Impact Analysis to Break the Cycle of Cyberbullying Using Machine Learning

Alka Singh[1], Jaydip Singh[2] & Amit[3]

ABSTRACT

In the era of Internet world, the development of the digital communication platforms has given the teens a unique chance to communicate with greater interactivity, but rather it has led to the cyberbullying to root even deeper now. This article introduces "Serenity," the problem of cyberbullying Chabot that benefits from NLP and ML techniques to proactively intercept and neutralize them in real-time. In contrast with the age-old lexicographer orientation, dynamic ML models being used by Serenity ensure. It is adaptable to the dynamic linguistic trends. The software not only identifies cyberbullying elements but also has built-in mechanisms to deter such scenarios, hence limiting the ease of access for such behaviors and ensuring the well-being of the user is safe. With this method of the parent involvement but not the violation of the rupture of privacy, Serenity guarantees the safety and freedom. This research indicates the need for high tech applications on social media platforms aimed at intervening in cases of distress to the vulnerable people and in this regard calls for the extension of such mechanisms to mainstream social media platforms. The presentation of these four design principles opens a significant room for policy makers, platform developers and educators as they contemplate on peer-involving strategies in building a more secure digital ecosystem for these young people.

Key words: cyberbullying, online safety, machine learning, natural language processing, human computer interaction

Dr. Alka Singh (Department of Geography, Banaras Hindu University)

Jaydip Singh (Student, Department of Engineering & Technology, Bharati Vidyapeeth)

ORCID ID: (jaydipsingh954@gmail.com for ORCID)

Amit (Student, Department of Engineering & Technology, Bharati Vidyapeeth)

1. Introduction

In the era of internet, information communication technologies have a far-reaching effect on the contemporary modern life. Instant messaging apps and social media platforms have generated a virtual space where people can stay and connected 24/7. This practices provides adolescents with a place to communicate with their peers without the interference of the adults where they

may feel more private. Unfortunately, these spaces also enable numerous forms of online harassment, such as cyberbullying and their impacts [Lee et al ., 2015]. Cyberbullying takes bullying and harassment to a domain where the perpetrator and victim are constantly available to each other on multiple platforms, regardless of geographic location. Many previous research have shown that 43.9% of adolescents between the ages of 11 and 18 have experienced harassment on current social media platforms [Wong et al ,. 2017]. Moreover, young people who especially experience cyberbullying are at a larger risk for self-harm and suicidal ideation [John et al ,. 2018], posing a real threat to their physical and mental health [Hinduja et al ,. 2010].

In the recent time, these issues are exacerbated by companies encouraging younger user groups to engage with these technologies. Instagram for Kids, for example, was designed as a version of Instagram targeted at children under the age of 13 [Gregg et al,. 2021]. The project received backlash from parents and policymakers [Chan et al ,. 2023], which aligns with a recent internal document revealing in which, Instagram generates a toxic environment for teenage girls [Hill et al ,. 2022]. While cyberbullying occurs on all social media platforms [Van et al ,. 2017], adolescents report experiencing cyberbullying on Instagram to the greatest extent, with 42% surveyed experiencing harassment [Hackett et al ,. 2017]. Likewise, Messenger Kids, an extension of Facebook's Messenger application, allows adolescents to interact with a set of the group of users as dictated by their guardians. However, Messenger Kids faced scrutiny as a bug that allowed children to chat with the unauthorized users using the group chat feature [Chan et al ,. 2023]. Facebook could not confirm how long the bug existed since the launch. This flaw could be legally sensitive; for instance, any application designed for children under 13 is subject to the Children's Online Privacy Protection Act (COPPA) in the United States.

Furthermore, newer social media applications such as now a days, TikTok reportedly possess userbases mainly consisting of children under the age of 14 [Barta et al,. 2021], which may put them in violation of their terms of service, stating that a user must be 13 and over. This platform encourages public video creation, distribution and has a high tolerance for anonymous user accounts, adolescents on TikTok. It is particularly susceptible to cyberbullying and grooming [Shutsko et al ,. 2020]. Even though, preventing adolescents from adopting new and potentially unsafe information communication technologies is challenging. Further work is needed to decrease the Impact Analysis of cyberbullying (IAC) on the population of given world.

