ali, I., Mujtaba, G., Chiroma, H., Khattak, H. A., & Gani, A. (2019). Predicting cyberbullying on social media in the big data era using machine learning algorithms: Review of literature and open challenges. *IEEE Access, 7*, 70701–70718. https://doi.org/10.1109/ACCESS.2019.2918354.

Anderson, M. (2018). *A majority of teens have experienced some form of cyberbullying* [Report]. https://www.pewresearch.org/internet/wp-content/uploads/sites/9/2018/09/PI_2018.09.27_teens-and-cyberbullying_FINAL.pdf.

Anderson, M., & Jiang, J. (2018). *Teens, social media & technology* [Report]. https://www.pewresearch.org/internet/wp-content/uploads/sites/9/2018/05/PI_2018.05.31_TeensTech_FINAL.pdf.

Balakrishnan, V., Khan, S., & Arabnia, H. R. (2019). Improving cyberbullying detection using Twitter users' psychological features and machine learning. *Computers & Security, 90*, 1-11. https://doi.org/10.1016/j.cose.2019.101710.

Bischoff, P. (2019). Almost 60 percent of parents with children aged 14 to 18 reported them being bullied. Retrieved from: https://www.comparitech.com/blog/vpn-privacy/boundless-bullies/.

Byers, D. S., Mishna, F., & Solo, C. (2021). Clinical practice with children and adolescents

involved in bullying and cyberbullying: Gleaning guidelines from the literature. *Clinical Social Work Journal*, *49*(1), 20-34.

Calvo-Morata, A., Alonso-Fernández, C., Freire, M., Martínez-Ortiz, I., & Fernández-Manjón, B. (2020). Serious games to prevent and detect bullying and cyberbullying: A systematic serious games and literature review. *Computers & Education*, *157*, 103958.

Camerini, A. L., Marciano, L., Carrara, A., & Schulz, P. J. (2020). Cyberbullying perpetration and victimization among children and adolescents: A systematic review of longitudinal studies. *Telematics and Informatics*, *49*, 1-13. https://doi.org/10.1016/j.tele.2020.101362

Chelmis, C., Zois, D., & Yao, M. (2017). Mining patterns of cyberbullying on Twitter. *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, 126-133, http://doi.org/10.1109/ICDMW.2017.22.

Cheng, L., Guo, R., & Liu, H. (2019). Robust cyberbullying detection with causal interpretation. *Companion Proceedings of The 2019 World Wide Web Conference*, 169–175. https://doi.org/10.1145/3308560.3316503.

Cheng, L., Guo, R., Silva, Y. N., Hall, D., Liu, H. (2019a). Hierarchical attention networks for cyberbullying detection on the Instagram social network. *The SIAM International Conference on Data Mining*, 235-243. https://doi.org/10.1137/1.9781611975673.27.

Cheng, L., Li, J., Silva, Y. N., Hall, D. L., & Liu, H. (2019b). PI-Bully: Personalized cyberbullying detection with peer influence. *International Joint Conference on Artificial Intelligence (IJCAI)*, 5829-5835.

Cheng, L., Guo, R., Silva, Y. N., Hall, D. L., & Liu, H. (2020). Modeling temporal patterns of cyberbullying with hierarchical attention networks. *ACM/IMS Transactions on Data Science*, 1-23. https://doi.org/10.1145/3441141.

Dani, H., Li, J., & Liu, H. (2017). Sentiment informed cyberbullying detection in social media. In: M. Ceci, J. Hollmén, L. Todorovski, C. Vens, & S. Džeroski (Eds.), *Machine learning and knowledge discovery in databases* (pp. 52-67). Springer. https://doi.org/10.1007/978-3-319-71249-9_4

De Choudhury, M., Mason, W. A., Hofman, J. M., & Watts, D. J. (2010). Inferring relevant social networks from interpersonal communication. *Proceedings of the 19th international conference on World wide web*, 301-310. https://doi.org/10.1145/1772690.1772722

Digman, J. M. (1990). Personality structure: Emergence of the five-factor model. *Annual Review of Psychology*, *41*(1), 417-440.

