2022

A Labeled Dataset for Investigating Cyberbullying Content Patterns in Instagram

Mara Hamlett
Arizona State University, jhamlet2@asu.edu

Grace Powell
Arizona State University, gpowell7@asu.edu

Yasin N. Silva
Loyola University Chicago, ysilva1@luc.edu

Deborah Hall
Arizona State University, d.hall@asu.edu

Recommended Citation

M. Hamlett, G. Powell, Y. N. Silva, D. L. Hall. A Labeled Dataset for Investigating Cyberbullying Content Patterns in Instagram. The 16th International AAAI Conference on Web and Social Media (ICWSM), Atlanta, GA, USA, 2022.

A Labeled Dataset for Investigating Cyberbullying Content Patterns in Instagram

Mara Hamlett[1], Grace Powell[1], Yasin N. Silva[2], Deborah Hall[1]

[1]Arizona State University, [2]Loyola University Chicago

jhamlet2@asu.edu, gpowell7@asu.edu, ysilva1@luc.edu, d.hall@asu.edu

Abstract

As online communication continues to become more prevalent, instances of cyberbullying have also become more common, particularly on social media sites. Previous research in this area has studied cyberbullying outcomes, predictors of cyberbullying victimization/perpetration, and computational detection models that rely on labeled datasets to identify the underlying patterns. However, there is a dearth of work examining the content of what is said when cyberbullying occurs and most of the available datasets include only basic labels (cyberbullying or not). This paper presents an annotated Instagram dataset with detailed labels about key cyberbullying properties, such as the content type, purpose, directionality, and co-occurrence with other phenomena, as well as demographic information about the individuals who performed the annotations. Additionally, results of an exploratory logistic regression analysis are reported to illustrate how new insights about cyberbullying and its automatic detection can be gained from this labeled dataset.

Introduction

As digital forms of communication and social networking sites have grown in use and popularity, so, too, have instances of cyberbullying and other negative online interactions. Cyberbullying is most often defined as aggressive behavior toward others through the use of electronic and online media with the intention to psychologically or emotionally harm another (Patchin & Hinduja, 2006; Hinduja & Patchin, 2008; Reekman & Cannard, 2009; Marcum et al., 2012). This form of harassment has been tied to increased rates of anxiety, depression, and suicide (Ybarra & Mitchell, 2004; Hoff & Mitchell, 2009; Kowalski et al. 2014; Hamm et al., 2015).

In psychology, extensive research has been conducted assessing the incidence patterns, negative outcomes of cyberbullying for victims, the modes by which perpetrators choose their targets, and even the occurrence rates across different social media platforms. Some of this work identified that females are more likely to be cyberbullying victims, while males are more likely to be perpetrators (Marcum et al., 2014), those with marginalized identities (i.e., sexual or gender minorities) or neuro-atypical statuses (i.e., those with ADHD) are disproportionately at risk to be involved in cyberbullying (Yen et al., 2014; Hamm et al., 2015; Abreu & Kenny, 2018), and that rates of cyberbullying vary considerably between different social media platforms (Ditch the Label, 2017). Yet crucial gaps in the literature remain. For instance, self-report survey data provides the basis for the majority of cyberbullying studies in psychology (Hamm et al., 2015; Abreau & Kenny, 2018), with few to our knowledge using datasets containing actual social media content. Furthermore, vital questions about how cyberbullying unfolds within social media interactions among different users and the extent to which it co-occurs with related phenomena have yet to be sufficiently explored.

In computer science, a number of cyberbullying detection models have been proposed. These models usually rely on machine learning algorithms that are trained using labeled datasets to identify whether a social media comment is cyberbullying or not (e.g., Al-Garadi et al., 2019; Muneer & Fati, 2020; Rosa et al., 2019; Salawu, He, & Lumsden, 2020). Despite the important work being done in this area, several crucial gaps remain. For example, most machine learning approaches to cyberbullying detection focus on binary classification tasks—i.e., on detecting whether a specific instance is cyberbullying or not (see Salawu et al., 2020). Comparatively fewer efforts have sought to detect or investigate more nuanced aspects of cyberbullying, such as the severity of a cyberbullying instance (cf., Hall et al., 2021) or how cyberbullying characteristics may differ