Previous research suggests about the increasing cyberbullying sensitivity through comprehensive social support and education. This technique allows victims to continue using

these technologies but navigate cyberbullying incidents more effectively [Akturk et al,. 2015]. Technical intervention may also decrease the rate of cyberbullying incidents. Cyberbullying exists in many forms, such as flaming, harassment, cyber-stalking, and masquerading. Existing technical solutions can detect certain types of cyberbullying. For example, Concepción-Sánchez et al. [Concepción-Sánchez et al,. 2017] this technique proposed a mobile application to detect harassment in instant messages. Incoming messages are separated into words and expressions and passed through a word repository to check for potential cyberbullying. Similarly, SafeChat is a plug-in designed to secure instant messages and filter offensive content [Fahrnberger et al,.2014]. Messages are encrypted, filtered using an illicit word dictionary, and then decrypted once the receiver is authenticated by the sender, which prevents masquerading behaviors. Weider et al. [Weider et al,. 2016] also proposed a controlled messaging mobile application to avoid cyberbullying incidents. Once the user's device is integrated with the application, the detection system will scan any messages sent or received. The detection system uses Machine Learning (ML) techniques, including Naïve Bayes, Bayesian Networks, and the Maximum Entropy Classifier, to rank a message's polarity (positive, neutral, or negative). The mobile application notifies the receiver and warns the sender if the message's polarity is beyond the negativity threshold. However, during their performance evaluation, the average execution time was 11 seconds, which is a significant delay in the context of real-time chat applications.

While ML has the potential to mitigate cyberbullying incidents in the context of instant messaging, current designs and prototypes need to be improved to address the real-time limitations of existing technological responses. In combination with the fact that existing solutions are either performing cyberbullying detection based on an outdated lexicon-approach, there is a need for a more efficient and effective technical solution to detect and prevent cyberbullying, which also overcomes the limitations of existing technological responses. In this paper, we propose Serenity, a chat application that aims to detect and deter cyberbullying in real time and decrease cyberbullying behaviors. Integrating ML, Serenity can detect conversations containing flaming and harassment in real time and create a flow of corresponding actions to create a safer cyber-space for adolescents.

2. Material and Methods

Natural Language Processing (NLP) has recently gained much media attention with tools like ChatGPT and its alternatives. NLP aims to apply computational techniques to understand textual data better [Choudhary et al,. 2020]. ML applications with NLP involve machine translation,

sentiment analysis, and cyberbullying detection. Unlike lexicon or dictionary-based solutions, where dictionaries must be maintained and updated, ML models can be automatically and continually retrained using online datasets.

Several ML models can detect cyberbullying to a high degree of accuracy. Noviantho and Ashianti [Noviantho et al,. 2017] train a set of Support Vector Machine-based ML models using textual data, achieving an average accuracy of 97.11% when detecting negative polarity in messages. Similarly, Mahlangu and Tu [Mahlangu et al,. 2016] propose an LSTM and stacked embedding model approach. Instead of analyzing the polarity of each word in a message, the model is trained to 'understand' each sentence. As such, cyberbullying detection is not limited to finding negative texts but extends to analyzing the underlying meaning of each sentence. Their approach achieves a 94.2% accuracy when detecting sentences with harmful intentions.

Recently, Aind et al. [Aind et al,. 2020] propose the Q-Bully framework based on the reinforcement learning technique. Reinforcement learning-based models require the developers to train a human-like agent using a set of textual data. Once trained, the agent can distinguish whether a message contains cyberbullying intentions. Their approach can also recognize misspelled words, words with repeated letters (e.g., hellloooo), and novel words which the agent has not seen before. Their method achieves an average accuracy of 88% on randomly sampled datasets.

Even though these three different ML techniques are benchmarked using various datasets, they achieved high accuracy when detecting cyberbullying content. More importantly, these ML techniques create an environment of continual automated improvements, a requirement for the current fast-paced landscape. However, they primarily focus on developing technological interventions and do not address the human side of cyberbullying. This paper intends to use technology to nudge the behavior of potential perpetrators and victims, aiming to detect, deter, prevent, and break the cycle of cyberbullying.