Dinakar, K., Jones, B., Havasi, C., Lieberman, H., & Picard, R. (2012). Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *2*(3), 1-30. http://doi.acm.org/10.1145/2362394.2362400.

Dinakar, K., Reichart, R., & Lieberman, H. (2011). Modeling the detection of textual cyberbullying. *Proceedings of the International AAAI Conference on Web and Social Media*, *5*(1), 11-17. https://ojs.aaai.org/index.php/ICWSM/article/view/14209.

Duggan, M. (2017). *Online harassment 2017* [Report]. https://www.pewresearch.org/internet/wp-content/uploads/sites/9/2017/07/PI_2017.07.11_Online-Harassment_FINAL.pdf.

Espelage, D. L., Holt, M. K., & Henkel, R. R. (2003). Examination of peer-group contextual effects on aggression during early adolescence. *Child Development*, *74*(1), 205–220. https://doi.org/10.1111/1467-8624.00531.

Festl, R., Scharkow, M., & Quandt, T. (2013). Peer influence, internet use and cyberbullying: A comparison of different context effects among German adolescents. *Journal of Children and Media, 7*(4), 446-462. https://doi.org/10.1080/17482798.2013.781514.

Goldberg, L. R. (1990). An alternative "description of personality": The Big-Five factor structure. *Journal of Personality and Social Psychology*, *59*, 1216–1229. https://doi.org/10.1037/0022-3514.59.6.1216.

Ghosh, A. K., Badillo-Urquiola, K., & Wisniewski, P. (2018). Examining the effects of parenting styles on offline and online adolescent peer problems. *Proceedings of the 2018 ACM Conference on Supporting Groupwork*, 150-153. https://doi.org/10.1145/3173574.3173698

Guo, S. (2016). A meta-analysis of the predictors of cyberbullying perpetration and victimization. *Psychology in the Schools, 53*(4), 432–453. https://doi.org/10.1002/pits.21914

Gupta, A., Yang, W., Sivakumar, D. P., Silva, Y., Hall, D., & Barioni, M. (2020). Temporal properties of cyberbullying on Instagram. *ACM CyberSafety: Computational Methods in Online Misbehavior*, 576-583. https://doi.org/10.1145/3366424.3385771.

Helfrich, E. L., Doty, J. L., Su, Y., Yourell, J. L., & Gabrielli, J. (2020). Parental views on preventing and minimizing negative effects of cyberbullying. *Children and Youth Services Review, 118*, 1-9. https://doi.org/10.1016/j.childyouth.2020.105377

Herkama, S., & Salmivalli, C. (2017). KiVa antibullying program. In M. Campbell & S. Bauman (Eds.) *Reducing Cyberbullying in Schools: International Evidence-Based Best Practices* (pp. 125-134). Academic Press.

Hinduja, S., & Patchin, J. W. (2013). Social influences on cyberbullying behaviors among

middle and high school students. *Journal of Youth and Adolescence*, *42*(5), 711–722.

Hinduja, S. & Patchin, J. W. (2020). Cyberbullying identification, prevention, and response. Cyberbullying Research Center. https://cyberbullying.org.

Hosseinmardi, H., Mattson, S. A., Ibn Rafiq, R., Han, R., Lv, Q., & Mishra, S. (2015). Analyzing labeled cyberbullying incidents on the Instagram social network. In T. Y. Liu, C. N. Scollon, & W. Zhu (Eds.), *Social Informatics* (Vol. 9471, pp. 49–66). Springer International Publishing. https://doi.org/10.1007/978-3-319-27433-1_4.

Hosseinmardi, H., Rafiq, R. I., Han, R., Lv, Q., & Mishra, S. (2016). Prediction of cyberbullying incidents in a media-based social network. *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 186–192. https://doi.org/10.1109/ASONAM.2016.7752233.