across motives or content areas (e.g., cyberbullying pertaining to physical appearance versus religious identity). Furthermore, only more recently have cyberbullying detection models begun to examine how cyberbullying interactions unfold over time through the continued interaction between different users online (see Potha & Maragoudakis, 2015; Soni & Sigh, 2018; Cheng et al., 2019). Finally, research is needed to identify ways to mitigate potential bias in machine learning cyberbullying detection models. A range of factors can contribute to algorithmic bias, including bias stemming from the individual characteristics and perspectives of the humans labeling training data. As argued by Kim and colleagues (2021), for instance, differing perspectives can alter the interpretation of cyberbullying comments by human annotators, with those having had similar experiences to ones mentioned in comments producing more false-positive cyberbullying labels and those who do being more conservative in their labeling of cyberbullying comments. Two complementary approaches for reducing and/or accounting for bias that can be introduced during data labeling are to: (1) recruit annotators who represent diverse demographic backgrounds and perspectives, and (2) record and report the demographic characteristics of the annotators who provide each label in a dataset.

The primary aims of this paper are to address these current gaps by:

- Introducing a new annotated dataset with detailed labels at the session (i.e., initial post and subsequent comments) and comment levels describing key cyberbullying properties, such as content *type* (gender identity, physical appearance, race, etc.), *purpose* (attack, defense), *directionality* (aimed at original post user or to others), and *co-occurrence* with other phenomena (e.g., depression, suicidality).
- Providing key demographic characteristics about the individuals who annotated the data to facilitate investigations of how bias introduced during the labeling process can be mitigated or accounted for.
- Discussing findings from an initial exploratory analysis of potential cyberbullying patterns, including descriptive and logistic regression results, to illustrate some of the novel insights about cyberbullying the dataset can help generate.
- Providing a dataset (available upon request at https://ysilva.cs.luc.edu/BullyBlocker/data) that will enable more nuanced and detailed analyses of actual social media, facilitate a better understanding of psychological aspects of cyberbullying, and yield more accurate machine learning cyberbullying detection models.

Labeled Instagram Dataset

The data used in our labeling process was adapted from a dataset of Instagram sessions collected by Hosseinmardi et al. (2015). The initial dataset (see Hosseinmardi et al., 2015) contained 2,218 Instagram sessions (i.e., initial posts along with all comments made in response to the initial post) that were labeled holistically, at the session level, as cyberbullying or not. Notably, Hosseinmardi and colleagues randomly selected the 2,218 Instagram sessions from a larger sample of 3,165,000 sessions identified as having at least one profanity in the comments on the post by a user other than the original poster. The sessions were obtained through a snowball sampling method using the Instagram API. From there, the final sessions were selected from public profiles– i.e., profiles of users whose settings allowed them to be seen by anyone on the platform (Hosseinmardi et al., 2015).

XXX	Why don't u all shut up and leave me alone! Byeee!
XXX	How about u shut up aye
XXX	Go to hell

Figure 1: Fragment of one of the labeled cyberbullying sessions. It includes a photo and three cyberbullying comments. Usernames of posters were removed for anonymity.

Following the work of Hosseinmardi and colleagues, Gupta et al. (2020) further annotated the same dataset in an investigation of temporal properties of cyberbullying. Specifically, they selected a subset of 100 sessions from the 2,218 previously labeled by Hosseinmardi et al., with an even distribution of session-level cyberbullying labels. That is, the subset of 100 Instagram sessions contained 50 cyberbullying and 50 non-cyberbullying sessions. As discussed in greater detail by Gupta et al. (2020), a cyberbullying label was then generated by members of a multidisciplinary research team for each comment in each of the 100 sessions. A primary contribution of the work by Gupta and colleagues was thus an annotated dataset of Instagram sessions that contained comment-level cyberbullying labels. To obtain access to the data initially collected and annotated by Hosseinmardi and colleagues, Gupta et al. completed a release form ensuring the privacy of the users whose posts comprised the Instagram sessions. They agreed to reference the study by Hosseinmardi et al. (2015) and to neither distribute the data to others, nor attempt to re-identify the users whose posts made up the Instagram sessions. The current study is an extension of Gupta et al. (2020)'s work, led by

members of the same research team, and follows the same privacy guidelines.