3. Results

3.1 impact analysis of Serenity

3.1 (a) *Conceptual Design*

In this paper, we propose Serenity, a chat application that aims to detect and prevent cyberbullying in real time among adolescents. To achieve this, Serenity should not only provide a novel communication channel where cyberbullying is detected but also be able to deter

cyberbullying behaviors from occurring. Akturk et al. [Akturk et al,. 2015] suggest that a higher sensitivity towards cyberbullying allows victims to manage cyberbullying incidents better. In addition, when they analyze the variables that influence adolescent cyberbullying sensitivity levels, perceived social support is the most significant variable that affects cyberbullying sensitivity. In other words, a lack of perceived social support may increase the risk of cyberbullying incidents resulting in more severe real-world consequences. Another critical factor is understanding the motives of perpetrators. A recent survey found that the leading motives include Intentions to Hurt, Secrecy and Namelessness, Freedom of Expression, and Unevenness of Power [Abbasiet et al,. 2018]. Among these motives, only Intentions to Hurt possesses clear intentions, while perpetrators driven by the other motives may not realize the effects of their actions. Combining the cyberbullying sensitivity level and perpetrator motives, we propose a causal map to illustrate the cyberbullying cycle (Figure 1).

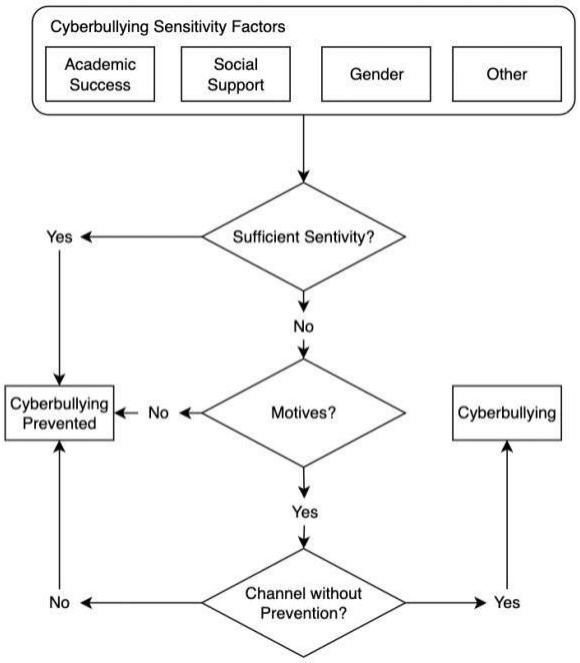

Fig 1. A causal map of cyberbullying

Based on our current understanding, there are three possible pathways to restrict cyberbullying in the context of digital communication among adolescents: 1) increasing cyberbullying sensitivity, 2) minimizing motives, and 3) providing a channel for cyberbullying prevention. Serenity will be designed to deliver a solution that can support all three of these pathways. It will have two types of account: an adolescent user account and a guardian account. Each adolescent user account must be linked with one or more guardian accounts. For each adolescent user account, Serenity analyzes all conversations made and provides an overall toxicity score in terms

of percentage. A user's toxicity score will increase if they continuously send messages with cyberbullying-related content. This toxicity score will be visible to the linked guardian account. The purpose of this toxicity score is to provide the guardian with a channel to prevent cyberbullying without invading the adolescent's privacy. In other words, the guardian could not read the content of any conversation of their children but they are informed about any initiation of potential cyberbullying behavior.

As mentioned earlier, most of the time the motive of a perpetrator could simply be an impulse of catharsis. Because of that, Serenity will try to prevent users from instantly sending messages that are detected with cyberbullying-related content. Instead, Serenity will alert the sender and give them two seconds to reconsider if they want to send the message. This type of nudges has been effective in reducing unwanted user behaviors [Caraban et al,. 2019]. By forcing the sender to reconsider their action, the objective is to reduce incidents driven by temperamental cyberbullying motives.

If the sender insists to send messages with cyberbullying- related content, Serenity will warn the receiver and hide those messages by default. While the receiver could choose to unhide the messages through certain interventions, the goal here is prioritize safety over convenience. In summary, Serenity is designed to cover all three pathways of mitigating cyberbullying. The corresponding casual map for Serenity is presented in Figure 2.