Hutson, E., Kelly, S., & Militello, L. K. (2018). Systematic review of cyberbullying interventions for youth and parents with implications for evidence- based practice. *Worldviews on evidence- based nursing*, *15*(1), 72-79.

https://doi.org/10.1111/wvn.12257.

Jacobs, G., Van Hee, C., & Hoste, V. (2020). Automatic classification of participant roles in cyberbullying: Can we detect victims, bullies, and bystanders in social media text? *Natural Language Engineering*, 1–26. https://doi.org/10.1017/S135132492000056X

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science, 349*(6245), 255-260. https://doi.org/10.1126/science.aaa8415.

Just, M. A., Pan, L., Cherkassky, V. L., McMakin, D. L., Cha, C., Nock, M. K., & Brent, D. (2017). Machine learning of neural representations of suicide and emotion concepts identifies suicidal youth. *Nature Human Behaviour, 1*(12), 911-919.

Kasper, D. (2007). Privacy as a social good. *Social Thought & Research, 28,* 165-189. http://www.jstor.org/stable/23252125

Koeze, E., & Popper, N. (2020, April 7). The virus changed the way we internet. *New York Times*. Retrieved from https://www.nytimes.com/interactive/2020/04/07/technology/coronavirus-internet-use.html.

Kokkinos, C. M., & Kipritsi, E. (2012). The relationship between bullying, victimization, trait emotional intelligence, self-efficacy and empathy among preadolescents. *Social Psychology of Education*, *15*(1), 41–58. https://doi.org/10.1007/s11218-011-9168-9.

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. https://doi.org/10.1037/a0035618

Lancaster, M. (2018). A systematic research synthesis on cyberbullying interventions in the United States. *Cyberpsychology, Behavior, and Social Networking*, *21*(10), 593-602. https://doi.org/10.1089/cyber.2018.0307.

Langos, C. (2012). Cyberbullying: The challenge to define. *Cyberpsychology, Behavior, and Social Networking*, *15*(6), 285-289.

Lenhart, A. (2007) Cyberbullying. Retrieved from https://www.pewresearch.org/internet/2007/06/27/cyberbullying/.

Lenhart, A., Purcell, K., Smith, A., & Zickuhr, K. (2010). *Social media and mobile internet use among teens and young adults* [Report]. https://files.eric.ed.gov/fulltext/ED525056.pdf.

Linthicum, K. P., Schafer, K. M., & Ribeiro, J. D. (2019). Machine learning in suicide science: Applications and ethics. *Behavioral Sciences & the Law*, *37*(3), 214-222.

https://doi.org/10.1002/bsl.2392.

Marciano, L., Schulz, P. J., & Camerini, A. L. (2020). Cyberbullying perpetration and victimization in youth: A meta-analysis of longitudinal studies. *Journal of Computer-Mediated Communication*, *25*(2), 163-181.

McCrae, R. R., & Costa, P. T., Jr. (2008). The five-factor theory of personality. In O. P. John, R. W. Robins, & L. A. Pervin (Eds.), *Handbook of personality: Theory and research* (pp. 159–181). Guilford Press.

Mesch, G. S. (2009). Parental mediation, online activities, and cyberbullying. *Cyberpsychology & Behavior, 12*(4), 387-393. https://doi.org/10.1089/cpb.2009.0068.

Mishna, F., Khoury-Kassabri, M., Gadalla, T., & Daciuk, J. (2012). Risk factors for involvement in cyber bullying: Victims, bullies and bully–victims. *Children and Youth Services Review*, *34*(1), 63-70. https://doi.org/10.1016/j.childyouth.2011.08.032.

Muneer, A., & Fati, S. M. (2020). A comparative analysis of machine learning techniques for cyberbullying detection on Twitter. *Future Internet*, *12*(11), 187. https://doi.org/10.3390/fi12110187.