For the present data labeling process, we used the 100 sessions labeled by Gupta et al (2020). For illustrative purposes, part of one of these sessions is presented in Figure 1. Across the 100 sessions, the minimum, maximum, and average number of comments (per session) were 14, 184, and 84.54, respectively. In addition to labeling each comment as being cyberbullying or not, to replicate the work of Gupta and colleagues, our team also labeled each cyberbullying comment along the following dimensions: (1) ***Type:*** whether cyberbullying content was related to sexual behavior (not identity-based), sexual orientation or gender identity, physical appearance, race or ethnicity, intelligence, religion, or general hate not pertaining to one of the previous content areas (e.g., threats, profanity, etc.); (2) ***Purpose*:** whether the comment was an attack against another user, was made in defense of oneself, or was made in defense of another user; (3) ***Directionality*:** whether the comment was directed at the user who made the original post or another user; and (4) ***Co-occurrence:*** whether the comment also included content pertaining to other related phenomena (e.g., depression, suicide, anxiety, and discrimination). The content types chosen for labeling were based on topics perceived to be relevant to the kinds of statements used in cyberbullying scenarios (Hamm et al., 2015; Ditch the Label, 2017).

Each session was labeled by two members of our team with expertise in computing and psychology. Specifically, the 100 sessions were first divided into 5 subsets containing 20 sessions each. Each of these 5 subsets were then annotated by two members of the research team. Consequently, each team member was responsible for labeling a total of 1 subset (i.e., 20 sessions) and each of the 5 subsets was labeled independently by two team members. The instructions for the labeling process included reading each comment in a session and first determining whether the comment reflected cyberbullying or not, and, if the comment was labeled as cyberbullying, to determine whether the comment did or did not fit the criteria for each of the added dimensions (type, purpose, directionality, co-occurrence). Labels were such that a value of 1 always denoted the affirmative (characteristic appears in this comment) and a value of 0 always denoted the negative (characteristic does not appear for this comment). For instance, cyberbullying comments were assigned a value of 1 and non-cyberbullying comments were assigned a value of 0; comments determined to reflect cyberbullying about physical appearance were assigned a label of 1 for this subcategory of the *type* dimension and cyberbullying comments unrelated to physical appearance were assigned a label of 0. *Purpose* had an additional number used when assessing the defensiveness of the comment, with a value of 1 denoting that a comment was in defense of the user making the comment (i.e., self-defense), a value of 2 denoting a comment in defense of another user in the thread, and a value of 0 denoting a comment that was not made in defense of oneself or another user. Non-cyberbullying comments were labeled for patterns of co-occurrence with related phenomena only.

Next, the detailed annotations for each of the two team members who labeled a specific subset of 20 sessions were compiled and compared. Discrepancies in the labels assigned by the two team members were then resolved by one of two additional team members who served as a third rater (i.e., a graduate student and a PhD researcher on the team).

Key demographic information for the team of annotators was assessed via an (anonymous) online survey. This was done to assess the diversity within the labeling team (presented in Table 1) and, crucially, to provide an indicator of potential bias in the labeling process that could be systematically accounted for by future researchers using this data. The mean age of the annotators was 24.77 years (SD = 6.19) and ranged from 19-43 years old. The annotators self-reported (via free response) their race/ethnicity as follows: 7 annotators identified as White (53.8%), 4 identified as Hispanic/Latiné (30.8%), and 2 identified as Asian (15.4%). When asked about their gender identity, 5 annotators reported they were cisgender men (38.5%), 5 annotators reported they were cisgender women (38.5%), 2 reported they were non-binary (15.4%), and one reported they were a

Coder ID	What is your current age?	What is your race/ethnicity?	What best describes your gender identity?	Would you consider yourself a member of the LGBTQ+ Community?	What is your current level of education?	Are you a first generation college student?	What is your area of study ?
C1	26	White	Cisgender Woman	Yes	Graduate 3+ Years	No	Psychology
C2	21	Hispanic	Cisgender Man	No	Undergraduate 4th Year	No	Computer Science
C3	22	Hispanic/Latino	Non-binary	Yes	Undergraduate 5+ Years	No	Computer Science
C4	26	White	Cisgender Woman	No	Graduate 3+ Years	Yes	Psychology
C5	19	White	Cisgender Woman	Yes	Undergraduate 3rd Year	No	Computer Science
C6	24	Asian	Cisgender Man	No	Graduate 2nd Year	Yes	Psychology
C7	27	Hispanic	Cisgender Man	No	Graduate 1st Year	Yes	Computer Science
C8	24	White	Cisgender Man	Yes	Undergraduate 5+ Years	No	Computer Science
C9	22	White	Cisgender Woman	No	Graduate 1st Year	No	Computer Science
C10	20	Asian	Cisgender Man	No	Undergraduate 3rd Year	Yes	Computer Science
C11	28	White	Transgender Woman	Yes	Graduate 3+ Years	No	Psychology
C12	43	Latino	Cisgender Woman	No	PhD Recipient	No	Computer Science
C13	20	White	Non-binary	Yes	Undergraduate 3rd Year	No	Psychology