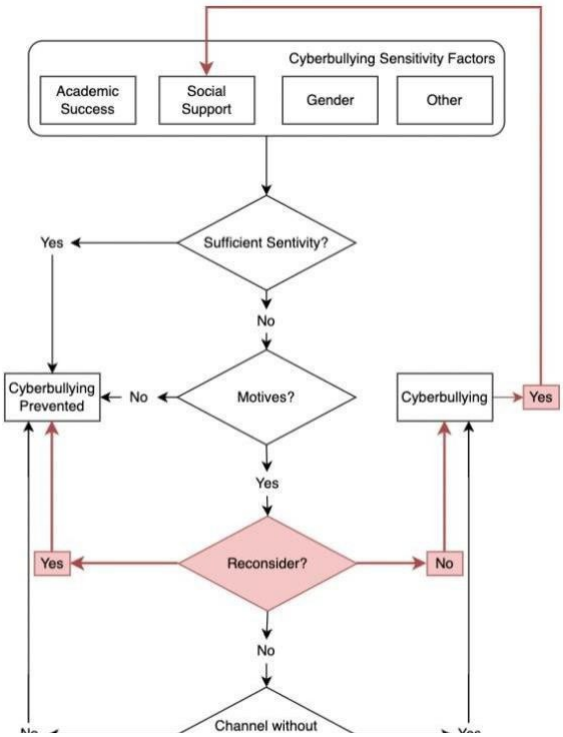

Fig 2. A causal map of cyberbullying with Serenity

3.1 (b) System Design

Serenity is developed using Ionic, a framework for cross- platform applications [Chan et al,. 2023]. There are two main stages in the construction of Ionic applications, code development and deployment. The code development is based on the Angular framework and React library. During this stage, developers do not need to specify the applications' targeted platforms; they only need to focus on writing functional code. Once code development is finalized, the developers can deploy the application to the web, Android, and iOS. In our prototype implementation, Serenity targets both Android and iOS. As for authentication and user data management, Serenity uses Google Firebase. An overview of Serenity's system architecture of both development and implementation is provided in Figure 3.

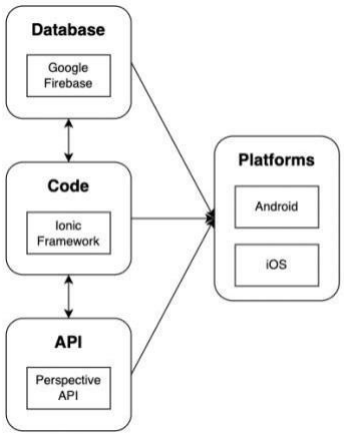

Fig 3. An overview of Serenity's system architecture

3.1 (c) Cyberbullying Detection

As mentioned above, ML is a powerful technique for detecting cyberbullying. Serenity integrates Google's Perspective API, which is explicitly designed to detect abuse or harassment in real time using novel ML models [Chan et al,. 2023]. Any message sent through Serenity is first sent to the Perspective API for content analysis. This process benefits from Google's high-performance ML servers as the entire process takes less than a second. Therefore, Serenity enables highly efficient cyberbullying detection and real-time user feedback. Once analysis is completed, Perspective API will send a toxicity score of the message to Serenity. Toxicity scores range from zero to one and indicate the likelihood that a message contains cyberbullying-related content. It

is formatted and displayed as a percentage. A score close to 0% suggests that the message does not have any cyberbullying-related content. In contrast, a score closes to 100% indicates that the corresponding message is highly likely to contain cyberbullying-related content (Figure 4 and 5).

Serenity users can set their own toxicity tolerance level (whereas the system level default is set to 0.75 or 75%) as shown in Figure 6. Therefore, adolescent users that are highly susceptible to the negative effects of cyberbullying can manually decrease their tolerance level, so they are less likely to be a victim of cyberbullying and prevent the cyberbullying cycle before it begins. If the message to be sent has a higher toxicity score than the receiver's toxicity tolerance level, the message is hidden and marked as a cyberbullying message (Figure 5).

Even though the ML models will update based on the latest conversations in cyberspace, there will still be a delay in delivering this to the user. Additionally, some words might be defined as safe by Perspective API, but they might be toxic to some specific users. To prevent this situation, Serenity enables users to build their own cyberbullying lexicon. Suppose the sent messages contain any words in the receiver's lexicon. In that case, the messages are automatically marked as cyberbullying messages with a toxicity score of one and hidden from the receiver's view. An overview of cyberbullying detection in Serenity (Figure 7).