Nelson, H. J., Kendall, G. E., Burns, S. K., Schonert- Reichl, K. A., & Kane, R. T. (2019). Measuring 8 to 12 year old children's self- report of power imbalance in relation to bullying: Development of the Scale of Perceived Power Imbalance. *BMC Public Health*, 19, 1–12.

O'Dea, B., Wan, S., Batterham, P. J., Calear, A. L., Paris, C., & Christensen, H. (2015). Detecting suicidality on Twitter. *Internet Interventions*, *2*(2), 183-188. https://doi.org/10.1016/j.invent.2015.03.005.

Padilla-Walker, L. M., Stockdale, L. A., Son, D., Coyne, S. M., & Stinnett, S. C. (2020).

Associations between parental media monitoring style, information management, and prosocial and aggressive behaviors. *Journal of Social and Personal Relationships, 37*(1), 180-200. https://doi.org/10.1177/0265407519859653

Pearl, J. (2000). *Causality: Models, reasoning, and inference.* Cambridge, UK: Cambridge University Press.

Pearl, J., & Bareinboim, E. (2011). Transportability of causal and statistical relations: A formal approach. *Proceedings of the AAAI Conference on Artificial Intelligence, 25*(1), 247-254. https://ojs.aaai.org/index.php/AAAI/article/view/7861.

Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). Linguistic inquiry and word count: LIWC 2001. Mahway, NJ: Lawrence Erlbaum Associates.

Perrin, A. (2015). *Social media usage: 2005-2015* [Report]. https://www.pewresearch.org/internet/wp-content/uploads/sites/9/2015/10/PI_2015-10-08_Social-Networking-Usage-2005-2015_FINAL.pdf.

Potha, N., & Maragoudakis, M. (2014). Cyberbullying detection using time series modeling. *2014 IEEE International Conference on Data Mining Workshop,* 373–382. https://doi.org/10.1109/ICDMW.2014.170.

Rosa, H., Pereira, N., Ribeiro, R., Ferreira, P. C., Carvalho, J. P., Oliveira, S., Coheur, L., Paulino, P., Veiga Simão, A. M., & Trancoso, I. (2019). Automatic cyberbullying detection: A systematic review. *Computers in Human Behavior, 93*, 333–345. https://doi.org/10.1016/j.chb.2018.12.021.

Rutter, M. (1988). Longitudinal data in the study of causal processes: Some uses and some pitfalls. In M. Rutter (Ed.), *Studies of psychosocial risk: The power of longitudinal data* (pp. 1–28). Cambridge University Press.

Safaria, T., Lubabin, F., Purwandari, E., Ratnaningsih, E. Z., Khairani, M., Saputra, N. E., ... & Mariyati, L. I. (2020). The role of dark triad personality on cyberbullying: Is it still a problem? *International Journal of Scientific & Technology Research*, *9*(2), 4256-4260.

Salawu, S., He, Y., & Lumsden, J. (2020). Approaches to automated detection of cyberbullying: A survey. *IEEE Transactions on Affective Computing*, *11*(1), 3–24. https://doi.org/10.1109/TAFFC.2017.2761757.

Salmivalli, C., & Poskiparta, E. (2012). KiVa antibullying program: Overview of evaluation studies based on a randomized controlled trial and national rollout in Finland. *International Journal of Conflict and Violence*, *6*(2), 293-301.

Sardá, T., Natale, S., Sotirakopoulos, N., & Monaghan, M. (2019). Understanding online anonymity. *Media, Culture & Society*, *41*(4), 557-564. https://doi.org/10.1177/0163443719842074.

Selig, J. P., & Little, T. D. (2012). Autoregressive and cross-lagged panel analysis for longitudinal data. In B. Laursen, T. D. Little, & N. A. Card (Eds.), *Handbook of developmental research methods* (pp. 265–278). Guilford Press.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Houghton Mifflin.