Table 1: Demographics of the annotators.

			Content Type							Purpose			Directionality		Cooccurrence			
	Total Comments	CB Comments	Sexual	Sexual Orientation/ Gender Identity	Physical Appearance	Race/ Ethnicity	Intellectual	Religious	General Hate	Attack	Defense of another	Defense of Self	Targets Original Post User	Targets Other Users	Depression	Suicide	Anxiety	Discrimination
CB Session	3,705	1045	226	101	193	130	194	40	663	973	86	78	394	619	21	22	8	118
Non-CB Sessions	4,749	20	5	5	2	0	3	0	5	17	2	0	9	9	3	1	2	1
Total	8,454	1065	231	106	195	130	197	40	668	990	88	78	403	628	24	23	10	119

Table 2: Core statistics of the labeled dataset.

transgender woman (7.7%). 6 annotators indicated that they were a member of the LGBTQ+ community (46.2%), whereas 7 annotators indicated that they were not (53.8%). Annotators' level of education ranged from currently pursuing an undergraduate (bachelor's) degree to holding a doctoral (Ph.D.) degree. In particular, 3 annotators reporting being a 3rd year undergraduate student (23.1%), 1 reporting being a 4th year undergraduate student (7.7%), 2 reporting being a 5th year or higher undergraduate student (15.4%), 2 reporting being a 1st year graduate student (15.4%), 1 reporting being a 2nd year graduate student (7.7%), 3 reporting being a 3rd year or higher graduate student (23.1%), and 1 reporting having received their Ph.D. (7.7%). 4 annotators (i.e., 30.8%) indicated that they were a first-generation college student, whereas 9 (69.2%) indicated that they were not. Lastly, 8 annotators were individuals whose primary area of academic study was computer science (61.5%) and 5 were individuals whose primary area of study was psychology (38.5%).

Analyses

Descriptive Results

As presented in Table 2, across all 100 sessions, 1,065 comments were labeled as cyberbullying and 7,389 as non-cyberbullying. The remaining details presented in Table 1 — e.g., content type, purpose, etc.— consider the case of cyberbullying comments. Among the cyberbullying comments, 231 (21.70%) were sexual comments, 106 (9.95%) were related to gender identity/sexual orientation, 195 (18.31%) were about physical appearance, 130 (12.21%) were about race/ethnicity, 197 (18.50%) were about intelligence, 40 (3.76%) were about religion, and 668 (62.72%) were considered general hate (see Table 1 for a breakdown of cyberbullying and non-cyberbullying sessions). Of all cyberbullying comments, 990 (92.96%) were considered attacks, 88 (8.26%) were in defense of another user, and 78 (7.32%) were in self-defense. 403 (37.8%) of the comments were directed at the user who made the initial post and 628 (59.0%) were directed at other users. Among the other phenomena assessed, across all of the cyberbullying comments, 21 (1.97%) pertained to depression, 22 (2.07%) to suicide, 8 (0.75%) to anxiety, and 118 (11.1%) to discrimination.

Logistic Regressions of Cyberbullying Comment Types

We performed binary logistic regression to shed light on how the purpose, directionality, and cooccurrence of cyberbullying comments might vary by content type. (General hate was not included in the analyses due to the broad nature of this category.) These analyses were exploratory in nature and intended to illustrate only a small handful of insights that our detailed annotations might help generate. One important consideration was the possibility of clustering effects due to the nested property of individual comments within each social media session.