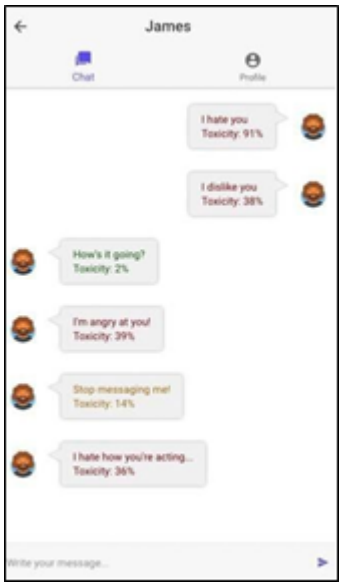

Fig 4. Displaying toxicity scores with messages

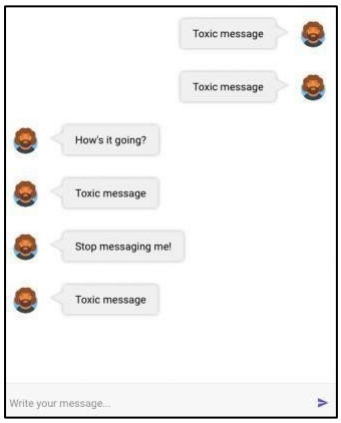

Fig 5. Hiding messages based on toxicity and tolerance level

Fig 6. User profile setting on toxicity tolerance level

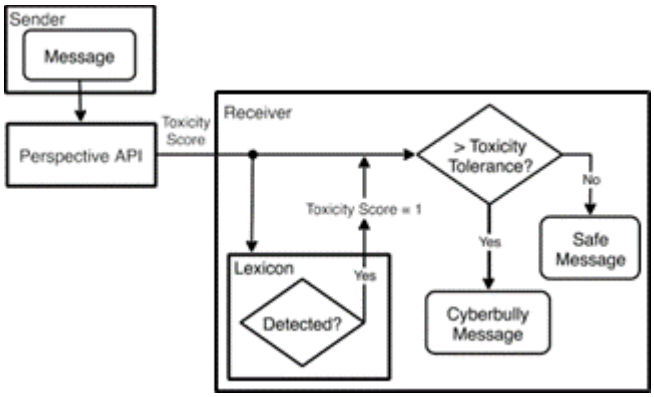

Fig 7. An overview of Serenity's cyberbullying detection process

4. Discussion

Inherently safe applications for online communications, moderated and managed by communities of guardians, teachers and social workers instead of just the profit- motivated organizations, are a crucial step towards reducing cyberbullying activities amongst adolescents. Serenity demonstrates the effectiveness of using pre-existing APIs to mitigate such activities in real-time. It goes beyond current technological design and instant messenger prototypes that do not perform in real-time. Exponential technological advancement has made it fast and relatively affordable to use these APIs to detect cyberbullying in a real-time instant messenger setting.

Alternatively, and more importantly, Serenity proposes design principles that should be evaluated and considered for integration into today's most popular social media platforms. Specifically, (Figures 4, 5, 6, and 8), the features of user-guardian accounts, nudges to reduce temperamental cyberbullying, toxicity scores to detect cyberbullying, and a self-managed toxicity tolerance score are design contributions that may benefit technologically mediated environments. Users, particularly adolescents, are more likely to use platforms where their peers are presence. On these platforms, adolescents need access to features that may deter and prevent cyberbullying. For example, while Facebook Messenger for Kids focuses on restricting who the user can message online, older adolescents can easily circumvent these restrictions by transitioning to unrestricted instant messenger such as the ones available through all major social media platforms [10]. Today, even non-traditional social media platforms such as TikTok have become environments where cyberbullying is highly likely due to the platform's contemporary nature and the acceptable levels of user anonymity [Shutsko et al ,. 2020].

As platforms encourage younger user groups to actively use their services and post more publicly available content online, these groups, particularly susceptible to cyberbullying, are exposing themselves to these possibilities to a greater extent than previous generations. Specifically, the most active population on social media, colloquially known as Gen Z, is more likely to engage in deeper forms of public content creation, such as posting pictures and videos of themselves where other potentially anonymous users may comment on their appearance [Jaconsen et al,. 2020]–[Vogels et al,. 2022]. Body image issues in this demographic are prevalent today, particularly amongst adolescent females, and have been linked to Instagram use [Chua et al,. 2016]. As anonymity is tolerated to a greater extent, persistent perpetrators can continually create new anonymous accounts to cyberbully their victims, particularly if they post content publicly or can be directly messaged from any account. A cyberbullying cycle is then established where the perpetrator is not held accountable for their actions even though they intend to cause harm.

Serenity solves this by automatically calculating and adjusting toxicity scores, enabling individual privacy and accountability.