Silva, Y., Hall, D., & Rich, C. (2018). BullyBlocker: Towards an interdisciplinary approach to identify cyberbullying. *Social Network Analysis and Mining*, *8*(18). https://doi.org/10.1007/s13278-018-0496-z.

Soni, D., & Singh, V. (2018). Time reveals all wounds: Modeling temporal characteristics of cyberbullying. *Proceedings of the International AAAI Conference on Web and Social Media, 12*(1), 684-687. https://ojs.aaai.org/index.php/ICWSM/article/view/15046

Squicciarini, A., Rajtmajer, S., Liu, Y., & Griffin, C. (2015). Identification and characterization of cyberbullying dynamics in an online social network. *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, 280–285. https://doi.org/10.1145/2808797.2809398.

Talpur, B. A., & O'Sullivan, D. (2020). Cyberbullying severity detection: A machine learning approach. *PLOS ONE, 15*(10), e0240924. https://doi.org/10.1371/journal.pone.0240924.

Topcu-Uzer, C., & Tanrıkulu, İ. (2018). Technological solutions for cyberbullying. In *Reducing Cyberbullying in Schools* (pp. 33-47). Academic Press.

Van Hee, C., Jacobs, G., Emmery, C., Desmet, B., Lefever, E., Verhoeven, B., De Pauw, G., Daelemans, W., & Hoste, V. (2018). Automatic detection of cyberbullying in social media text. *PLoS ONE, 13*(10), e0203794. https://doi.org/10.1371/journal.pone.0203794.

Walsh, C. G., Ribeiro, J. D., & Franklin, J. C. (2017). Predicting risk of suicide attempts over time through machine learning. *Clinical Psychological Science, 5*(3), 457-469.

Walsh, C. G., Riberio, J. D., & Franklin, J. C. (2018). Predicting suicide attempts in adolescents with longitudinal clinical data and machine learning. *The Journal of Child Psychology and Psychiatry, 59*(12), 1261-1270. https://doi.org/10.1111/jcpp.12916.

Wang, M. J., Yogeeswaran, K., Andrews, N. P., Hawi, D. R., & Sibley, C. G. (2019). How common is cyberbullying among adults? Exploring gender, ethnic, and age differences in the prevalence of cyberbullying. *Cyberpsychology, Behavior, and Social Networking, 22*(11), 736-741.

Xu, J., Jun, K., Zhu, X., & Bellmore, A. (2012). Learning from bullying traces in social media. *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* Association for

Computational Linguistics, 656-666.

Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., & Hovy, E. (2016). Hierarchical attention networks for document classification. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1480–1489. https://doi.org/10.18653/v1/N16-1174.

Zarnoufi, R., & Abik, M. (2019). Big five personality traits and ensemble machine learning to detect cyber-violence in social media. In M. Serrhini, C. Silva, & S. Aljahdali (Eds.), *Innovation in Information Systems and Technologies to Support Learning Research* (Vol. 7, pp. 194–202). Springer International Publishing. https://doi.org/10.1007/978-3-030-36778-7_21.

Zhang, D., Huebner, E. S., & Tian, L. (2020a). Longitudinal associations among neuroticism, depression, and cyberbullying in early adolescents. *Computers in Human Behavior*, *112*, 106475. https://doi.org/10.1016/j.chb.2020.106475.

Zhang, D., Huebner, E. S., & Tian, L. (2020b). Neuroticism and cyberbullying among elementary school students: A latent growth curve modeling approach. *Personality and Individual Differences*, 110472. https://doi.org/10.1016/j.paid.2020.110472.

Zhao, R., & Mao, K. (2016). Cyberbullying detection based on semantic-enhanced marginalized denoising auto-encoder. *IEEE Transactions on Affective Computing*, *8*(3), 328–339. https://doi.org/10.1109/TAFFC.2016.2531682.

Ziems, C., Vigfusson, Y., & Morstatter, F. (2020). Aggressive, repetitive, intentional, visible, and imbalanced: Refining representations for cyberbullying classification. In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 14, pp.

808-819).