To investigate the extent to which this nested structure might be problematic for binary logistic regression, we tested a series of unconditional mean models following Sommet and Morselli (2017). For each model, the intraclass correlation coefficient (ICC) was calculated using the formula: $var(u_{0j}) / (var(u_{0j}) + (\pi^2/3))$. As shown in Table 3, the models for gender identity/sexual orientation, physical appearance, race/ethnicity, and discrimination had ICCs that were significantly different from 0—indicated by 95% confidence intervals that did not include 0. Given evidence of clustering effects for these specific models, we omitted them from the binary logistic regression analyses. Although beyond the scope of this paper, future analyses that employ multilevel logistic regression might yield additional insights (see Sommet and Morselli, 2017).

Each model was run separately comparing the cyberbullying content type with the aspects of purpose, directionality, and co-occurrence with related phenomena. The logistic regression models are reported in Table 4. Below we briefly discuss results that were significant at the level of $p \leq .05$.

Content Type and Attacks

Cyberbullying comments pertaining to religion were 0.42 times as likely to be labeled an attack (versus not)—that is, they were significantly *less* likely to be labeled an attack.

	95% Confidence Interval				95% Confidence Interval		
	Intraclass Correlation Coefficient	Lower Bound	Upper Bound		Intraclass Correlation Coefficient	Lower Bound	Upper Bound
Type				**Directionality**			
Sexual Content	<.001	–	–	Directed at Others	<.001	–	–
Gender Identity/ Sexual Orientation	0.24	0.45	23.89	Directed at Original Poster	<.001	–	–
Physical Appearance	0.09	0.002	62.94	**Co-Occurence**			
Race/Ethnicity	0.48	0.05	199.13	Depression	<.001	–	–
Intelligence	<.001	–	–	Suicide	<.001	–	–
Religion	<.001	–	–	Discrimination	0.11	0.01	25.54
Purpose							
Attack/Insult	<.001	–	–				
Defense of Self	<.001	–	–				
Defense of Another	<.001	–	–				

Table 3: ICCs and 95% confidence intervals.

Content Type and Defensive Comments

Cyberbullying comments about intelligence were 2.10 times more likely to be labeled as self-defense and cyberbullying comments about religion were 3.50 times more likely to be labeled self-defense. (The likelihood of cyberbullying comments being in defense of another user did not vary by content type.)

Content Type and Comments Directed at User with Initial Post

Cyberbullying comments about sexual content, intelligence, and religion were significantly *less* likely to be directed at the user who made the initial post, reflected in odds ratios of 0.56, 0.17, and 0.08, respectively.

Content Type and Comments Directed at Others

Cyberbullying comments about intelligence were 6.13 times more likely to be directed at other users (versus not) and cyberbullying comments about sexual content were 2.16 times more likely to be directed at other users (versus not).

Content Type and Co-Occurrence

Due to the low number of cyberbullying comments labeled as depression, suicide, and anxiety, no significant differences in the likelihood of these labels were observed between content type. In fact, these models failed to converge after 700 iterations of the maximum likelihood estimation procedure for estimating the logistic regression parameters. (As a result, the findings for depression, suicide, and anxiety are not included in Table 4.)

Discussion

In the present paper, we discuss a detailed annotation procedure that our research team employed to facilitate a more nuanced understanding of cyberbullying on social media. Building on data collected and the session-level cyberbullying labels generated by Hosseinmardi et al. (2015), our team created labels to shed light on the content type, purpose, directionality, and co-occurrence with related phenomena of

	Attack/Insult					Defense of Self				
	b	SE	Exp(*B*)	$\chi 2$	*p*	*b*	SE	Exp(*B*)	$\chi 2$	*p*
Sexual Content	0.27	0.29	1.31	0.91	0.34	-0.53	0.36	0.59	2.25	0.13
Intelligence	0.52	0.33	1.68	2.26	0.11	0.74	0.28	2.1	7.05	0.01
Religion	-0.87	0.47	0.42	3.51	0.06	1.26	0.42	3.53	8.86	0.003

	Defense of Another					Directed at Original Poster				
	b	SE	Exp(*B*)	$\chi 2$	*p*	*b*	SE	Exp(*B*)	$\chi 2$	*p*
Sexual Content	-0.21	0.38	0.81	0.32	0.57	-0.58	0.18	0.56	11.04	0.001
Intelligence	-0.45	0.43	0.64	1.08	0.3	-1.76	0.25	0.17	49.14	<.001
Religion	0.17	0.78	1.18	0.05	0.83	-2.52	1.03	0.08	6	0.01