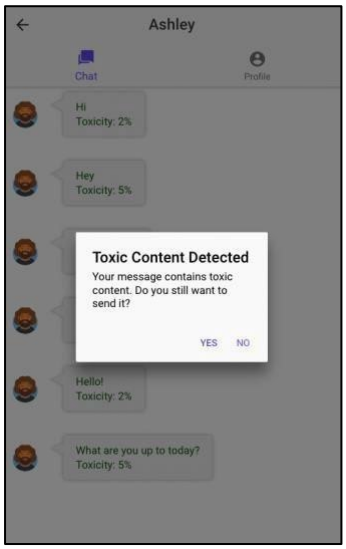

Fig 8. Displaying toxic content detection nudge

Addressing cyberbullying through multiple contexts, such as instant messaging, inappropriately posting about others, and inappropriately commenting on other users' posts, is a complex issue. Adolescents are inherently more likely to engage in risky activities and, therefore, are more exposed to cyberbullying. Thus, this group needs an environment where they can safely engage with others, consuming and creating content online.

5. Conclusion

In this written work, we advanced the prototype of a mobile messaging application which is baptized Serenity and is accounted with the anti-cyberbullying features and which targets high school youth. Serenity aims to reduce the initiation and iteration of cyberbullying events through three aspects: awareness of the users' online bullying issues; detection of the people who offensively act and confronting them with the persuasion to re-think their actions; and blocking the toxic material from offensive users to restrain from the potential victims. The detection of cyclick does not exclude integration with Google's Perspective API as a mechanism. Through collaboration with Google, high-speed servers and clever new methods of machine learning, Serenity carries out real-time instant recognition of bullying with considerable accuracy. Toxicity also allows for the user to set up his/her personal lexicon database for toxic words and to log his/her tolerance level which is customized for him/her.

What a particular person puts into his/her lexicon will be registered also as a cyberbullying candidate by the mechanism for the cyberbullying identification of Serenity.

This work should be viewed considering its intrinsic limitations. Primarily, as we are dealing with children's private information, implementations of Serenity must comply with protective legislation such as COPPA [Chan et al ,. 2023]. This is, perhaps, the most challenging task when implementing Serenity. However, future work will focus on evaluating the

importance of such a solution and its design principles rather than aiming for a production-level implementation. We hope all apps and tools serving young people with messaging functions out there would consider adopting some of the proposed design principles for the sake of protecting our next generation; and we wish parents, teachers, counsellors, social workers and policymakers to be inspired and to recognize what could be done more.

Acknowledgement

The datasets analyzed in this study were combined with funding from Bharati Vidyapeeth Deemed University Department of Engineering & Technology, Navi Mumbai. We extend our gratitude to all the organizations and individuals involved in this research work and the collection of these data.

Disclosure statement

No potential conflict of interest was reported by the authors.

ORCID: Jaydip Singh (https://orcid.org/0009-0000-9987-2949)

Data availability statement

Data will be made available if asked and required.

References

Lee, So-Hyun, and Hee-Woong Kim. "Why people post benevolent and malicious comments online." *Communications of the ACM* 58, no. 11 (2015): 74-79.

Chan, Johnny, Shohil Kishore, and Xin Yang. "Breaking the Cyberbullying Cycle with Machine Learning." In *10th IEEE Asia Pacific Conference on Computer Science and Data Engineering*. 2023

John, Ann, Alexander Charles Glendenning, Amanda Marchant, Paul Montgomery, Anne Stewart, Sophie Wood, Keith Lloyd, and Keith Hawton. "Self-harm, suicidal behaviours, and cyberbullying in children and young people: Systematic review." *Journal of medical internet research* 20, no. 4 (2018): e9044.

Hinduja, Sameer, and Justin W. Patchin. "Bullying, cyberbullying, and suicide." *Archives of suicide research* 14, no. 3 (2010): 206-221.

Gregg, Aaron. "Facebook is hitting pause on Instagram Kids app." *The Washington Post* (2021): NA-NA.

Chan, Johnny, Shohil Kishore, and Xin Yang. "Breaking the Cyberbullying Cycle with Machine Learning." In *10th IEEE Asia Pacific Conference on Computer Science and Data Engineering*. 2023.

Hill, Paige K. "INSTAGRAM IDEALS: COLLEGE WOMEN'S BODY IMAGE AND SOCIAL COMPARISON." (2022).