	Directed at Others				
	b	SE	Exp(*B*)	$\chi 2$	*p*
Sexual Content	0.77	0.18	2.16	19.22	<.001
Intelligence	1.81	0.24	6.13	59.5	<.001
Religion	0.26	0.44	1.29	0.34	0.56

Table 4: Logistic regression results.

individual cyberbullying comments in a dataset of 100 Instagram sessions. Moreover, we collected demographic data from our team of annotators, which included individuals from a range of sociodemographic backgrounds, that is reported with the annotations, themselves. By doing so, our aim was to allow researchers accessing this data to account for and potentially mitigate bias that may have resulted from the characteristics and perspectives of our annotators. Finally, the results of a series of exploratory logistic regression analyses demonstrate how these more nuanced cyberbullying labels can provide new insights about the nature of cyberbullying.

Interestingly, several "either/or" patterns emerged, such that content types that were more likely to reflect certain kinds of purpose and directionality were less likely to occur in other circumstances. For example, cyberbullying comments about sexual content and about intelligence were more likely to be directed at other commenters and less likely to be directed at the user who made the initial post.

As previously mentioned, there was a general hate (content type) category that was used to label any cyberbullying comment that did not fall into one of the other content type categories. The high frequency of cyberbullying comments in this broad category reflects that cyberbullying may most typically involve general hostility in the absence of a more specific topic (e.g., race/ethnicity, religion). Because of this high frequency, comments in the general hate category were not included in the exploratory logistic regression analyses we performed. We note, however, that this marks an important avenue for future research examining cyberbullying content on social media.

Additionally, the logistic regressions reported in this paper represent preliminary analysis at the comment-level intended to provide initial insights into cyberbullying content patterns. Future work should take into account a range of additional factors that may impact these patterns. One such factor, for example, is the extent to which users repeatedly post within the same session–a characteristic viewed by some (see Hamm et al., 2015) to be a defining element of cyberbullying. In some sessions, the multiple comments made by a single user and unique patterns of repetition may impact cyberbullying behaviors in ways not optimally captured through comment-level investigations solely.

Ethical Considerations

A number of ethical considerations guided the labeling process and procedures. First, efforts were made to ensure sociodemographic and disciplinary diversity among the team of annotators. As mentioned previously, our team of annotators comprised individuals with variability in level and area of academic training, age, and multiple dimensions of identity (e.g., race, ethnicity, gender identity, membership in the LGBTQ+ community). The collection of annotator demographic data was performed in a way that minimized the number of team members able to access this information; the survey was anonymous, with each annotator receiving only a numerical ID, and only one member of the research team (the first author) had full access to the survey. Privacy of the Instagram users whose comments comprised the dataset was also prioritized, as no attempts were made to discover the identities of the users at any point.

Our main goal for the dataset and this paper is to facilitate future work that uses psychological frameworks to better understand cyberbullying in social media data and inform the development of more effective cyberbullying detection models. A vital secondary goal, however, is to maintain the privacy of those included in the dataset. To this end, researchers who request access to the dataset through the researchers' project website (https://ysilva.cs.luc.edu/BullyBlocker/data) will be required to maintain the privacy of the users and annotators who contributed to the labeled dataset and ethicality in the use of the data.

References

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.

Abreu, R. L., & Kenny, M. C. (2017). Cyberbullying and LGBTQ youth: A systematic literature review and recommendations for prevention and intervention. *Journal of Childhood Adolescent Trauma*, 11, 81-97. https://doi.org/10.1007/s40653-017-0175-7.

Cheng, L., Guo R., & Liu, H. (2019). Robust Cyberbullying Detection with Causal Interpretation. *In Companion Proceedings of The 2019 World Wide Web Conference (WWW '19)*, 169–175. doi:https://doi.org/10.1145/3308560.3316503

Cheng, L., Li, J., Silva, Y., Hall, D., & Liu, H. (2019). PI-Bully: Personalized Cyberbullying Detection with Peer Influence. *The 28th International Joint Conference on Artificial Intelligence (IJCAI)*. https://doi.org/10.24963/ijcai.2019/808

Cheng, L., Li, J., Silva, Y., Hall, D., & Liu, H. (2019). Xbully: Cyberbullying detection within a multi-modal context. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining* (pp. 339-347). https://doi.org/10.1145/3289600.3291037

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

Gupta, A., Yang, W., Sivakumar, D., Silva, Y. N., Hall, D. L., & Nardini Barioni, M. C. (2020). Temporal properties of cyberbullying on Instagram. *WWW '20: Companion Proceedings of the Web Conference 2020*, 576-583. https://doi.org/10.1145/3366424.3385771.

Hall, D., Silva, Y., Wheeler, B., Cheng, L., Baumel, K. (2021). Harnessing the Power of Interdisciplinary Research with Psychology-Informed Cyberbullying Detection Models. *Int Journal of Bullying Prevention*. https://doi.org/10.1007/s42380-021-00107-5

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. https://doi.org/10.1001/jamapediatrics.2015.0944.

Hinduja, S., & Patchin, J. (2008). Cyberbullying: An exploratory analysis of factors related to offending and victimization. *Journal of Deviant Behavior*, *29*(2), 129-156. https://doi.org/10.1080/01639620701457816.

Hoff, D. L., & Mitchell, S. N. (2009). Cyberbullying: Causes, effects, and remedies. *Journal of Educational Administration*, *47*(5), 652-665. https://doi.org/10.1108/09578230910981107.

Hosseinmardi, H., Mattson, S. A., Ibn Rafiq, R., Han, R., Lv, Q., & Mishra, S. (2015). Analyzing labeled cyberbullying incidents on the Instagram social network. *International Conference on Social Informatics*, *9471*, 49-66. https://doi.org/10.1007/978-3-319-27433-1_4.

Kim, S., Razi, A., Stringhini, G., Wisniewski, P. J., & De Choudhury, M. (2021). You Don't Know How I Feel: Insider-Outsider Perspective Gaps in Cyberbullying Risk Detection. *Proceedings of the International AAAI Conference on Web and Social Media*, *15*(1), 290-302. https://ojs.aaai.org/index.php/ICWSM/article/view/18061

Koo, T. K., & Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of chiropractic medicine*, *15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

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.

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

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.

Nektaria, P., & Manolis, M. (2015). Time series forecasting in cyberbullying data. *Proceedings of the Engineering Applications of Neural Networks - 16th International Conference, EANN 2015*, *517*, Springer, 289– 303. https://doi.org/10.1007/978-3-319-23983-5_27

Patchin, J. W., & Hinduja, S. (2006). Bullies move beyond the schoolyard: A preliminary look at cyberbullying. *Journal of Youth Violence and Juvenile Justice*, *4*(2), 148-169. https://doi.org/10.1177/1541204006286288.

Reeckman, B., & Cannard, L. (2009). Cyberbullying: A TAFE perspective. *Youth Studies Australia*, *28*(2), 41-49.

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.

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.

Sommet, N. and Morselli, D. (2017). Keep Calm and Learn Multilevel Logistic Modeling: A Simplified Three-Step Procedure Using Stata, R, Mplus, and SPSS. *International Review of Social Psychology*, *30*(1), 203–218, doi: https://doi.org/10.5334/irsp.90

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). https://ojs.aaai.org/index.php/ICWSM/article/view/15046

The annual bullying survey 2017 (Ditch the Label, Comp.). (2017). Ditch the Label.

Whittaker, E., & Kowalski, R. M. (2015). Cyberbullying via social media. *Journal of School Violence*, *14*, 11-29. https://doi.org/10.1080/15388220.2014.949377.

Ybarra, M. L., & Mitchell, K. J. (2004). Online aggressor/targets, aggressors, and targets: A comparison of associated youth characteristics. *Journal of Child Psychology and Psychiatry*, *45*(7), 1308-1316. https://doi.org/10.1111/j.1469-7610.2004.00328.x.

Yen, C.-F., Chou, W.-J., Liu, T.-L., Ko, C.-H., Yang, P., & Hu, H.-F. (2014). Cyberbullying among male adolescents with attention-deficit/hyperactivity disorder: Prevalence, correlates, and association with poor mental health status. *Research in Developmental Disabilities*, *35*(12), 3543-3553. https://doi.org/10.1016/j.ridd.2014.08.035.