Van Royen, Kathleen, Karolien Poels, Heidi Vandebosch, and Philippe Adam. "“Thinking before posting?” Reducing cyber harassment on social networking sites through a reflective message." *Computers in human behavior* 66 (2017): 345-352.

Hackett, Liam, S. Jones, and Y. Yakovlev. "The annual bullying survey." (2017).

Chan, Johnny, Shohil Kishore, and Xin Yang. "Breaking the Cyberbullying Cycle with Machine Learning." In *10th IEEE Asia Pacific Conference on Computer Science and Data Engineering*. 2023.

Barta, Kristen, and Nazanin Andalibi. "Constructing authenticity on TikTok: Social norms and social support on the" Fun" Platform." *Proceedings of the ACM on Human-Computer Interaction* 5, no. CSCW2 (2021): 1-29.

Shutsko, Aliaksandra. "User-generated short video content in social media. A case study of TikTok." In *Social Computing and Social Media. Participation, User Experience, Consumer Experience, and Applications of Social Computing: 12th International Conference, SCSM 2020, Held as Part of the 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings, Part II 22*, pp. 108-125. Springer International Publishing, 2020.

Akturk, Ahmet Oguz. "Analysis of cyberbullying sensitivity levels of high school students and their perceived social support levels." *Interactive Technology and Smart Education* 12, no. 1 (2015): 44-61.

Concepción-Sánchez, José Á., Pino Caballero-Gil, and Jezabel Molina-Gil. "Application Based on Fuzzy Logic to Detect and Prevent Cyberbullying Through Smartphones." *Computer Science & Information Technology* 11 (2017).

Fahrnberger, Günter, Deveeshree Nayak, Venkata Swamy Martha, and Srini Ramaswamy. "SafeChat: A tool to shield children's communication from explicit messages." In *2014 14th International Conference on Innovations for Community Services (I4CS)*, pp. 80-86. IEEE, 2014.

Weider, D. Yu, Maithili Gole, Nishanth Prabhuswamy, Sowmya Prakash, and Vidya Gowdru Shankaramurthy. "An approach to design and analyze the framework for preventing cyberbullying." In *2016 IEEE International Conference on Services Computing (SCC)*, pp. 864-867. IEEE, 2016.

Chowdhary, KR1442, and K. R. Chowdhary. "Natural language processing." *Fundamentals of artificial intelligence* (2020): 603-649.

Noviantho, SM Isa, and Livia Ashianti. "Cyberbullying classification using text mining." In *2017 1st International Conference on Informatics and Computational Sciences (ICICoS)*, pp. 241-246. 2017.

Mahlangu, Thabo, and Chunling Tu. "Deep learning cyberbullying detection using stacked embbedings approach." In *2019 6th International Conference on Soft Computing & Machine*

Aind, Alwin T., Akashdeep Ramnaney, and Divyashikha Sethia. "Q-bully: a reinforcement learning based cyberbullying detection framework." In *2020 International conference for emerging technology*

Abbasi, Sidra, Adnan Naseem, Azra Shamim, and Muhammad Ahsan Qureshi. "An empirical investigation of motives, nature and online sources of cyberbullying." In *2018 14th International Conference on Emerging Technologies (ICET)*, pp. 1-6. IEEE, 2018.

Caraban, Ana, Evangelos Karapanos, Daniel Gonçalves, and Pedro Campos. "23 ways to nudge: A review of technology-mediated nudging in human-computer interaction." In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pp. 1-15. 2019.

Chan, Johnny, Shohil Kishore, and Xin Yang. "Breaking the Cyberbullying Cycle with Machine Learning." In *10th IEEE Asia Pacific Conference on Computer Science and Data Engineering*. 2023.

Jacobsen, Stephanie L., and Nora Ganim Barnes. "Social media, gen Z and consumer misbehavior: Instagram made me do it." *Journal of Marketing Development and Competitiveness* 14, no. 3 (2020).

Shearer, Elisa, and Amy Mitchell. "News use across social media platforms in 2020." (2021).

Vogels, Emily A., Risa Gelles-Watnick, and Navid Massarat. "Teens, social media and technology 2022." (2022).

Chua, Trudy Hui Hui, and Leanne Chang. "Follow me and like my beautiful selfies: Singapore teenage girls' engagement in self-presentation and peer comparison on social media." *Computers in human behavior* 55 (2016): 190-197.